Lecture 15: Dynamic Memory (cont.)

Instructor: Mitch Neilsen
Office: N219D

Quote of the Day

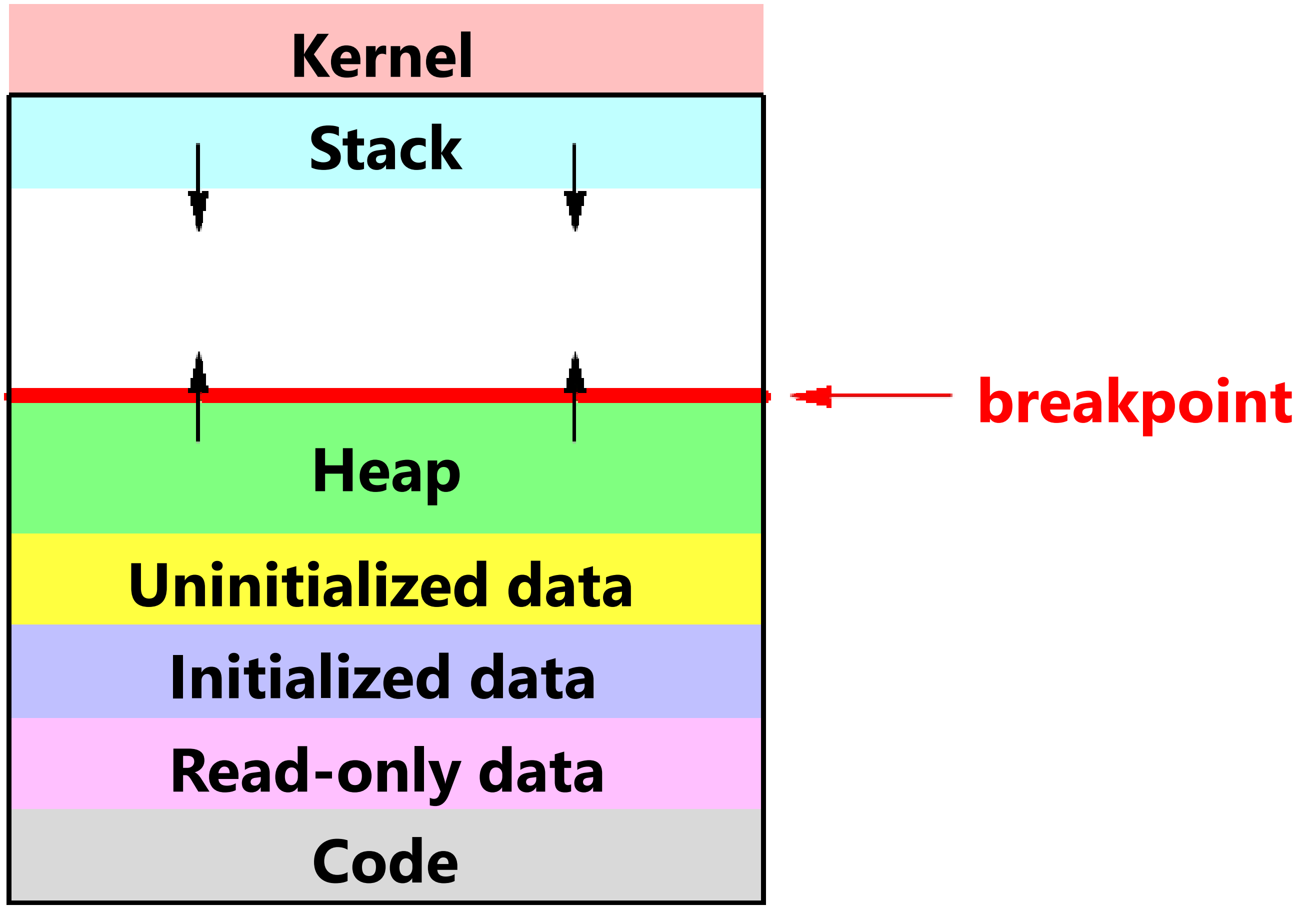
"I never did a day's work in my life. It was all fun."

-- Thomas Edison

Chapter 9: Virtual Memory

- Background
- Demand Paging
- Copy-on-Write
- Page Replacement
- Allocation of Frames
- Thrashing
- **Memory-Mapped Files**
- **Allocating Kernel Memory**
- **Other Considerations**
- **Operating-System Examples**

Recall typical virtual address space



- **Dynamically allocated memory goes in heap**
- **Top of heap called *breakpoint***
  - Addresses between breakpoint and stack all invalid

Early VM system calls

- **OS keeps "Breakpoint" – top of heap**
  - Memory regions between breakpoint & stack fault on access
- char *brk (const char *addr);
  - Set and return new value of breakpoint
- char *sbrk (int incr);
  - Increment value of the breakpoint & return old value
- **Can implement** malloc **in terms of** sbrk
  - But hard to "give back" physical memory to system

Example (memtest.c)

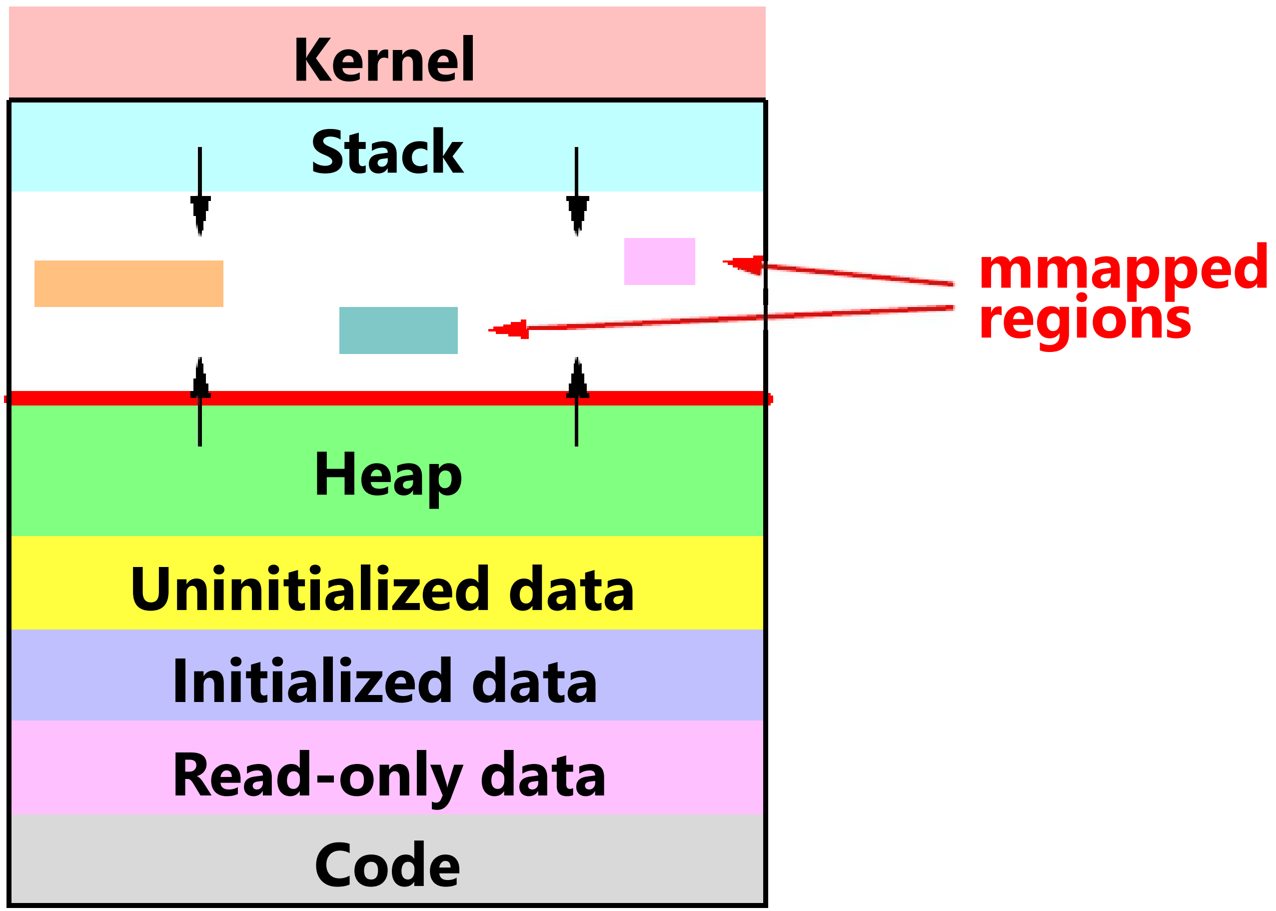
cislinux.cis.ksu.edu - PuTTY

```
   i = sbrk(1024);
   printf("Old top of heap = %16X\n", (unsigned) i);
   i = sbrk(1024);
   printf("Old top of heap = %16X\n", (unsigned) i);

   fd = open(argv[1], O_RDWR);
   fstat(fd, &sb);
   printf("File size: %lu \n",(uint64_t)sb.st_size);
```

...

```
neilsen@cougar:/pub/CIS520/programs$ ./memtest test.txt
Old top of heap =          17AF000
Old top of heap =          17AF400
File size: 27
[0]=63 [1]=64 [2]=65 [3]=66 [4]=67 [5]=68 [6]=69 [7]=6A [8]=6B [9]=6C
[0]=64 [1]=65 [2]=66 [3]=67 [4]=68 [5]=69 [6]=6A [7]=6B [8]=6C [9]=6D
neilsen@cougar:/pub/CIS520/programs$ ./memtest test.txt
Old top of heap =           E57000
Old top of heap =           E57400
File size: 27
[0]=64 [1]=65 [2]=66 [3]=67 [4]=68 [5]=69 [6]=6A [7]=6B [8]=6C [9]=6D
[0]=65 [1]=66 [2]=67 [3]=68 [4]=69 [5]=6A [6]=6B [7]=6C [8]=6D [9]=6E
```

Memory mapped files



- Other memory objects between heap and stack

mmap **system call**

- void *mmap (void *addr, size t len, int prot, int flags, int fd, off t_offset)
  - Map file specified by fd at virtual address addr
  - If addr is NULL, let kernel choose the address
- prot **– protection of region**
  - OR of PROT_EXEC, PROT_READ, PROT_WRITE, PROT_NONE
- flags
  - MAP_ANON – anonymous memory (fd should be -1)
  - MAP_PRIVATE – modifications are private
  - MAP_SHARED – modifications seen by everyone

More VM system calls

- int msync(void *addr, size_t len, int flags);
  - Flush changes of mmapped file to backing store
- int munmap(void *addr, size_t len)
  - Removes memory-mapped object
- int mprotect(void *addr, size_t len, int prot)
  - Changes protection on pages
- int mincore(void *addr, size_t len, char *vec)
  - Returns in vec which pages are present

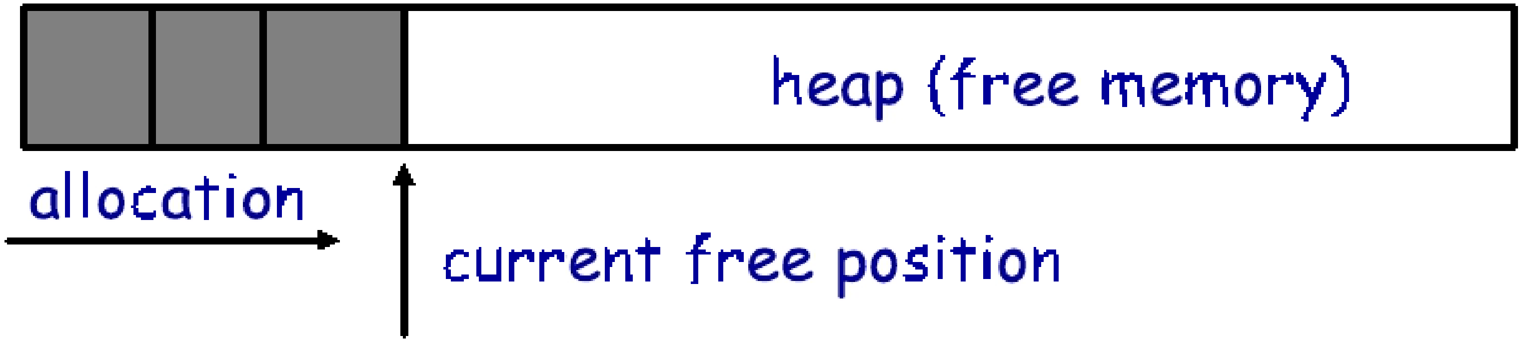
```
    fd = open(argv[1], O_RDWR);
    fstat(fd, &sb);
    printf("File size: %lu \n",(uint64_t)sb.st_size);

    memblock = mmap(NULL, sb.st_size, PROT_READ|PROT_WRITE, MAP_SHARED, fd, 0);
    if (memblock == MAP_FAILED) handle_error("mmap");

    for(i=0; i<10; i++)
    {
      printf("[%lu]=%X ", i, memblock[i]);
      memblock[i]++;
    }
    printf("\n");

    for(i=0; i<10; i++)
    {
      printf("[%lu]=%X ", i, memblock[i]);
    }
    printf("\n");

    if (msync(memblock, sb.st_size, MS_SYNC) == -1) handle_error("msync");
    if (munmap(memblock, sb.st_size) == -1) handle_error("munmap");
    close(fd);
    return(0);
}
neilsen@cougar:/pub/CIS520/programs$ ./memtest test.txt
Old top of heap =          17AF000
Old top of heap =          17AF400
File size: 27
[0]=63 [1]=64 [2]=65 [3]=66 [4]=67 [5]=68 [6]=69 [7]=6A [8]=6B [9]=6C
[0]=64 [1]=65 [2]=66 [3]=67 [4]=68 [5]=69 [6]=6A [7]=6B [8]=6C [9]=6D
neilsen@cougar:/pub/CIS520/programs$ ./memtest test.txt
Old top of heap =           E57000
Old top of heap =           E57400
File size: 27
[0]=64 [1]=65 [2]=66 [3]=67 [4]=68 [5]=69 [6]=6A [7]=6B [8]=6C [9]=6D
[0]=65 [1]=66 [2]=67 [3]=68 [4]=69 [5]=6A [6]=6B [7]=6C [8]=6D [9]=6E
```

Dynamic memory allocation

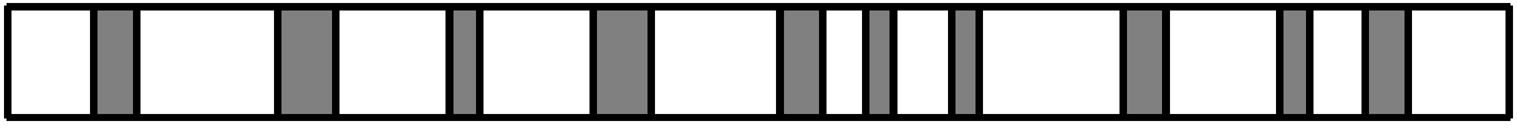
- **Almost every useful program uses it**
  - Gives wonderful functionality benefits
    - ⊲ Don't have to statically specify complex data structures
    - ⊲ Can have data grow as a function of input size
    - ⊲ Allows recursive procedures (stack growth)
  - But, can have a huge impact on performance
- **Today: how to implement it**
  - Lecture draws on [Wilson] (good survey from 1995)
- **Some interesting facts:**
  - Two or three line code change can have huge, non-obvious impact on how well an allocator works (examples to come)
  - Proven: impossible to construct an "always good" allocator
  - Surprising result: after 35 years, memory management still poorly understood

Why is it hard?

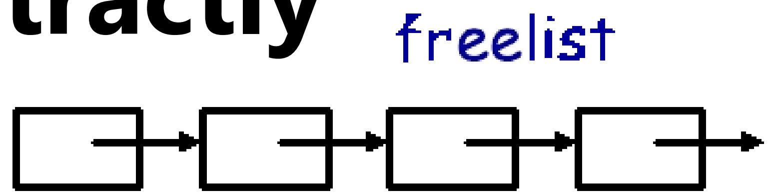
- Satisfy arbitrary set of allocation and free's.
- Easy without free: set a pointer to the beginning of some big chunk of memory ("heap") and increment on each allocation:



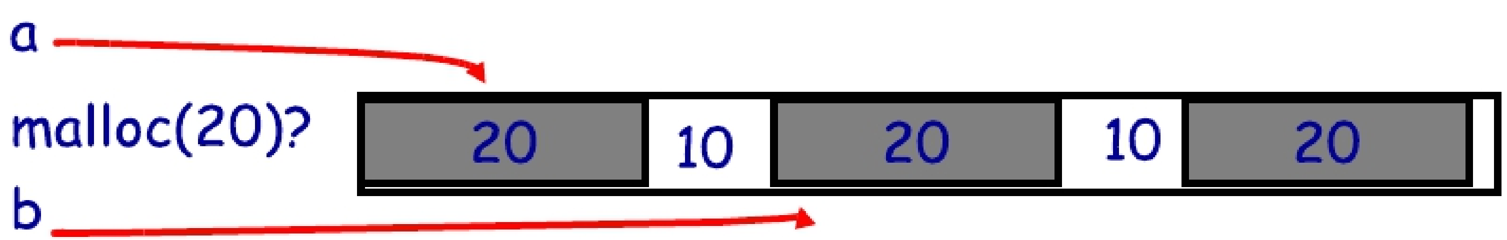
- Problem: free creates holes ("fragmentation") Result? Lots of free space but cannot satisfy request!



More abstractly



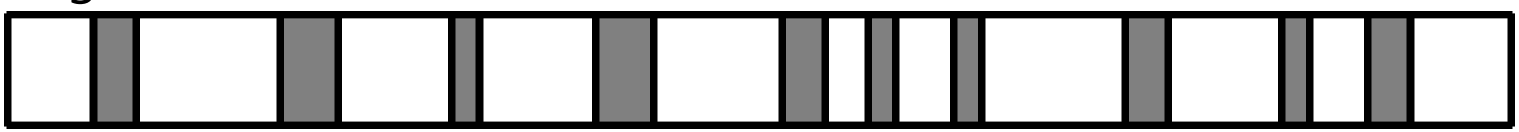
- **What an allocator must do:**
  - Track which parts of memory in use, which parts are free
  - Ideal: no wasted space, no time overhead
- **What the allocator cannot do:**
  - Control order of the number and size of requested blocks
  - Change user ptrs $\Longrightarrow$ (bad) placement decisions permanent



- **The core fight: minimize fragmentation**
  - App frees blocks in any order, creating holes in “heap”
  - Holes too small? cannot satisfy future requests

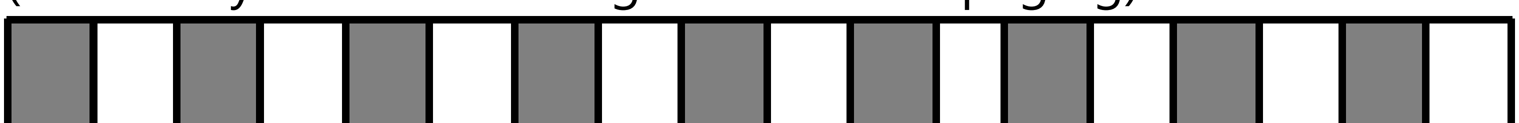
What is fragmentation really?

- **Inability to use memory that is free**
- **Two factors required for fragmentation**
  - Different lifetimes—if adjacent objects die at different times, then fragmentation:



  - If they die at the same time, then no fragmentation:

  - Different sizes: If all requests the same size, then no fragmentation (that’s why no external fragmentation w. paging):

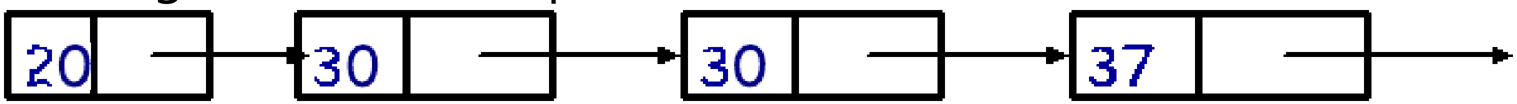


Impossible to "solve" fragmentation

- **If you read allocation papers to find the best allocator**
  - All discussions revolve around tradeoffs
  - The reason? There cannot be a best allocator
- **Theoretical result:**
  - For any possible allocation algorithm, there exist streams of allocation and deallocation requests that defeat the allocator and force it into severe fragmentation.
- **How much fragmentation should we tolerate?**
  - Let M = bytes of live data, $n_{\min}$ = smallest allocation, $n_{\max}$ = largest – How much gross memory required?
  - Bad allocator: $M \cdot (n_{\max} / n_{\min})$ (only ever uses a memory location for a single size)
  - Good allocator: $\sim M \cdot \log(n_{\max} / n_{\min})$

Best fit

- **Strategy: minimize fragmentation by allocating space from block that leaves smallest fragment**
  - Data structure: heap is a list of free blocks, each has a header holding block size and pointers to next

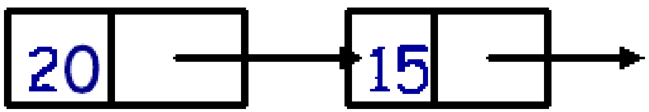


  - Code: Search freelist for block closest in size to the request. (Exact match is ideal)
  - During free (usually) coalesce adjacent blocks
- **Problem: Sawdust**
  - Remainder so small that over time left with "sawdust" everywhere
  - Fortunately not a problem in practice

First fit

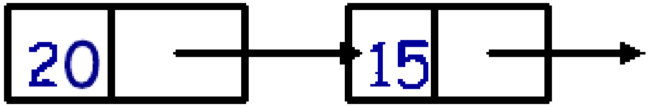
- **Strategy: pick the first block that fits**
  - Data structure: free list, sorted lifo, fifo, or by address
  - Code: scan list, take the first one
- **LIFO: put free object on front of list.**
  - Simple, but causes higher fragmentation
  - Potentially good for cache locality
- **Address sort: order free blocks by address**
  - Makes coalescing easy (just check if next block is free)
  - Also preserves empty/idle space (locality good when paging)
- **FIFO: put free object at end of list**
  - Gives similar fragmentation as address sort, but unclear why

First fit: Nuances

- **First fit sorted by address order, in practice:**
  - Blocks at front preferentially split, ones at back only split when no larger one found before them
  - Result? Seems to roughly sort free list by size
  - So? Makes first fit operationally similar to best fit: a first fit of a sorted list = best fit!
- **Problem: sawdust at beginning of the list**
  - Sorting of list forces a large requests to skip over many small blocks. Need to use a scalable heap organization
- **Suppose memory has free blocks:** 

  - If allocation ops are 10 then 20, best fit wins
  - When is FF better than best fit?

First fit: Nuances

- **First fit sorted by address order, in practice:**
  - Blocks at front preferentially split, ones at back only split when no larger one found before them
  - Result? Seems to roughly sort free list by size
  - So? Makes first fit operationally similar to best fit: a first fit of a sorted list = best fit!
- **Problem: sawdust at beginning of the list**
  - Sorting of list forces a large requests to skip over many small blocks. Need to use a scalable heap organization
- **Suppose memory has free blocks:** 

  - If allocation ops are 10 then 20, best fit wins
  - When is FF better than best fit?
  - **Suppose allocation ops are 8, 12, then 12 ⇒ first fit wins**

First/best fit: weird parallels

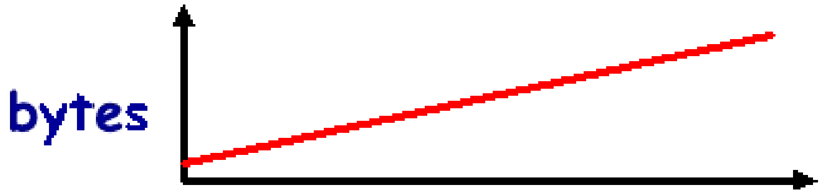
- **Both seem to perform roughly equivalently**
- **In fact the placement decisions of both are roughly identical under both randomized and real workloads!**
  - No one knows why
  - Pretty strange since they seem pretty different
- **Possible explanations:**
  - First fit like best fit because over time its free list becomes sorted by size: the beginning of the free list accumulates small objects and so fits tend to be close to best
  - Both have implicit "open space heuristic" try not to cut into large open spaces: large blocks at end only used when have to be (e.g., first fit: skips over all smaller blocks)

Some worse ideas

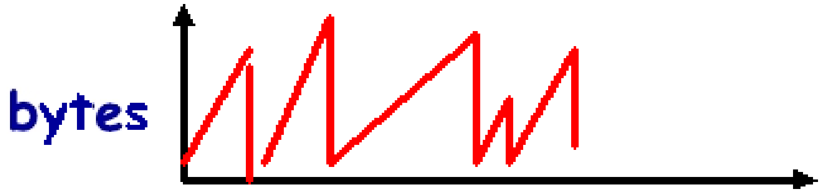
- **Worst-fit:**
  - Strategy: fight against sawdust by splitting blocks to maximize leftover size
  - In real life seems to ensure that no large blocks around
- **Next fit:**
  - Strategy: use first fit, but remember where we found the last thing and start searching from there
  - Seems like a good idea, but tends to break down entire list
- **Buddy systems:**
  - Round up allocations to power of 2 to make management faster
  - Result? Heavy internal fragmentation

Known patterns of real programs

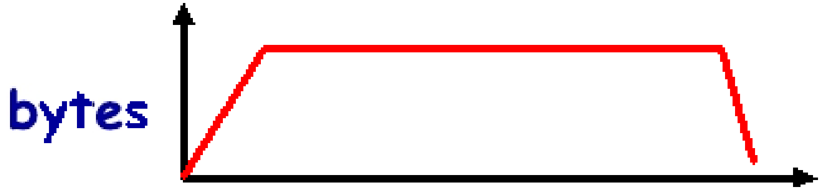
- **So far we've treated programs as black boxes.**
- **Most real programs exhibit 1 or 2 (or all 3) of the following patterns of alloc/dealloc:**
  - *Ramps*: accumulate data monotonically over time



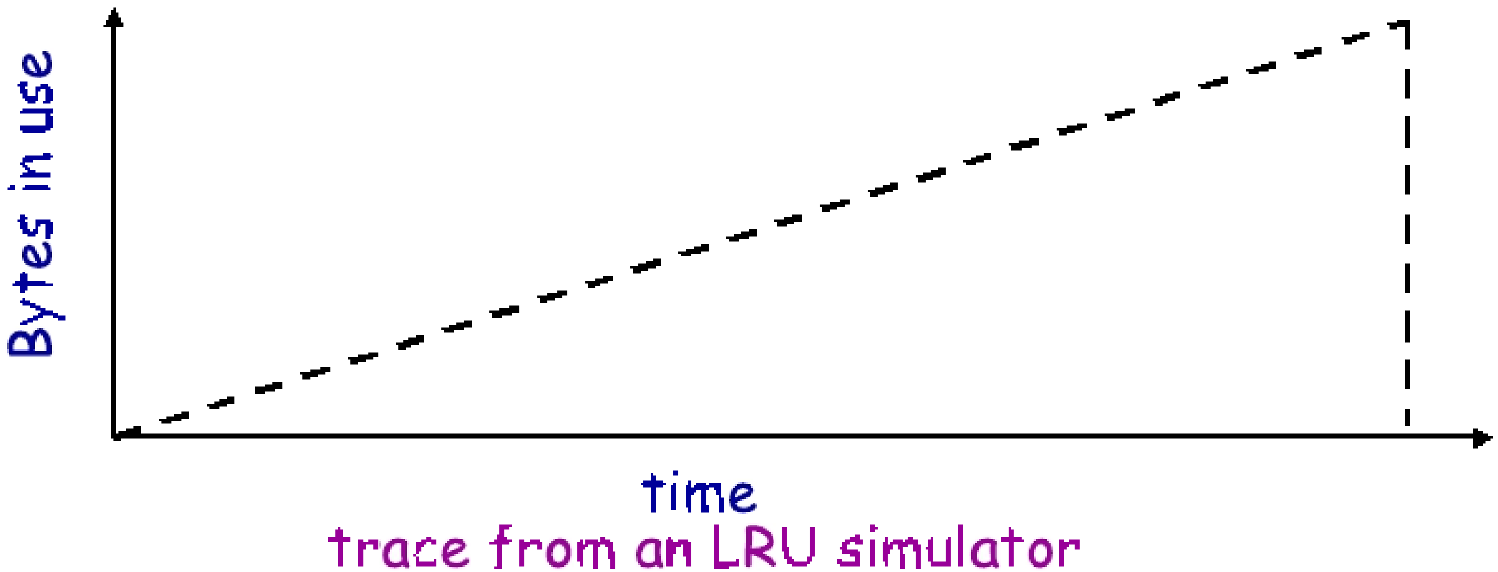
  - *Peaks*: allocate many objects, use briefly, then free all



  - *Plateaus*: allocate many objects, use for a long time

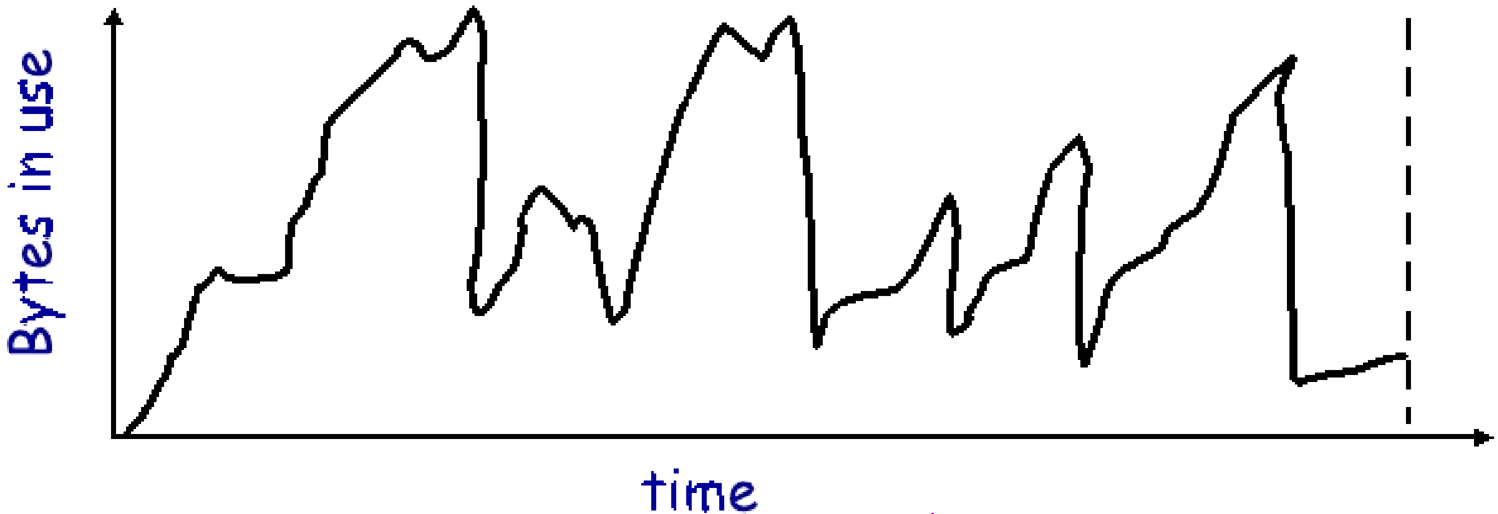


Pattern 1: ramps



- **In a practical sense: ramp = no free!**
  - Implication for fragmentation?
  - What happens if you evaluate allocator with ramp programs only?

Pattern 2: peaks

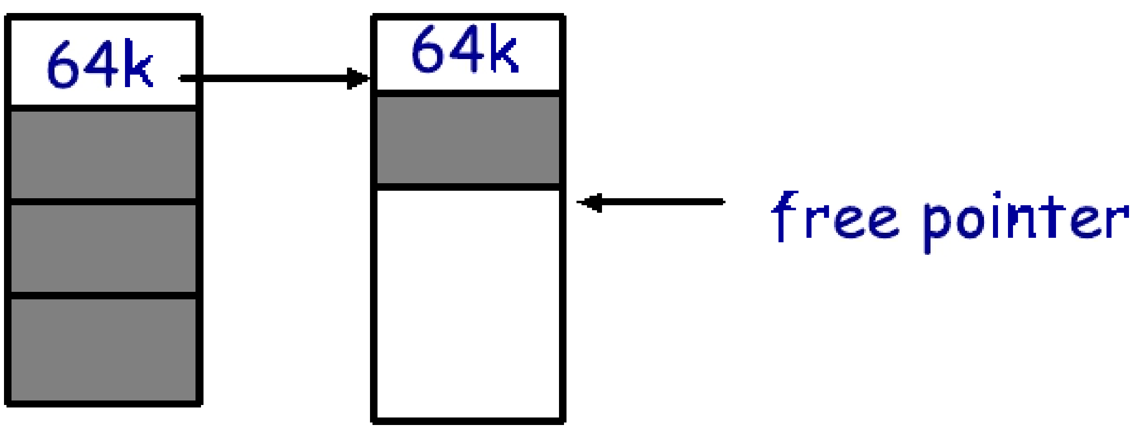


trace of gcc compiling with full optimization

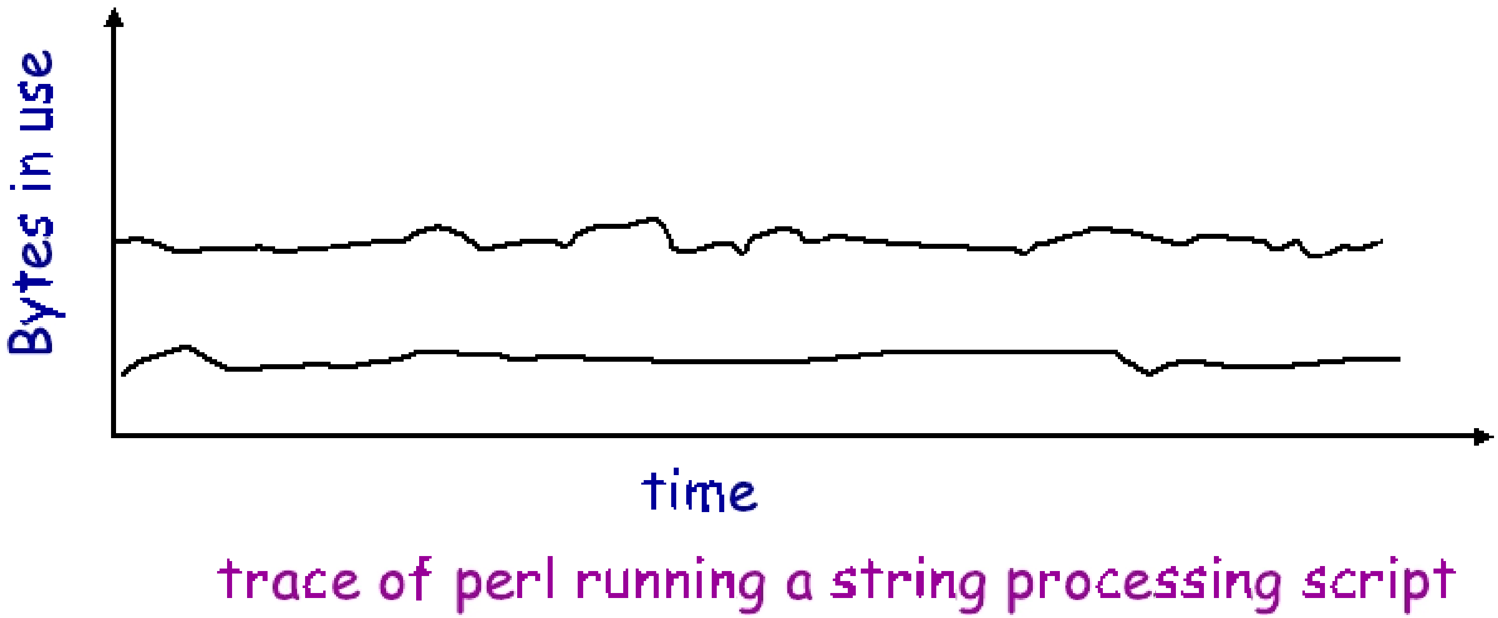
- **Peaks: allocate many objects, use briefly, then free all**
  - Fragmentation a real danger
  - What happens if peak allocated from contiguous memory?
  - Interleave peak & ramp? Interleave two different peaks?

Exploiting peaks

- **Peak phases: alloc a lot, then free everything**
  - So have new allocation interface: alloc as before, but only support free of everything
  - Called "arena allocation", "obstack" (object stack), or alloca/procedure call (by compiler people)
- **Arena = a linked list of large chunks of memory**
  - Advantages: alloc is a pointer increment, free is "free"
    No wasted space for tags or list pointers



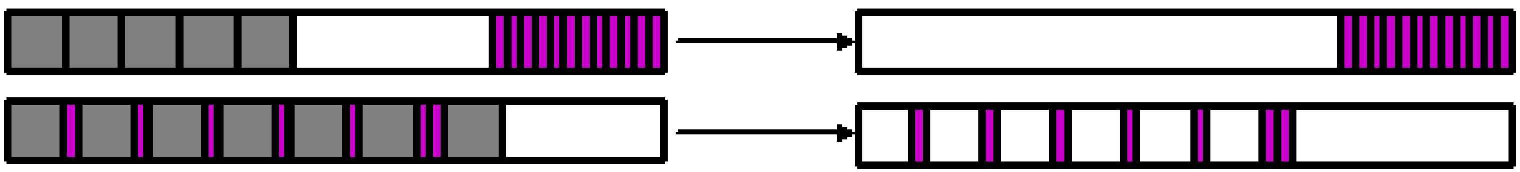
Pattern 3: Plateaus



- **Plateaus: allocate many objects, use for a long time**
  - What happens if overlap with peak or different plateau?

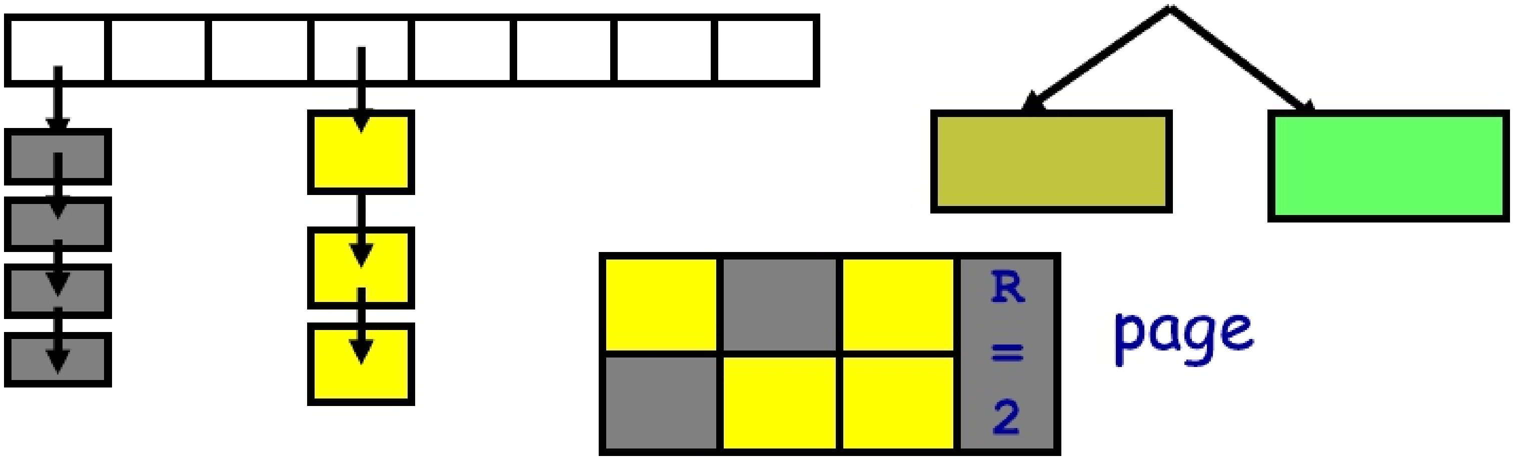
Fighting fragmentation

- **Segregation = reduced fragmentation:**
  - Allocated at same time ~ freed at same time
  - Different type ~ freed at different time



- **Implementation observations:**
  - Programs allocate small number of different sizes
  - Fragmentation at peak use more important than at low
  - Most allocations small (< 10 words)
  - Work done with allocated memory increases with size
  - Implications?

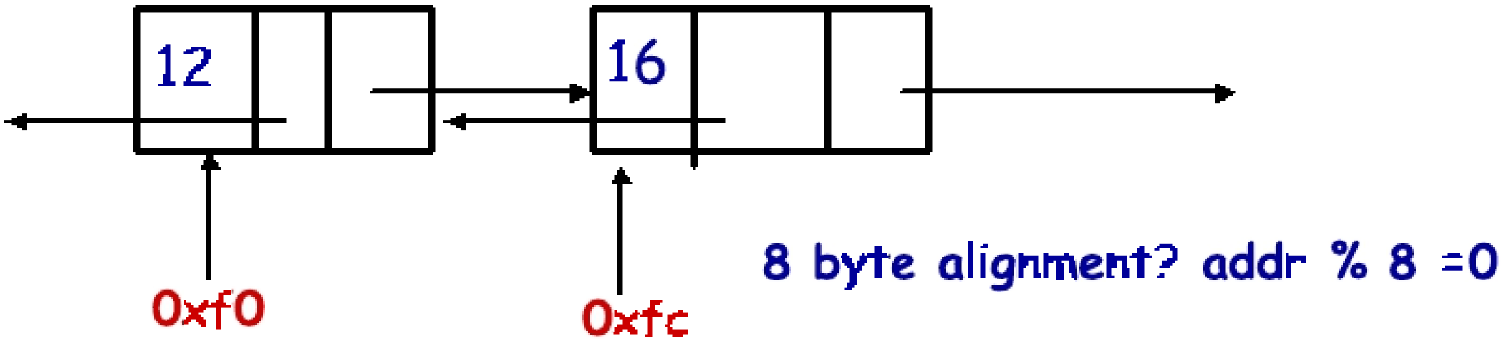
Simple, fast segregated free lists



- **Array of free lists for small sizes, tree for larger**
  - Place blocks of same size on same page
  - Have count of allocated blocks: if goes to zero, can return page
- **Pro: segregate sizes, no size tag, fast small alloc**
- **Con: worst case waste: 1 page per size even w/o free, after pessimal free waste 1 page per object**

Typical space overheads

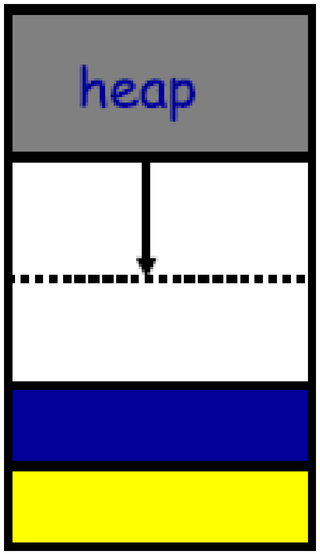
- **Free list bookkeeping + alignment determine minimum allocatable size:**
  - Store size of block
  - Pointers to next and previous freelist element



  - Machine enforced overhead: alignment. Allocator doesn't know type. Must align memory to conservative boundary
  - Minimum allocation unit? Space overhead when allocated?

Getting more space from OS

- **On Unix, can use** sbrk
  - E.g., to activate a new zero-filled page:



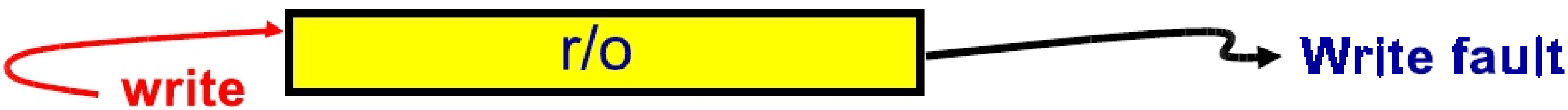

```
sbrk(4096)

/* add nbytes of valid virtual address space */
void *get_free_space(unsigned nbytes) {
   void *p;
   if(!(p = sbrk(nbytes)))
         error("virtual memory exhausted");
   return p;
}
```

- **For large allocations,** sbrk **a bad idea**
  - May want to give memory back to OS
  - Can't with sbrk unless big chunk last thing allocated
  - So allocate large chunk using mmap's MAP_ANON

Faults + resumption = power

- **Resuming after fault lets us emulate many things**
  - "every problem can be solved with layer of indirection"
- **Example: sub-page protection**
- **To protect sub-page region in paging system:**

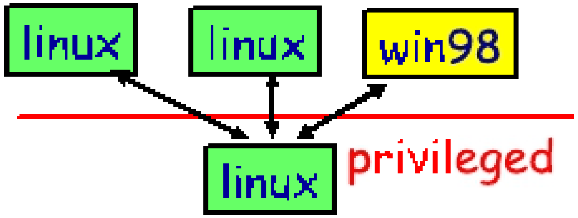
  - Set entire page to weakest permission; record in PT



  - Any access that violates perm will cause an access fault
  - Fault handler checks if page special, and if so, if access allowed. Continue or raise error, as appropriate

More fault resumption examples

- **Emulate accessed bits:**
  - Set page permissions to “invalid”.
  - On any access will get a fault: Mark as accessed
- **Avoid save/restore of FP registers**
  - Make first FP operation fault to detect usage
- **Emulate non-existent instructions:**
  - Give inst an illegal opcode; OS fault handler detects and emulates fake instruction
- **Run OS on top of another OS!**

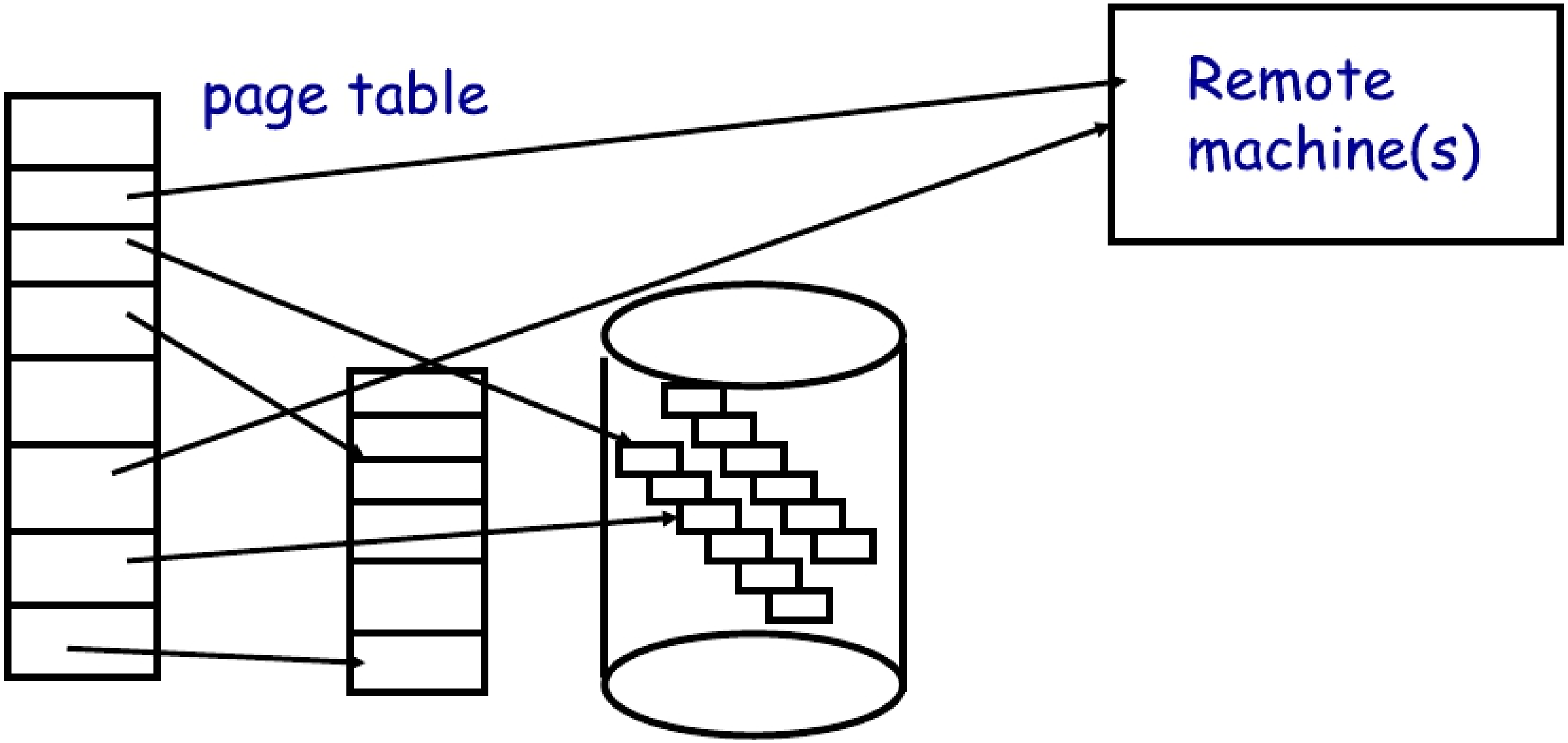


  - Slam OS into normal process
  - When does something “privileged,” real OS gets woken up with a fault.
  - If op allowed, do it, otherwise kill.
  - IBM’s VM/370. VMware (sort of)

Not just for kernels

- **User-level code can resume after faults, too**
- mprotect **– protects memory**
- sigaction **– catches signal after page fault**
  - Return from signal handler restarts faulting instruction
- **Many applications detailed by Appel & Li**
- **Example: concurrent snapshotting of process**
  - Mark all of process's memory read-only w. mprotect
  - One thread starts writing all of memory to disk
  - Other thread keeps executing
  - On fault – write that page to disk, make writable, resume

Distributed shared memory



- **Virtual memory allows us to go to memory or disk**
  - But, can use the same idea to go anywhere! Even to another computer. Page across network rather than to disk. Faster, and allows network of workstations (NOW)

Persistent stores

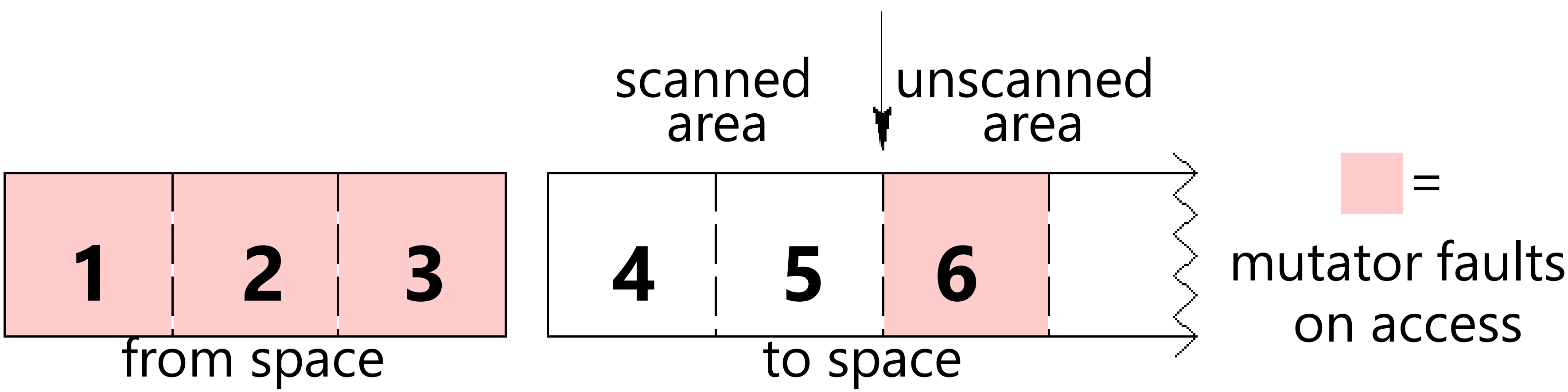
- **Idea: Objects that persist across program invocations**
  - E.g., object-oriented database; useful for CAD/CAM type apps
- **Achieve by memory-mapping a file**
- **But only write changes to file at end if commit**
  - Use dirty bits to detect which pages must be written out
  - Or emulate dirty bits with *mprotect*/*sigaction* (using write faults)
- **On 32-bit machine, store can be larger than memory**
  - But single run of program won't access > 4GB of objects
  - Keep mapping betw. 32-bit mem ptrs and 64-bit disk offsets
  - Use faults to bring in pages from disk as necessary
  - After reading page, translate pointers—known as *swizzling*

Garbage collection

- **In safe languages, run time knows about all pointers**
  - So can move an object if you change all the pointers
- **What memory locations might a program access?**
  - Any objects whose pointers are currently in registers
  - Recursively, any pointers in objects it might access
  - Anything else is *unreachable*, or *garbage*; memory can be re-used
- **Example: stop-and-copy garbage collection**
  - Memory full? Temporarily pause program, allocate new heap
  - Copy all objects pointed to by registers into new heap
    - Mark old copied objects as copied, record new location
  - Start scanning through new heap. For each pointer:
    - Copied already? Adjust pointer to new location
    - Not copied? Then copy it and adjust pointer
  - Free old heap—program will never access it—and continue

Concurrent garbage collection

- **Idea: Stop & copy, but without the stop**
  - *Mutator* thread runs program, *collector* concurrently does GC
- **When collector invoked:**
  - Protect from space & unscanned to space from mutator
  - Copy objects in registers into *to space*, resume mutator
  - All pointers in scanned *to space* point to *to space*
  - If mutator accesses unscanned area, fault, scan page, resume



Heap overflow detection

- **Many GCed languages need fast allocation**
  - E.g., in lisp, constantly allocating cons cells
  - Allocation can be as often as every 50 instructions
- **Fast allocation is just to bump a pointer**

```
char  *next_free;
char  *heap_limit;

void  *alloc  (unsigned  size)  {
  if  (next_free  +  size  >  heap_limit)          /*  1  */
     invoke_garbage_collector  ();                 /*  2  */
   char  *ret  =  next_free;
   next_free  +=  size;
   return  ret;
}
```

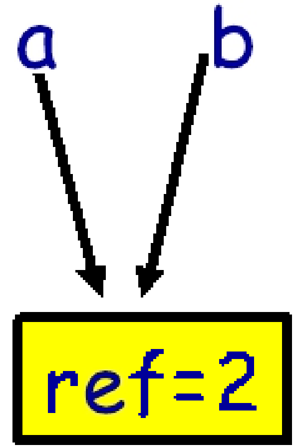
- **But would be even faster to eliminate lines 1 & 2!**

Heap overflow detection 2

- **Mark page at end of heap inaccessible**
  - mprotect (heap_limit, PAGE_SIZE, PROT_NONE);
- **Program will allocate memory beyond end of heap**
- **Program will use memory and fault**
  - Note: Depends on specifics of language
  - But many languages will touch allocated memory immediately
- **Invoke garbage collector**
  - Must now put just allocated object into new heap
- **Note: requires more than just resumption**
  - Faulting instruction must be resumed
  - But must resume with different target virtual address
  - Doable on most architectures since GC updates registers

Reference counting

- **Seemingly simpler GC scheme:**
  - Each object has “ref count” of pointers to it
  - Increment when pointer set to it
  - Decremented when pointer killed (C++ destructors handy for such “smart pointers”)

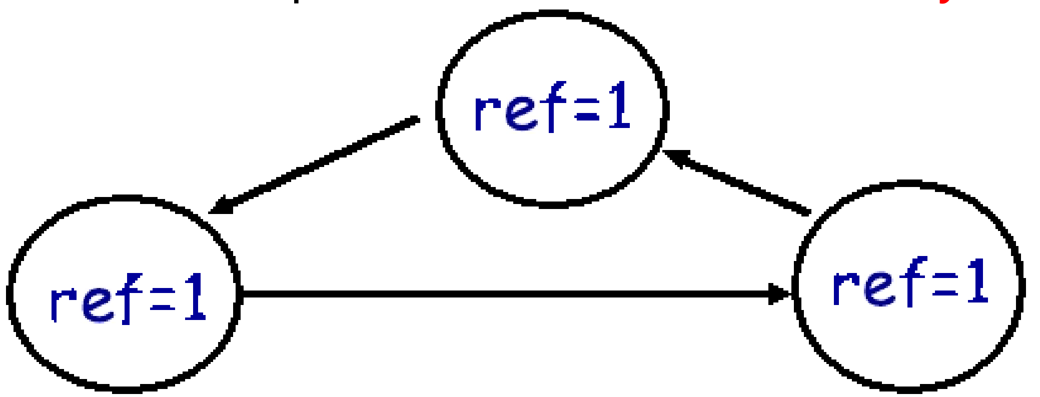



```
void foo(bar c) {
    bar a, b;
    a = c;              c->refcnt++;
    b = a;              a->refcnt++;
    a = 0;              c->refcnt--;
    return;             b->refcnt--;
}
```

  - ref count == 0? Free object
- **Works well for hierarchical data structures**
  - E.g., pages of physical memory

Reference counting pros/cons

- **Circular data structures always have ref count > 0**
  - No external pointers means lost memory



- **Can do manually w/o PL support, but error-prone**
- **Potentially more efficient than real GC**
  - No need to halt program to run collector
  - Avoids weird unpredictable latencies
- **Potentially less efficient than real GC**
  - With real GC, copying a pointer is cheap
  - With reference counting, must write ref count each time

Summary

- Read Ch. 1-8
- Processes and Threads (Ch. 4)
- Process Scheduling (Ch. 5)
- Synchronization (Ch. 6)
- Deadlock (Ch. 7)
- Memory Management (Ch. 8)
- Virtual Memory (Ch. 9)
- Project #2 – System Calls and User-Level Processes