ECE437/CS481

M07D: VIRTUAL MEMORY

CHAPTER 9.5-9.4

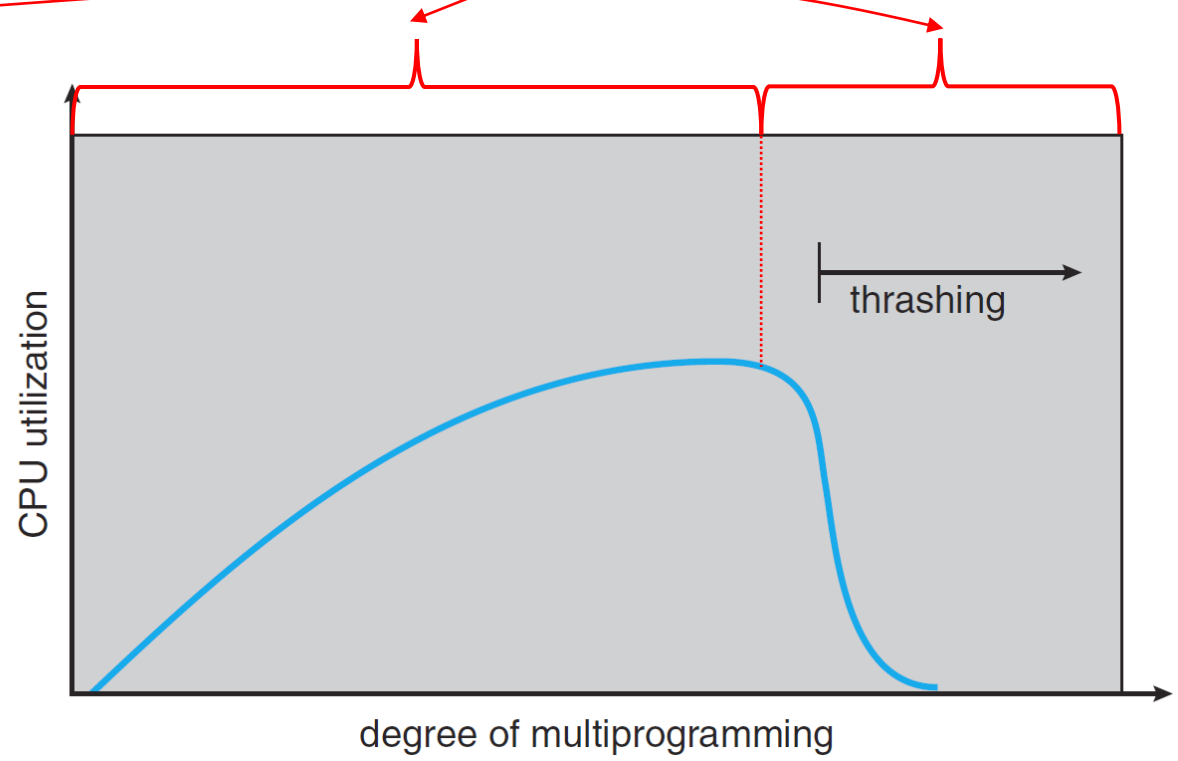
Xiang Sun

The University of New Mexico

Thrashing

- As the degree of multiprogramming increases, more processes will be brought into the main memory
  - The average CPU utilization may increase.
  - The size of memory space allocated to each process may reduce.
    - → If reduced, more pages of a process will be stored in the hard drive.
    - →The page fault rate increases.
    - →The OS spends more time on swapping in/out pages.
    - → Reduce the CPU utilization.

- Thrashing ➔ A process is busy in swapping pages in and out.

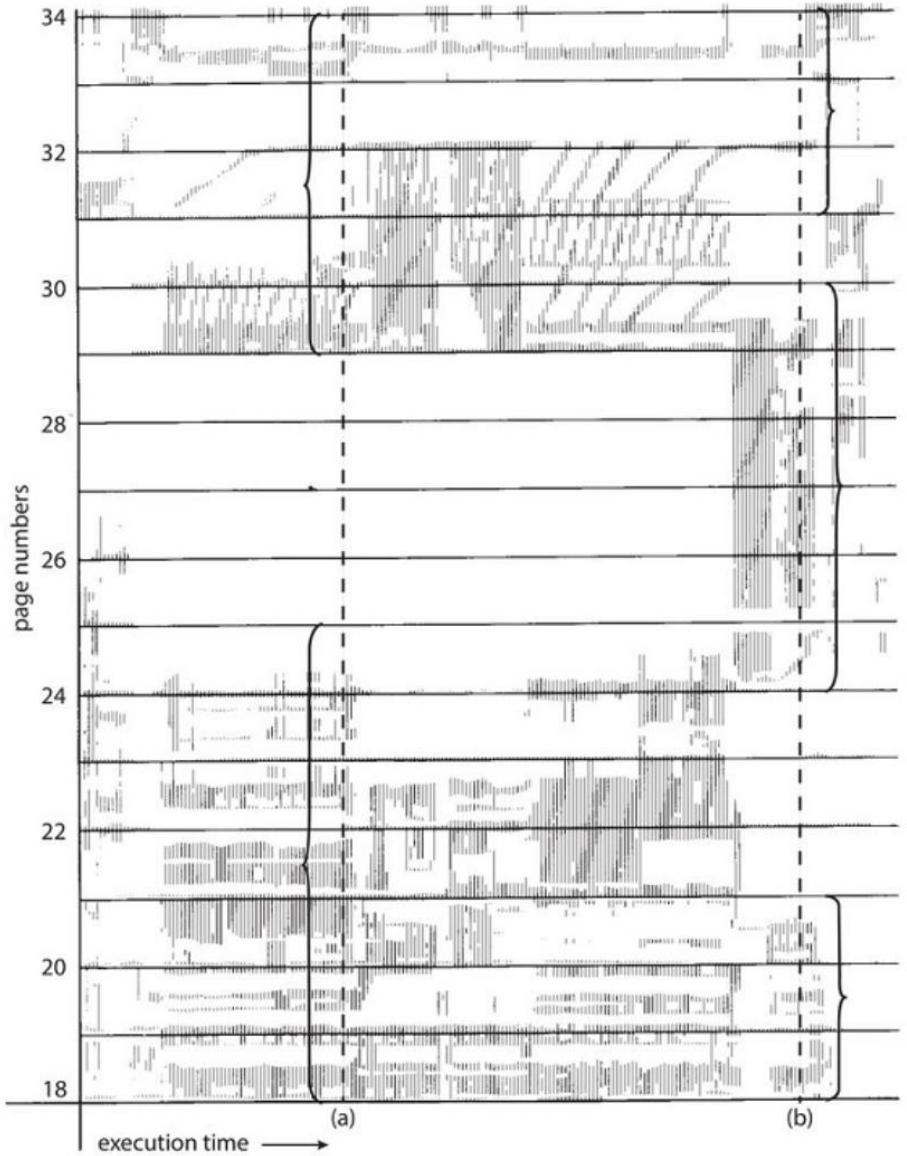


Thrashing

- How to mitigate process thrashing?
  - Basic idea: 1) prefetch the memory pages of a process that will be referenced in the near future; 2) if the memory space is not enough, suspend the process.
    - ✓ Figuring out the memory pages of a process to be referenced soon is difficult.
  - Locality feature of a process—a set of pages for a process are actively used together
  - Two types of locality
    - ✓ Temporal locality
      - A process's pages, which are referenced recently, will likely be referenced in the near future.
      - E.g., looping, counting/reduction variables, etc.
    - ✓ Spatial locality
      - One page of a process is referenced, and its nearby pages will be referenced in the near future.
      - E.g., array, sequential code.

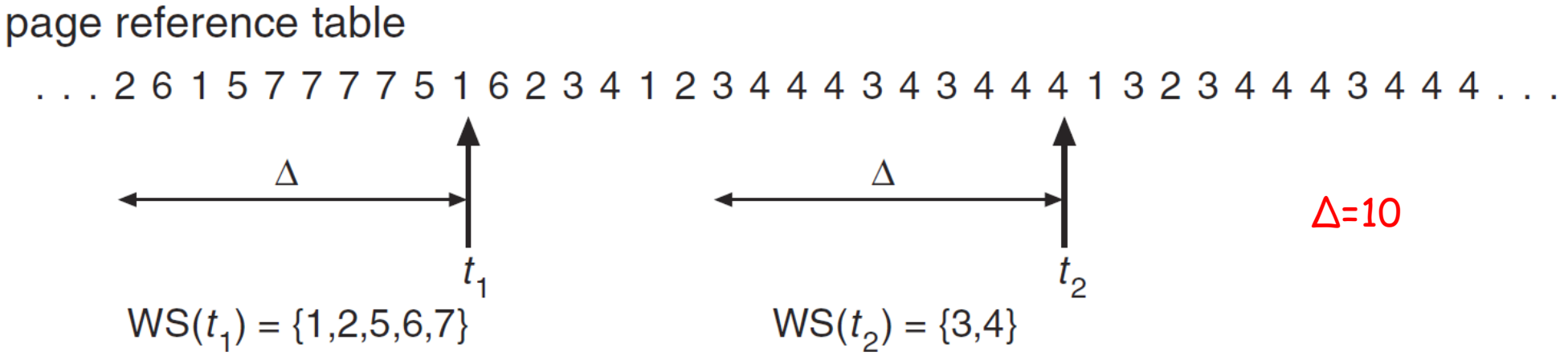
Thrashing

- Locality in a memory-reference pattern
  - In time slot (a), the spatial locality is [18,19,20,21,22,23,24,29,30,33]
  - In time slot (b), the spatial locality is [18, 19, 20, 24, 25, 26, 27, 28, 29, 31, 32, 33].
  - The locality may change over time.



Thrashing

- Based on the locality feature, the working set model is designed to figure out the need of a process in the near future.
  - One parameter, working set window Δ, is used to monitor the recent referenced pages.



  - The accuracy of the working set depends on the value of Δ.
    - ✓ If Δ is too small, it will not cover the entire locality.
    - ✓ If Δ is too large, it may overlap several localities.
    - ✓ If Δ is infinite, the working set includes all the pages during the process execution.

Thrashing

- Procedures of using working set to mitigate the process thrashing.
  - The OS monitors the working set of each running process and tries to allocate the frames to each running process, i.e., # of frames allocated to process $i = |WS_i|$, where $|WS_i|$ indicates the size of the working set for process i.
  - Check if the total number of free frames (denoted as D) is larger than the requirements, i.e., D > $\sum_i |WSi|$?
    - ✓ If yes, another process can be initiated (by allocating frames to the process).
    - ✓ If no, the OS selects a process(es) to suspend.

Thrashing

- Relationship between the working set of a process and its page fault rate.
  - As mentioned before, the locality/working set of a process may vary over time.
  - The page fault rate of a process transits between peaks and valleys over time. Here,
    - ✓ The peaks indicate the pages in working set changes, and thus the OS has to bring the pages from the hard disks into memory (demand paging)
    - ✓ The valleys indicate all the pages in the working set have swapped in the memory.