

Scheduling Algorithms

- **Assign priorities to processes** Varying complexity
- **May allocate CPU time blocks**
- **Conflicting goals**
 - High CPU utilization
 - Fast response time (reduced waiting)
 - Low overhead

First-Come, First-Served

- **Perhaps simplest algorithm**
- **Process queue**
 - Ready processes entered at end
- **CPU allocation**
 - Next process from head of queue
 - Process runs till blocked/completed

First-Come, First-Served

Grocery checkout (one lane)



User with smallest demand may wait longest

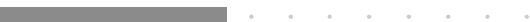
Queuing theory (L. Kleinrock)

First-Come, First-Served FCFS

P ₁ (24 ms)	P ₂ (3 ms)	P ₃ (3 ms)
------------------------	-----------------------	-----------------------

P	Wait
P ₁	0
P ₂	24
P ₃	27

Average wait = 17 ms



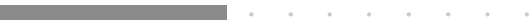
First-Come, First-Served

Different arrival order

P ₂ (3 ms)	P ₃ (3 ms)	P ₁ (24 ms)
-----------------------	-----------------------	------------------------

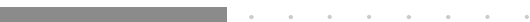
P	Wait
P ₁	6
P ₂	0
P ₃	3

Average wait = 3 ms



SJF Scheduling

- **Shortest job first**
 - Like second case of FCFS
- **Minimum average wait**
 - For a given set of processes
- **Only one problem**
 - Requires future knowledge!
 - Length of next burst



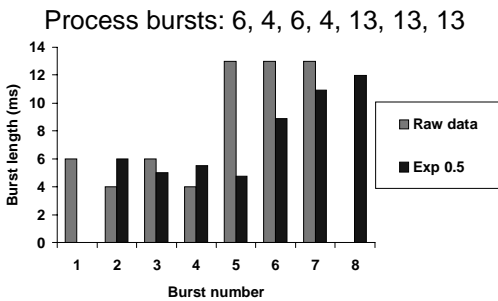
Burst Length Prediction

- **Estimate future from past**
 - Track previous process bursts
- **But what if process varies?**
 - Interactive I/O with computation
- **Derive overall estimate**
 - Average of previous history?

History Averaging

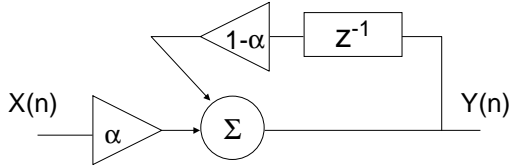
- **Moving average**
 - Average of last N samples (storage)
 - Or fewer if not enough samples
- **Exponential smoothing**
 - $\tau_{n+1} = \alpha t_n + (1-\alpha)\tau_n$
 - Implementation for binary α ?

History Averaging

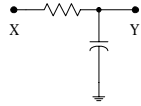


Exponential Smoothing

$$Y(n) = \alpha X(n) + (1 - \alpha)Y(n - 1)$$



Digital filtering: stability?



Priority Scheduling

- **Priority set for each process**
 - In SJF, $p \propto 1/\tau$
- **Run highest priority first**
 - Often integer priority values
 - In fixed range
 - High priority may be low integer

Priority Determination

- **System resource demands**
 - Memory, disk, maximum time limit
 - Priority varies inversely?
- **External factors**
 - Importance of process function
 - Or of "owner" of process?

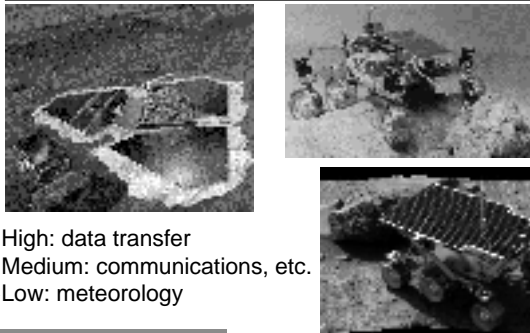
Starvation

- **Problem with priority**
 - If system is always busy
- **Some processes may never run**
 - If priority is lower than others
- **Possible solution - aging**
 - Raise priority of long-waiting processes

Priority Inversion

- **High-priority process**
 - Needs access to resource or data
- **Low-priority process**
 - Has resource or is producing data
- **So: high-priority task waits**
 - Solution(?): temporarily boost "low"
 - If "high" waits on "low's" resource

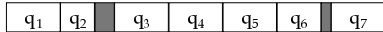
Priority Inversion on Mars



High: data transfer
 Medium: communications, etc.
 Low: meteorology

Round-Robin Scheduling

- Adds preemption to FCFS
- Each process limited in time
 - Executes for one "quantum"
 - Also known as "time slicing"
- If burst > quantum, preempt
 - Current process to end of queue



Quantum Selection

- Large quantum
 - Degenerates to FCFS (long waits?)
- Small quantum
 - Context switch overhead significant
- Adaptive
 - May vary by process

Multilevel Queues

- Process of different types
 - Interactive - I/O, response time
 - Batch - computation, throughput
- Schedule differently
 - Interactive - high priority, small quantum
 - Batch - low priority, larger quantum

Multilevel with Feedback

- **Does OS know process type?**
 - Could be specified by user
- **Categorize based on behavior**
 - Process exceeds quantum
 - Increase quantum, decrease priority
 - Process blocks early (unused quantum)
 - Decrease quantum, increase priority

Real-Time Scheduling

- **Interrupt latency**
 - Time from interrupt to service
 - Often critical in real-time OS
- **Make all processes responsive?**
 - Not all need it
- **Special facilities?**

Scheduling Overhead

- **Simple algorithms**
 - May have poor performance
- **Complex algorithms**
 - Require time/resources
 - Better to use resources for real processes?
- **Attempt to achieve balance**