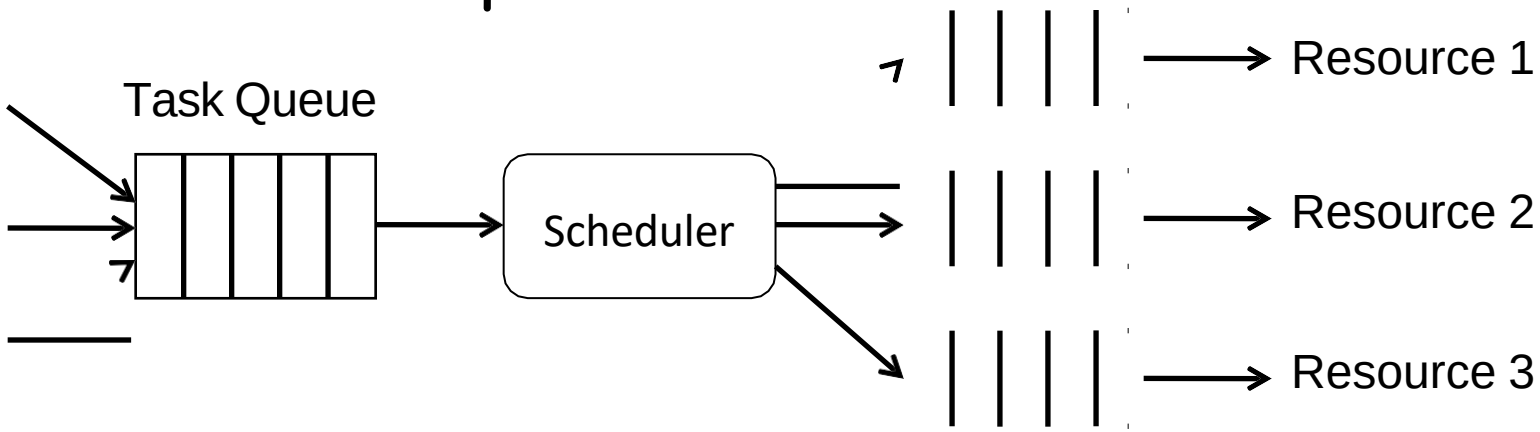
# Scheduling

- Fundamentals
  - The Scheduler Model
  - Terminology
  - Goals
- Multi-/Many-core Scheduling

# The Scheduler Model

- Scheduling: determine the order in which tasks will be performed



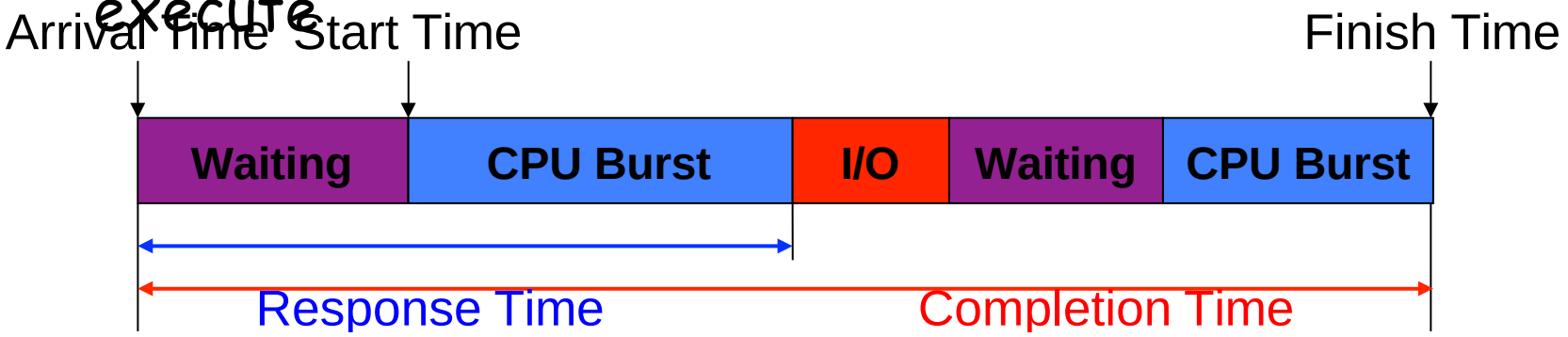

- Tasks are of various characteristics, e.g. CPU-intensive, memory-intensive
- Tasks may have different QoS requirements

- Scheduler can be distributed or centralized
- Its overhead must be low

- Resources are of many kinds
- Even for the resources of the same kind, they might also differ
- Resources are managed centrally or distributed

# Terminology

- Arrival time: time when job arrives
- Start time: time when job actually starts
- Finish time: time when job is done
- Completion time (aka Turn-around time)
  - Finish time – Arrival time
- Response time
  - Time when user sees response – Arrival time
- Execution time (aka cost): time a job needs to execute



# Goals

- Minimize response time
  - elapsed time to do an operation (or job)
  - Time perceived by the user
- Maximize throughput
  - operations (or jobs) per second
  - Minimize overhead/Efficient use of system resources
- Fairness
  - share CPU among users in some equitable way

# Difficulties

- Goals are conflicting
  - Latency vs. throughput
  - Fairness vs. low overhead
- Incomplete knowledge
  - Execution time may not be known
  - I/O device use may not be known
- Huge solution space
  - Scheduler must make decision fast

# Status

- Many policies/algorithms for CPU scheduling since 1970s with assumptions
  - One program per user
  - One thread per program
  - Programs are independent
- Open issues: unrealistic assumptions
  - Interleaving CPU/IO burst
  - Programs with dependency
  - Batch processing
  - …