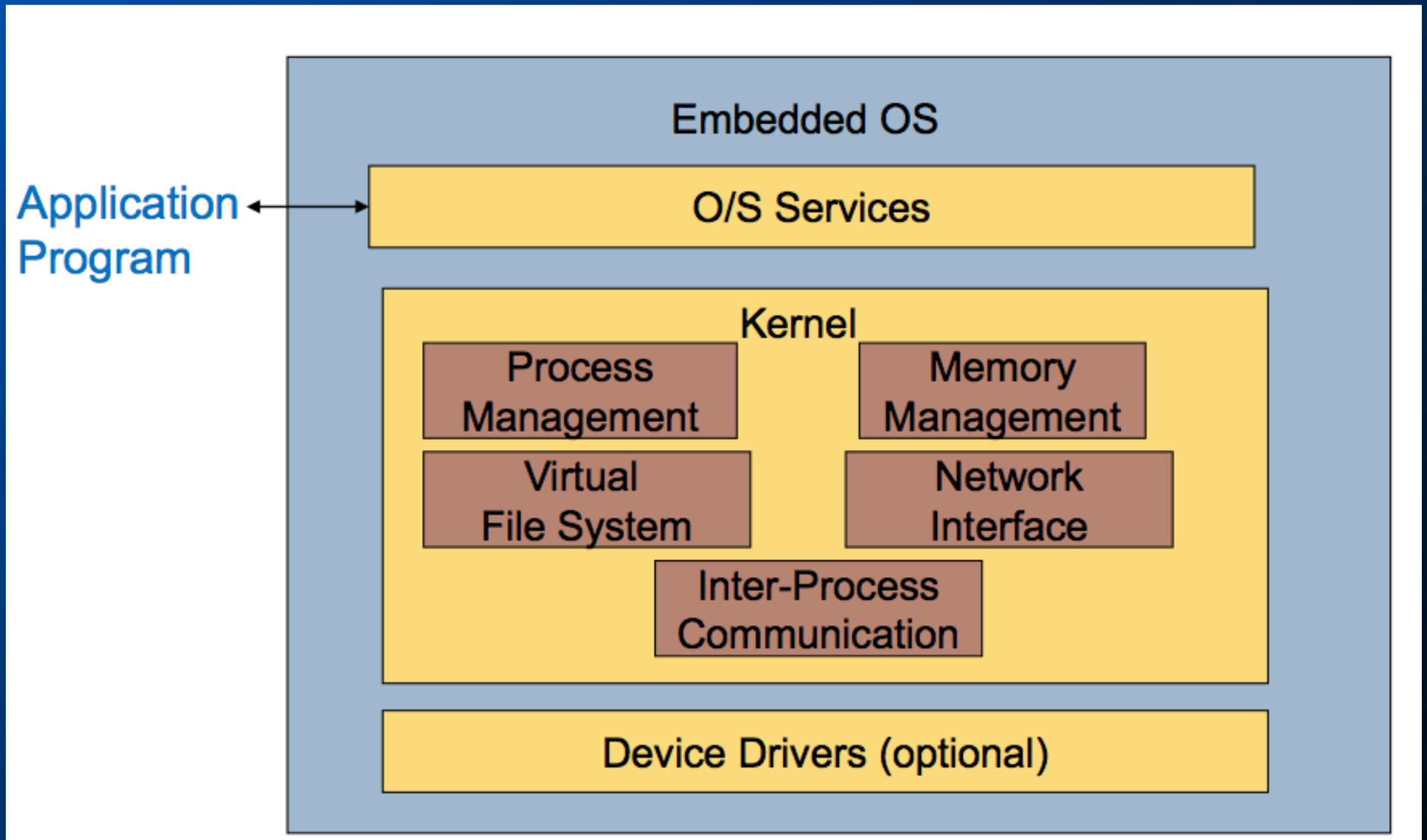
Real Time Operating Systems (RTOS)

General OS model (Linux-like)



Operating systems

- The operating system controls resources:
  - who gets the CPU;
  - when I/O takes place;
  - how much memory is allocated.
  - how processes communicate.
- The most important resource is the CPU itself.
  - CPU access controlled by the scheduler.

RTOS and GPOS

- 相似的功能
  - 多任务级别
  - 软件和硬件资源管理
  - 为应用提供基本的OS服务
  - 从软件应用抽象硬件
-

RTOS从GPOS中分离出来的不同功能

- 嵌入式应用上下文中具有更好的可靠性
- 满足应用需要的剪裁能力
- 更快的特性
- 减少内存需求
- 为实时嵌入式系统提供可剪裁的调度策略
- 支持无盘化嵌入式系统，允许从ROM或RAM上引导和运行
- 对不同硬件平台具有更好的可移植性

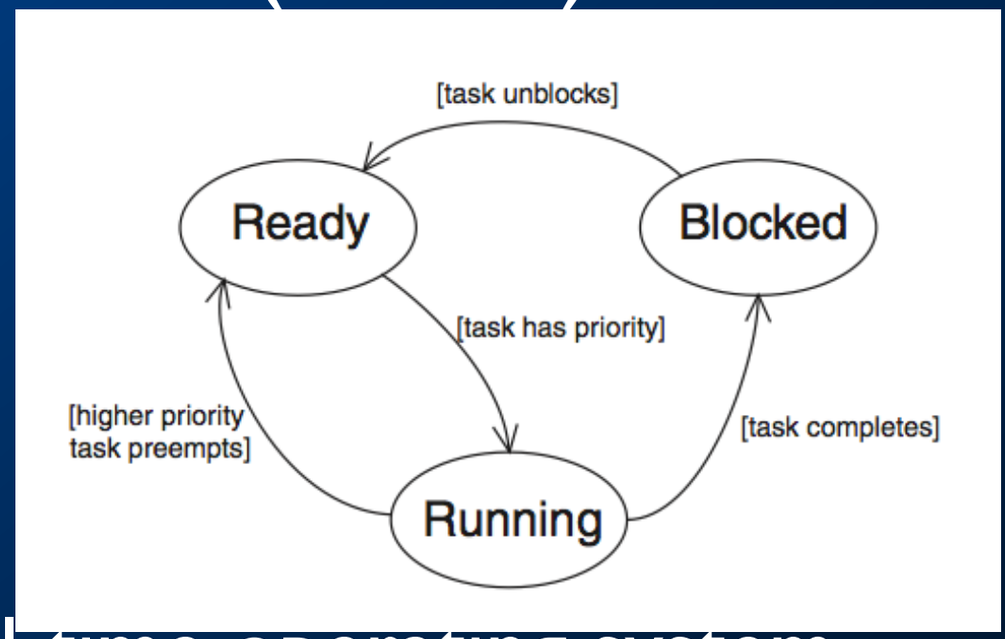
Real-time operating system (RTOS) features

- reliability
- predictability, deterministic
- performance
- compactness
- scalability

Commercial RTOSs (partial)

- AMX (KADAK)
- C Executive (JMI Software)
- RTX (CMX Systems)
- eCos (Red Hat)
- INTEGRITY (Green Hills Software)
- LynxOS (LynuxWorks)
- µC/OS-II (Micrium)
- Neutrino (QNX Software Systems)
- Nucleus (Mentor Graphics)
- POSIX (IEEE Standard)
- FreeRTOS.org
- RTOS-32 (OnTime Software)
- OS-9 (Microware)
- OSE (OSE Systems)
- pSOSystem (Wind River)
- QNX (QNX Software Systems)
- Quadros (RTXC)
- RTEMS (OAR)
- ThreadX (Express Logic)
- Linux/RT (TimeSys)
- VRTX (Mentor Graphics)
- VxWorks (Wind River)

Real-time Operating System (RTOS)



- Why use one?
  - flexibility
  - response time
- The elemental component of a real-time operating system is a task, and it's straightforward to add new tasks or delete obsolete ones because there is no main loop: The RTOS schedules when each task is to run based on its priority.
- The part of the RTOS called a scheduler keeps track of the state of each task, and decides which one should be running.

实时系统

- 一个实时系统是指计算的正确性不仅取决于程序的逻辑正确性，也取决于结果产生的时间，如果系统的时间约束条件得不到满足，将会发生系统出错。
- 实时系统（Real-time system,RTS）的正确性不仅依赖系统计算的逻辑结果，还依赖于产生这个结果的时间。

术语

- 确定性（Determinism）：如果一个系统始终会为某个已知输入产生相同的输出，则该系统是确定性的。非确定性系统的输出具有随机变化特征。
- 截止时限（Deadline）：截止时限就是必须完成某项任务的有限时间窗口。指明计算何时必须结束。
- 服务质量（QoS，Quality of Service）：指一个网络的整体性能，包括带宽、吞吐量、可用性、抖动（jitter）、延迟和错误率等因素。
- 到达时刻：任务实例所属任务的起始时刻称为该任务实例的到达时刻。
- 释放时刻：任务实例被置为就绪态的时刻称为该任务实例的释放时刻。

http://design.ros2.org/articles/realtime_background.html

- 硬实时（Hard real-time）软件系统有一组严格的截止时间，且错过一个截止时间就会认为系统失败。硬实时系统的示例包括：飞机传感器和自动驾驶系统、航天器和行星探测器。
- 软实时（Soft real-time）系统会试图满足截止时间要求，但如果错过了某个截止时间也不会认为系统失败。但是，在这样一个事件中，软实时系统可能会降低其服务质量以改进其响应能力。软实时系统的示例包括：用于娱乐的音频和视频传输软件（延迟是不可取的，但不是灾难性的）。
- 准实时（Firm real-time）系统会将截止时间之后交付的信息/计算视为无效。与软实时系统一样，准实时系统在错过某个截止时间后不会认为系统失败，并且如果错过了某个截止时间，准实时系统可能会降低服务质量（QoS）。准实时系统的示例包括：财务预测系统、机器人装配线。

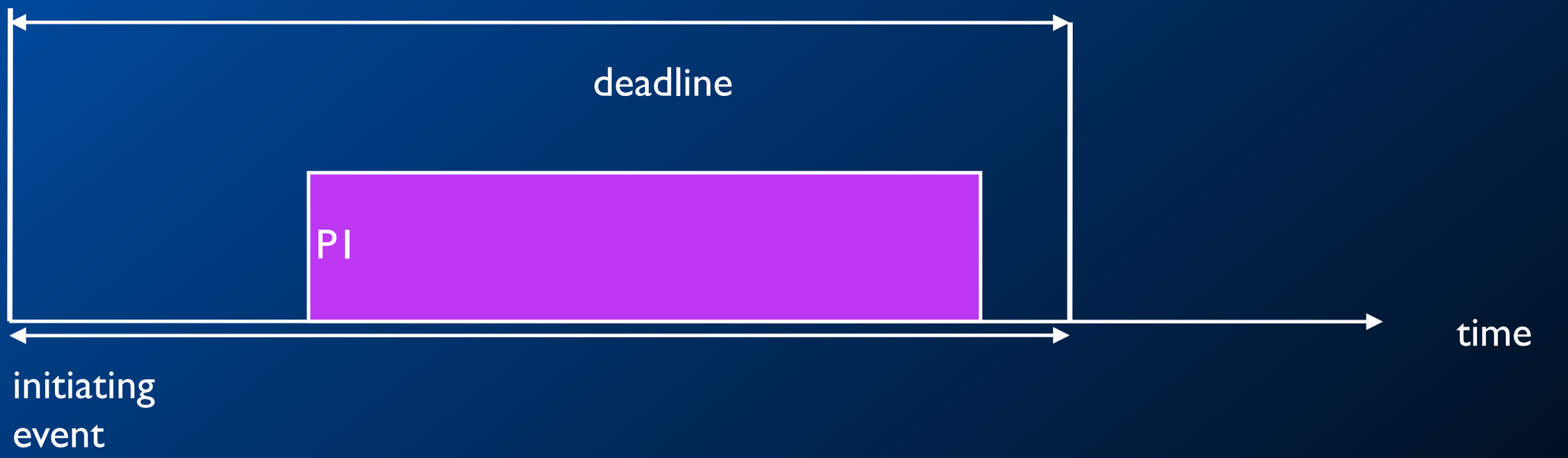
任务的周期

- 周期任务：周期任务是指按一定周期达到并请求运行，每次请求称为任务的一个任务实例，任务实例所属任务的起始时刻称为该任务实例的到达时刻，任务实例被置为就绪态的时刻称为该任务实例的释放时刻。
- 偶发任务：在偶发任务中，虽然其任务实例的到达时刻不是严格周期的，但相邻任务实例到达时刻的时间间隔一定大于等于某个最小值，即偶发任务的各任务实例按照不高于某个值的速率到达。因此在实际应用中，偶发任务经常被当作周期任务进行处理，其周期为相邻任务实例到达时刻的最小时间时隔。
- 非周期任务：非周期任务是指随机到达系统的任务。

实时环境的示例

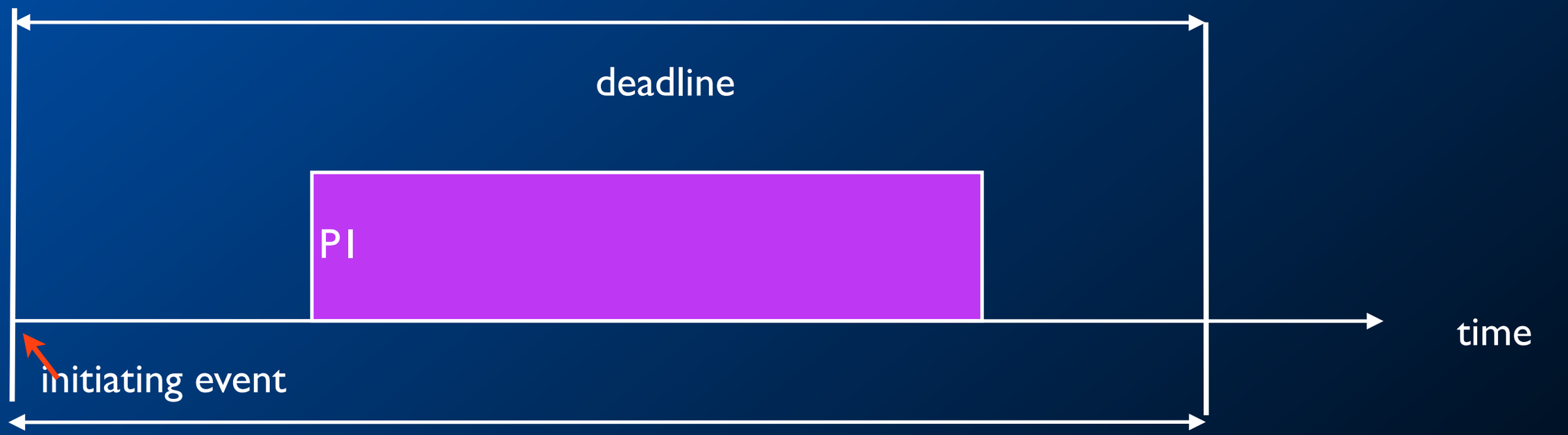
- 实时计算机系统同时需要实时运行的操作系统和提供确定性执行的用户代码。非实时操作系统上的确定性用户代码和实时操作系统上的非确定性用户代码都不会产生实时性能。
- 实时环境的一些示例包括：
  - RT_PREEMPT Linux内核补丁，该内核补丁会将Linux的调度器修改为完全可抢占。
  - Xenomai，一个符合POSIX标准的协同内核（或管理程序），是一个可以提供与Linux内核协作的实时内核。Linux内核会被视为实时内核调度器的空闲任务（最低优先级任务）。
  - RTAI，一个协同内核（co-kernel）的替代解决方案。
  - QNX Neutrino，一个符合POSIX标准、适用于关键任务系统的实时操作系统。

Release times and deadlines



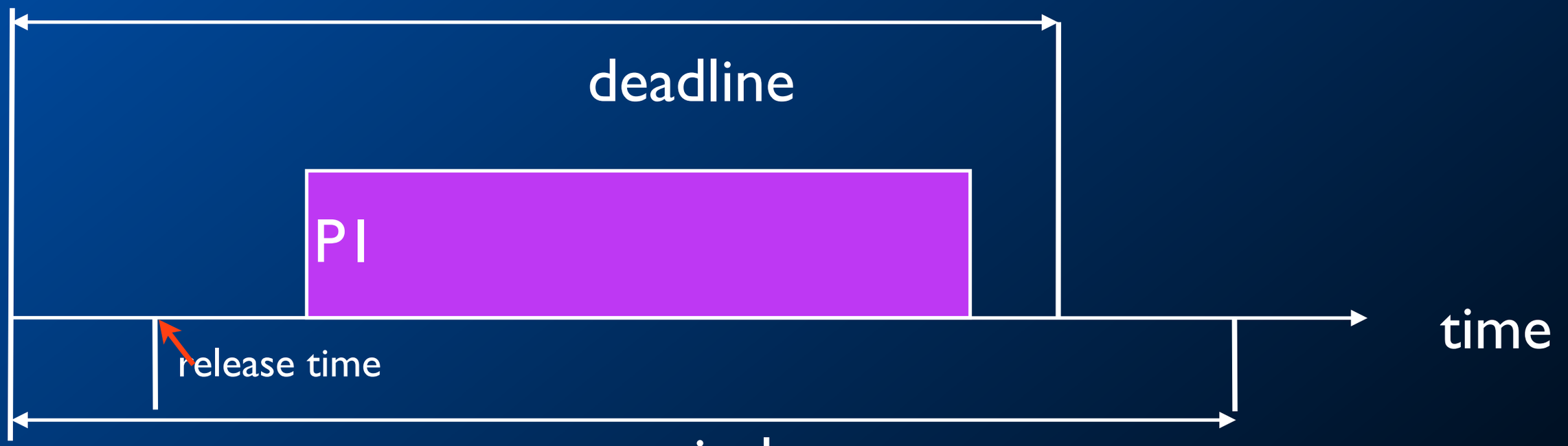
aperiodic process

Release times and deadlines



periodic process initiated
at start of period

Release times and deadlines



periodic process initiated by event

Rate requirements on processes

- Period: interval between process activations.
- Rate: reciprocal of period.
- Initiation rate may be higher than period---several copies of process run at once.

Timing violations

- What happens if a process doesn't finish by its deadline?
  - Hard deadline: system fails if missed.
  - Soft deadline: user may notice, but system doesn't necessarily fail.

Example: Space Shuttle software error

- Space Shuttle's first launch was delayed by a software timing error:
  - Primary control system PASS and backup system BFS.
  - BFS failed to synchronize with PASS.
  - Change to one routine added delay that threw off start time calculation.
  - 1 in 67 chance of timing problem.

Process execution characteristics

- Process execution time Ti.
  - Execution time in absence of preemption.
  - Possible time units: seconds, clock cycles.
  - Worst-case, best-case execution time may be useful in some cases.
- Sources of variation:
  - Data dependencies.
  - Memory system.
  - CPU pipeline.

Utilization

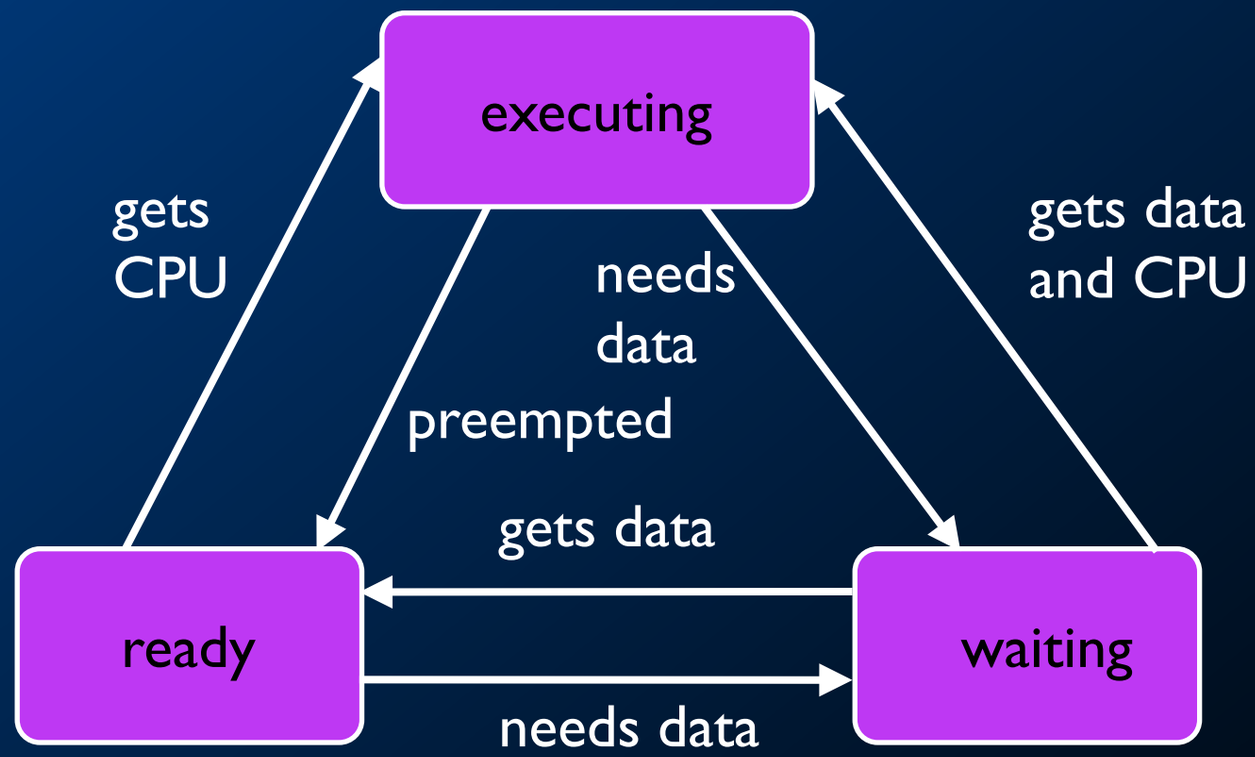
- CPU utilization:
  - Fraction of the CPU that is doing useful work.
  - Often calculated assuming no scheduling overhead.
- Utilization:
  - U = (CPU time for useful work)/ (total available CPU time)

    $= [\Sigma_{t1 \le t \le t2} T(t)] / [t2 - t1]$

    $= T/t$

State of a process

- A process can be in one of three states:
  - executing on the CPU;
  - ready to run;
  - waiting for data.



The scheduling problem

- Can we meet all deadlines?
  - Must be able to meet deadlines in all cases.
- How much CPU horsepower do we need to meet our deadlines?

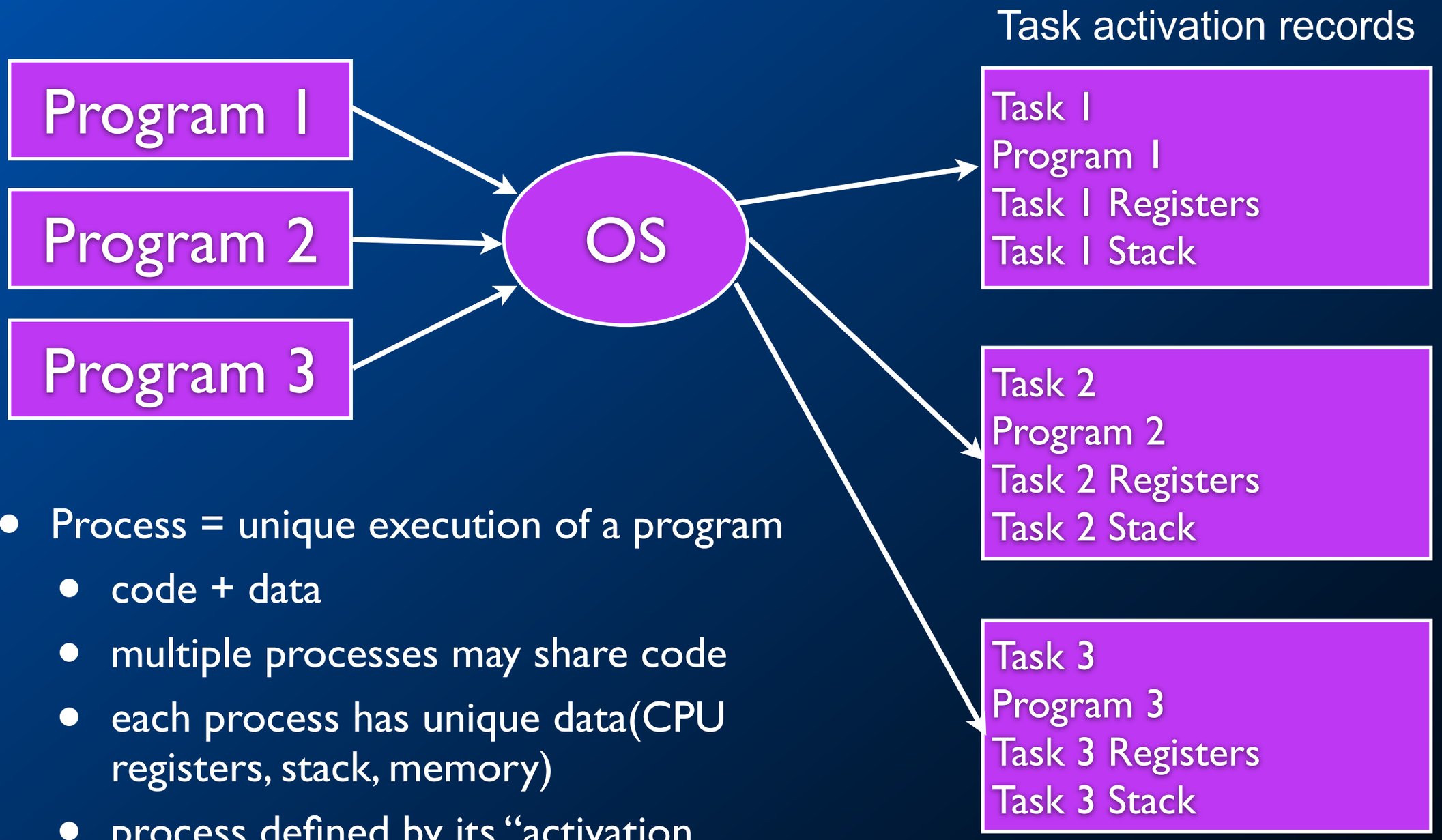
Embedded vs. general-purpose scheduling

- Workstations try to avoid starving processes of CPU access.
  - Fairness = access to CPU.
- Embedded systems must meet deadlines.
  - Low-priority processes may not run for a long time.

OS process management

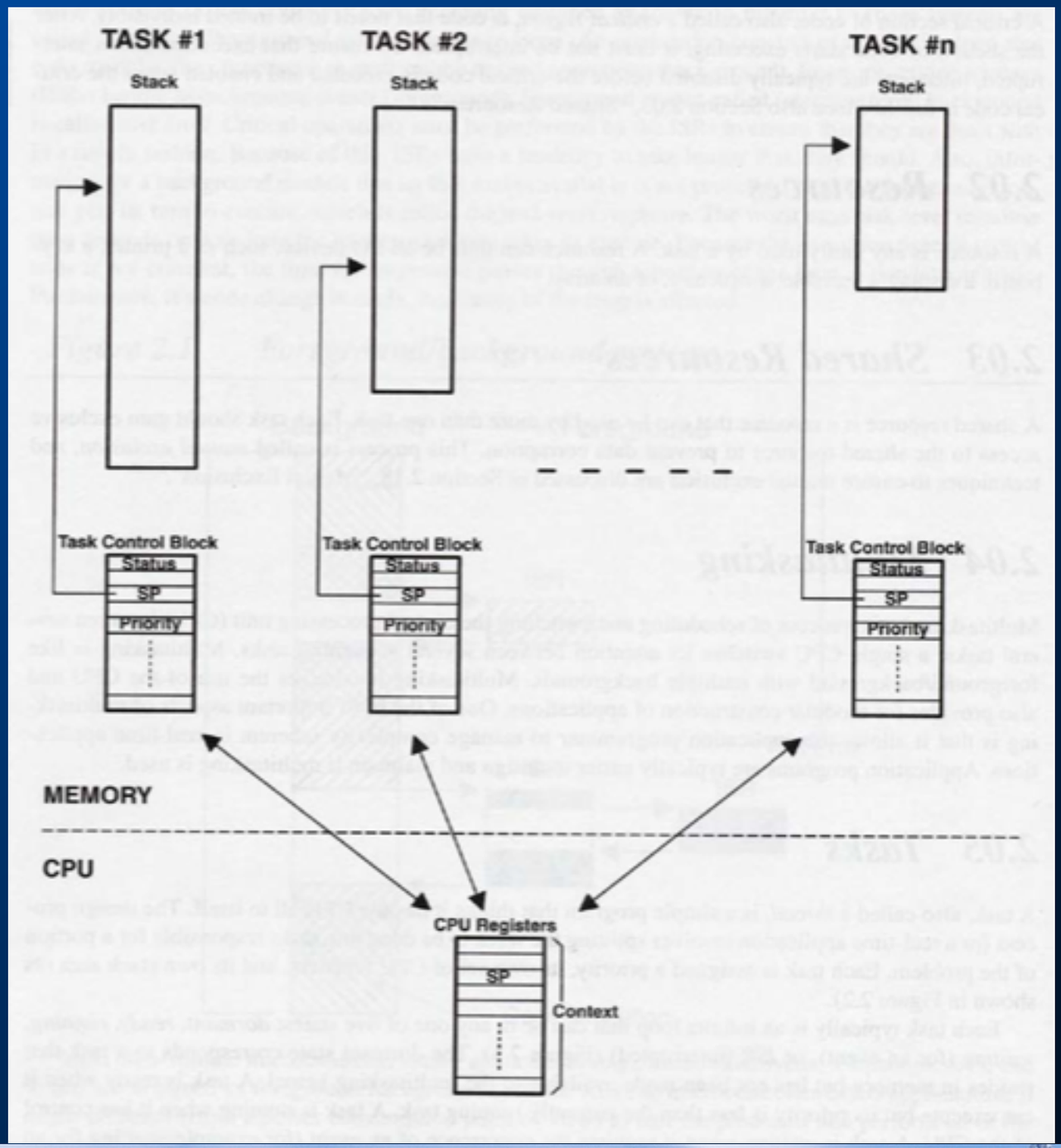
- OS needs to keep track of:
  - process priorities;
  - scheduling state;
  - process activation record.
- Processes may be created:
  - statically before system starts;
  - dynamically during execution.
    - Example: incoming telephone call

Multitasking OS

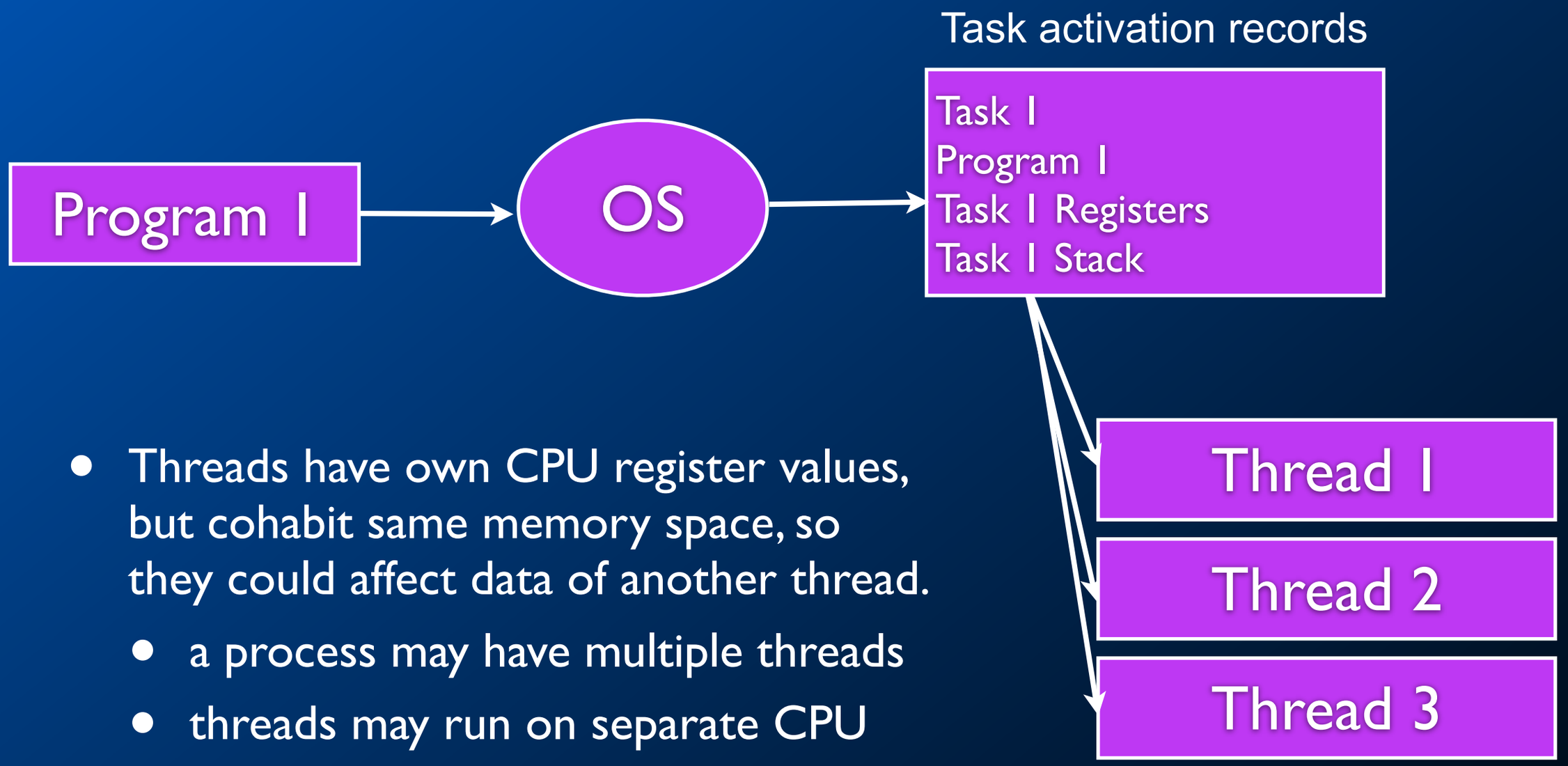


- Process = unique execution of a program
  - code + data
  - multiple processes may share code
  - each process has unique data(CPU registers, stack, memory)
  - process defined by its “activation record”

Multitasking OS



Process threads (lightweight processes)



- Threads have own CPU register values, but cohabit same memory space, so they could affect data of another thread.
  - a process may have multiple threads
  - threads may run on separate CPU cores

Typical process/task activation records (task control blocks)

- Task ID
- Task state (running, ready, blocked)
- Task priority
- Task starting address
- Task stack
- Task CPU registers
- Task data pointer
- Task time (ticks)

When can a new thread be dispatched?

- Under non-preemptive scheduling:
  - When the current thread completes.
- Under Preemptive scheduling:
  - Upon a timer interrupt
  - Upon an I/O interrupt (possibly)
  - When a new thread is created, or one completes.
  - When the current thread blocks on or releases a mutex
  - When the current thread blocks on a semaphore
  - When a semaphore state is changed
  - When the current thread makes any OS call
    - file system access
    - network access
    - ...

The Focus Today:
How to decide which thread to schedule?

Considerations:

- Preemptive vs. non-preemptive scheduling
- Periodic vs. aperiodic tasks
- Fixed priority vs. dynamic priority
- Priority inversion anomalies
- Other scheduling anomalies

Metrics

- How do we evaluate a scheduling policy?
  - Ability to satisfy all deadlines.
  - CPU utilization---percentage of time devoted to useful work.
  - Scheduling overhead---time required to make scheduling decision.
  - lateness.
  - total completion time.

Preemptive Scheduling

Assume all threads have priorities

- either statically assigned (constant for the duration of the thread)
- or dynamically assigned (can vary).

Assume further that the kernel keeps track of which threads are "enabled" (able to execute, e.g. not blocked waiting for a semaphore or a mutex or for a time to expire).

Preemptive scheduling:

- At any instant, the enabled thread with the highest priority is executing.
- Whenever any thread changes priority or enabled status, the kernel can dispatch a new thread.

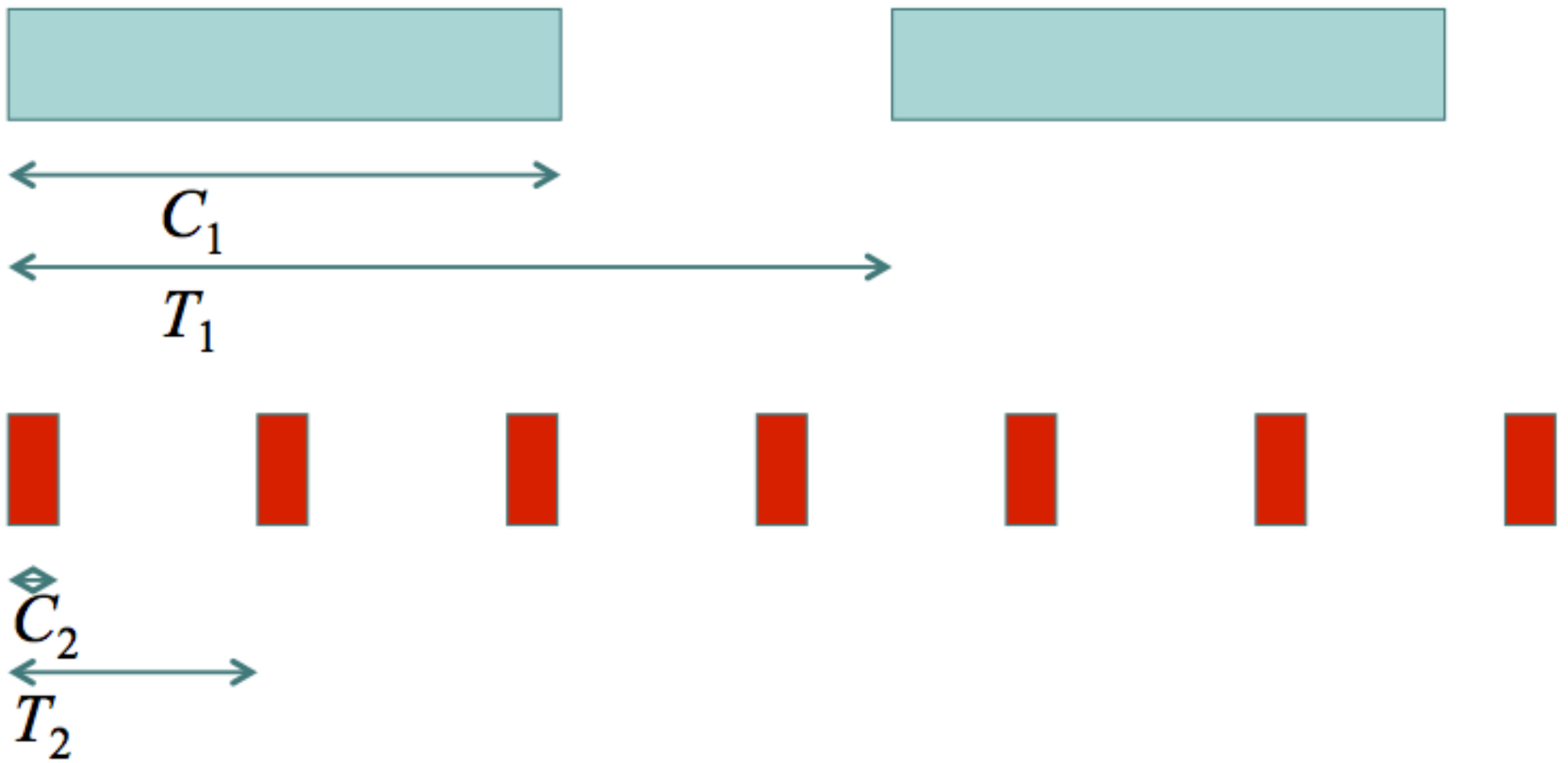
Rate Monotonic Scheduling

- Assume n tasks invoked periodically with:
  - periods T1, ... ,Tn (impose real-time constraints)
  - All tasks are independent.
  - worst-case execution times (WCET) C1, ... ,Cn
    - assumes no mutexes, semaphores, or blocking I/O
  - no precedence constraints
    - fixed priorities
    - The time required for context switching is negligible
    - preemptive scheduling
- Theorem: If any priority assignment yields a feasible schedule, then priorities ordered by period (smallest period has the highest priority) also yields a feasible schedule.
- RMS is optimal in the sense of feasibility.
- Liu and Leland, "Scheduling algorithms for multiprogramming in a hard-real-time environment," J. ACM, 20(1), 1973.

Feasibility for RMS

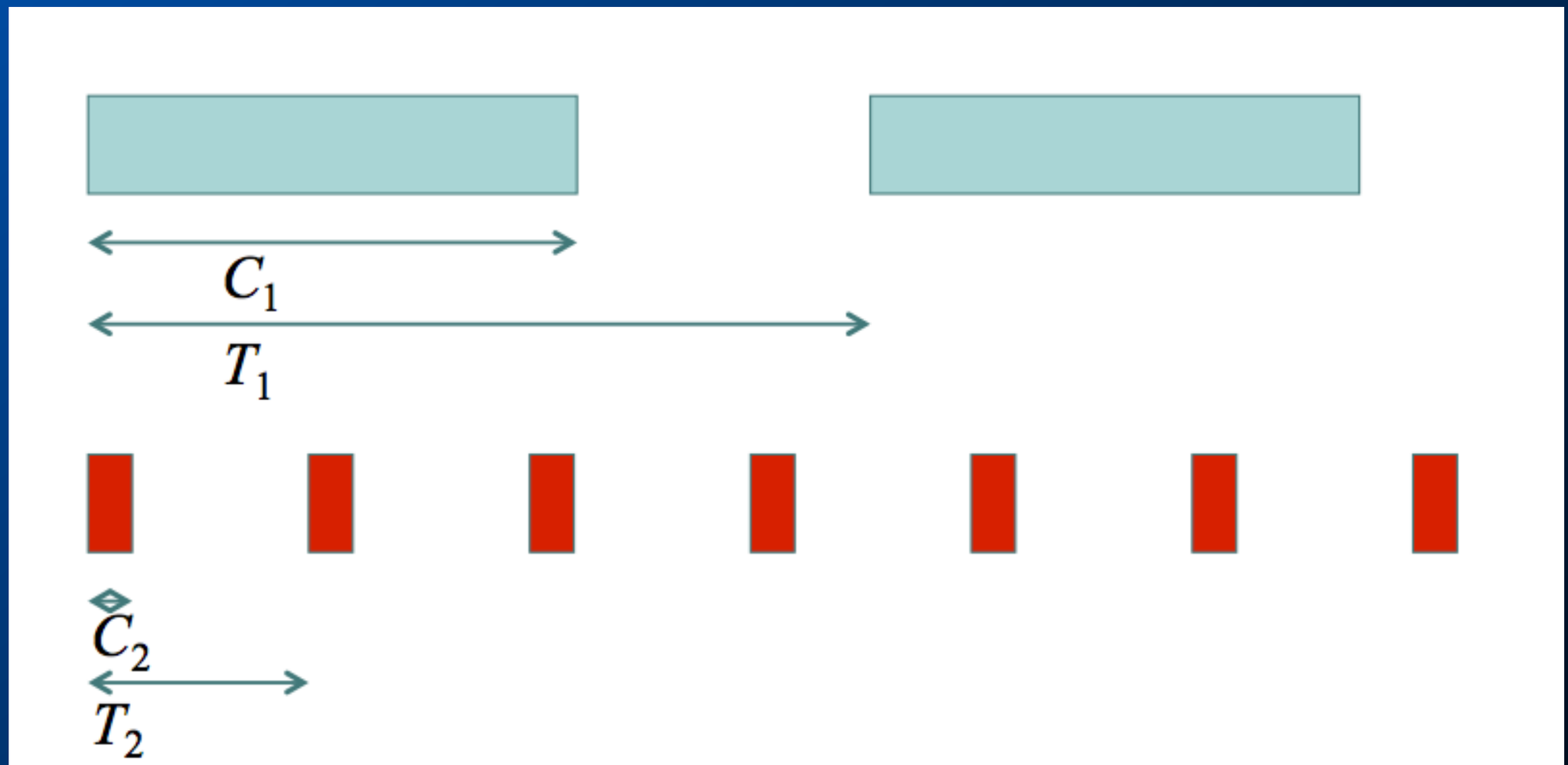
- Feasibility is defined for RMS to mean that every task executes to completion once within its designated period.

Showing Optimality of RMS:
Consider two tasks with different periods



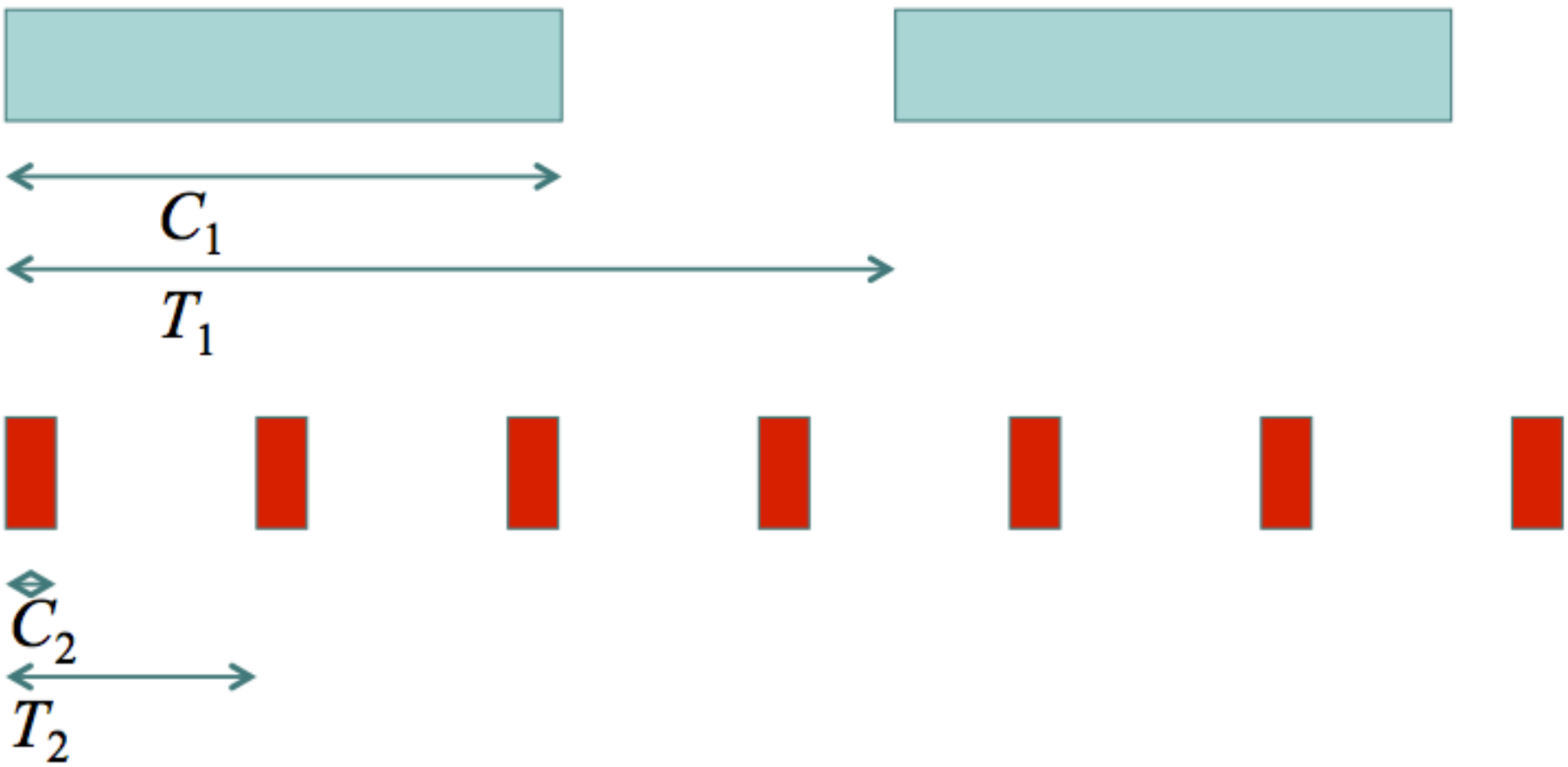
Is a non-preemptive schedule feasible?

Showing Optimality of RMS: Consider two tasks with different periods



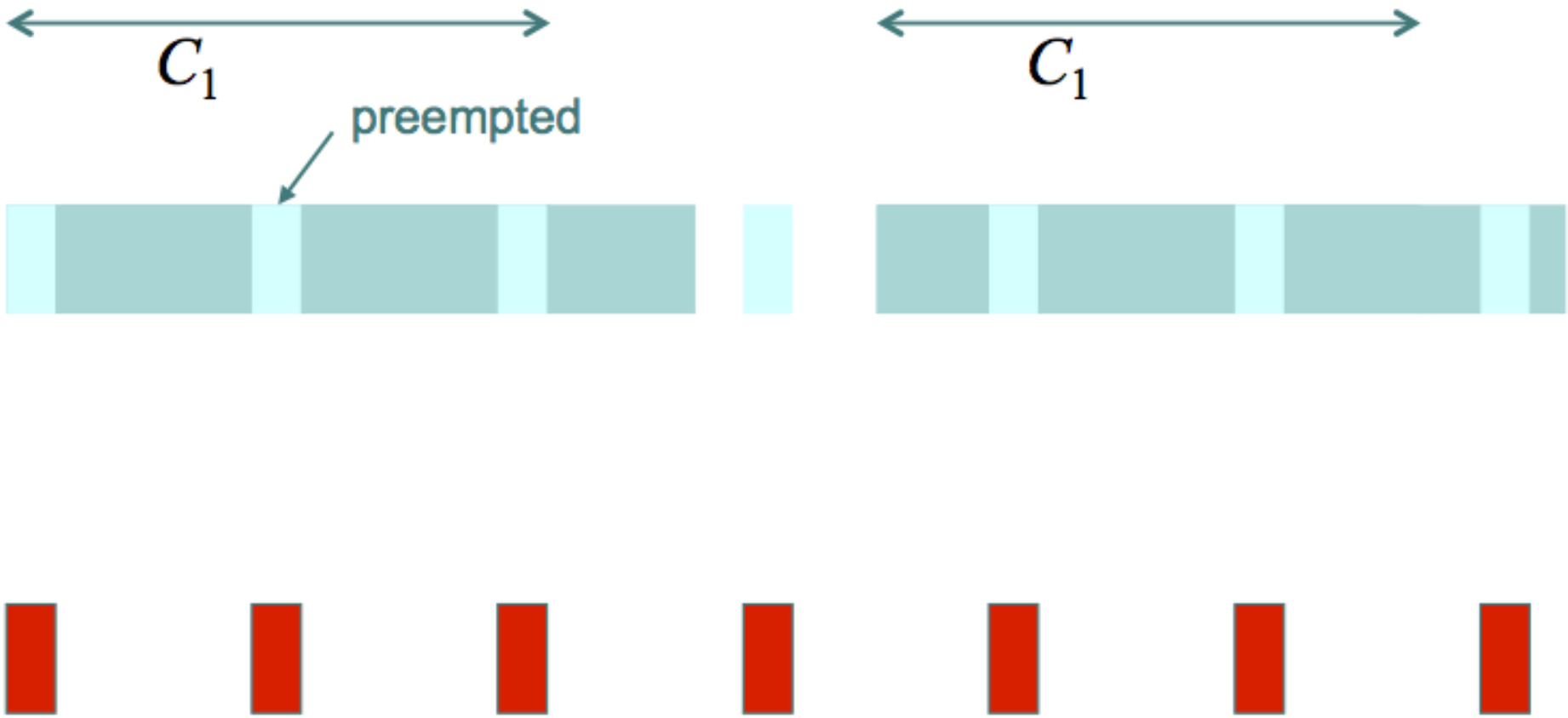
Non-preemptive schedule is not feasible. Some instance of the Red Task (2) will not finish within its period if we do non-preemptive scheduling.

Showing Optimality of RMS: Consider two tasks with different periods



What if we had a preemptive scheduling with higher priority for red task?

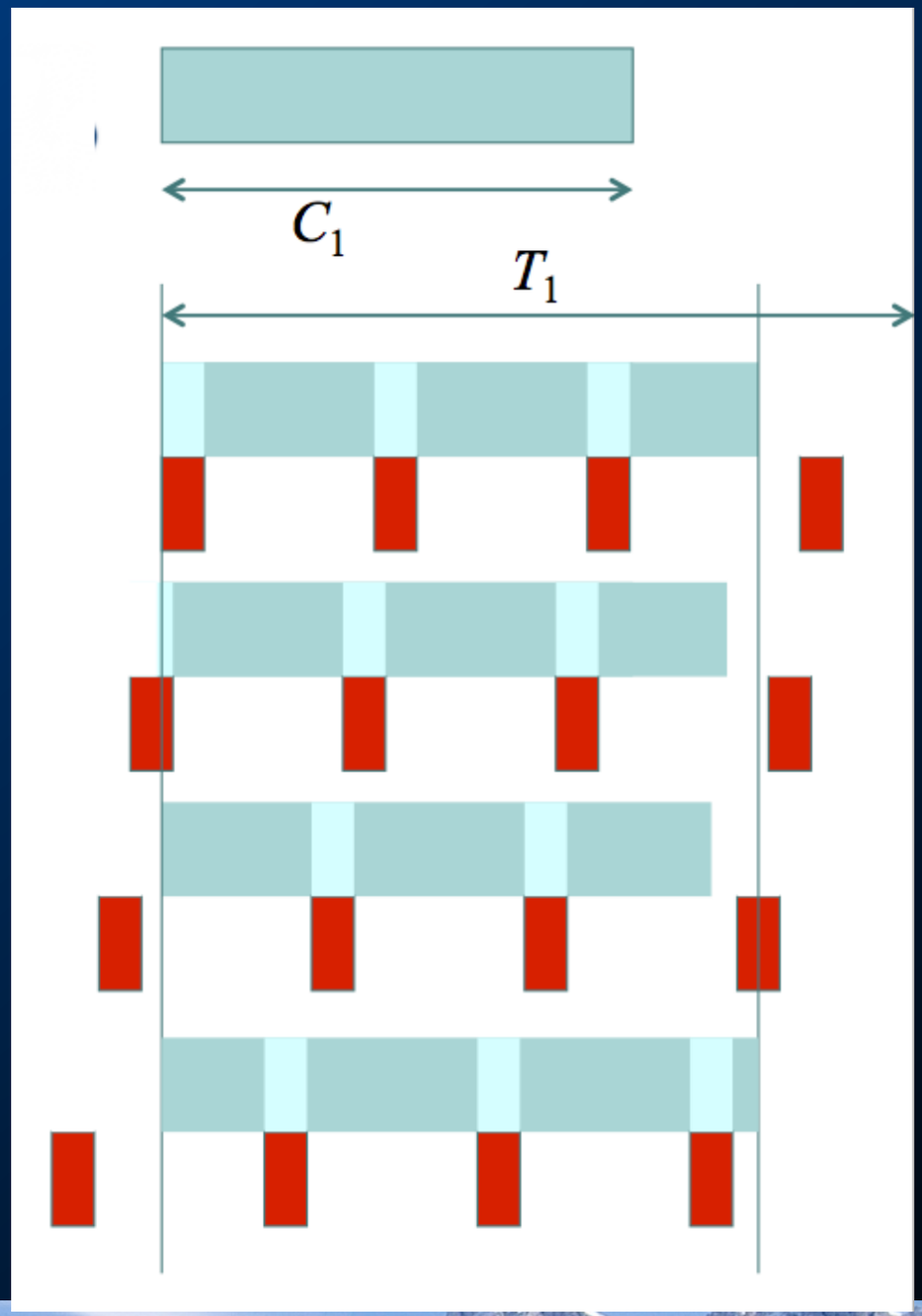
Showing Optimality of RMS: Consider two tasks with different periods



Preemptive schedule with the red task having higher priority is feasible.
Note that preemption of the blue task extends its completion time.

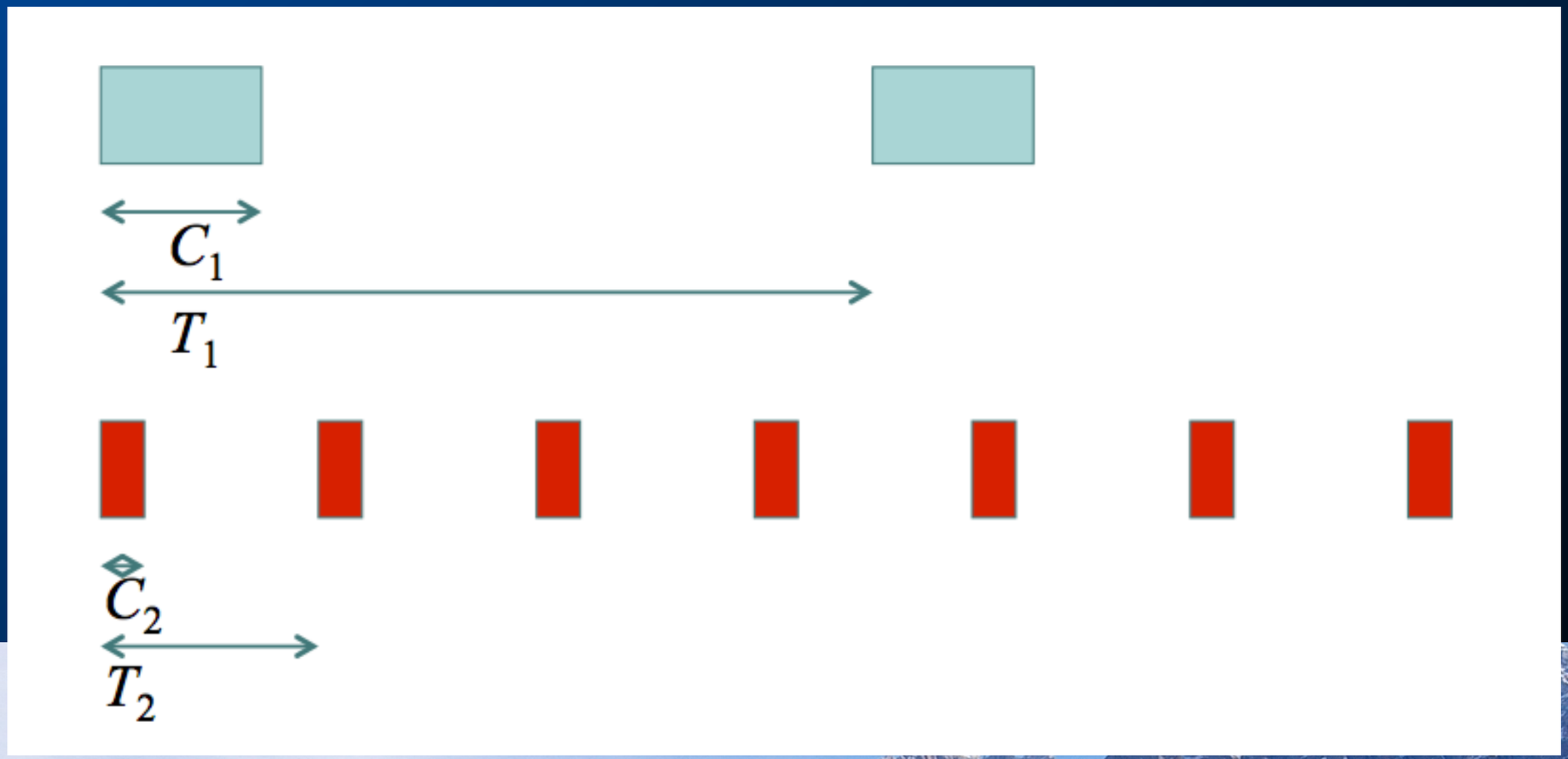
Showing Optimality of RMS: Alignment of tasks

- Completion time of the lower priority task is worst when its starting phase matches that of higher priority tasks.
- Thus, when checking schedule feasibility, it is sufficient to consider only the worst case: All tasks start their cycles at the same time.



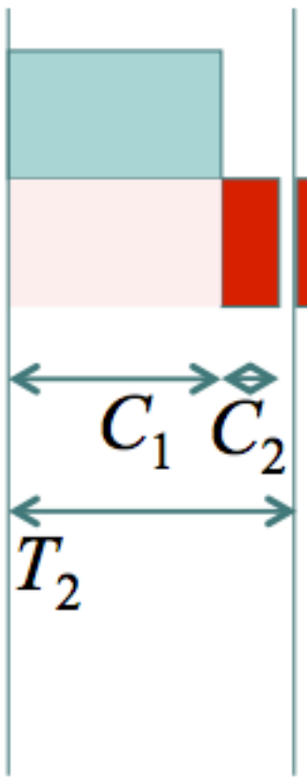
Showing Optimality of RMS: (for two tasks)

- It is sufficient to show that if a non-RMS schedule is feasible, then the RMS schedule is feasible.
- Consider two tasks as follows:



Showing Optimality of RMS: (for two tasks)

- The non-RMS, fixed priority schedule looks like this:

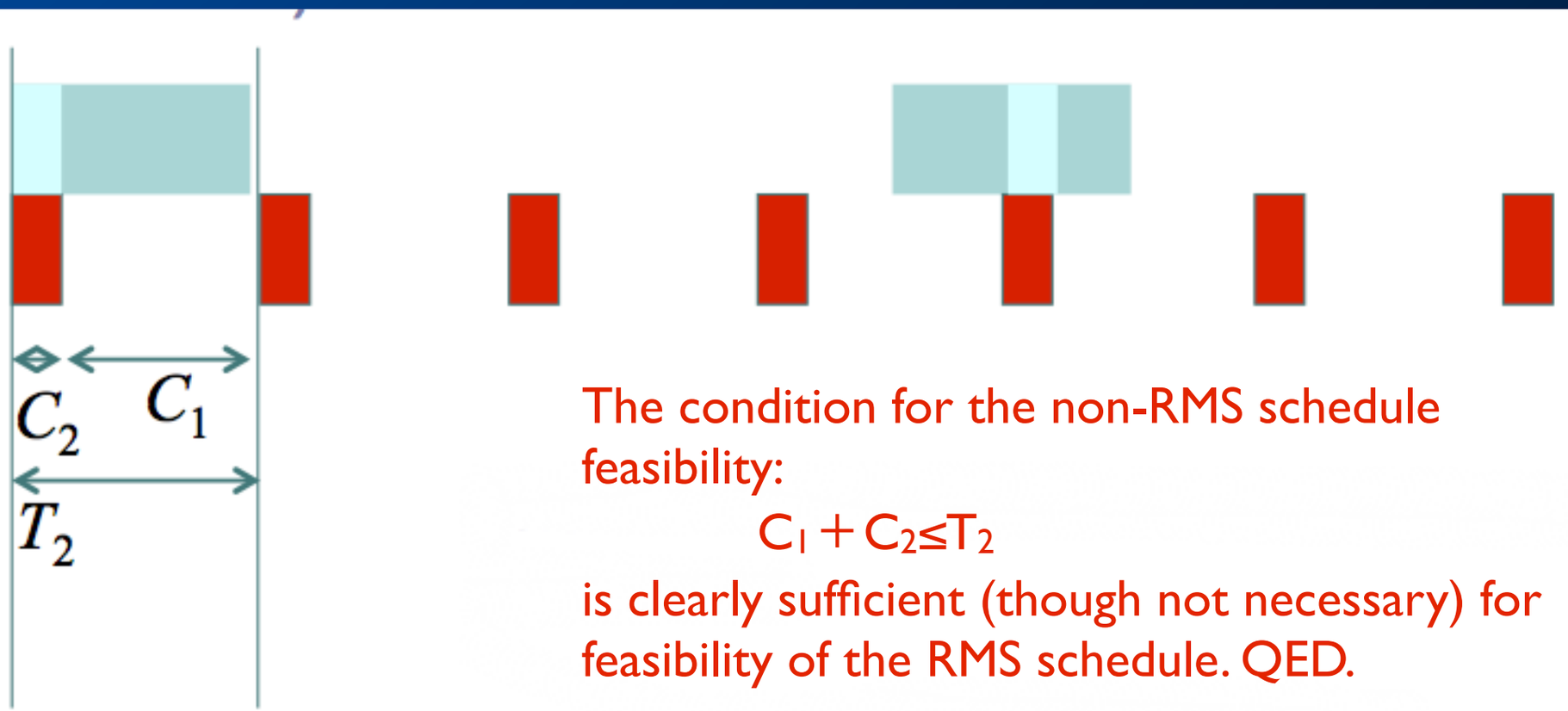


From this, we can see that the non-RMS schedule is feasible if and only if

$$C_1 + C_2 \leq T_2$$

We can then show that this condition implies that the RMS schedule is feasible.

Showing Optimality of RMS: (for two tasks)

- The RMS schedule looks like this: (task with smaller period moves earlier)
- 

The condition for the non-RMS schedule feasibility:

$$C_1 + C_2 \leq T_2$$

is clearly sufficient (though not necessary) for feasibility of the RMS schedule. QED.

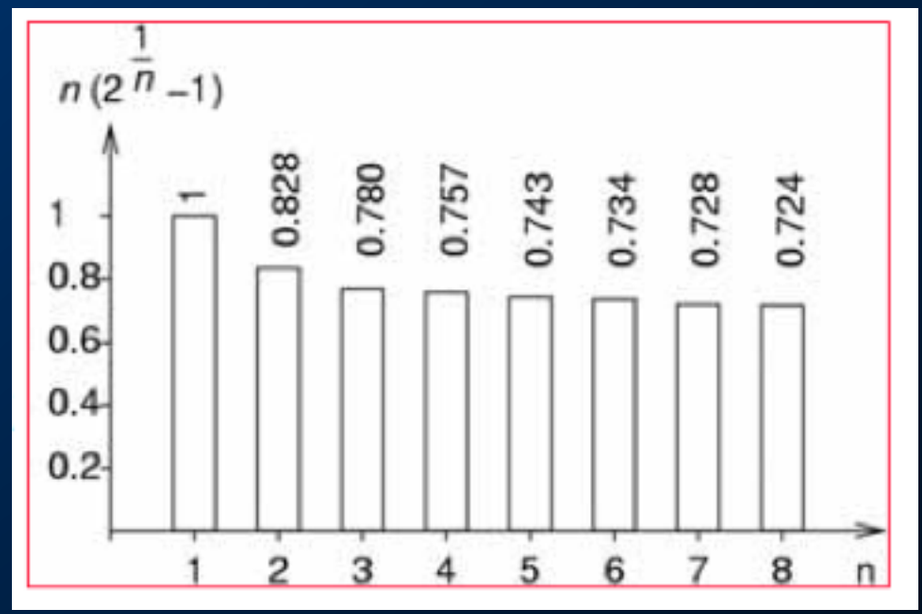
Comments

- This proof can be extended to an arbitrary number of tasks (though it gets much more tedious).
- This proof gives optimality only w.r.t. feasibility. It says nothing about other optimality criteria.
- Practical implementation:
  - Timer interrupt at greatest common divisor of the periods
  - Multiple timers

RMS

- Schedulability analysis: A set of periodic tasks is schedulable with RM if

$$\sum_{i=1}^{n} \frac{C_i}{T_i} \leq n\left(2^{1/n} - 1\right)$$



- This condition is sufficient but not necessary.
- The term

$$U = \sum_{i=1}^{n} \frac{C_i}{T_i}$$

- denotes the processor utilization factor U which is the fraction of processor time spent in the execution of the task set.

Deadline Driven Scheduling: 1. Jackson's Algorithm: EDD (1955)

- Given n independent one-time tasks with deadlines $d_1, \ldots, d_n$, schedule them to minimize the maximum lateness, defined as

$$L_{\max} = \max_{1 \le i \le n} \{ f_i - d_i \}$$

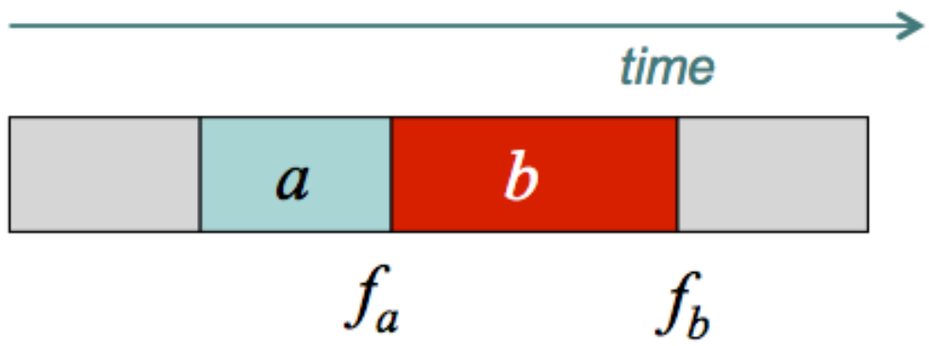
  where f_i is the finishing time of task i. Note that this is negative iff all deadlines are met.
- Earliest Due Date (EDD) algorithm: Execute them in order of non-decreasing deadlines.
- Note that this does not require preemption.

Theorem: EDD is Optimal in the Sense of Minimizing Maximum Lateness

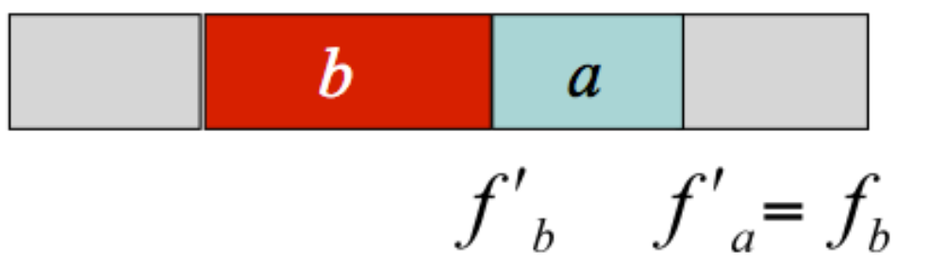
- To prove, use an interchange argument. Given a schedule *S* that is not EDD, there must be tasks *a* and *b* where *a* immediately precedes *b* in the schedule but $d_a > d_b$. Why?
- We can prove that this schedule can be improved by interchanging *a* and *b*. Thus, no non-EDD schedule is achieves smaller max lateness than EDD, so the EDD schedule must be optimal.

Consider a non-EDD Schedule *S*

- There must be tasks *a* and *b* where a immediately precedes *b* in the schedule but $d_a > d_b$



$$L_{\max} = \max\{f_a - d_a, f_b - d_b\} = f_b - d_b$$



$$L'_{\max} = \max\{f'_a - d_a, f'_b - d_b\}$$

Theorem: $L'_{max} \leq L_{max}$.
Hence, S' is no worse than S.

Case 1: $f'_a - d_a > f'_b - d_b$.
Then: $L'_{max} \leq f'_a - d_a = f_b - d_a \leq L_{max}$
(because: $d_a > d_b$).

Case 2: $f'_a - d_a \leq f'_b - d_b$.
Then: $L'_{max} \leq f'_b - d_b \leq L_{max}$
(because: $f'_b < f_b$).

Deadline Driven Scheduling: 1. Horn's algorithm: EDF (1974)

- Extend EDD by allowing tasks to "arrive" (become ready) at any time.
- Earliest deadline first (EDF): Given a set of n independent tasks with arbitrary arrival times, any algorithm that at any instant executes the task with the earliest absolute deadline among all arrived tasks is optimal w.r.t. minimizing the maximum lateness.
- Proof uses a similar interchange argument.

Using EDF for Periodic Tasks

- The EDF algorithm can be applied to periodic tasks as well as aperiodic tasks.
  - Simplest use: Deadline is the end of the period.
  - Alternative use: Separately specify deadline (relative to the period start time) and period.

RMS vs. EDF? Which one is better?

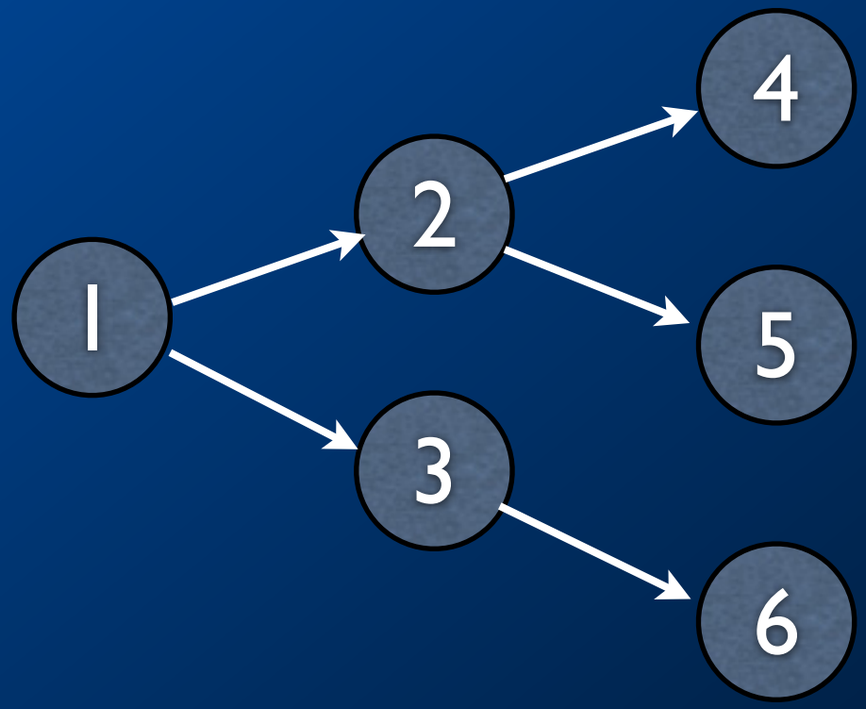
- What are the pros and cons of each?

Comparison of EDF and RMS

- Favoring RMS
  - Scheduling decisions are simpler (fixed priorities vs. the dynamic priorities required by EDF. EDF scheduler must maintain a list of ready tasks that is sorted by priority.)
- Favoring EDF
  - Since EDF is optimal w.r.t. maximum lateness, it is also optimal w.r.t. feasibility. RMS is only optimal w.r.t. feasibility. For infeasible schedules, RMS completely blocks lower priority tasks, resulting in unbounded maximum lateness.
  - EDF can achieve full utilization where RMS fails to do that
  - EDF results in fewer preemptions in practice, and hence less overhead for context switching.
  - Deadlines can be different from the period.

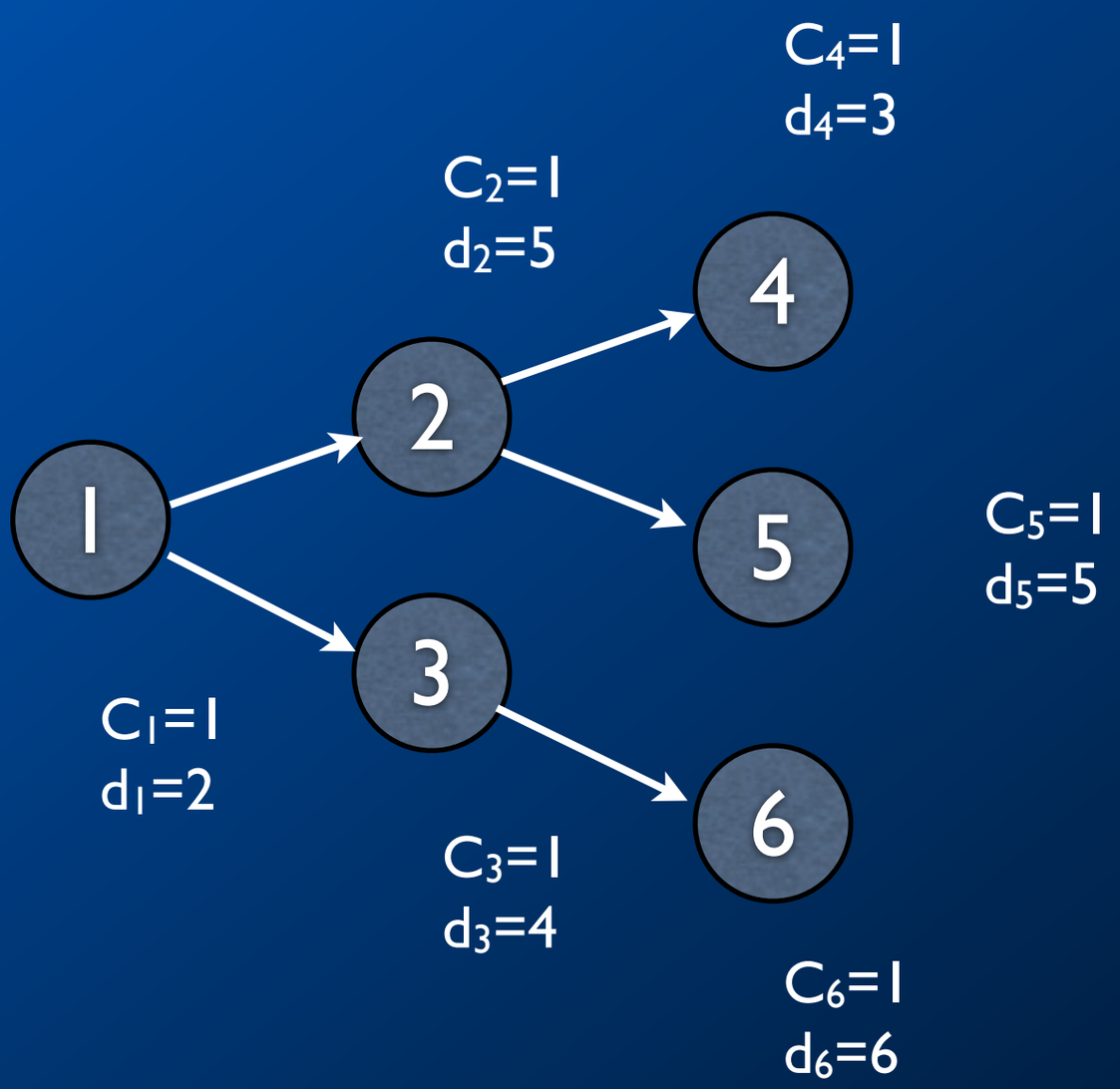
Precedence Constraints

- A directed acyclic graph (DAG) shows precedences, which indicate which tasks must complete before other tasks start.



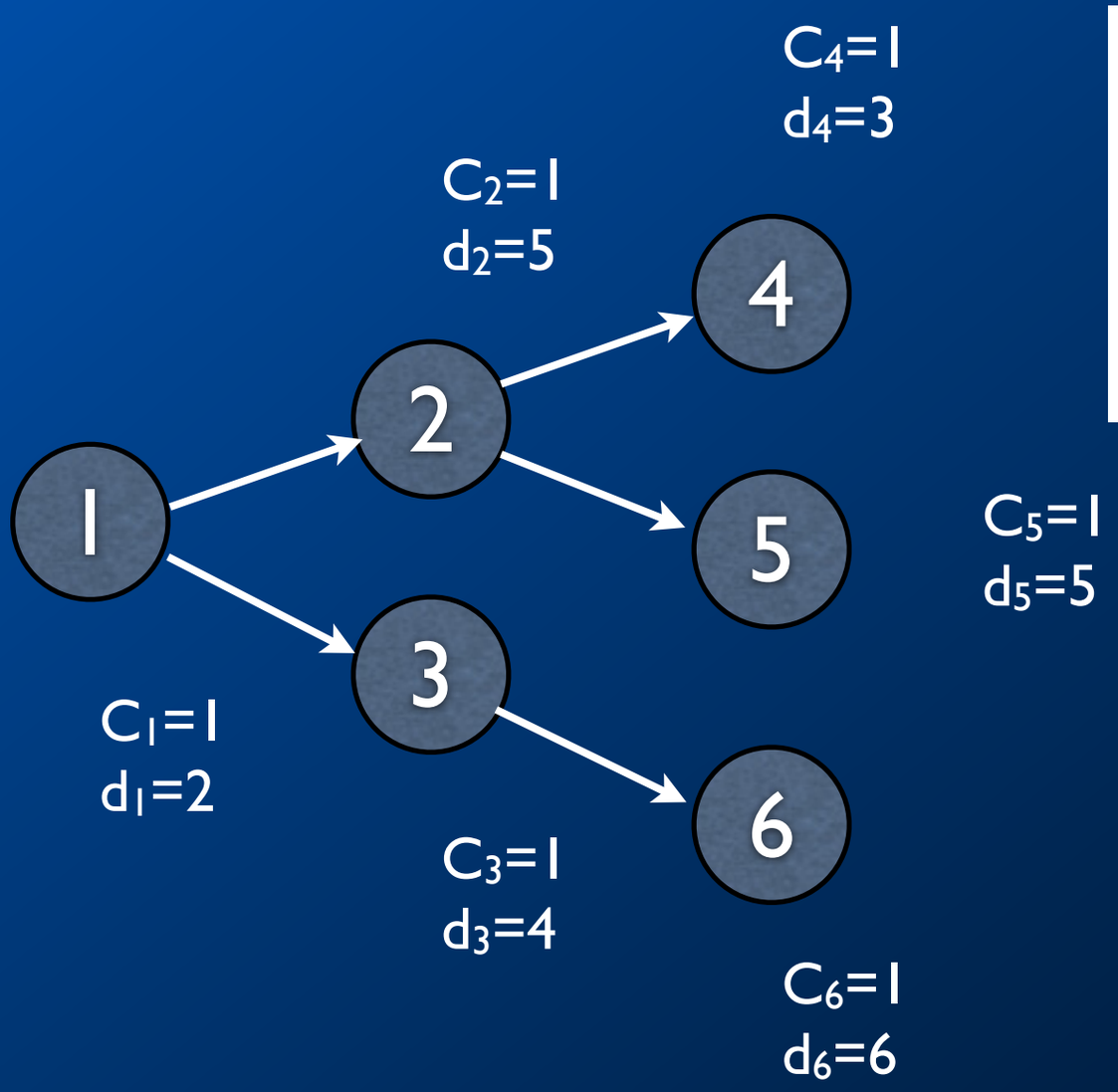
DAG, showing that task 1 must complete before tasks 2 and 3 can be started, etc.

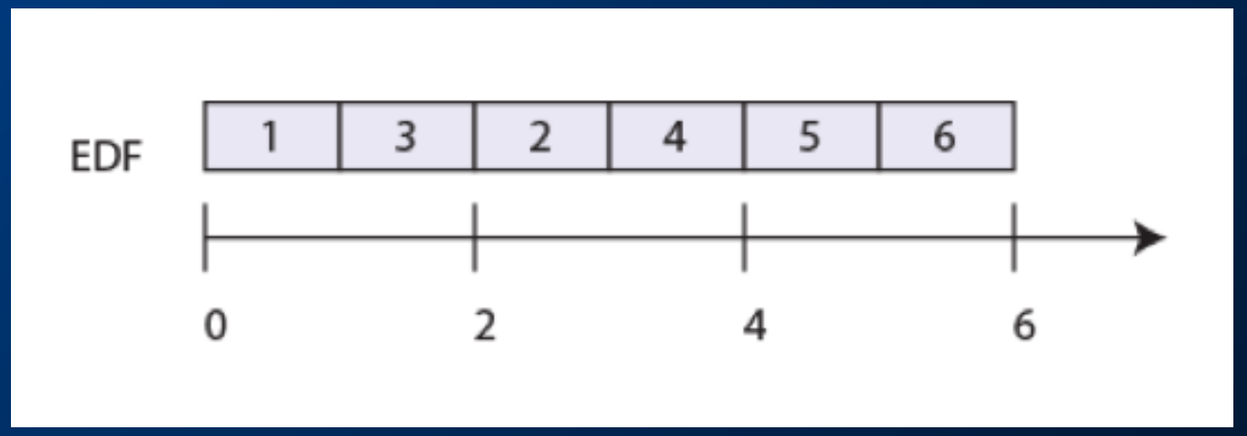
Example: EDF Schedule



Is this feasible? Is it optimal?

EDF is not optimal under precedence constraints

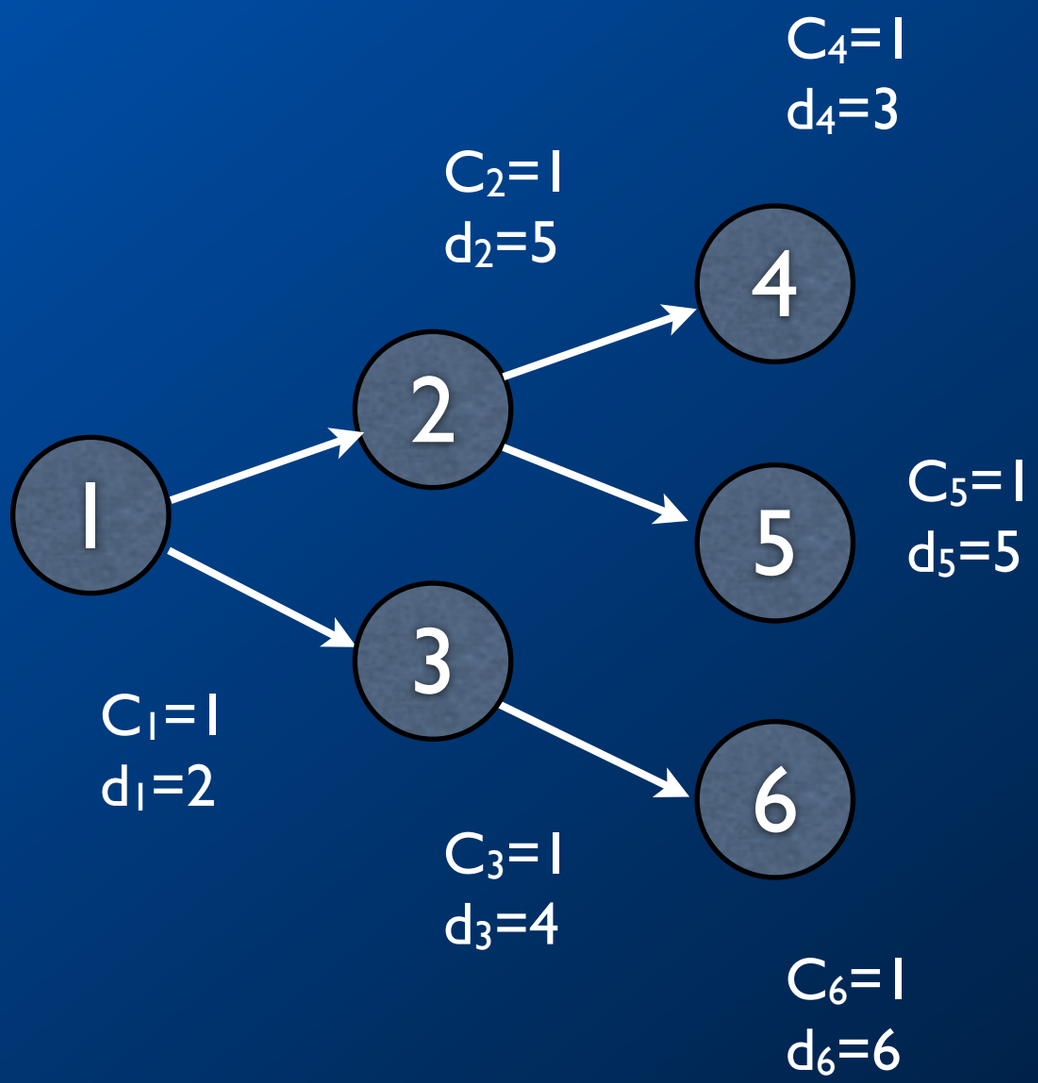




The EDF schedule chooses task 3 at time 1 because it has an earlier deadline. This choice results in task 4 missing its deadline.

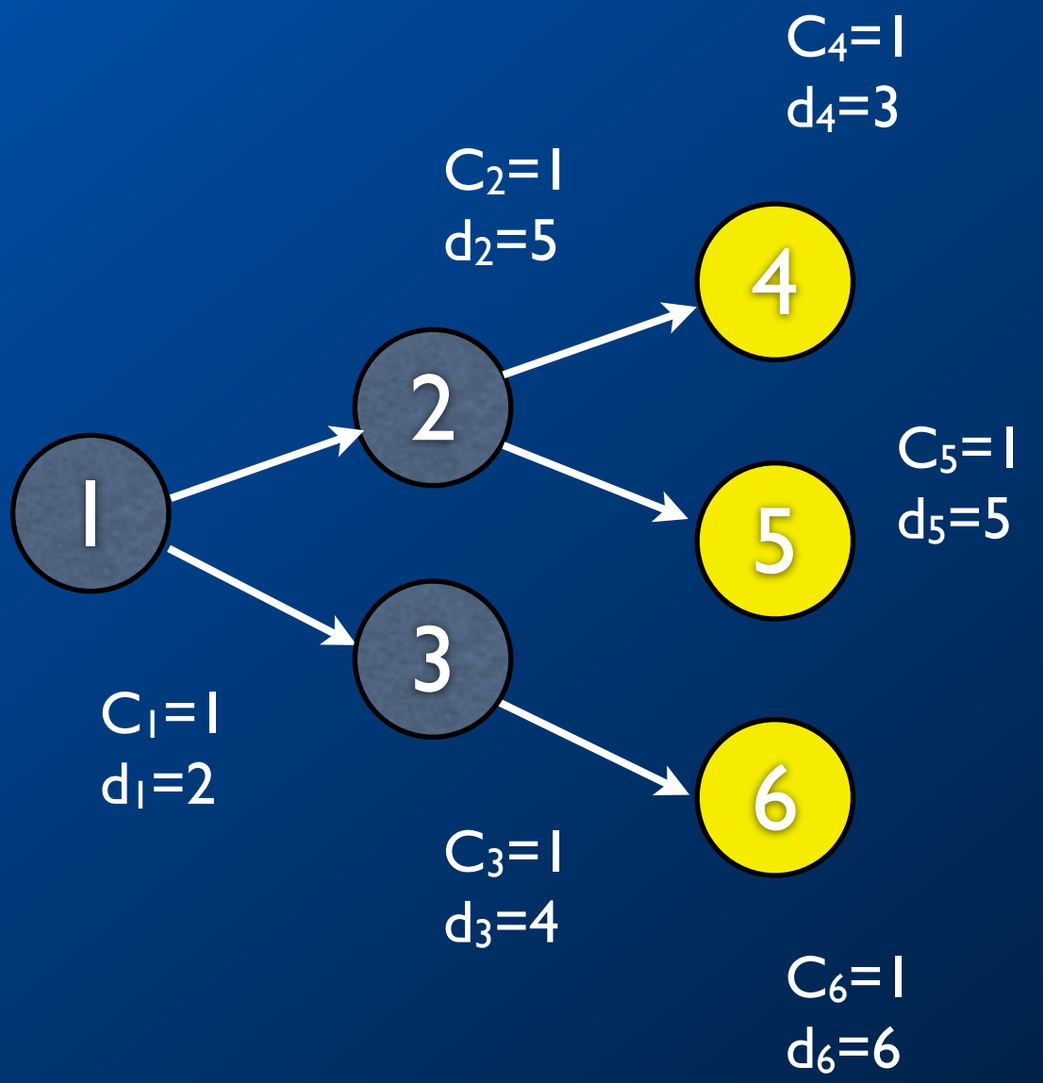
Is there a feasible schedule?

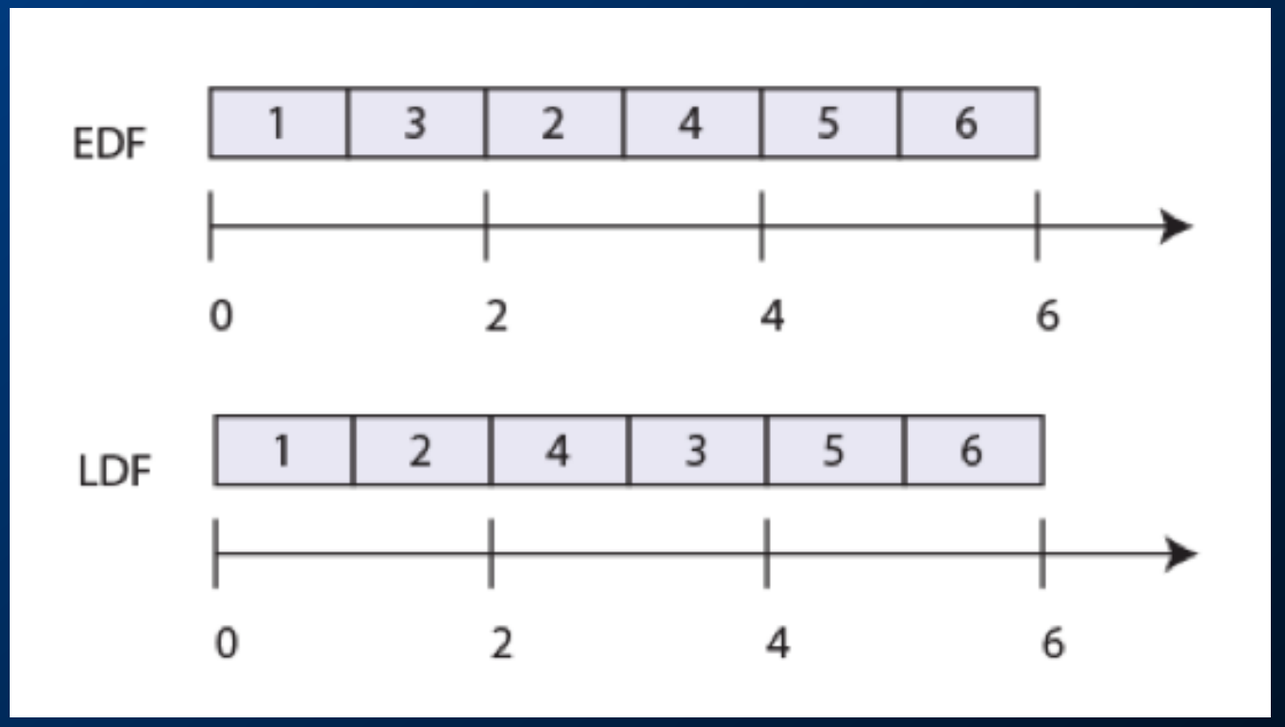
LDF is optimal under precedence constraints



The LDF schedule shown at the bottom respects all precedences and meets all deadlines.

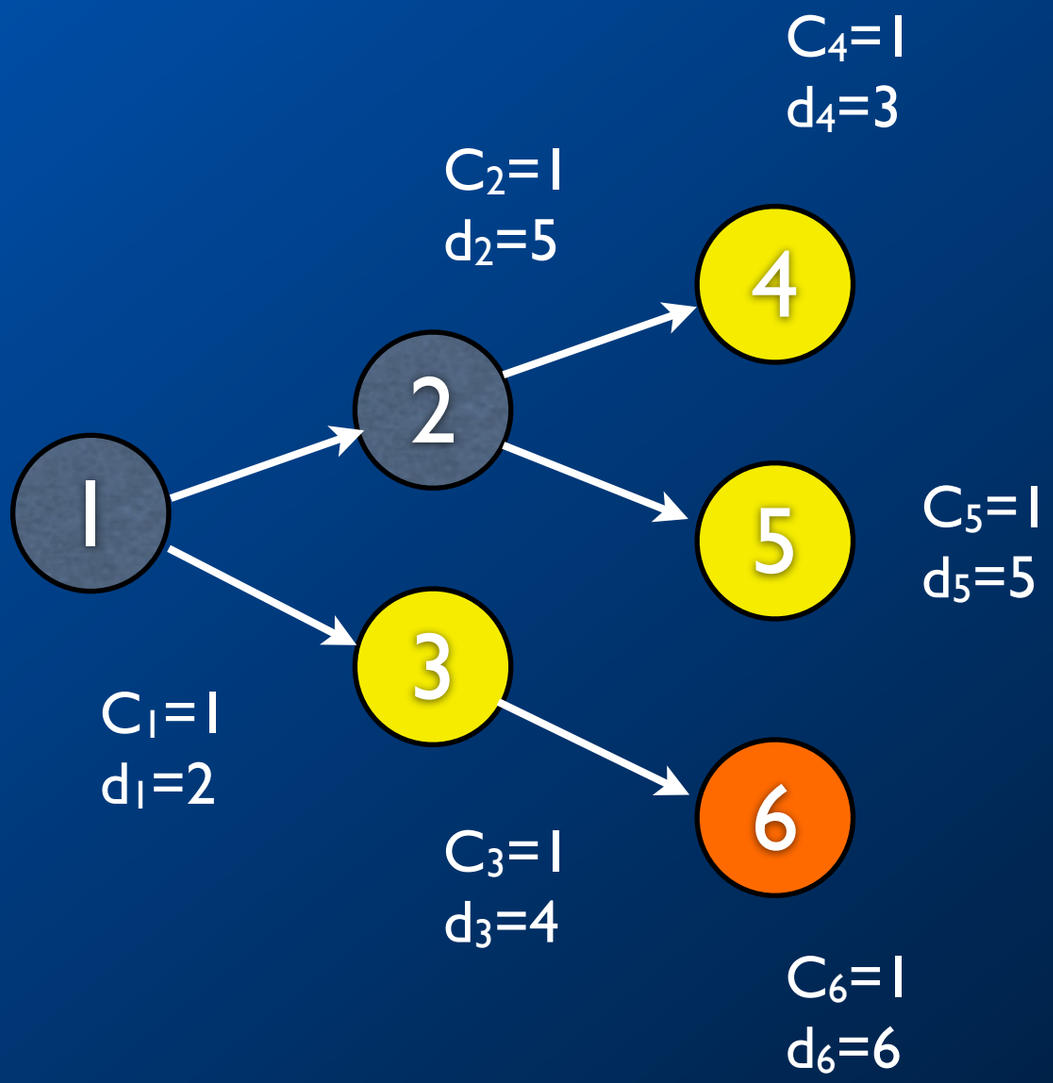
Latest Deadline First (LDF)(Lawler, 1973)

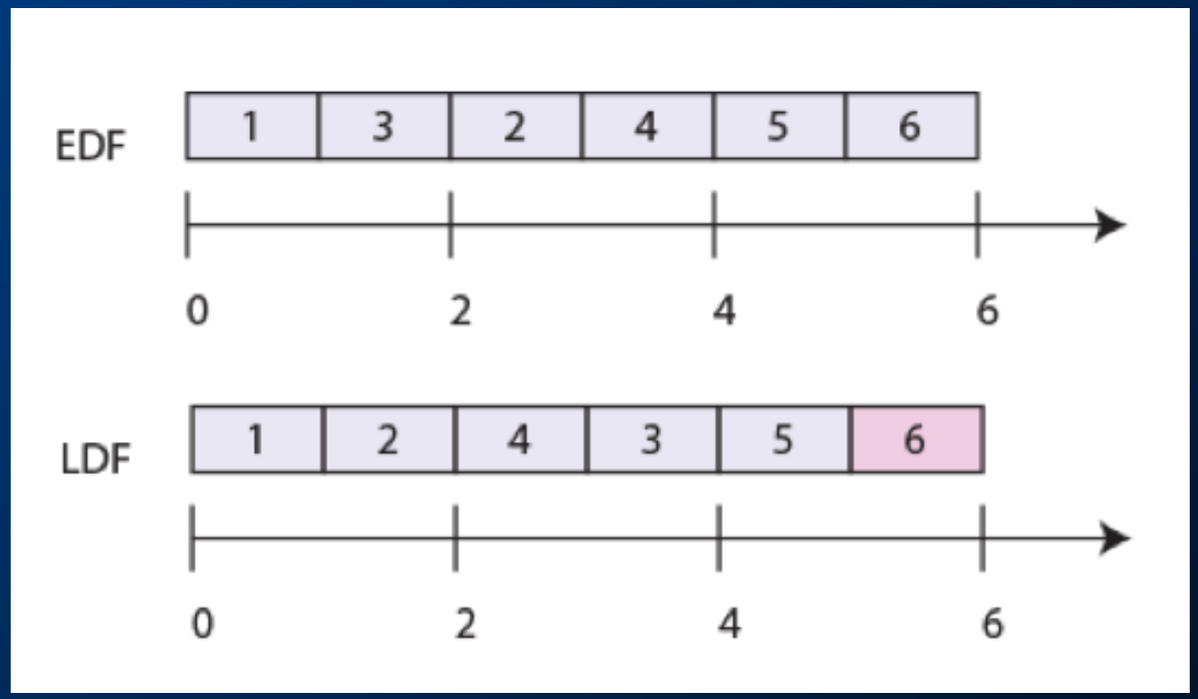




The LDF scheduling strategy builds a schedule backwards. Given a DAG, choose the leaf node with the latest deadline to be scheduled last, and work backwards.

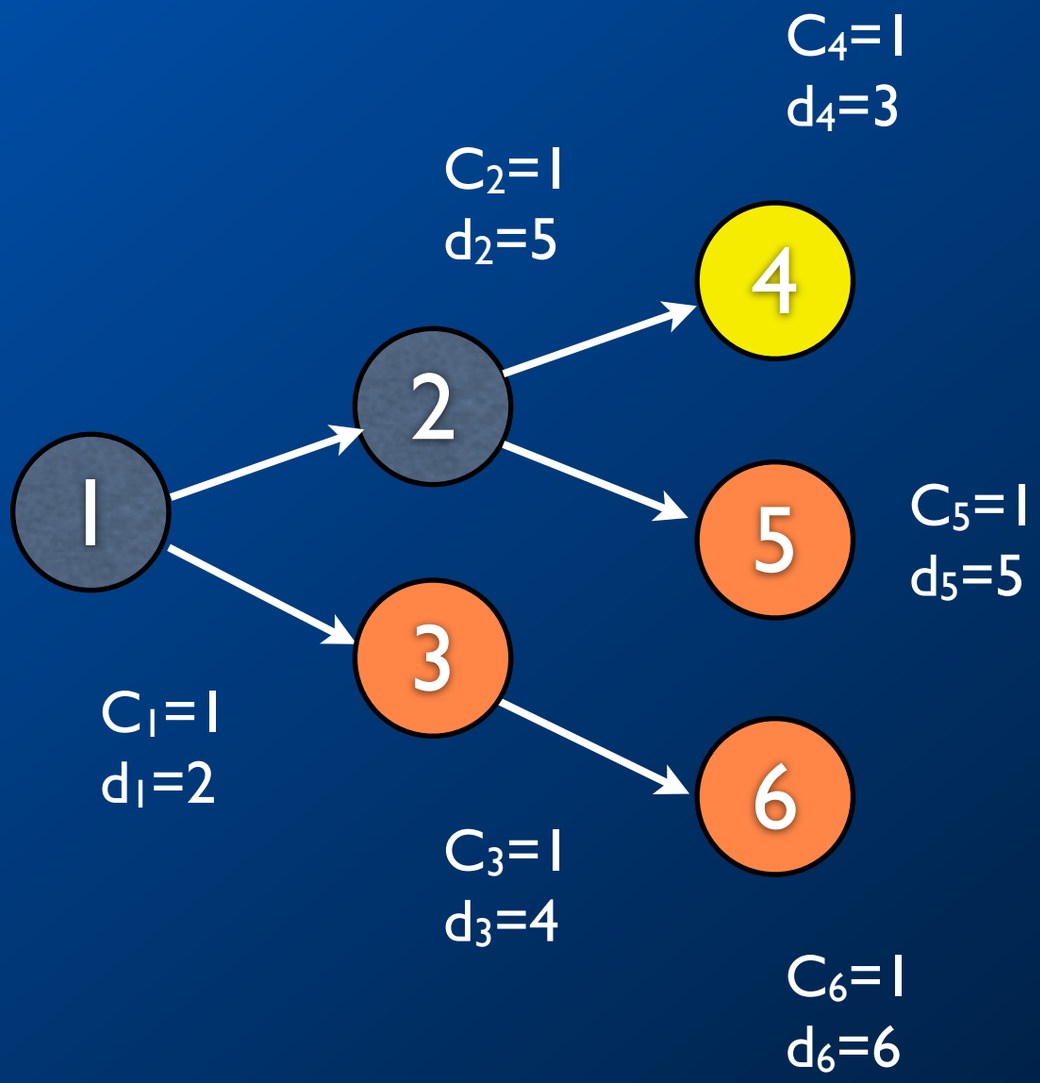
Latest Deadline First (LDF)(Lawler, 1973)

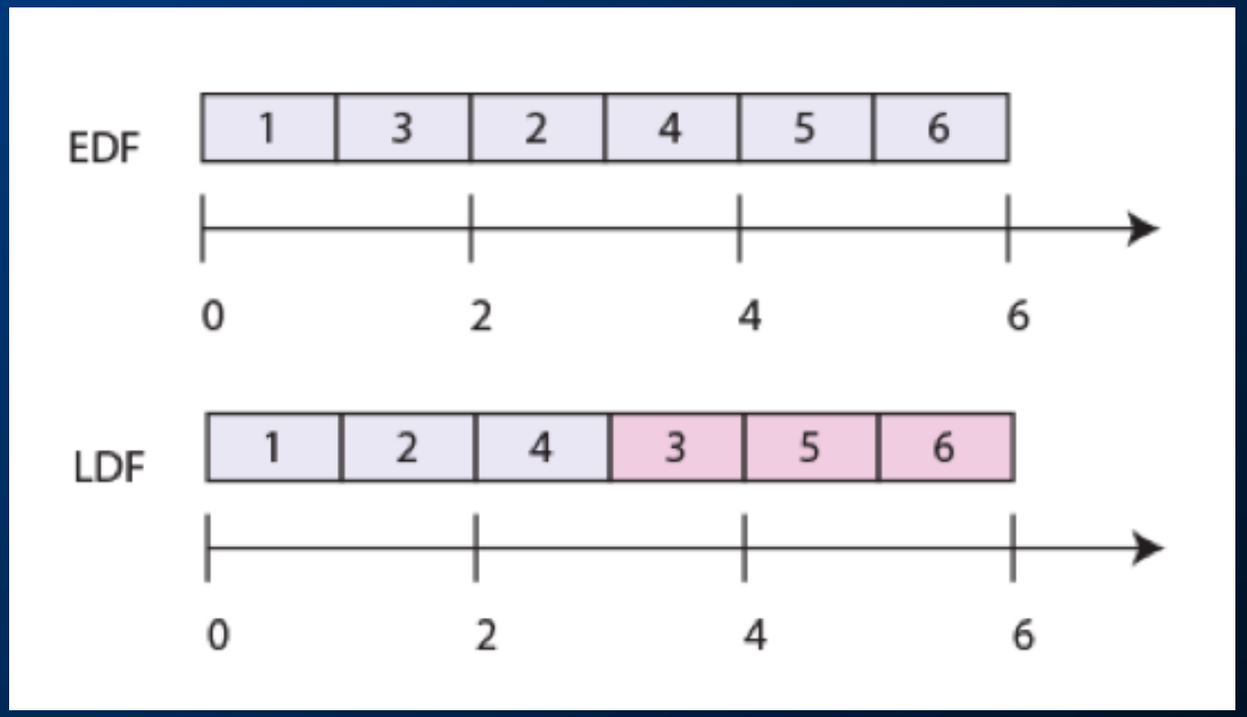




The LDF scheduling strategy builds a schedule backwards. Given a DAG, choose the leaf node with the latest deadline to be scheduled last, and work backwards.

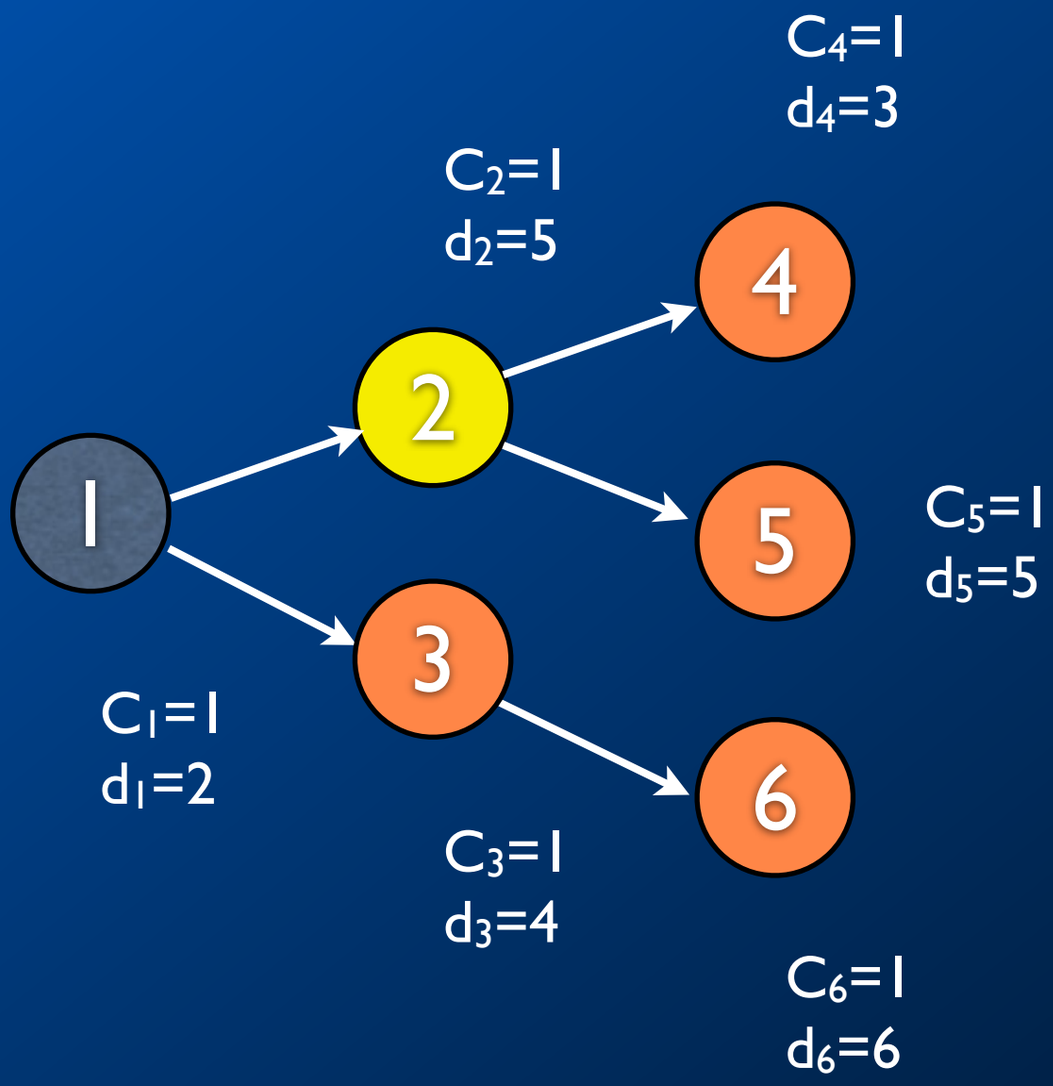
Latest Deadline First (LDF)(Lawler, 1973)

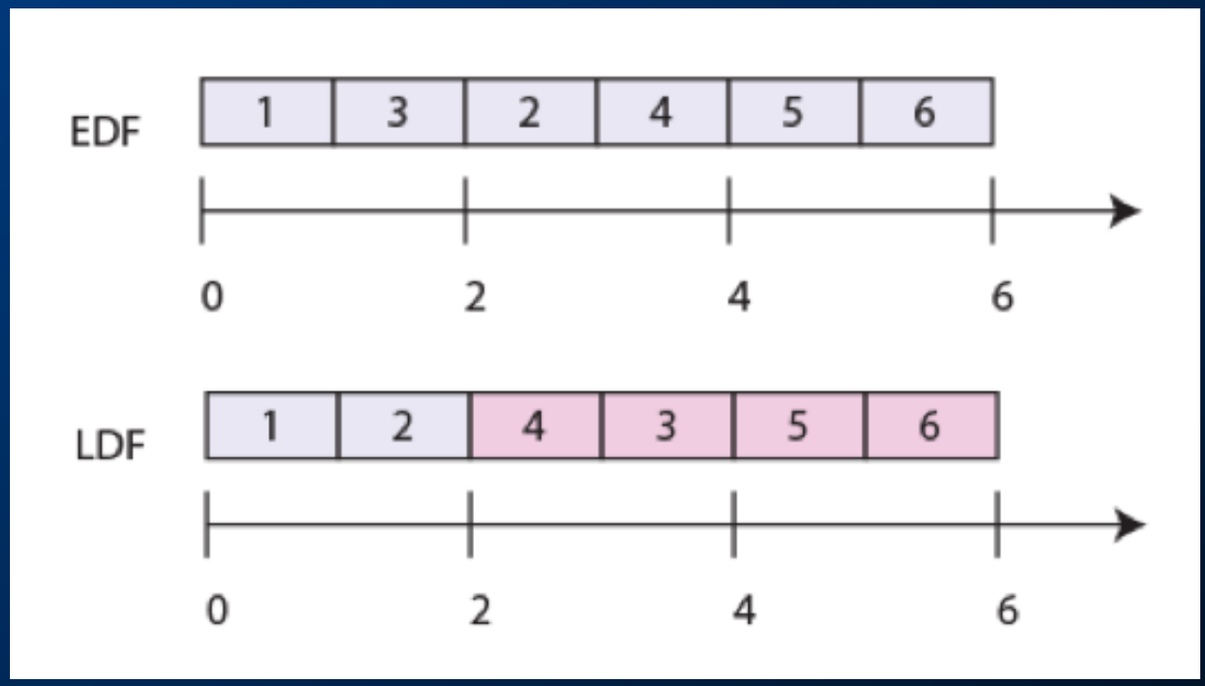




The LDF scheduling strategy builds a schedule backwards. Given a DAG, choose the leaf node with the latest deadline to be scheduled last, and work backwards.

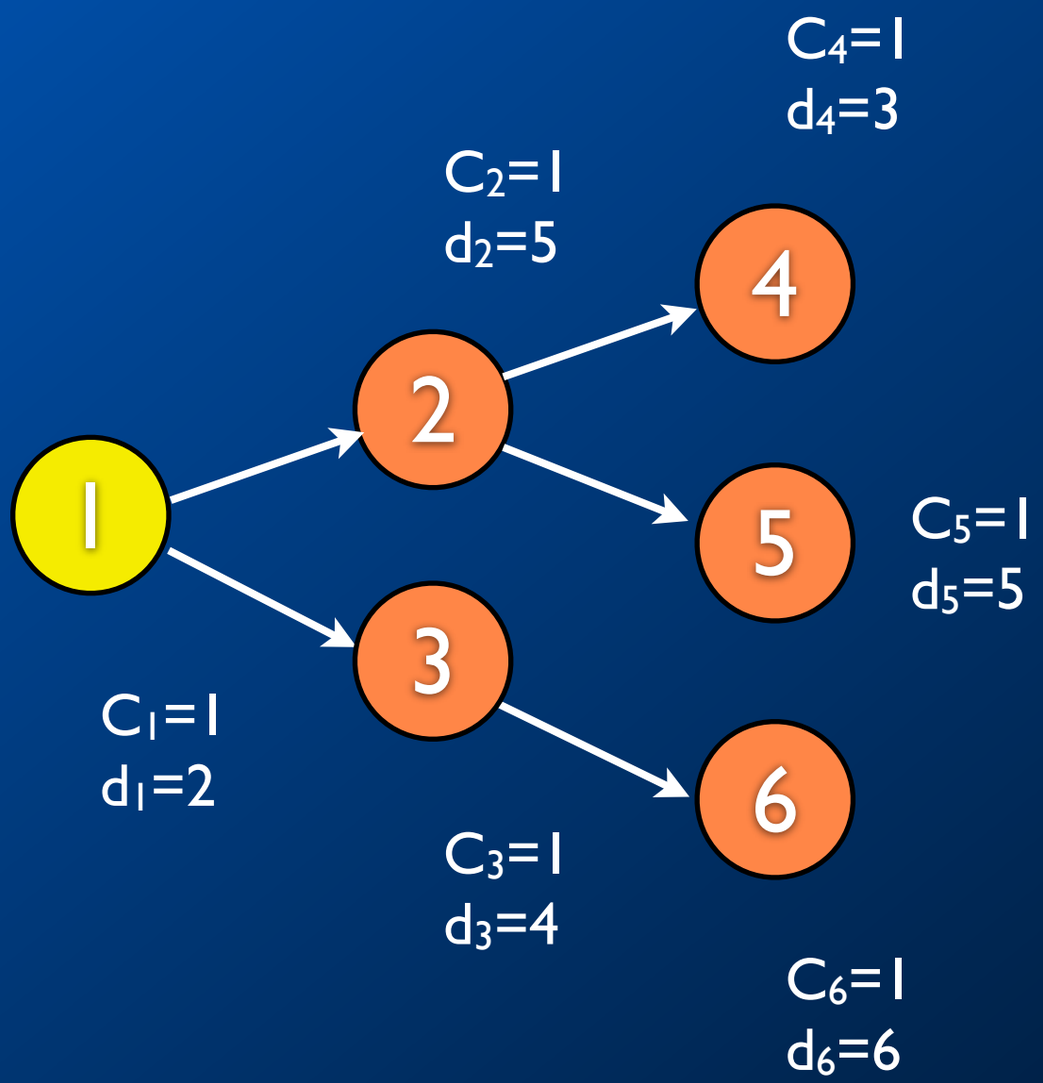
Latest Deadline First (LDF)(Lawler, 1973)

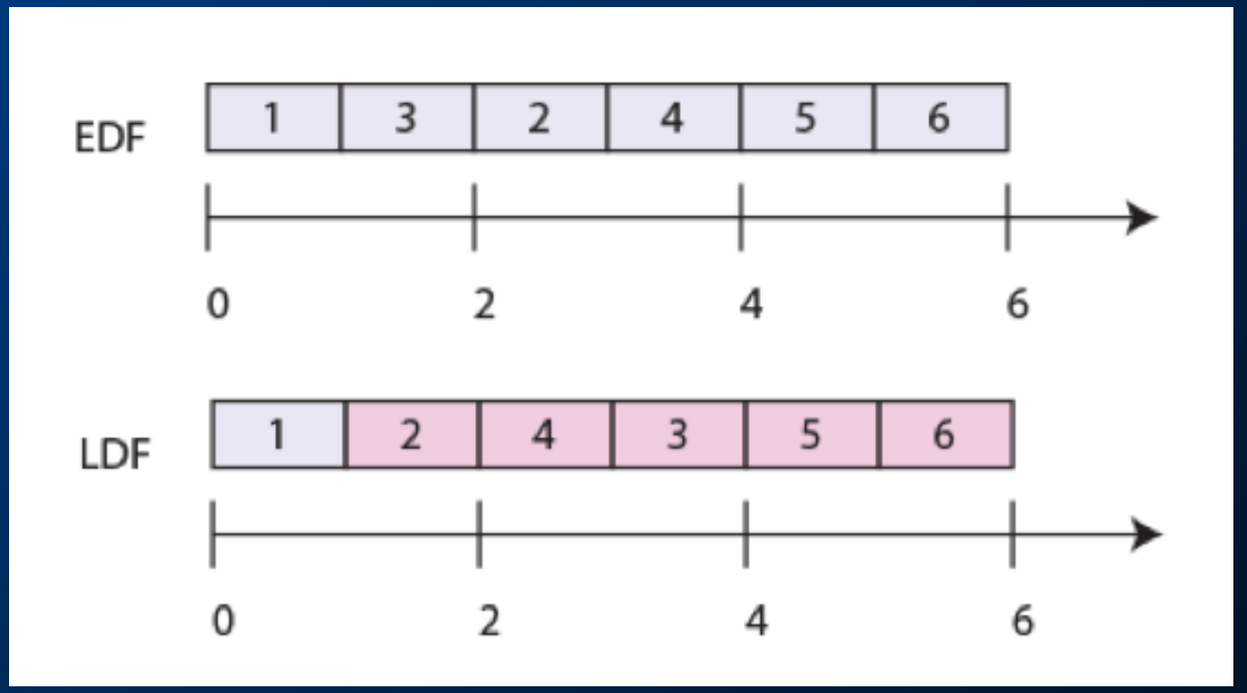




The LDF scheduling strategy builds a schedule backwards. Given a DAG, choose the leaf node with the latest deadline to be scheduled last, and work backwards.

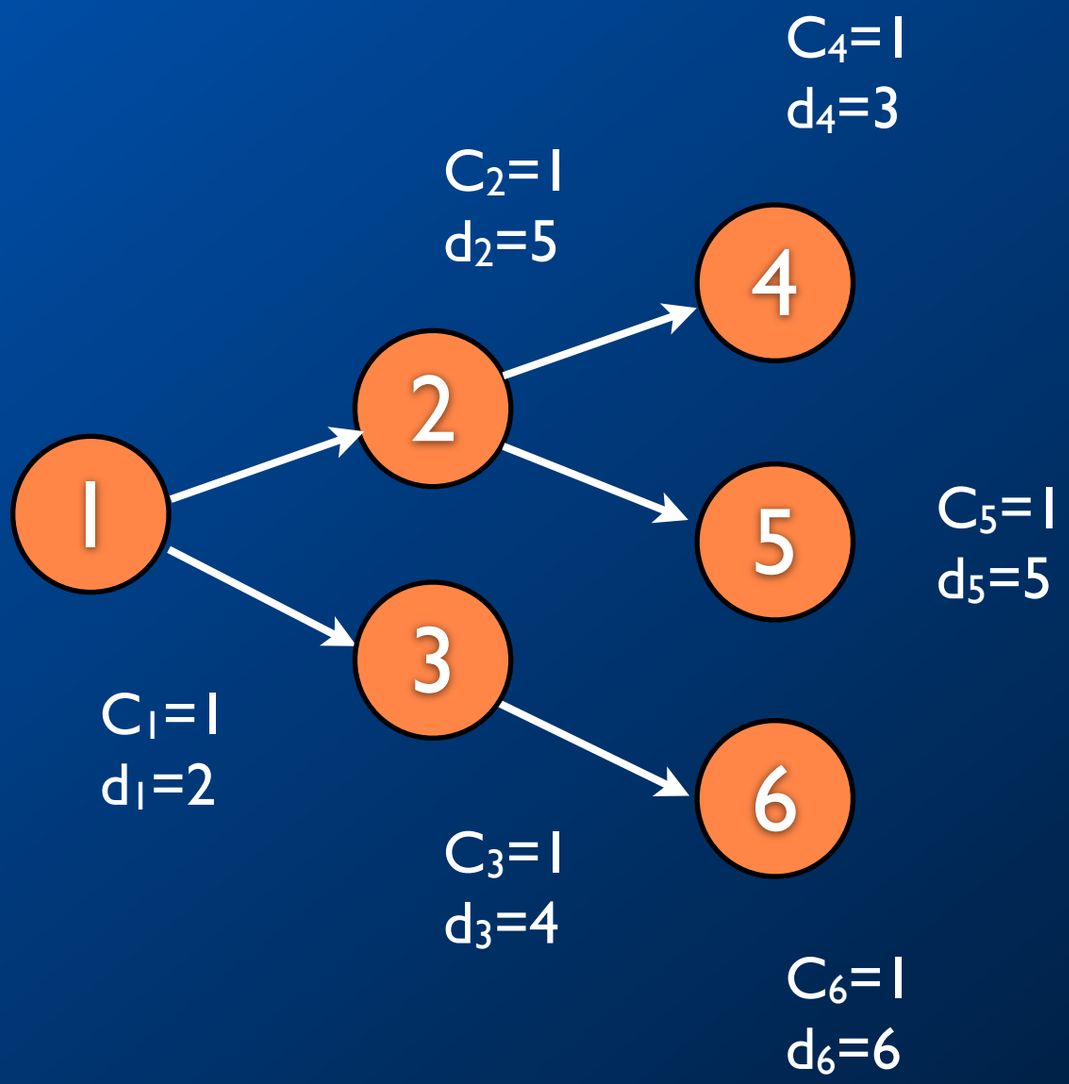
Latest Deadline First (LDF)(Lawler, 1973)

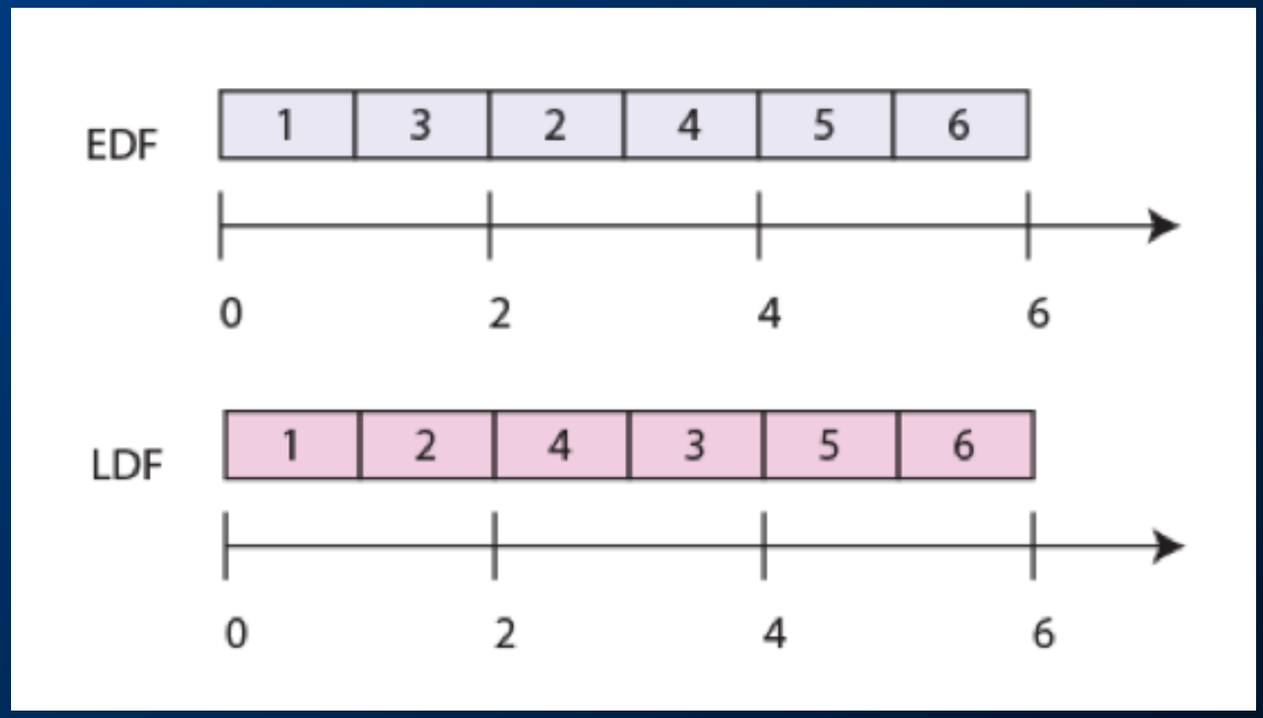




The LDF scheduling strategy builds a schedule backwards. Given a DAG, choose the leaf node with the latest deadline to be scheduled last, and work backwards.

Latest Deadline First (LDF)(Lawler, 1973)





The LDF scheduling strategy builds a schedule backwards. Given a DAG, choose the leaf node with the latest deadline to be scheduled last, and work backwards.

Latest Deadline First (LDF) (Lawler, 1973)

- LDF is optimal in the sense that it minimizes the maximum lateness.
- It does not require preemption. (We'll see that EDF does.)
- However, it requires that all tasks be available and their precedences known before any task is executed.

EDF*

- The problem of scheduling a set of n tasks with precedence constraints (concurrent activation) can be solved in polynomial time complexity if tasks are preemptable.
- The EDF* algorithm determines a feasible schedule in the case of tasks with precedence constraints if there exists one.
- By the modification it is guaranteed that if there exists a valid schedule at all then
  - a task starts execution not earlier than its release time and not earlier than the finishing times of its predecessors (a task cannot preempt any predecessor)
  - all tasks finish their execution within their deadlines

EDF*

- Modification of deadlines:
  - Task must finish the execution time within its deadline
  - Task must not finish the execution later than the maximum start time of its successor
- Solution: task j depends on task i

$$d_i' = \min(d_i, d_j' - C_j)$$

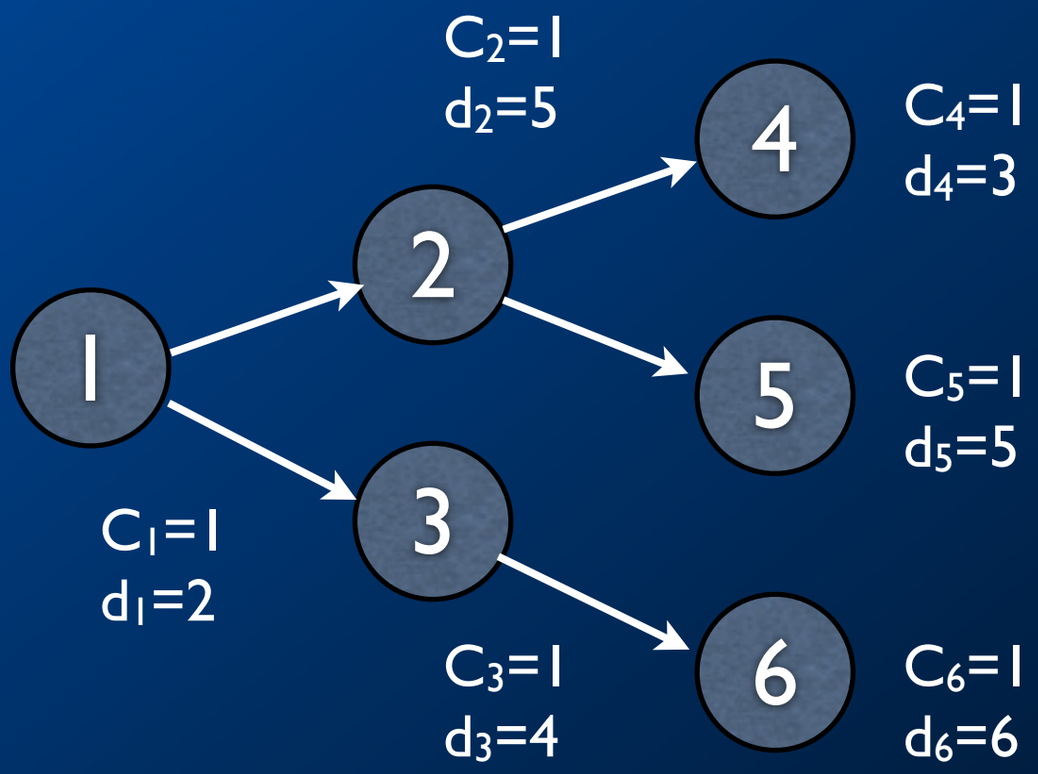
EDF*

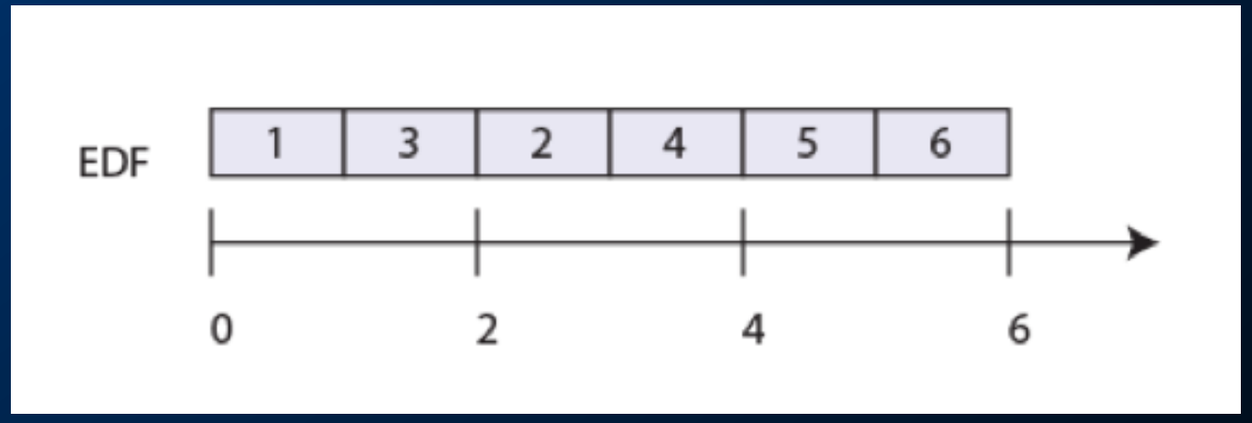
- Modification of release times:
  - Task must start the execution not earlier than its release time.
  - Task must not start the execution earlier than the minimum finishing time of its predecessor.
- Solution: task j depends on task i

$$r_j' = \max(r_j, r_i + C_i)$$

EDF with Precedences

With a preemptive scheduler, EDF can be modified to account for precedences and to allow tasks to arrive at arbitrary times. Simply adjust the deadlines and arrival times according to the precedences.





Recall that for the tasks at the left, EDF yields the schedule above, where task 4 misses its deadline.

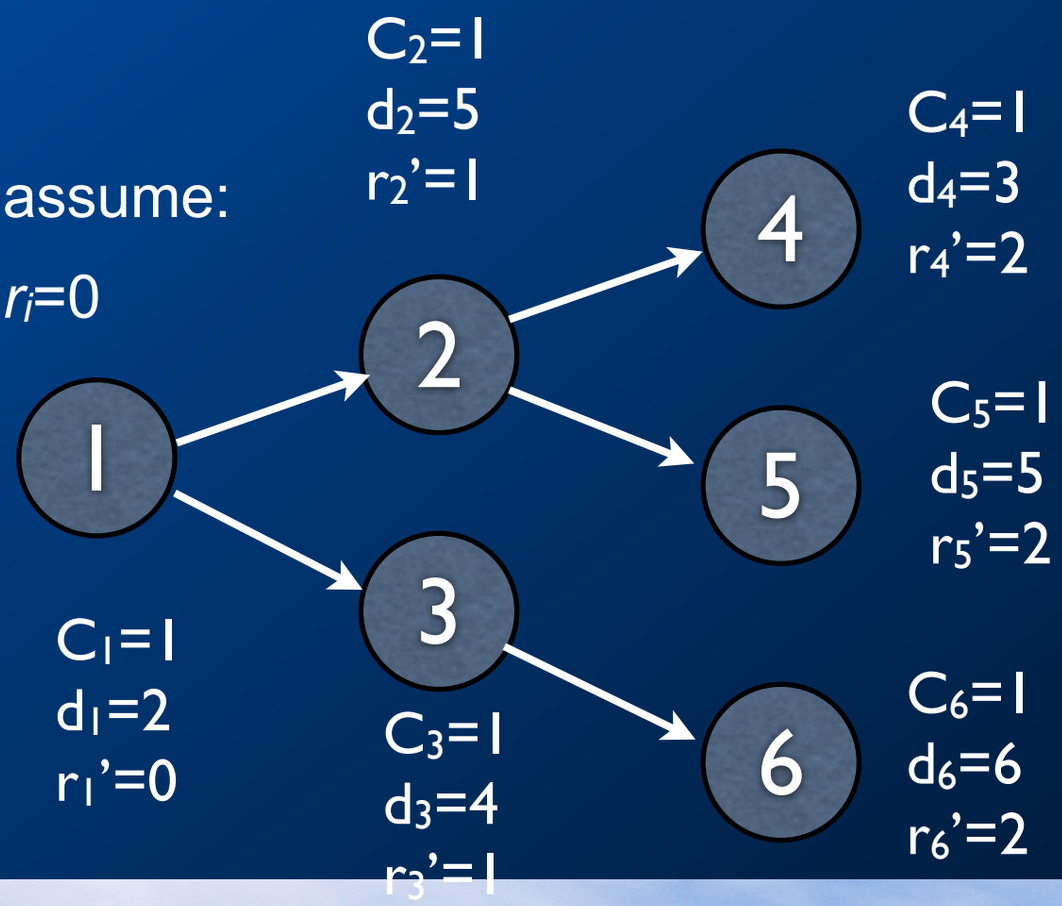
EDF with Precedences
Modifying release times

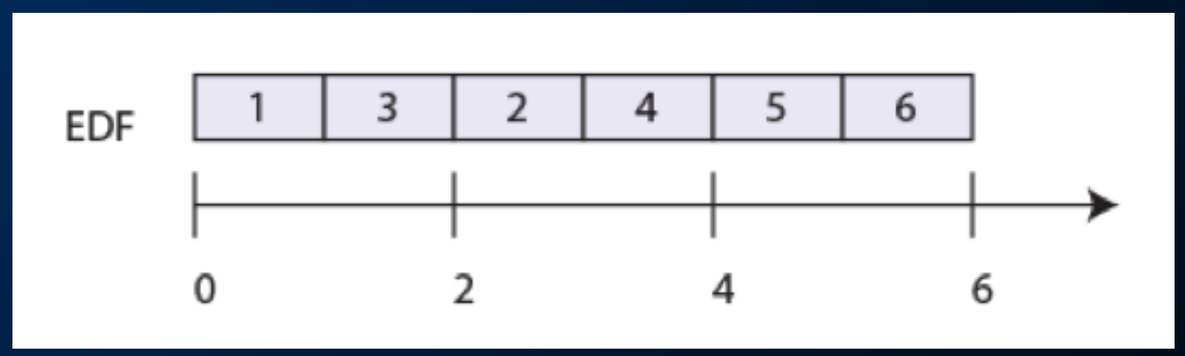
Given *n* tasks with precedences and release times r_i, if task *i* immediately precedes task *j*, then modify the release times as follows:

$$r'_j = \max(r_j, r_i + C_i)$$

assume:

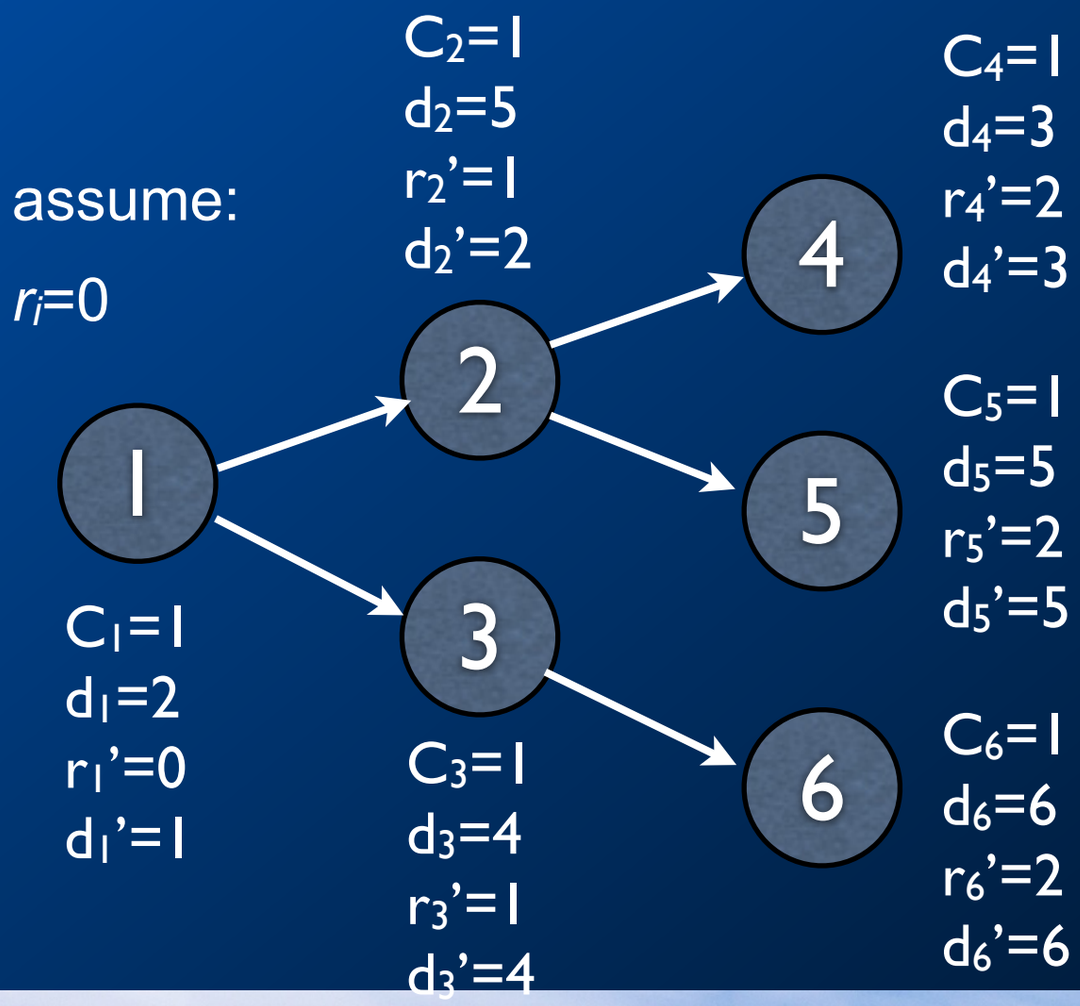
r_i=0



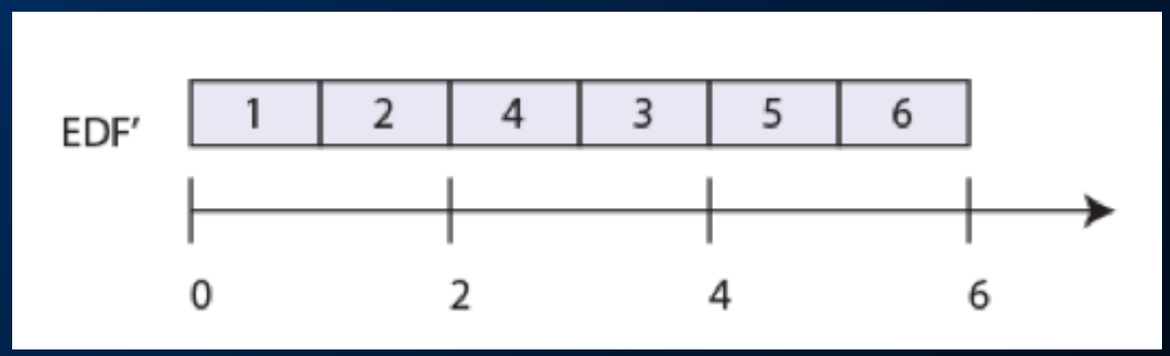


EDF with Precedences
Modifying deadlines

Given *n* tasks with precedences and deadlines d_i, if task i immediately precedes task *j*, then modify the deadlines as follows:



$$d_i' = \min(d_i, d_j' - C_j)$$



Using the revised release times and deadlines, the above EDF schedule is optimal and meets all deadlines.

Optimality

- EDF with precedences is optimal in the sense of minimizing the maximum lateness.

How

- 如果进程组不可调度，又需要保证在截至时限前完成
  - 使用更快的cpu。不改变周期就可以减少执行时间。但cpu利用率更低。
  - 重新设计进程以减少执行时间。需要了解代码。
  - 重写规格说明以改变截至时限。

scheduling anomalies

- Mutual exclusion
  - Priority inversion
  - Priority inheritance
  - Priority ceiling
- Multiprocessor scheduling
  - Richard's anomalies

Accounting for Mutual Exclusion

- When threads access shared resources, they need to use mutexes to ensure data integrity.
- Mutexes can also complicate scheduling.

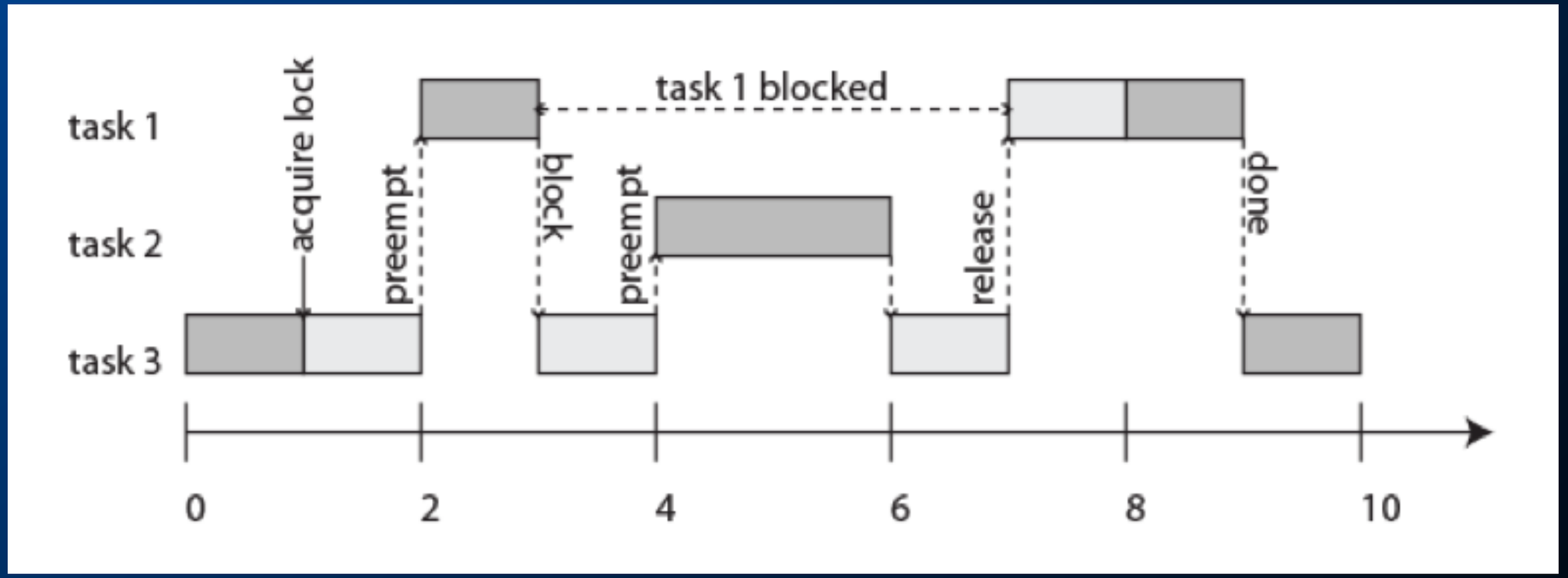
Recall mutual exclusion mechanism in pthreads

```
#include <pthread.h>
...
pthread_mutex_t lock;
void* addListener(notify listener) {
        pthread_mutex_lock(&lock);
        ...
        pthread_mutex_unlock(&lock);
}
void* update(int newValue) {
        pthread_mutex_lock(&lock);
        value = newValue;
        elementType* element = head;
        while (element != 0) {
            (*(element->listener))(newValue);
            element = element->next;
        }
        pthread_mutex_unlock(&lock);
  }
int main(void) {
  pthread_mutex_init(&lock, NULL);
  ...
}
```

Whenever a data structure is shared across threads, access to the data structure must usually be atomic. This is enforced using mutexes, or mutual exclusion locks. The code executed while holding a lock is called a *critical section.*

Priority Inversion: A Hazard with Mutexes

- Task 1 has highest priority, task 3 lowest. Task 3 acquires a lock on a shared object, entering a critical section. It gets preempted by task 1, which then tries to acquire the lock and blocks. Task 2 preempts task 3 at time 4, keeping the higher priority task 1 blocked for an unbounded amount of time. In effect, the priorities of tasks 1 and 2 get inverted, since task 2 can keep task 1 waiting arbitrarily long.



The MARS Pathfinder problem (1)

"But a few days into the mission, not long after Pathfinder started gathering meteorological data, the spacecraft began experiencing total system resets, each resulting in losses of data.

The MARS Pathfinder problem (2)

- “VxWorks provides preemptive priority scheduling of threads. Tasks on the Pathfinder spacecraft were executed as threads with priorities that were assigned in the usual manner reflecting the relative urgency of these tasks.”
- “Pathfinder contained an "information bus", which you can think of as a shared memory area used for passing information between different components of the spacecraft.”
- A bus management task ran frequently with high priority to move certain kinds of data in and out of the information bus. Access to the bus was synchronized with mutual exclusion locks (mutexes).”

The MARS Pathfinder problem (3)

- The meteorological data gathering task ran as an infrequent, low priority thread, ... When publishing its data, it would acquire a mutex, do writes to the bus, and release the mutex.
- The spacecraft also contained a communications task that ran with medium priority."

High priority: retrieval of data from shared memory
Medium priority: communications task
Low priority: thread collecting meteorological data

The MARS Pathfinder problem (4)

"Most of the time this combination worked fine. However, very infrequently it was possible for an interrupt to occur that caused the (medium priority) communications task to be scheduled during the short interval while the (high priority) information bus thread was blocked waiting for the (low priority) meteorological data thread. In this case, the long-running communications task, having higher priority than the meteorological task, would prevent it from running, consequently preventing the blocked information bus task from running. After some time had passed, a watchdog timer would go off, notice that the data bus task had not been executed for some time, conclude that something had gone drastically wrong, and initiate a total system reset. This scenario is a classic case of priority inversion."

Priority inversion on Mars

- Priority inheritance also solved the Mars Pathfinder problem: the VxWorks operating system used in the pathfinder implements a flag for the calls to mutex primitives. This flag allows priority inheritance to be set to "on". When the software was shipped, it was set to "off".

The problem on Mars was corrected by using the debugging facilities of VxWorks to change the flag to "on", while the Pathfinder was already on the Mars [Jones, 1997].

Priority Inheritance Protocol (PIP) (Sha, Rajkumar, Lehoczky, 1990)

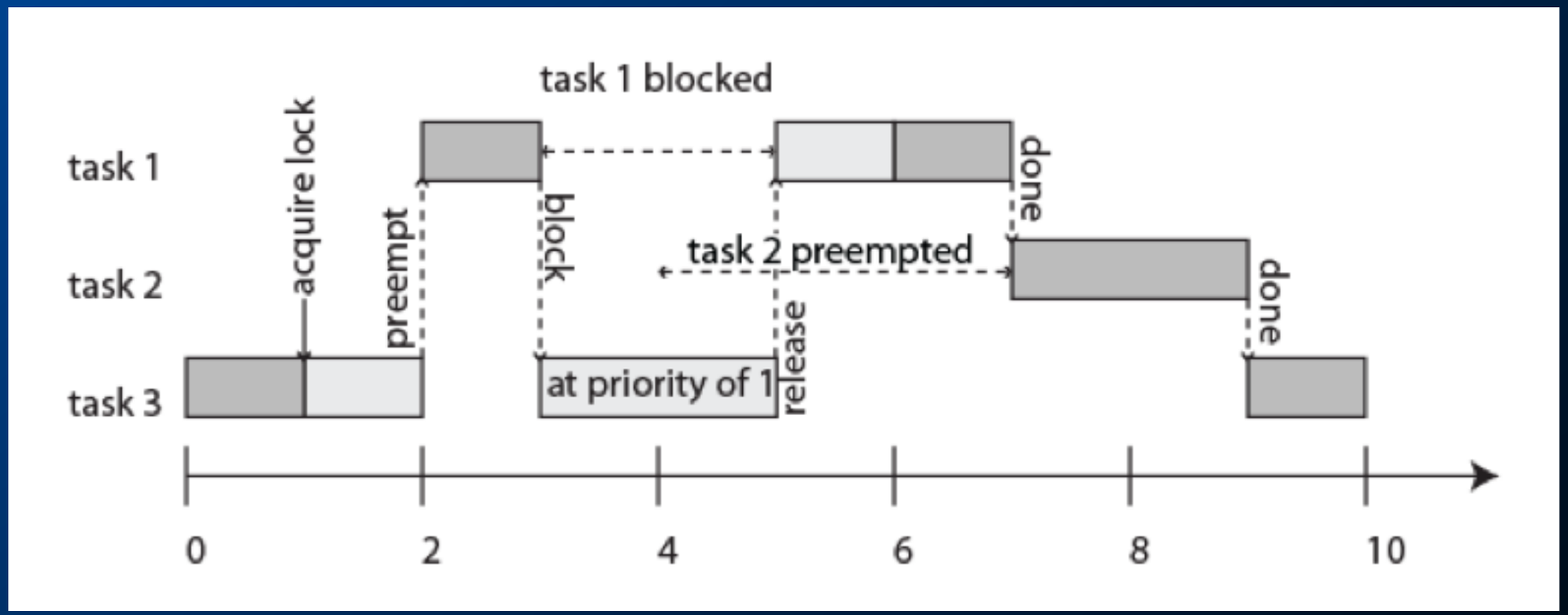
- Assumptions:
  - *n* tasks which cooperate through *m* shared resources;
  - fixed priorities, all critical sections on a resource begin with a wait(*Si*) and end with a signal(*Si*) operation.
- Basic idea:
  - When a task *Ji* blocks one or more higher priority tasks, it temporarily assumes (inherits) the highest priority of the blocked tasks.
- Terms:
  - We distinguish a fixed nominal priority *Pi* and an active priority *pi* larger or equal to *Pi*. Jobs J_1, ...J_n are ordered with respect to nominal priority where J_1 has highest priority. Jobs do not suspend themselves.

Priority Inheritance Protocol (PIP)

- Algorithm:
  - Jobs are scheduled based on their active priorities. Jobs with the same priority are executed in a FCFS discipline.
  - When a job J_i tries to enter a critical section and the resource is blocked by a lower priority job, the job J_i is blocked. Otherwise it enters the critical section.
  - When a job J_i is blocked, it transmits its active priority to the job J_k that holds the semaphore. J_k resumes and executes the rest of its critical section with a priority $p_k=p_i$ (it inherits the priority of the highest priority of the jobs blocked by it).
  - When J_k exits a critical section, it unlocks the semaphore and the highest priority job blocked on that semaphore is awakened. If no other jobs are blocked by J_k, then p_k is set to P_k, otherwise it is set to the highest priority of the jobs blocked by J_k.
  - Priority inheritance is transitive, i.e. if 1 is blocked by 2 and 2 is blocked by 3, then 3 inherits the priority of 1 via 2.

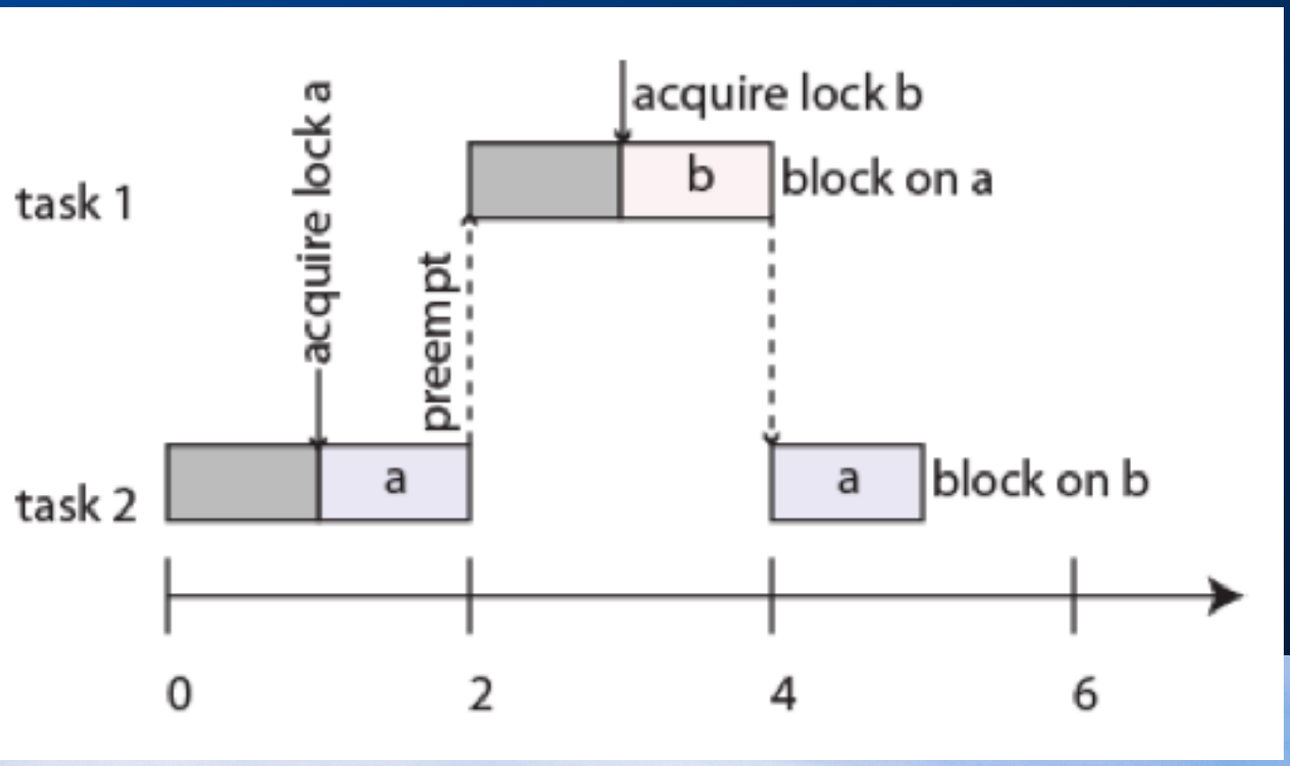
Priority Inheritance Protocol (PIP) - example

- Task 1 has highest priority, task 3 lowest. Task 3 acquires a lock on a shared object, entering a critical section. It gets preempted by task 1, which then tries to acquire the lock and blocks. Task 3 inherits the priority of task 1, preventing preemption by task 2.



Deadlock

The lower priority task starts first and acquires lock a, then gets preempted by the higher priority task, which acquires lock b and then blocks trying to acquire lock a. The lower priority task then blocks trying to acquire lock b, and no further progress is possible.




```
#include <pthread.h>
...
pthread_mutex_t lock_a, lock_b;
void* thread_1_function(void* arg) {
    pthread_mutex_lock(&lock_b);
    ...
    pthread_mutex_lock(&lock_a);
    ...
    pthread_mutex_unlock(&lock_a);
    ...
    pthread_mutex_unlock(&lock_b);
    ...
}
void* thread_2_function(void* arg) {
    pthread_mutex_lock(&lock_a);
    ...
    pthread_mutex_lock(&lock_b);
    ...
    pthread_mutex_unlock(&lock_b);
    ...
    pthread_mutex_unlock(&lock_a);
    ...
}
```

Priority Ceiling Protocol (PCP) (Sha, Rajkumar, Lehoczky, 1990)

- Every lock or semaphore is assigned a priority ceiling equal to the priority of the highest-priority task that can lock it.
  - Can one automatically compute the priority ceiling?
- A task T can acquire a lock only if the task's priority is strictly higher than the priority ceilings of all locks currently held by other tasks
  - Locks that are not held by any task don't affect the task
- This prevents deadlocks
- There are extensions supporting dynamic priorities and dynamic creations of locks (stack resource policy)

OCPP and ICPP

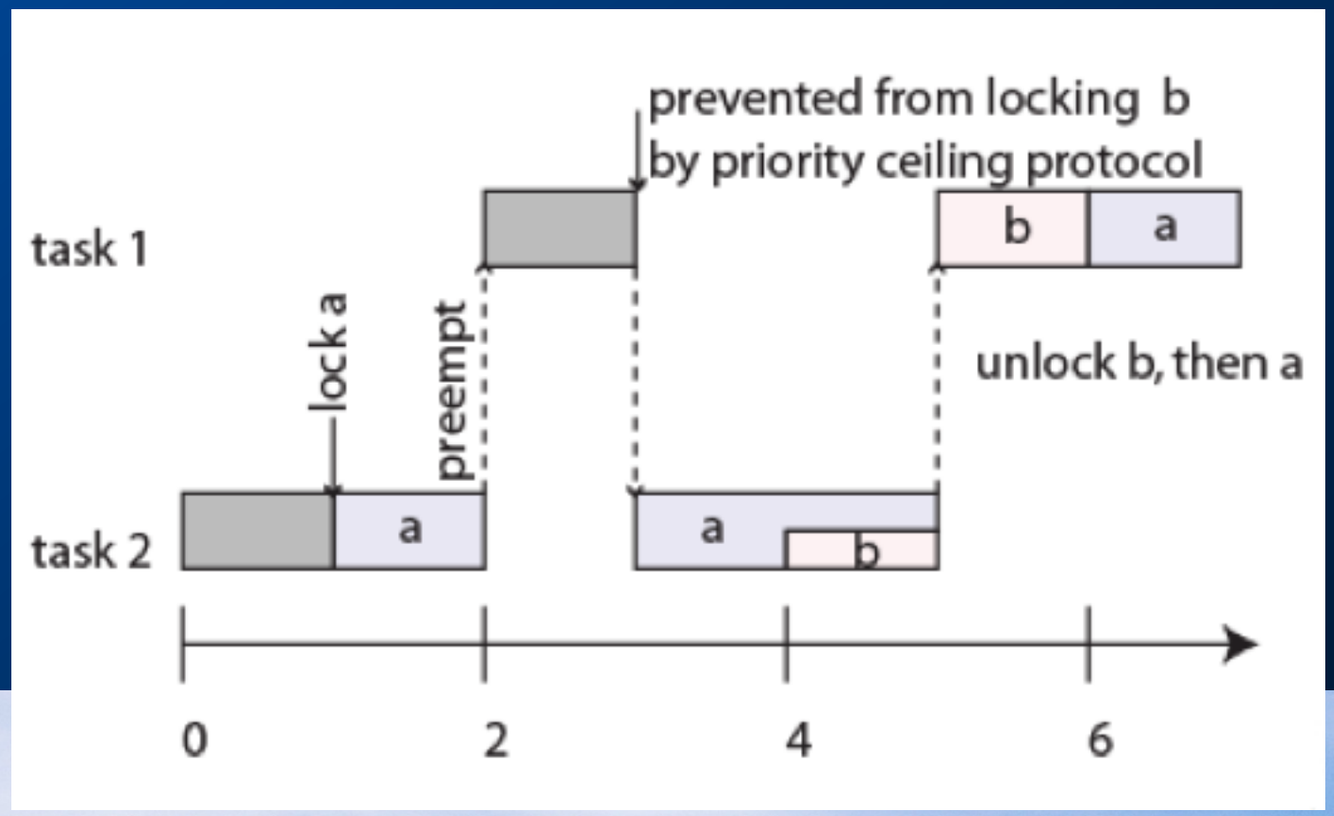
- In OCPP, a task X's priority is raised when a higher-priority task Y tries to acquire a resource that X has locked. The task's priority is then raised to the priority ceiling of the resource, ensuring that task X quickly finishes its critical section, unlocking the resource. A task is only allowed to lock a resource if its dynamic priority is higher than the priority ceilings of all resources locked by other tasks. Otherwise the task becomes blocked, waiting for the resource.
- In ICPP, a task's priority is immediately raised when it locks a resource. The task's priority is set to the priority ceiling of the resource, thus no task that may lock the resource is able to get scheduled.

ICPP versus OCPP

- The worst-case behaviour of the two ceiling schemes is identical from a scheduling view point. However, there are some differences:
  - ICPP is easier to implement than OCPP, as blocking relationships need not be monitored
  - ICPP leads to fewer context switches as blocking is prior to first execution
  - ICPP requires more priority movements as this happens with all resource usage
  - OCPP changes priority only if an actual block has occurred

Priority Ceiling Protocol

In this version, locks a and b have priority ceilings equal to the priority of task 1. At time 3, task 1 attempts to lock b, but it can't because task 2 currently holds lock a, which has priority ceiling equal to the priority of task 1.

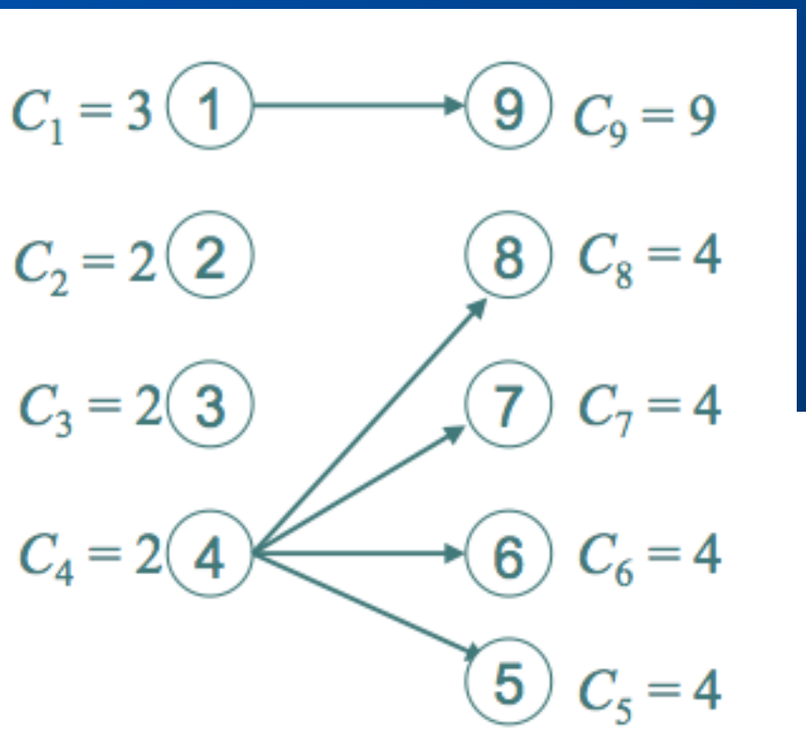



```
#include <pthread.h>
...
pthread_mutex_t lock_a, lock_b;
void* thread_1_function(void* arg) {
    pthread_mutex_lock(&lock_b);
    ...
    pthread_mutex_lock(&lock_a);
    ...
    pthread_mutex_unlock(&lock_a);
    ...
    pthread_mutex_unlock(&lock_b);
    ...
}
void* thread_2_function(void* arg) {
    pthread_mutex_lock(&lock_a);
    ...
    pthread_mutex_lock(&lock_b);
    ...
    pthread_mutex_unlock(&lock_b);
    ...
    pthread_mutex_unlock(&lock_a);
    ...
}
```

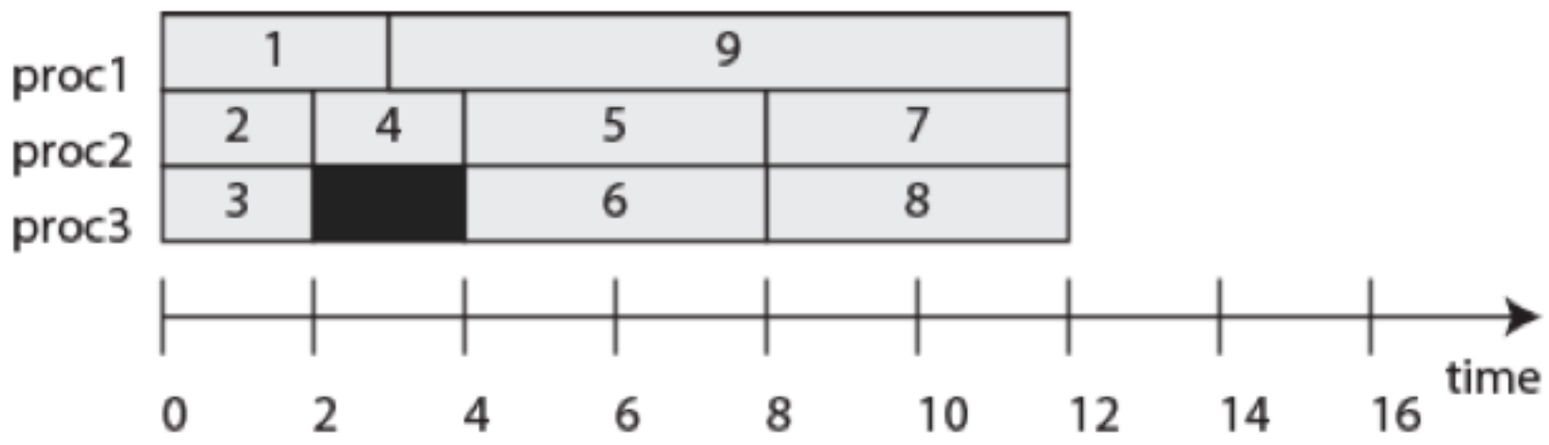
Brittleness

- In general, all thread scheduling algorithms are brittle: Small changes can have big, unexpected consequences.
- I will illustrate this with multiprocessor (or multicore) schedules.
- *Theorem (Richard Graham, 1976): If a task set with fixed priorities, execution times, and precedence constraints is scheduled according to priorities on a fixed number of processors, then increasing the number of processors, reducing execution times, or weakening precedence constraints can increase the schedule length.*

Richard's Anomalies

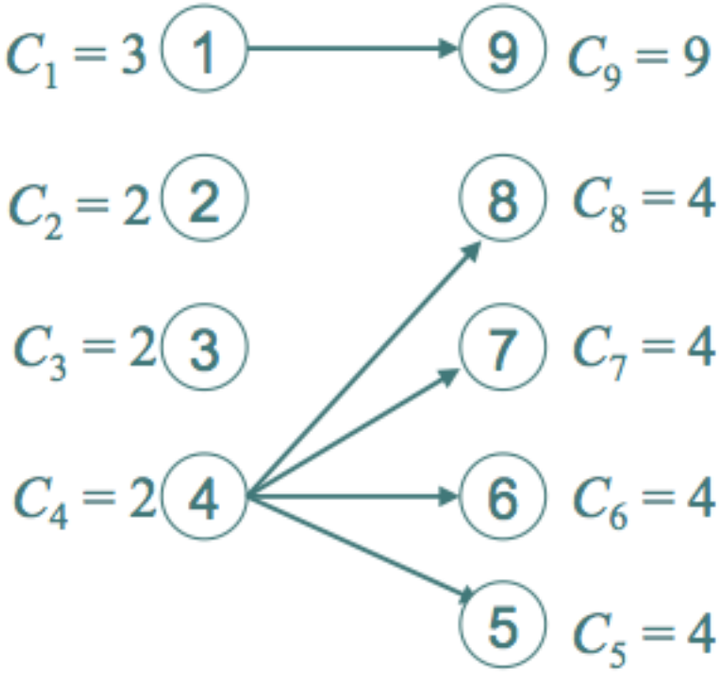


9 tasks with precedences and the shown execution times, where lower numbered tasks have higher priority than higher numbered tasks. Priority-based 3 processor schedule:

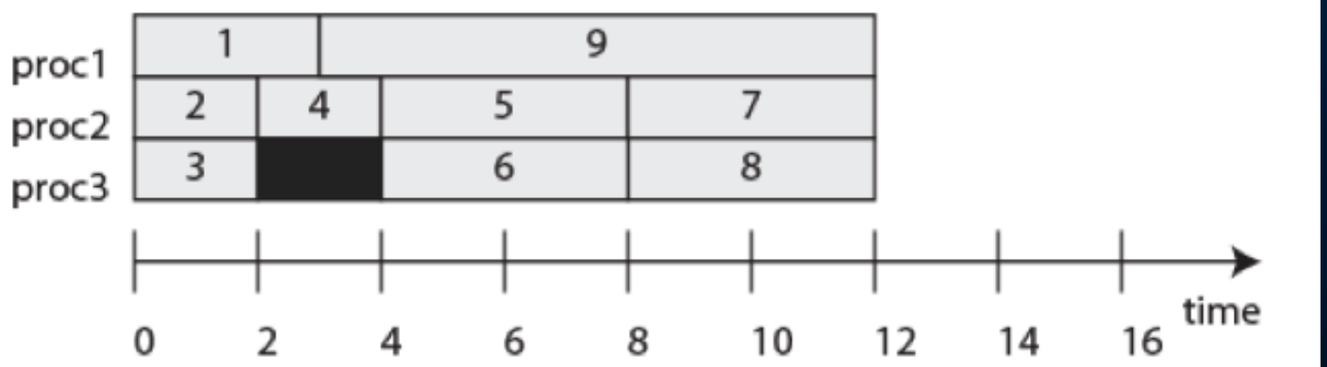


What happens if you increase the number of processors to four?

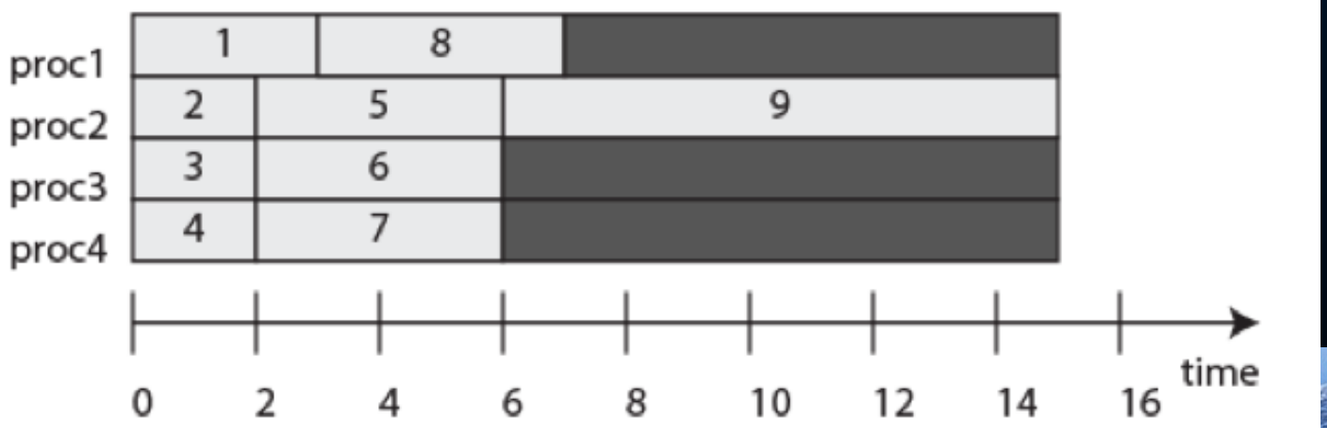
Richard's Anomalies: Increasing the number of processors



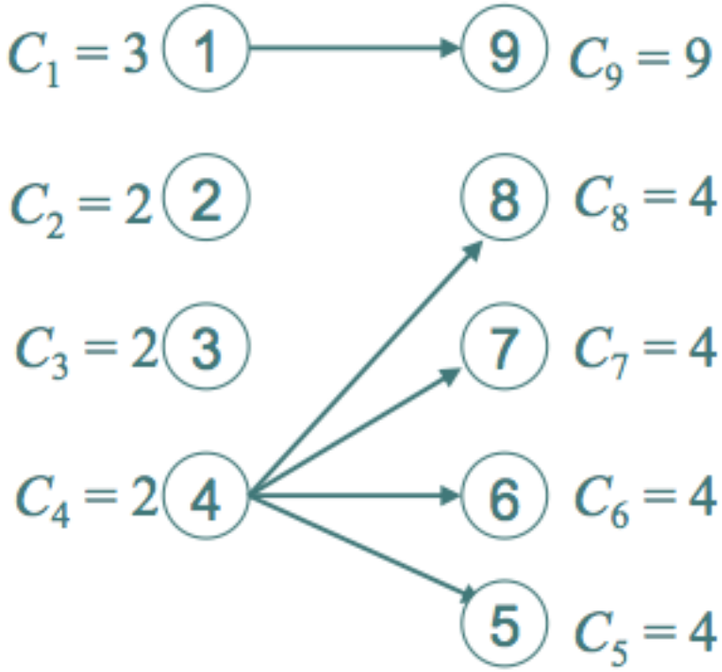
9 tasks with precedences and the shown execution times, where lower numbered tasks have higher priority than higher numbered tasks. Priority-based 3 processor schedule:



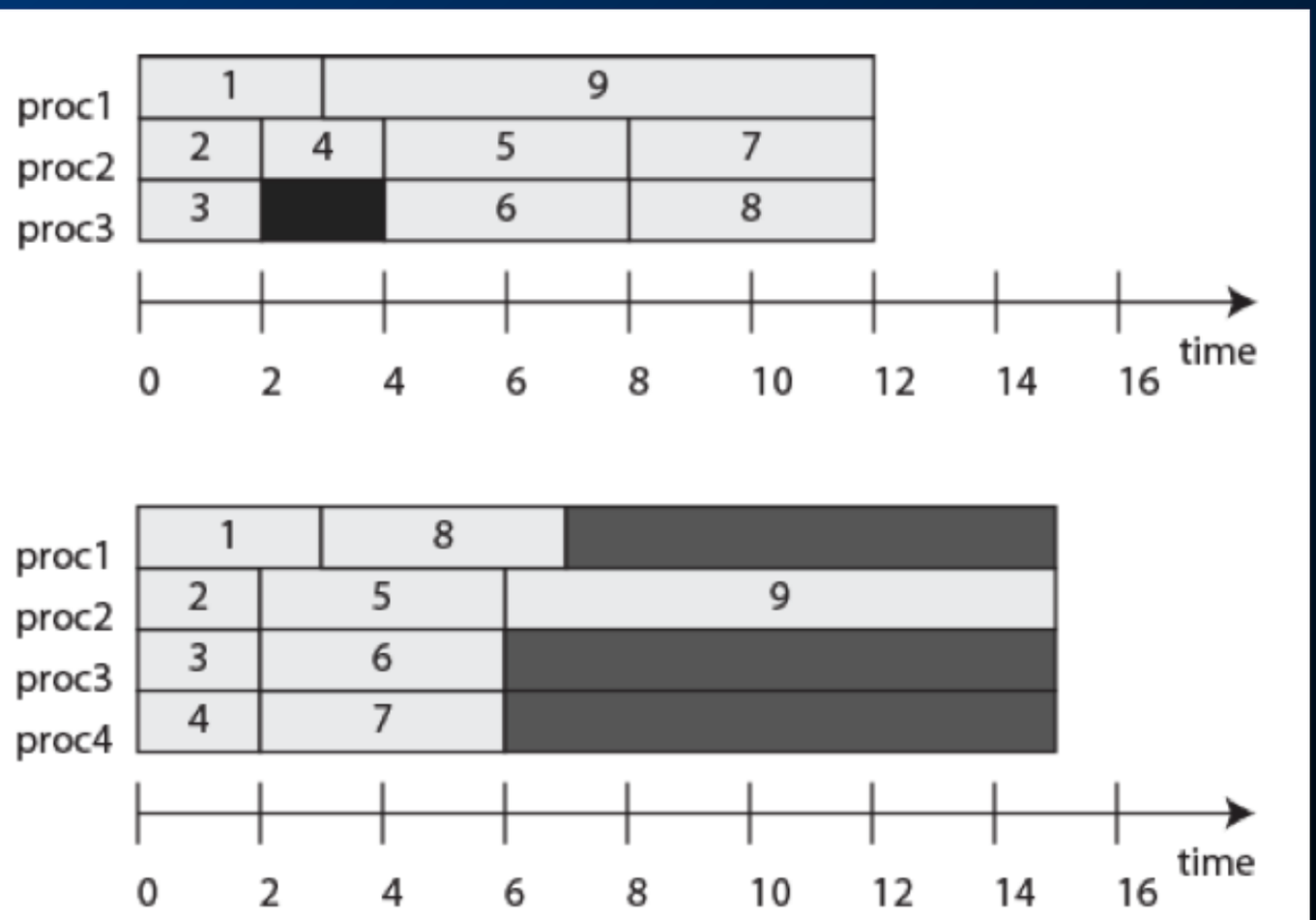
The priority-based schedule with four processors has a longer execution time.



Greedy Scheduling

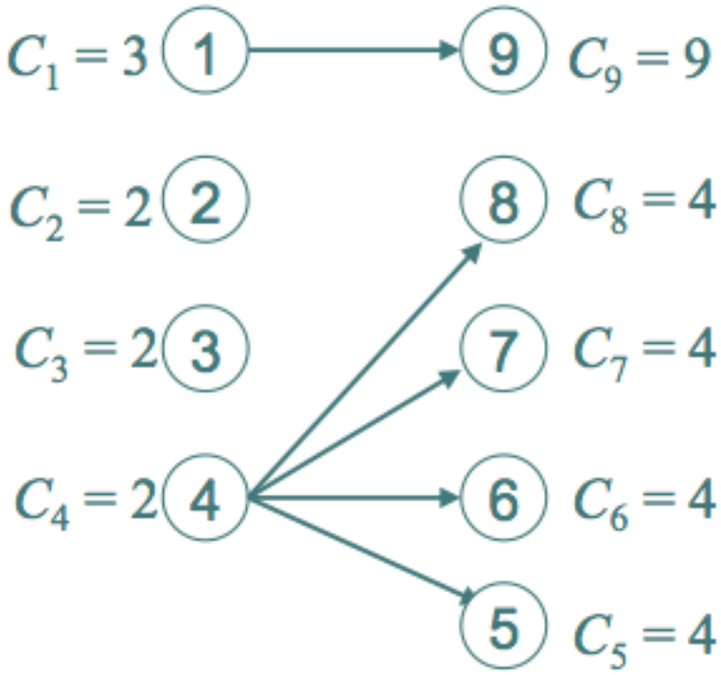


9 tasks with precedences and the shown execution times, where lower numbered tasks have higher priority than higher numbered tasks. Priority-based 3 processor schedule:

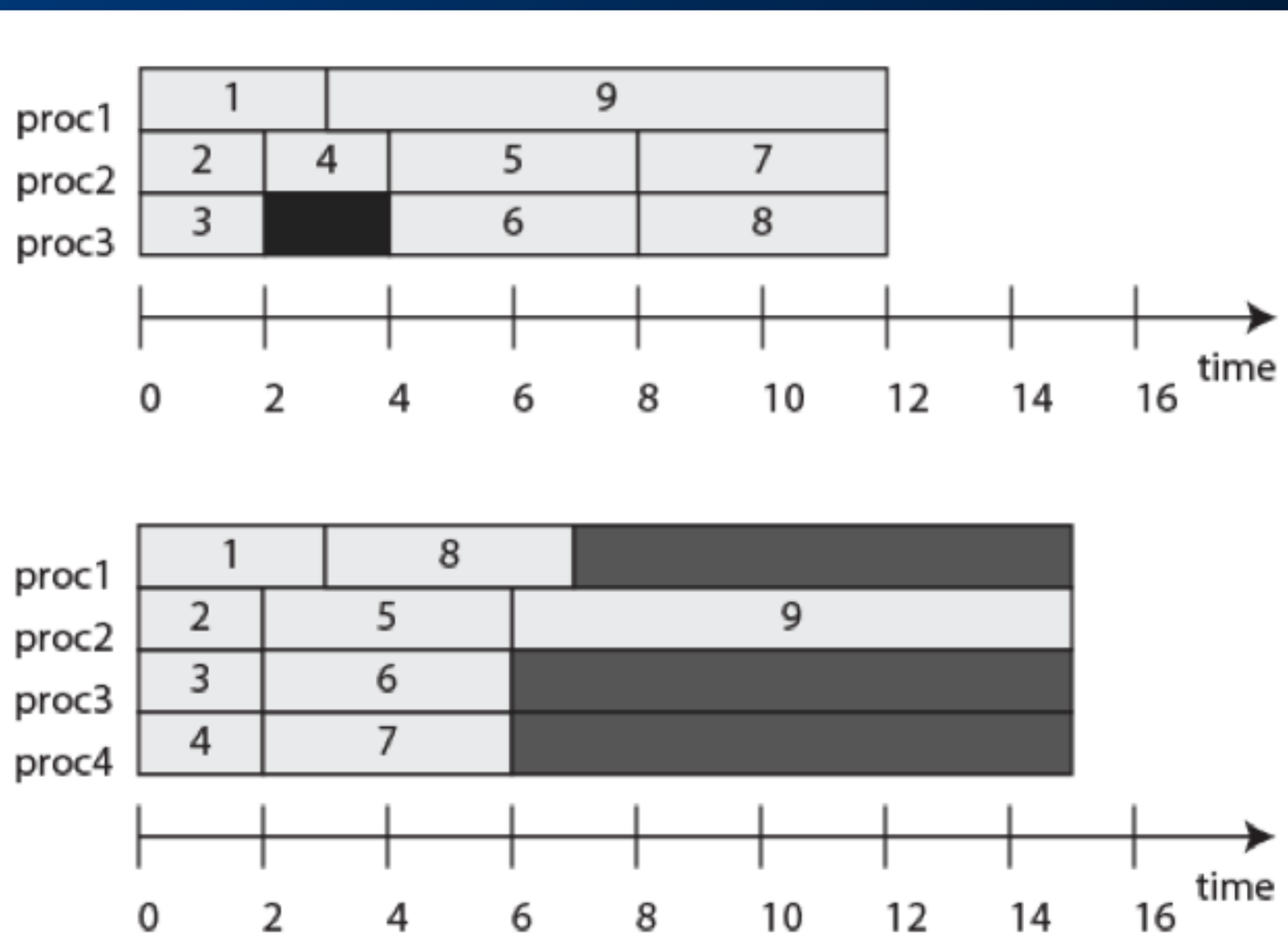


Priority-based scheduling is "greedy." A smarter scheduler for this example could hold off scheduling 5, 6, or 7, leaving a processor idle for one time unit.

Greedy scheduling may be the only practical option.

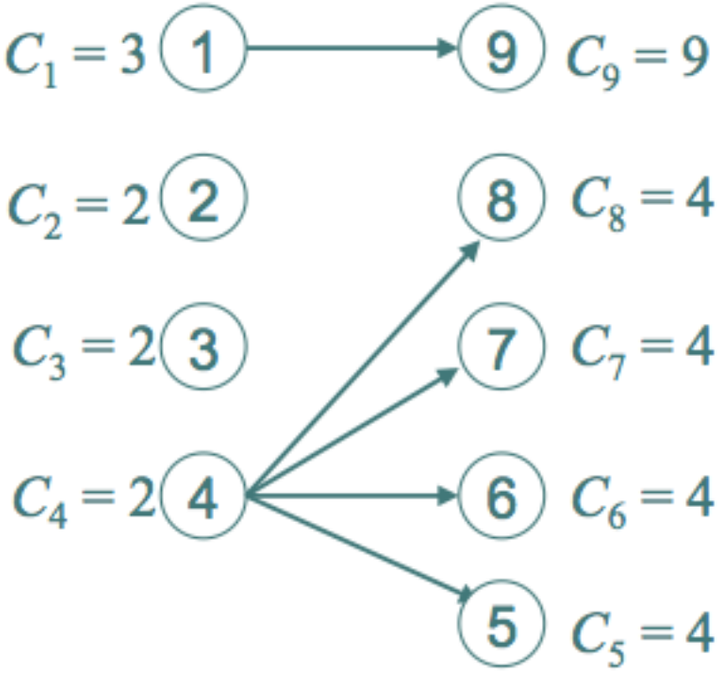


9 tasks with precedences and the shown execution times, where lower numbered tasks have higher priority than higher numbered tasks. Priority-based 3 processor schedule:

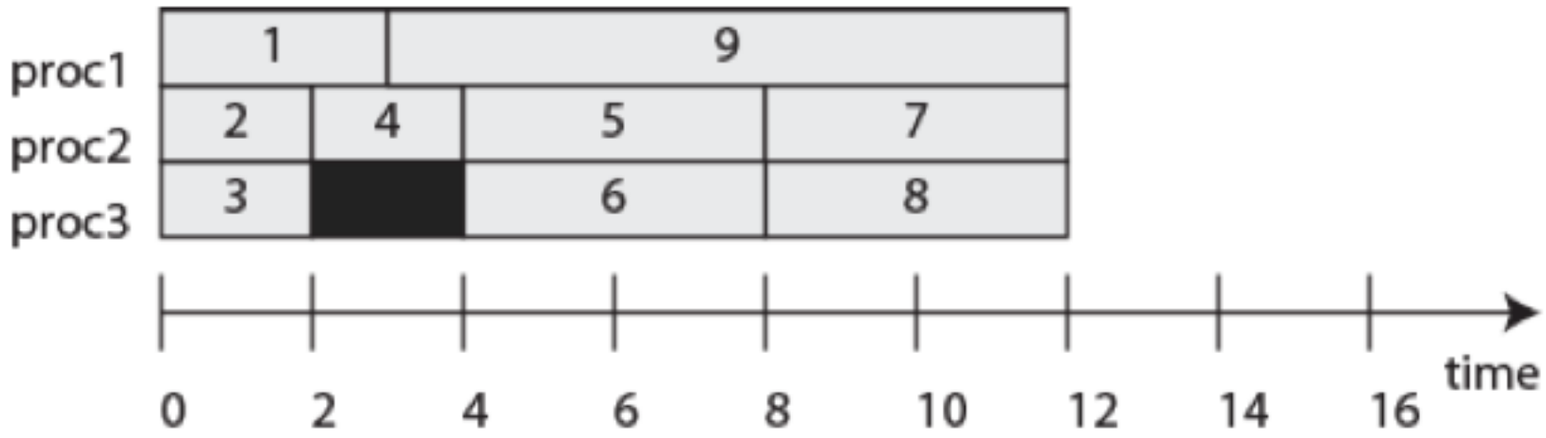


If tasks “arrive” (become known to the scheduler) only after their predecessor completes, then greedy scheduling may be the only practical option.

Richard's Anomalies

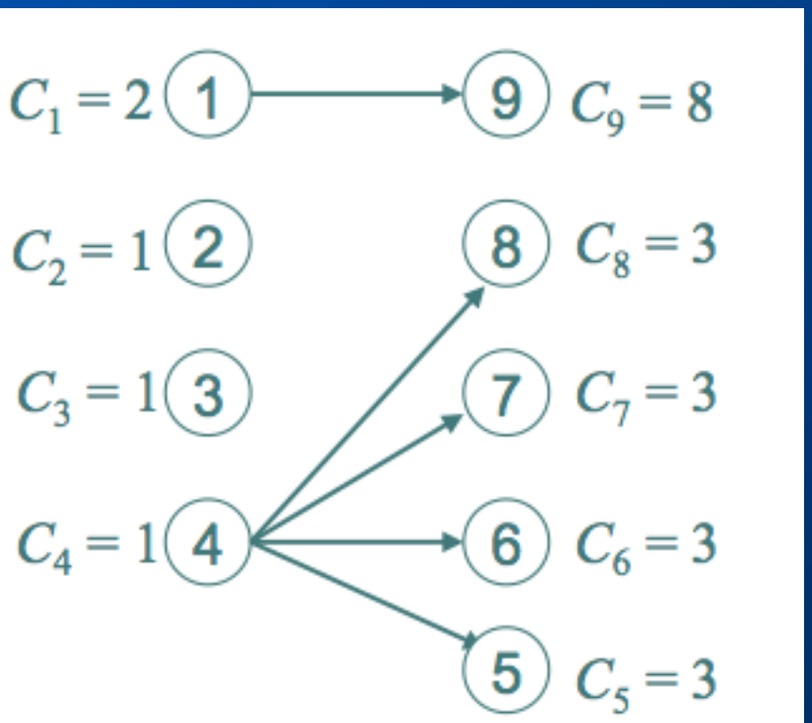


9 tasks with precedences and the shown execution times, where lower numbered tasks have higher priority than higher numbered tasks. Priority-based 3 processor schedule:

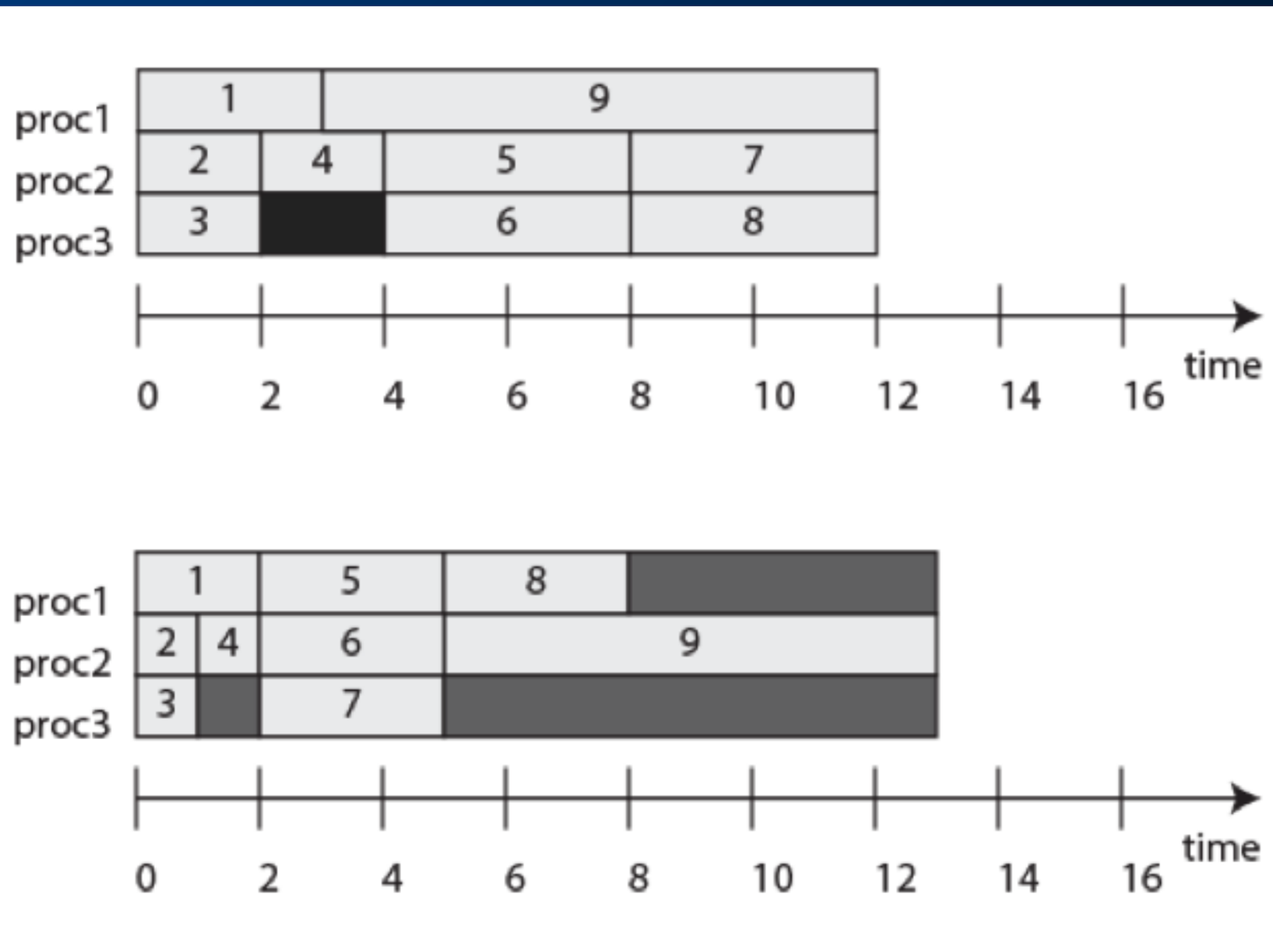


What happens if you reduce all computation times by 1?

Richard's Anomalies: Reducing computation times

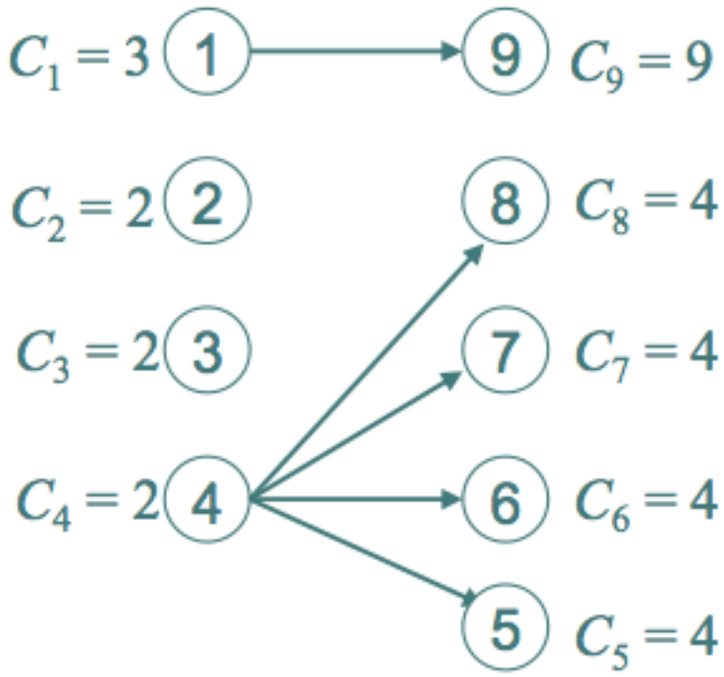


9 tasks with precedences and the shown execution times, where lower numbered tasks have higher priority than higher numbered tasks. Priority-based 3 processor schedule:

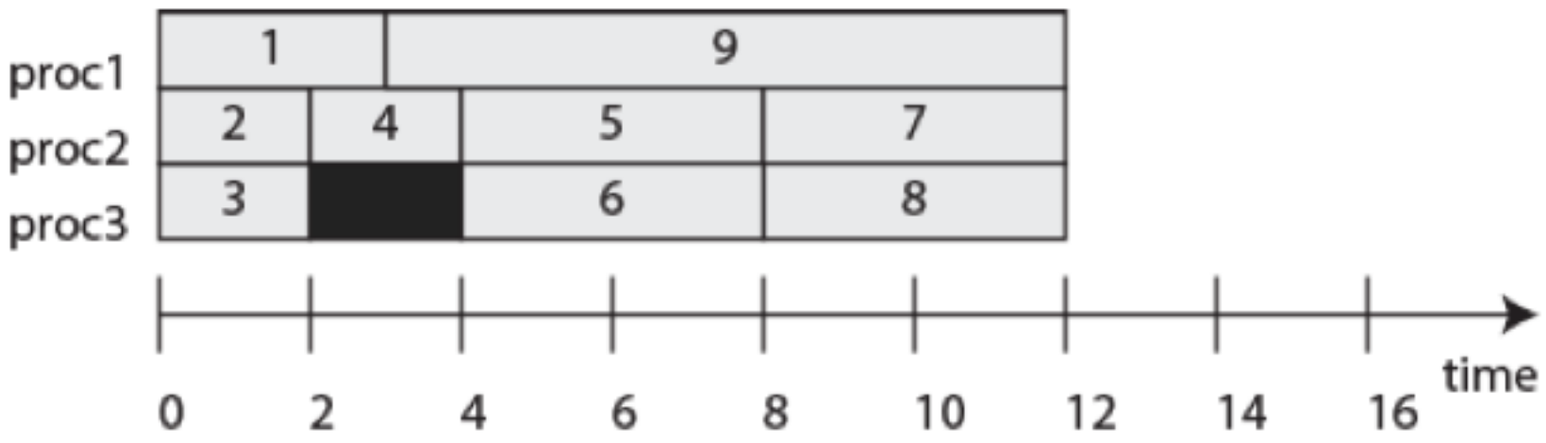


Reducing the computation times by 1 also results in a longer execution time.

Richard's Anomalies

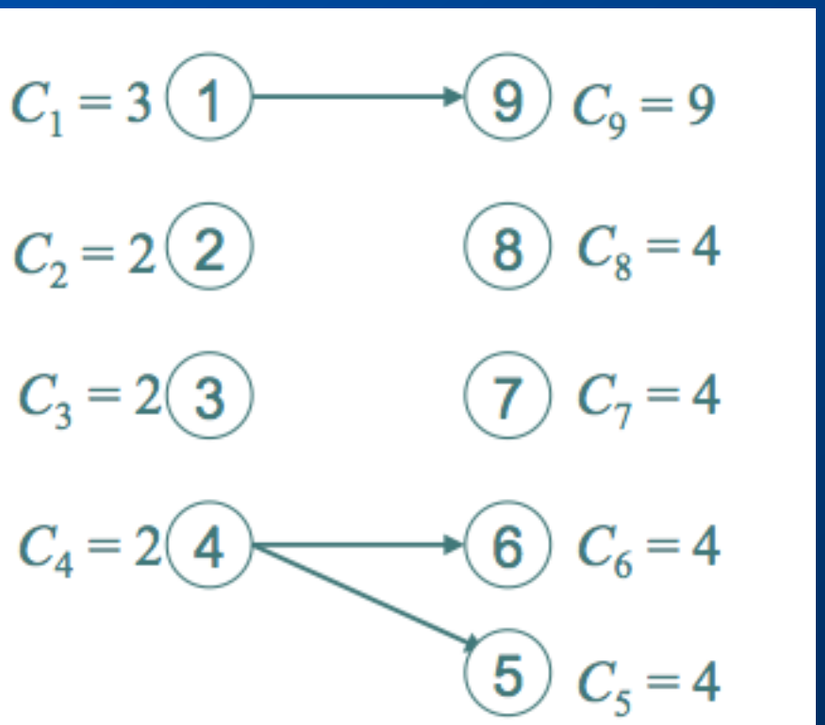


9 tasks with precedences and the shown execution times, where lower numbered tasks have higher priority than higher numbered tasks. Priority-based 3 processor schedule:

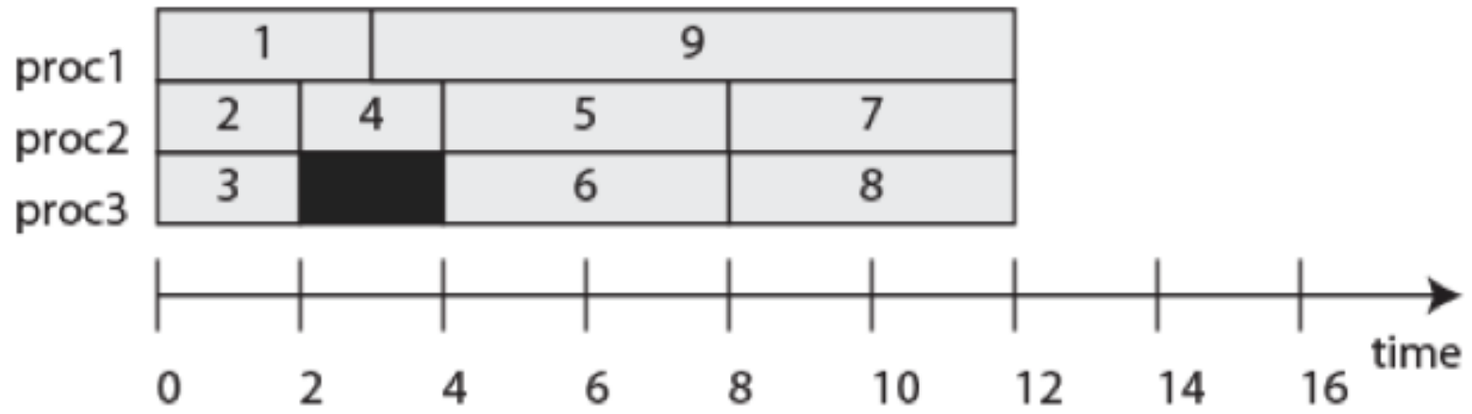


What happens if you remove the precedence constraints (4,8) and (4,7)?

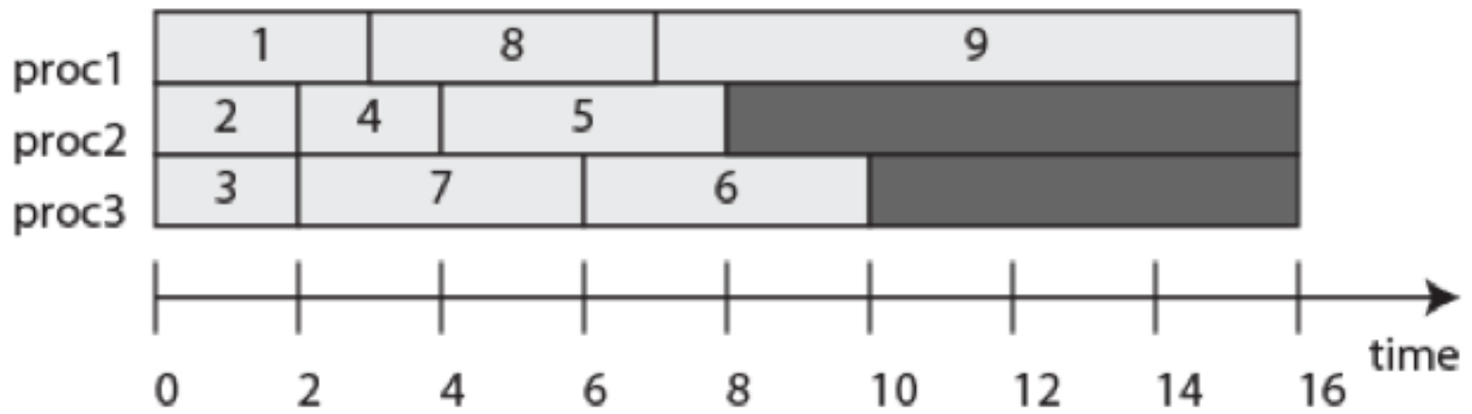
Richard's Anomalies: Weakening the precedence constraints



9 tasks with precedences and the shown execution times, where lower numbered tasks have higher priority than higher numbered tasks. Priority-based 3 processor schedule:

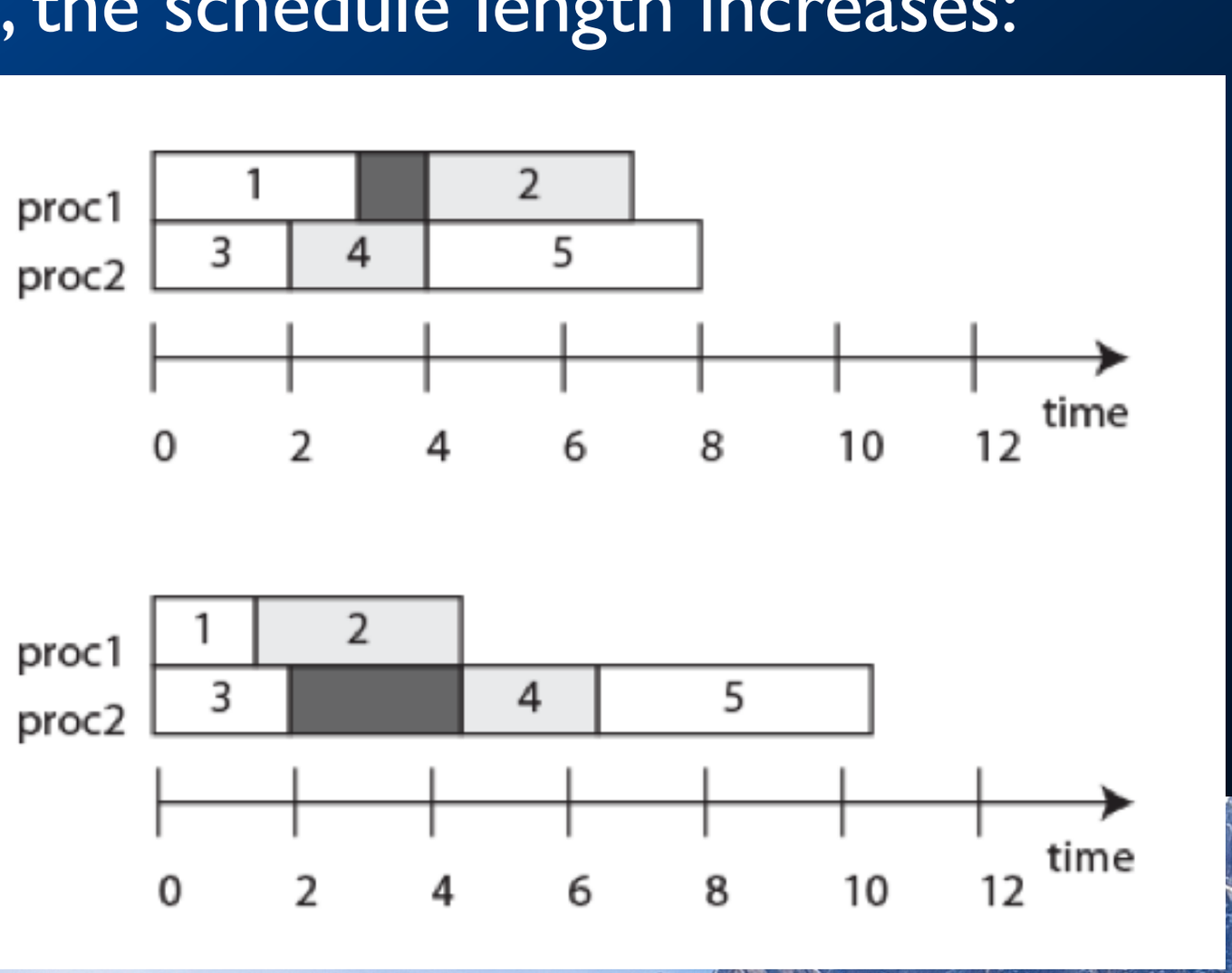


Weakening precedence constraints can also result in a longer schedule.



Richard's Anomalies with Mutexes: Reducing Execution Time

- Assume tasks 2 and 4 share the same resource in exclusive mode, and tasks are statically allocated to processors. Then if the execution time of task 1 is reduced, the schedule length increases:



Conclusion

- Timing behavior under all known task scheduling strategies is brittle. Small changes can have big (and unexpected) consequences.
- Unfortunately, since execution times are so hard to predict, such brittleness can result in unexpected system failures.

References

- Edward Ashford Lee, Sanjit Arunkumar Seshia. Introduction to Embedded Systems: A Cyber-Physical Systems Approach, chapter 11. Lulu.com. (嵌入式系统导论：CPS方法)

扩展阅读

- https://gitee.com/openharmony/docs/blob/master/zh-cn/OpenHarmony-Overview_zh.md
- https://www.openeuler.org/zh/learn/mooc/detail/?id=1
- https://playground.harmonyos.com/#/cn/onlineDemo

Thank you!