Computer Architecture

04. 流水线基础和性能分析

内容来自张晨曦教材

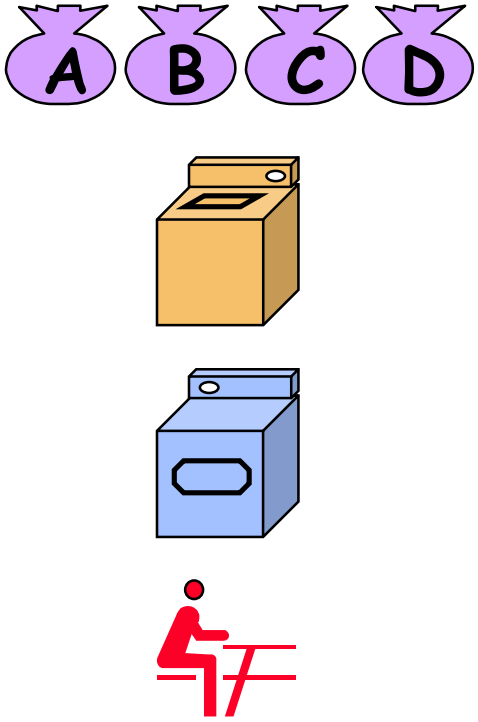
Jianhua Li
College of Computer and Information
Hefei University of Technology

内容概要

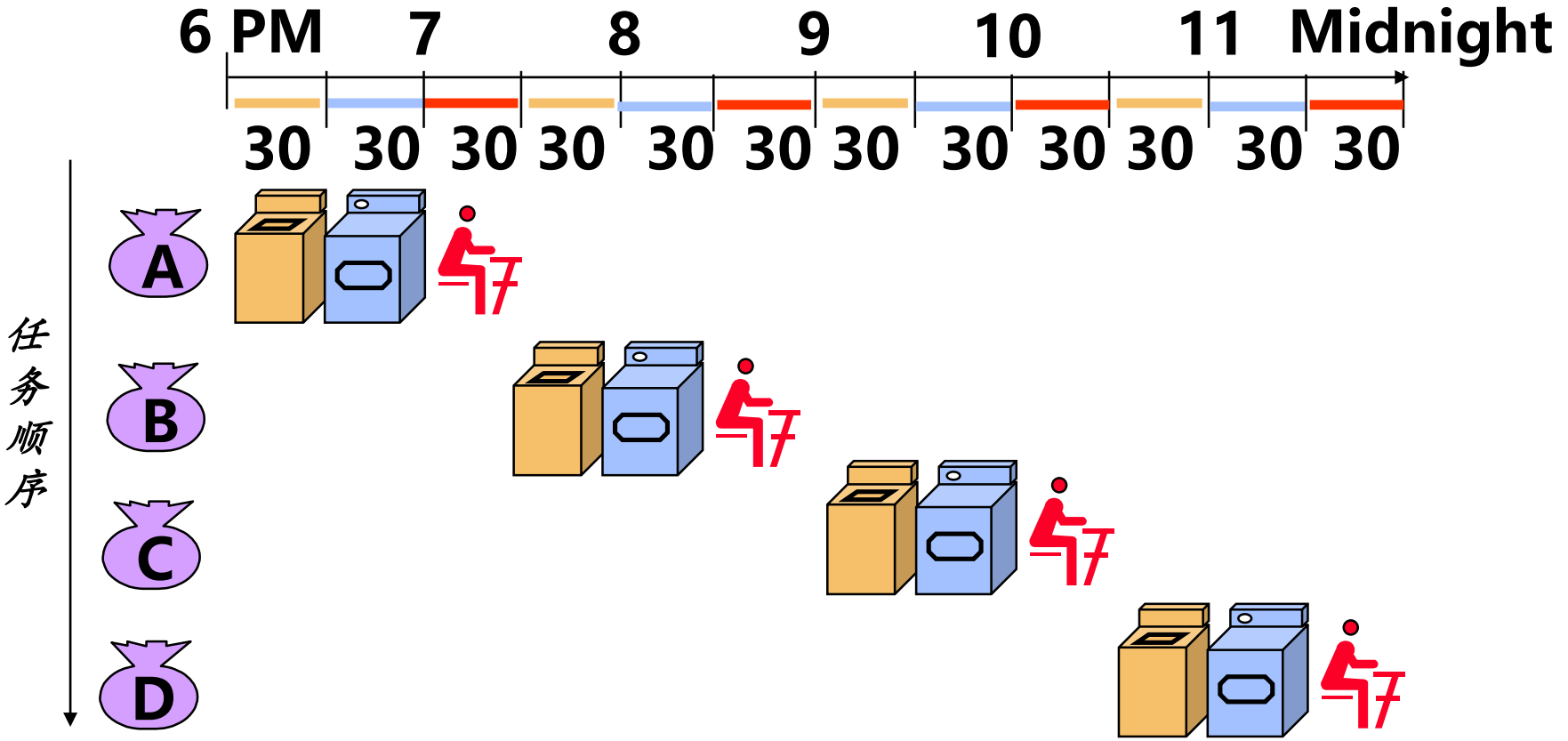
- 流水线基础
  - 流水线概要
  - 时空图表示
  - 流水线分类
- 流水线的性能分析
  - 流水线的吞吐率
  - 流水线的加速比
  - 流水线的效率

流水线基础：部件耗时相等的洗衣店

- A, B, C, D need to wash, dry, and fold clothes
- Washer takes 30 minutes
- Dryer takes 30 minutes
- Folder takes 30 minutes

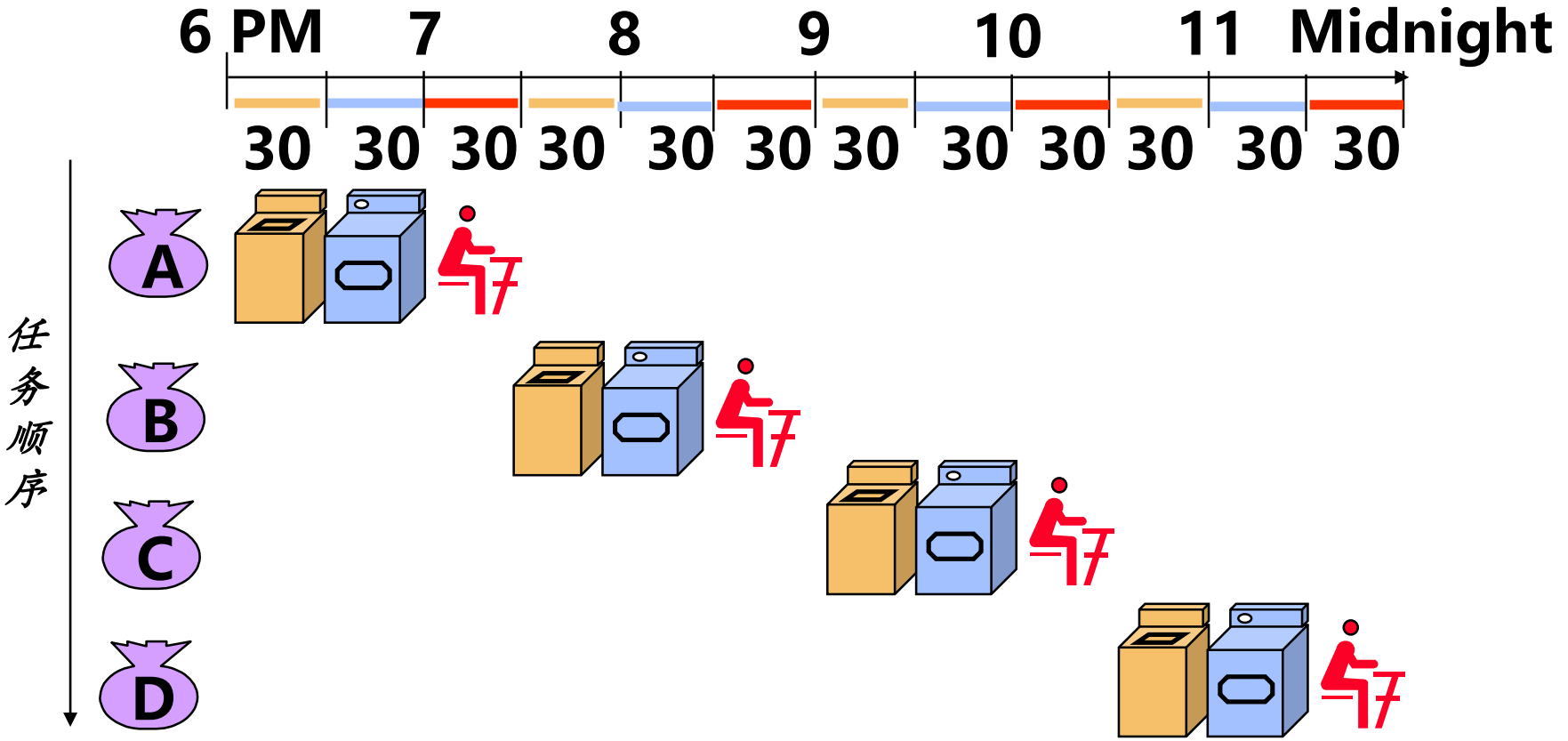


串行工作的洗衣店



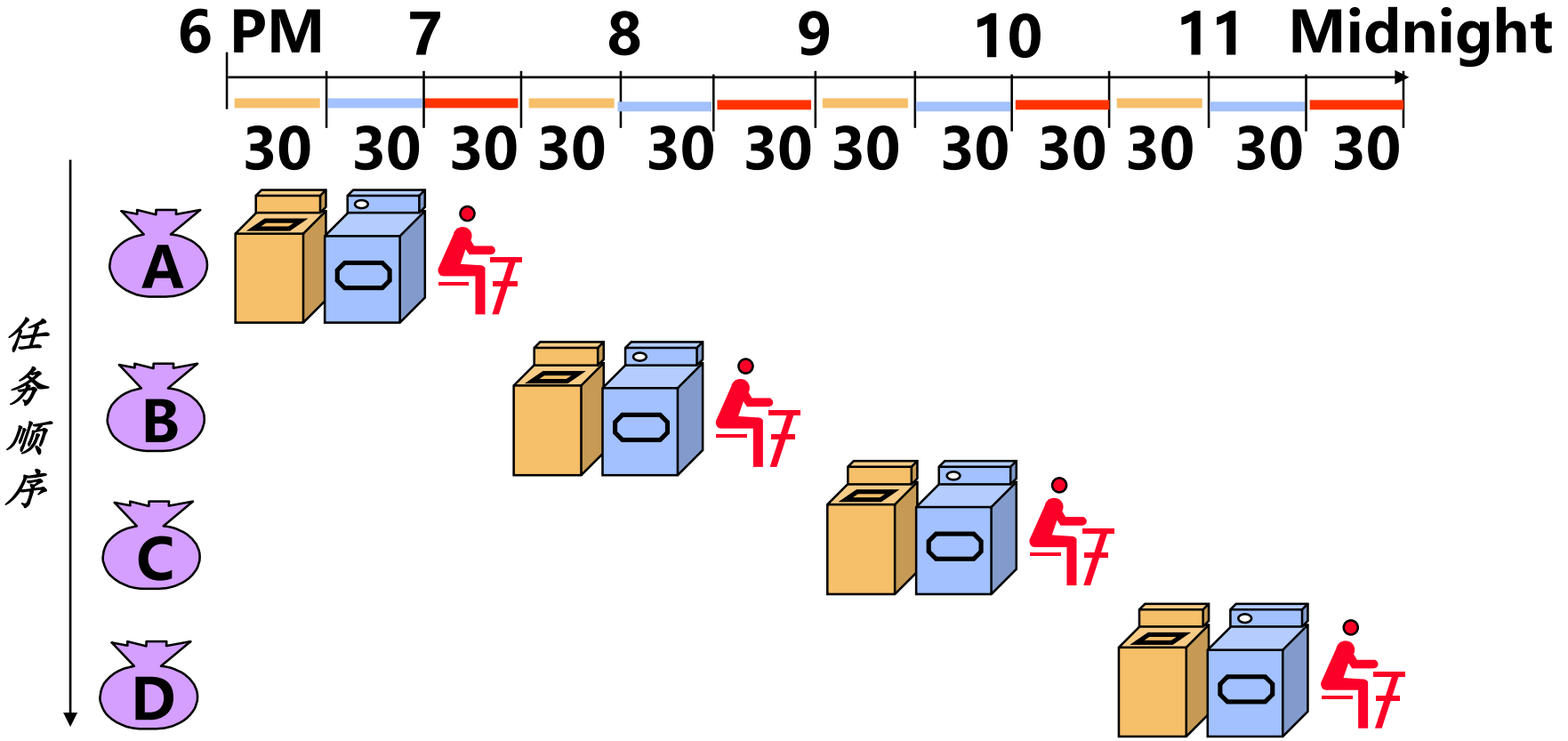
- 洗衣店用 6小时完成了4个任务（0.67t/h）；

串行工作的洗衣店



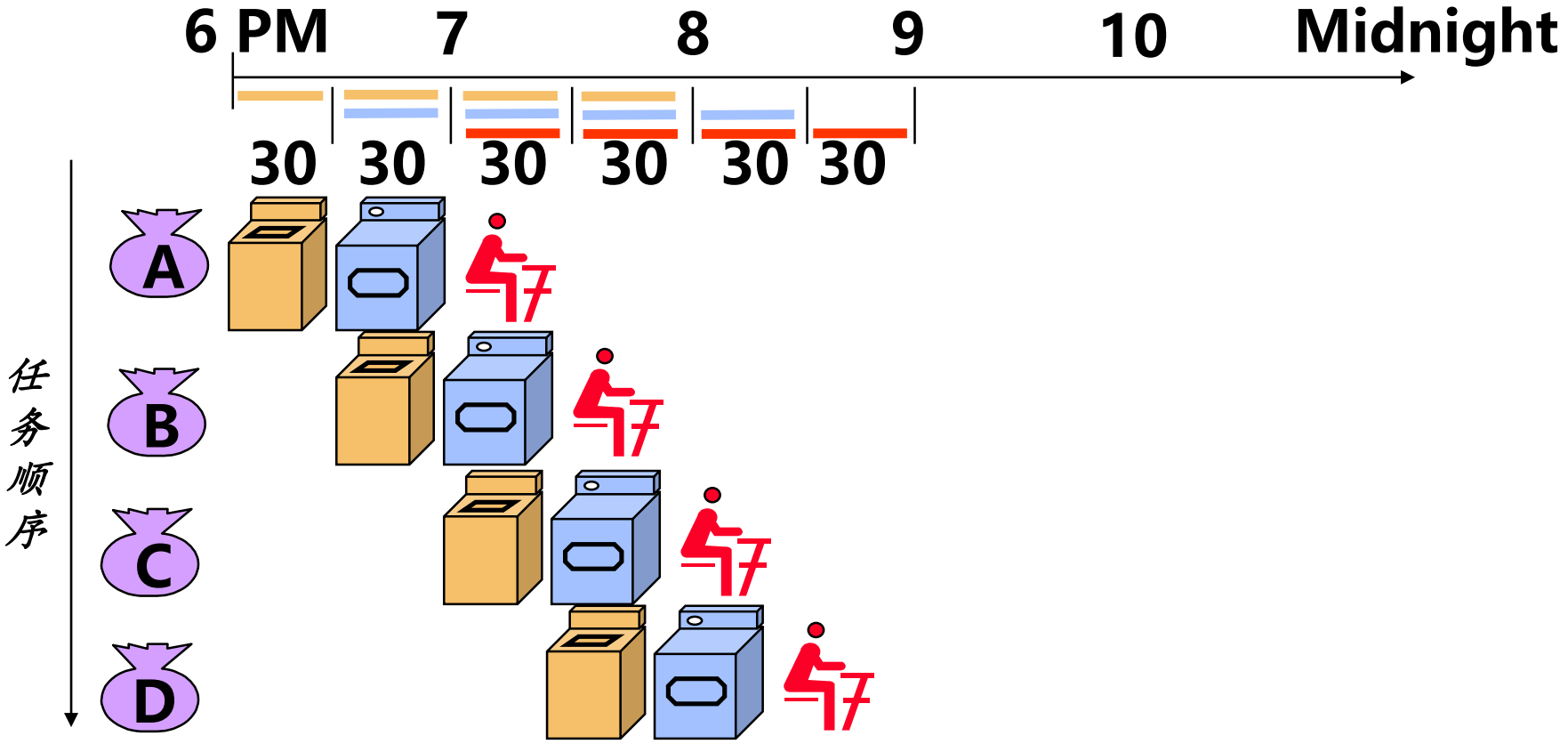
- 洗衣店用 6小时完成了4个任务（0.67t/h）；
- 4个同学各等待了1.5小时；

串行工作的洗衣店



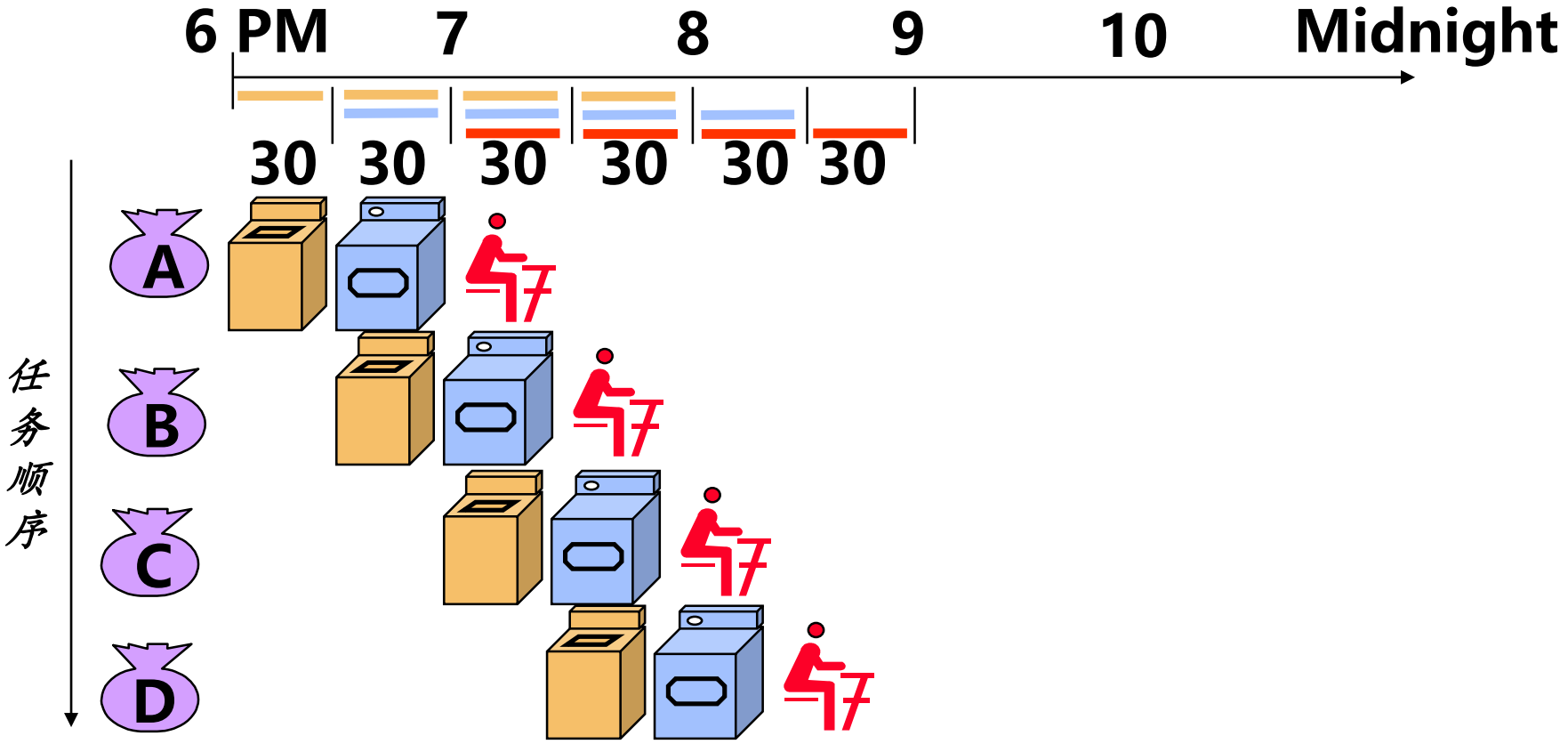
- 洗衣店用 6小时完成了4个任务（0.67t/h）；
- 4个同学各等待了1.5小时；
- Washer使用2小时；Dryer使用2小时；Folder使用2小时；

流水工作的洗衣店



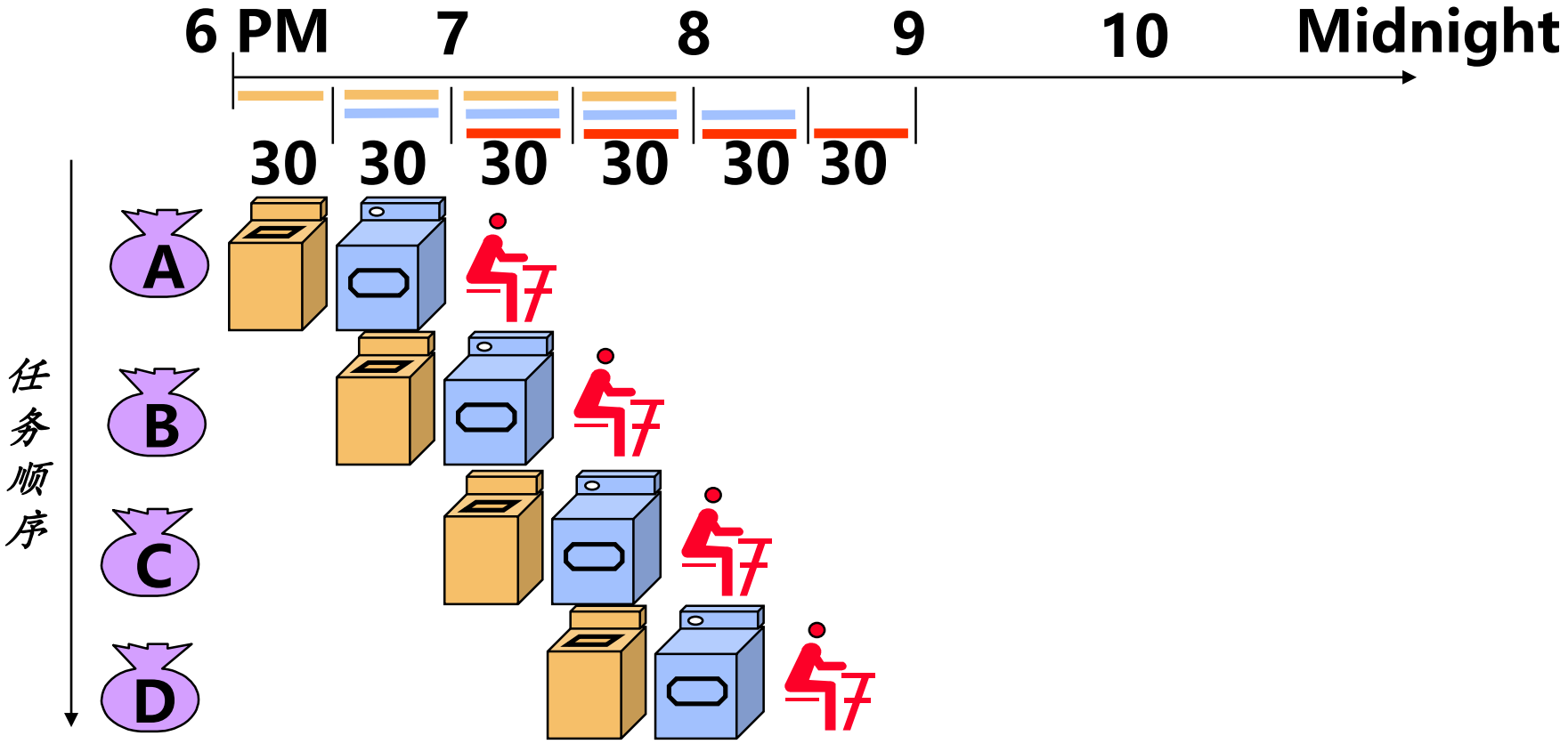
- 洗衣店用3小时完成了4个任务（1.33t/h）；

流水工作的洗衣店



- 洗衣店用3小时完成了4个任务（1.33t/h）；
- 4个同学各等待了1.5小时；

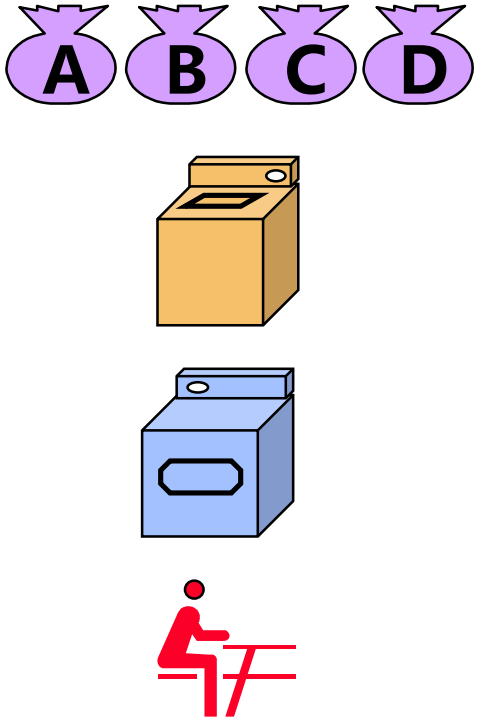
流水工作的洗衣店



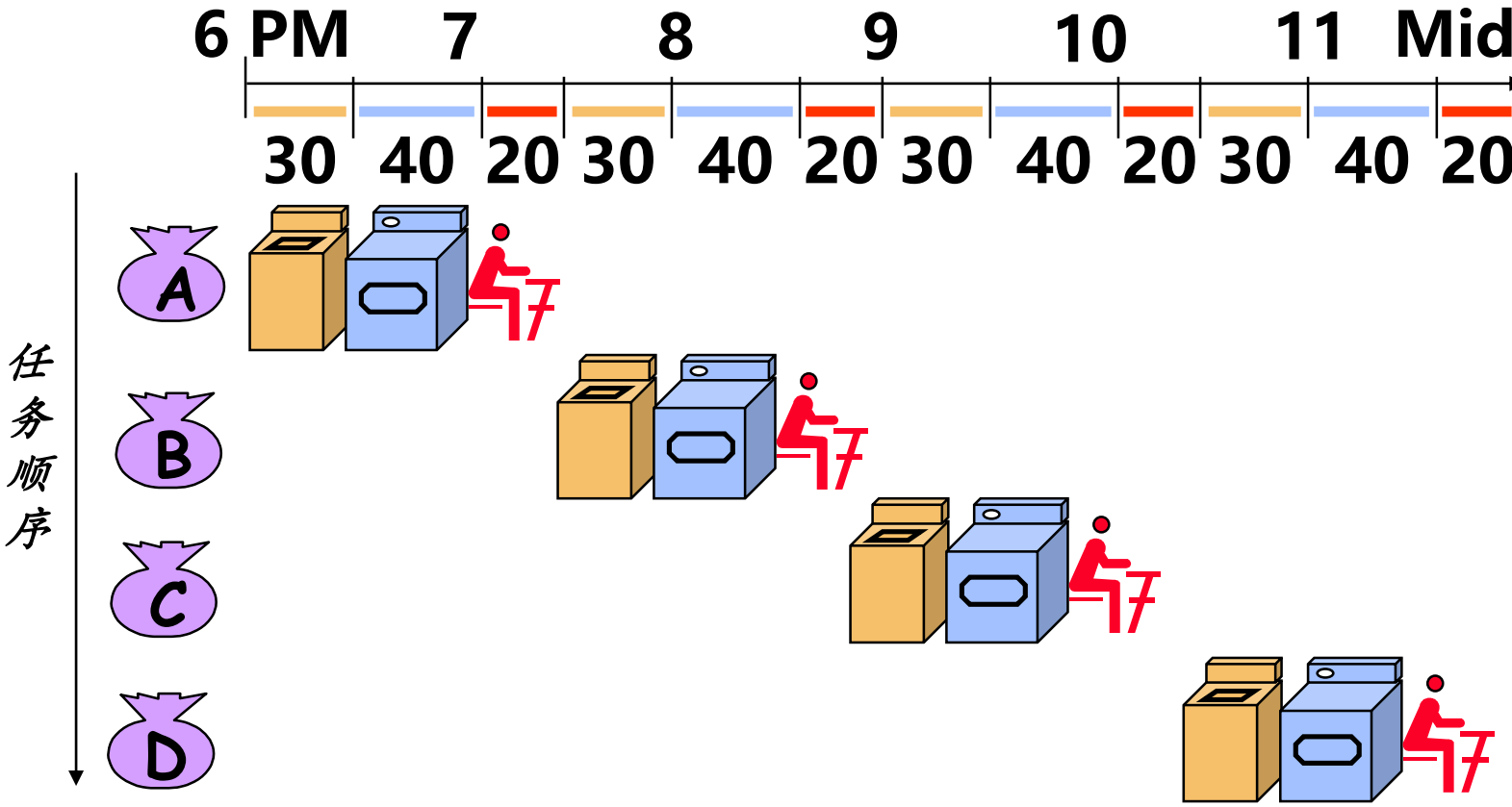
- 洗衣店用3小时完成了4个任务（1.33t/h）；
- 4个同学各等待了1.5小时；
- Washer使用2小时；Dryer使用2小时；Folder使用2小时；

流水线基础：部件耗时不等的洗衣店

- A, B, C, D need to wash, dry, and fold;
- Washer takes 30 minutes
- Dryer takes 40 minutes
- Folder takes 20 minutes

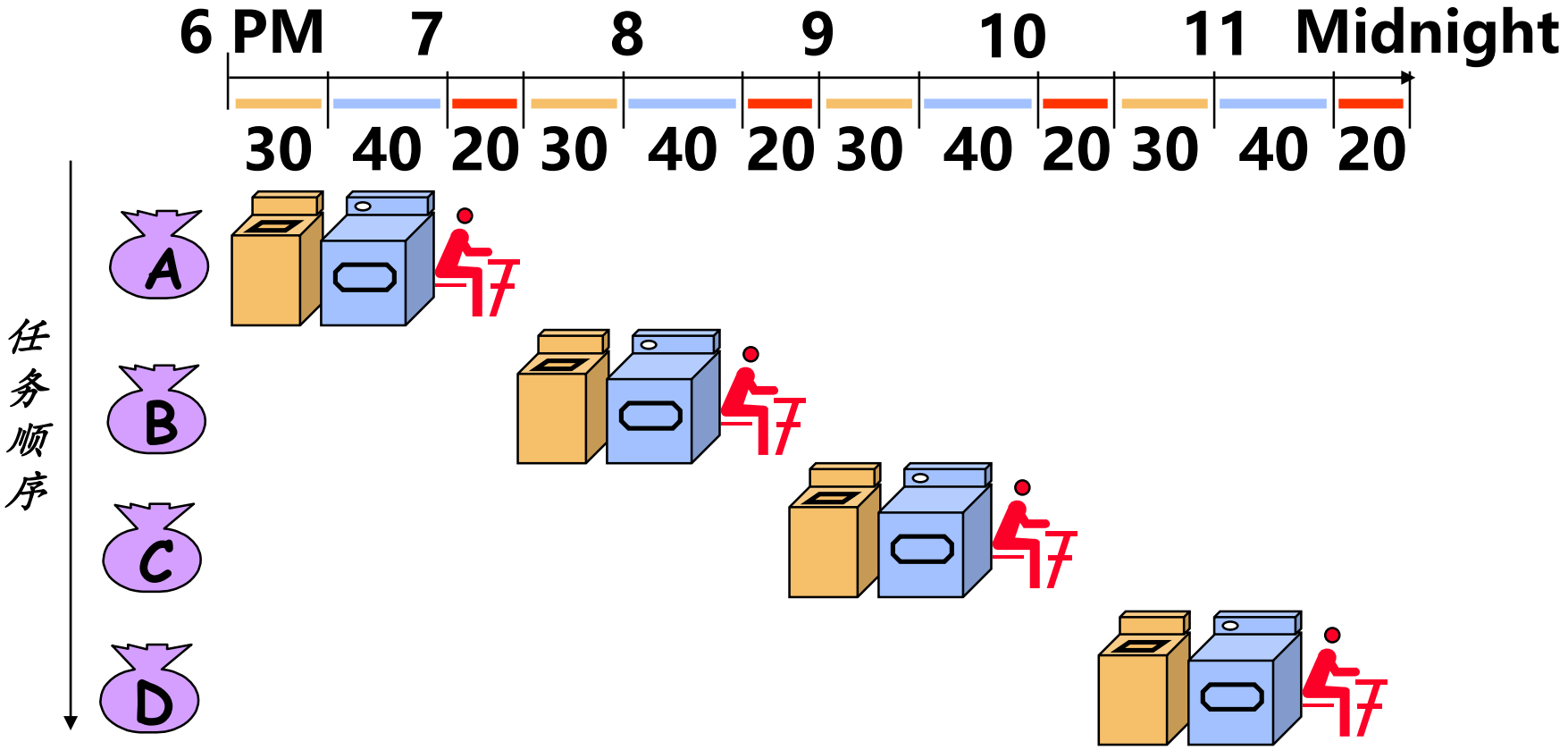


串行工作的洗衣店



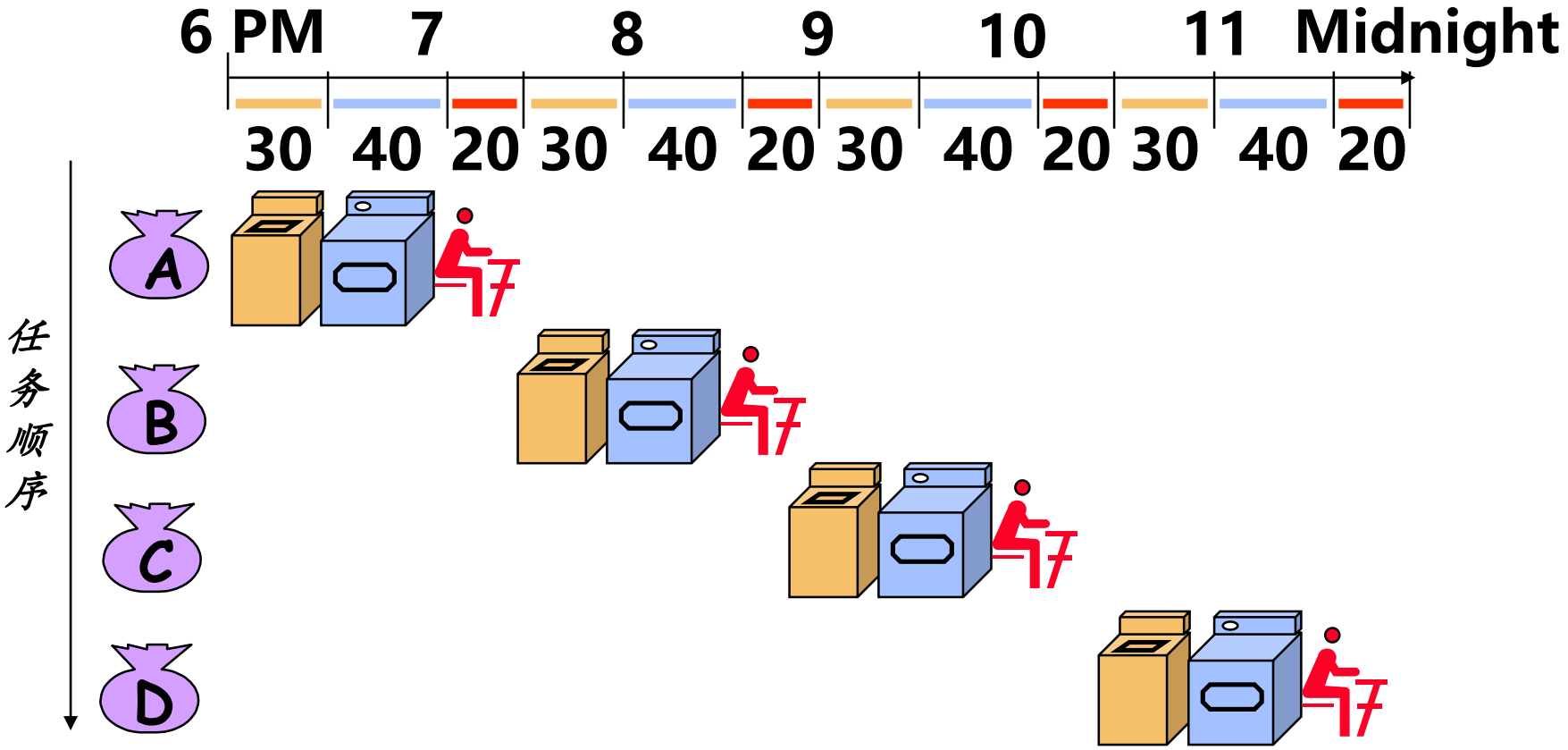
- 洗衣店用 6小时完成了4个任务（0.67t/h）；

串行工作的洗衣店



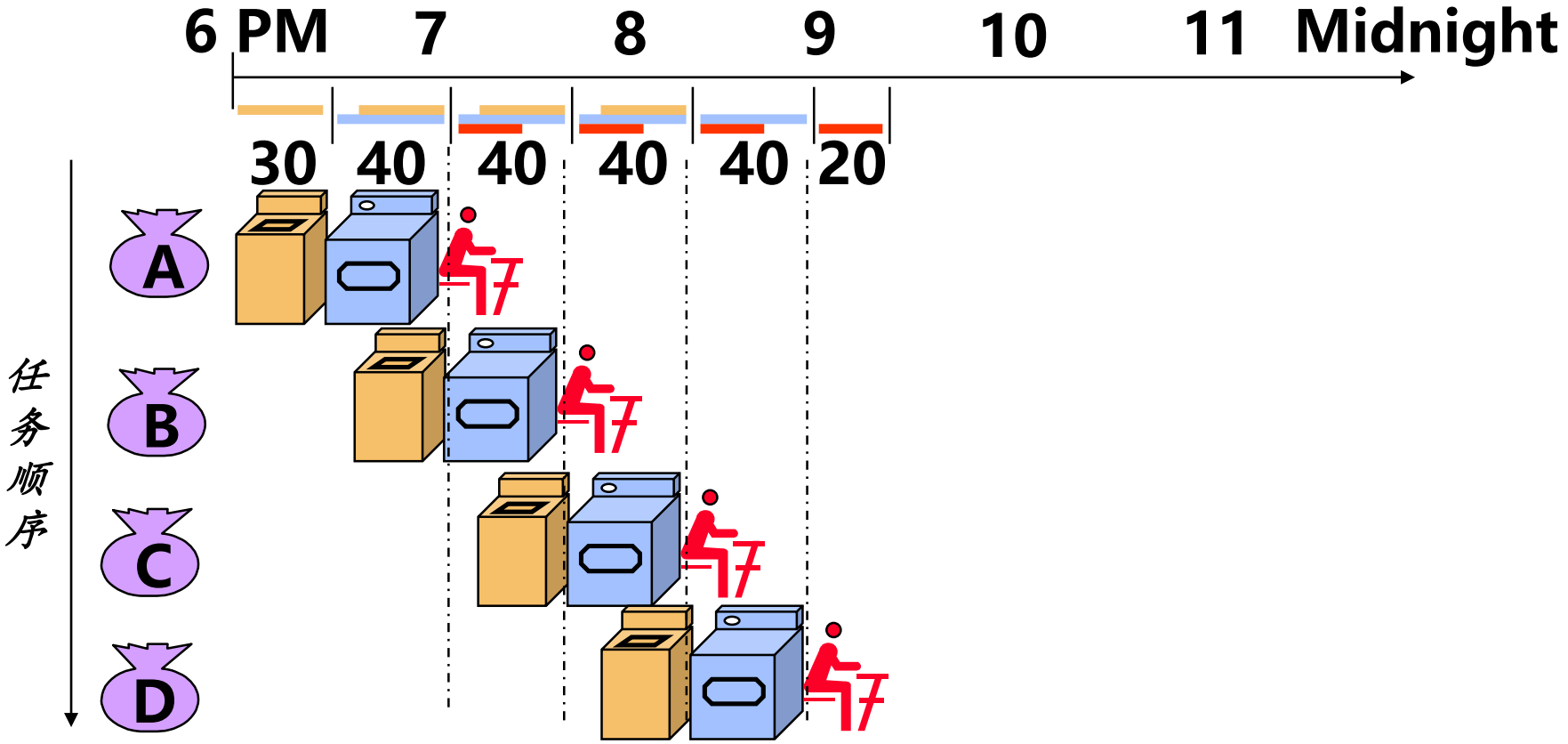
- 洗衣店用 6小时完成了4个任务（0.67t/h）；
- 4个同学各等待了1.5小时；

串行工作的洗衣店



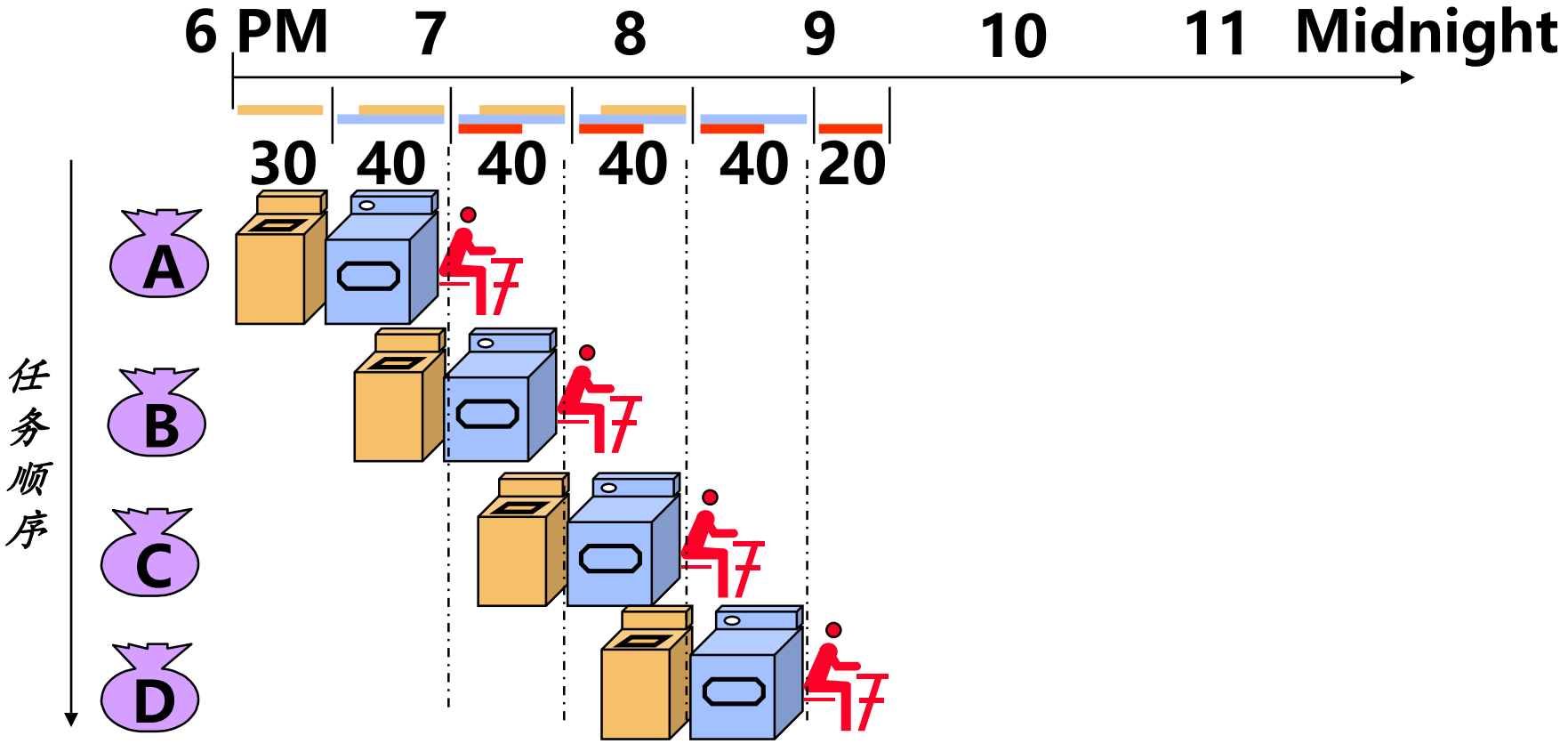
- 洗衣店用 6小时完成了4个任务（0.67t/h）；
- 4个同学各等待了1.5小时；
- Washer使用2小时；Dryer使用2小时40分；Folder使用1小时20分；

流水工作的洗衣店



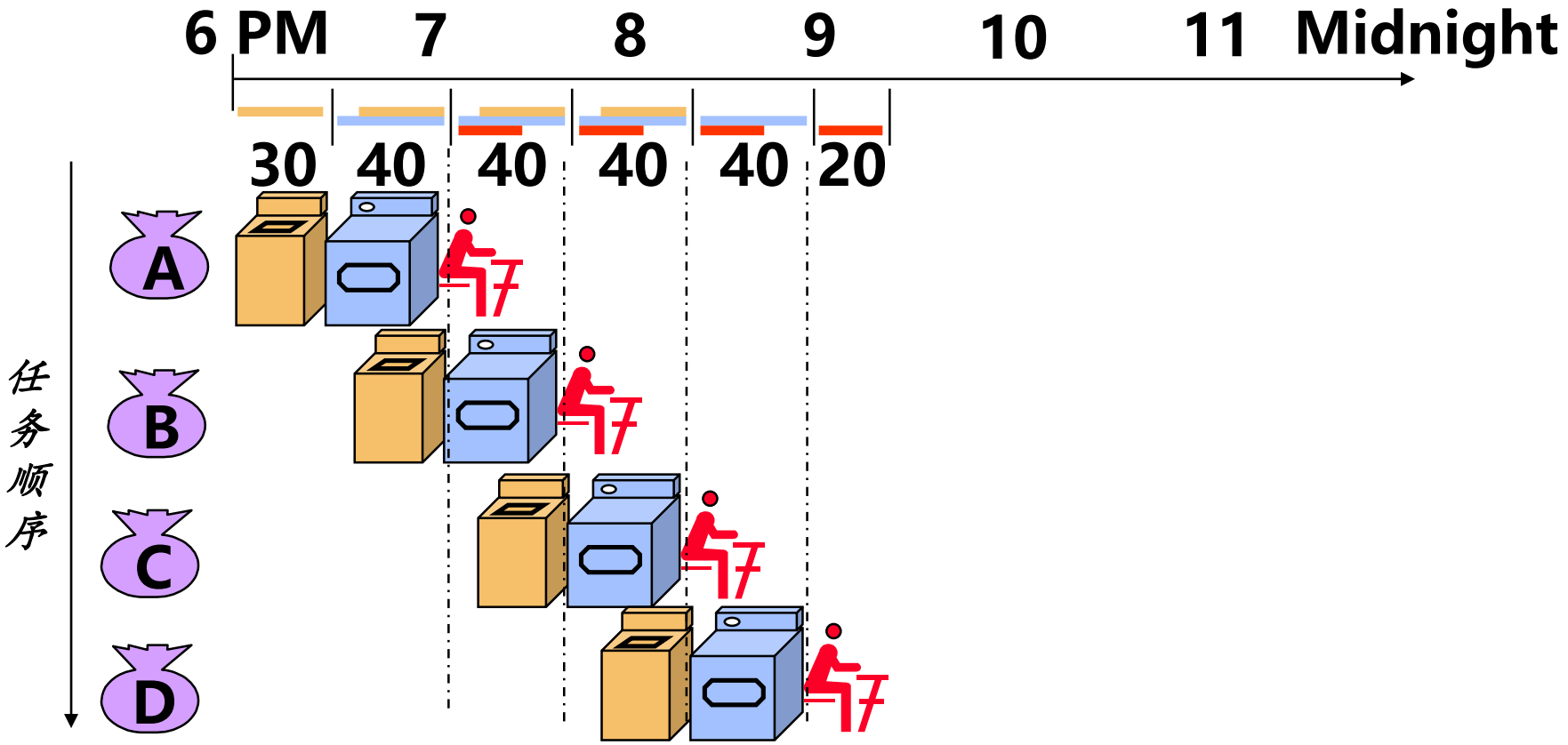
- 洗衣店用3.5小时完成了4个任务（1.14t/h）；

流水工作的洗衣店



- 洗衣店用3.5小时完成了4个任务（1.14t/h）；
- 4个同学各等待了1.5小时；

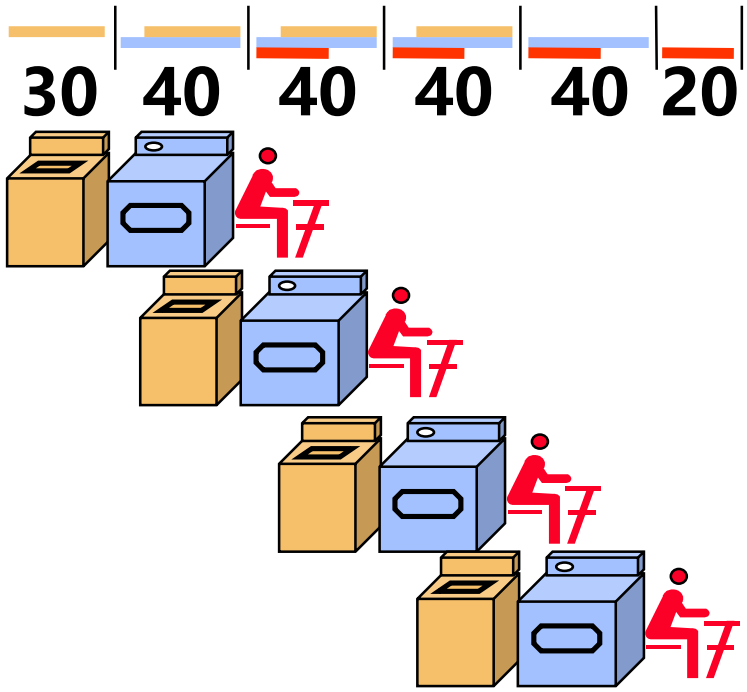
流水工作的洗衣店



- 洗衣店用3.5小时完成了4个任务（1.14t/h）；
- 4个同学各等待了1.5小时；
- Washer使用2小时；Dryer使用2小时40分；Folder使用1小时20分；

流水线基础：洗衣店的结论

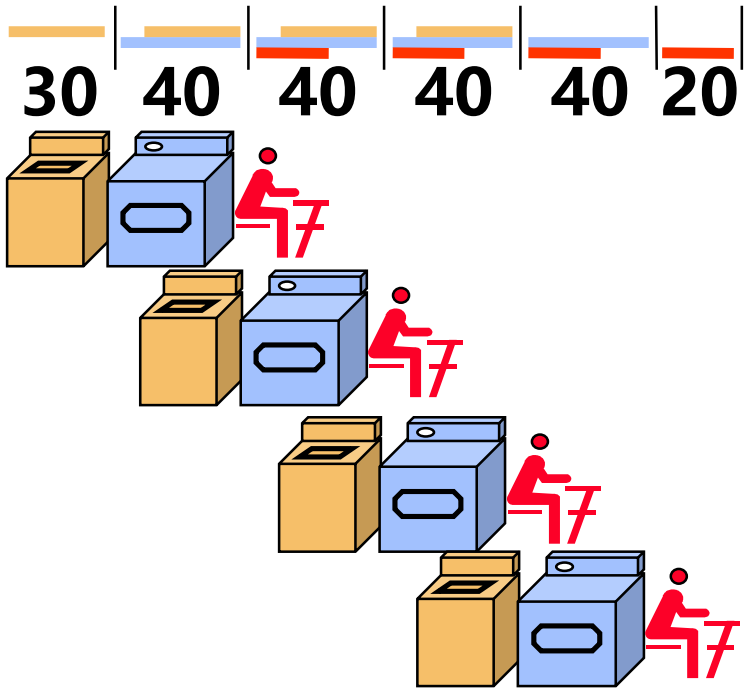
I. 流水线不能缩短单个任务的响应时间，但可以提高吞吐率；



流水线基础：洗衣店的结论

I. 流水线不能缩短单个任务的响应时间，但可以提高吞吐率；

II. 流水线速度受限于最慢流水站的速度；

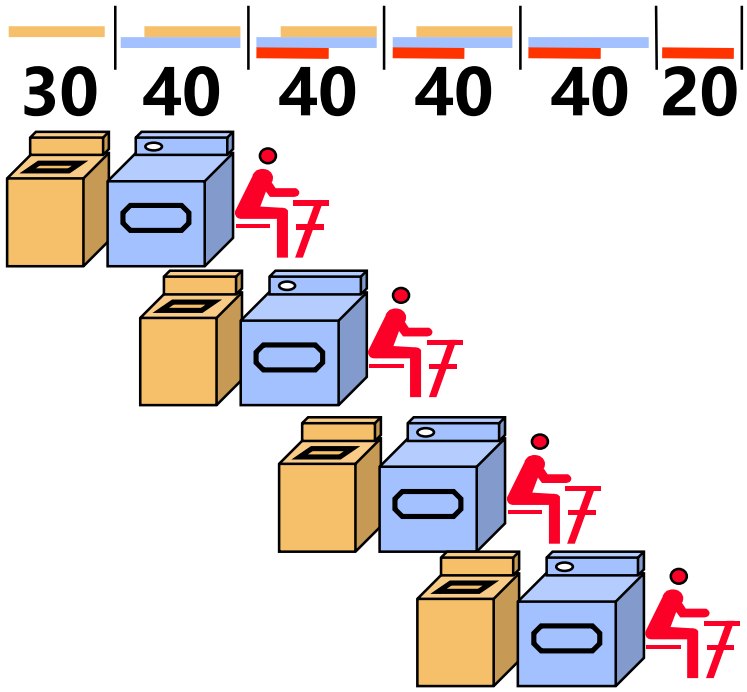


流水线基础：洗衣店的结论

I. 流水线不能缩短单个任务的响应时间，但可以提高吞吐率；

II. 流水线速度受限于最慢流水站的速度；

III. 流水线中多个任务是并行处理的；



流水线基础：洗衣店的结论

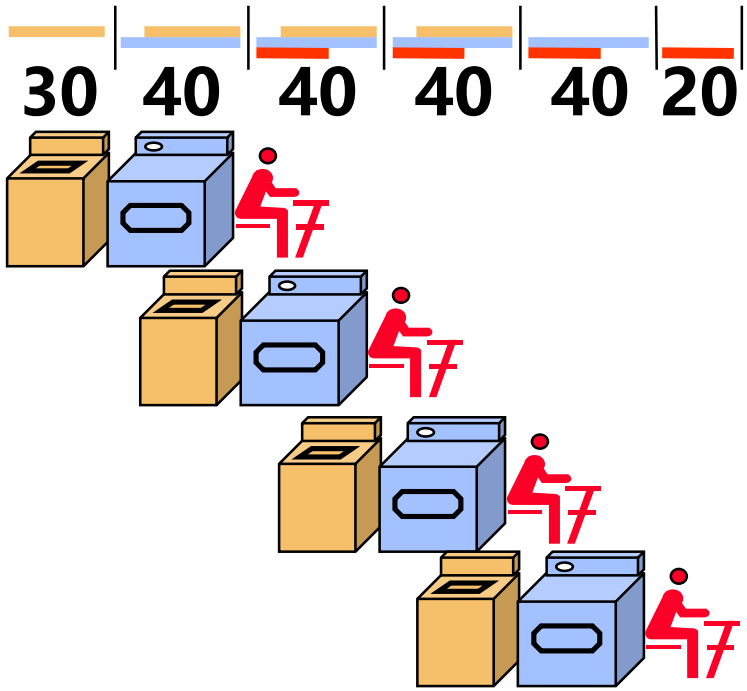
I. 流水线不能缩短单个任务的响应时间，但可以提高吞吐率；

II. 流水线速度受限于最慢流水站的速度；

III. 流水线中多个任务是并行处理的；

IV. 最大加速比 = 流水站数

- 流水站速度不匹配
- 流水线“填充”和“排空”时间



流水线基础：计算机中的流水线

- **MIPS 5-stage pipeline**

Instr. No.	Pipeline Stage						
1	IF	ID	EX	MEM	WB		
2		IF	ID	EX	MEM	WB	
3			IF	ID	EX	MEM	WB
4				IF	ID	EX	MEM
5					IF	ID	EX
Clock Cycle	1	2	3	4	5	6	7

流水技术及时空图

- 流水技术
  - 将一重复的时序过程分解为若干子过程，每个子过程都可有效地在其**专用功能段**上与其它子过程**同时执行**，这种技术称为流水技术。

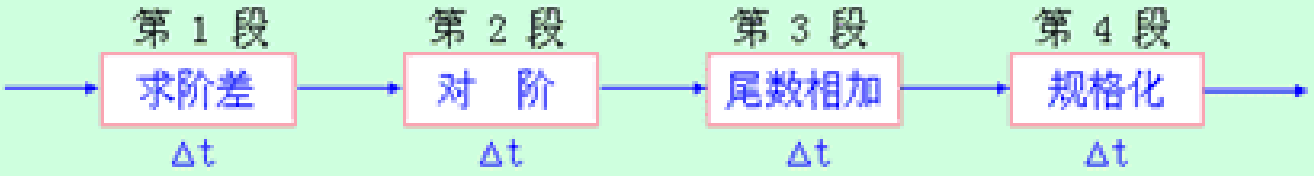
流水技术及时空图

- 流水技术
  - 将一重复的时序过程分解为若干子过程，每个子过程都可有效地在其**专用功能段**上与其它子过程**同时执行**，这种技术称为流水技术。
- 流水线的表示方式：时空图
  - 从时间和空间两个方面描述流水线的工作过程
    - 横坐标表示时间
    - 纵坐标表示各流水段

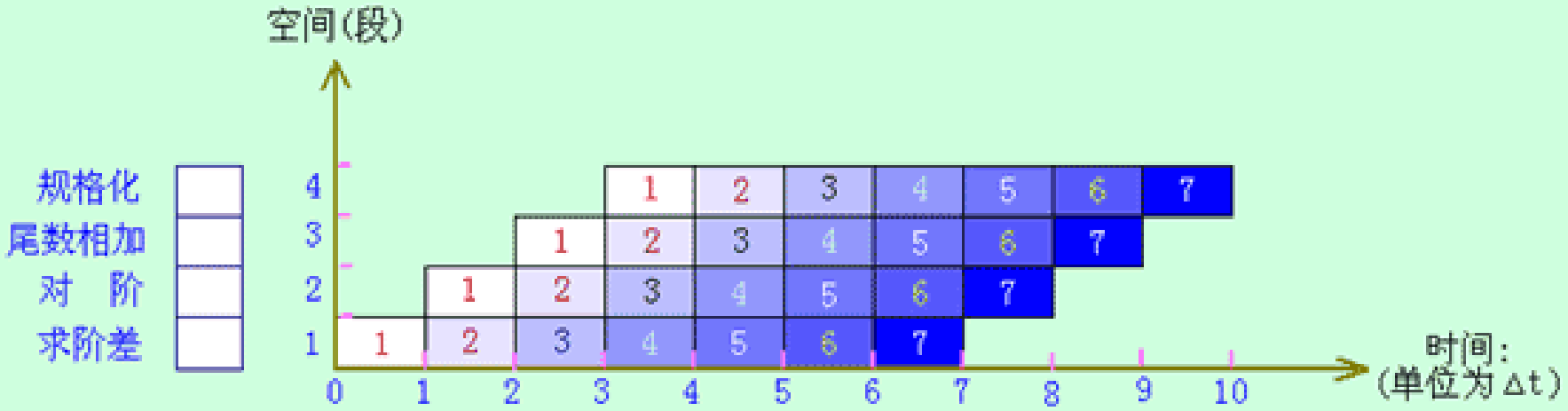
时空图示例

流水线的时-空图

流水线的工作过程常用时(间)-空(间)图来描述。下面以浮点加法流水线的时-空图为例来说明。



浮点加法流水线



流水线基础：流水线的特点

- 流水过程由多个相关的子过程组成，这些子过程称为流水线的“级”或“段”。段的数目称为流水线的“深度”。

流水线基础：流水线的特点

- 流水过程由多个相关的子过程组成，这些子过程称为流水线的“级”或“段”。段的数目称为流水线的“深度”。
- 每个子过程由专用的功能段实现，各功能段的时间应基本相等，通常为1个时钟周期。

流水线基础：流水线的特点

- 流水过程由多个相关的子过程组成，这些子过程称为流水线的“级”或“段”。段的数目称为流水线的“深度”。
- 每个子过程由专用的功能段实现，各功能段的时间应基本相等，通常为1个时钟周期。
- 流水线需要经过一定的通过时间才能稳定。

流水线基础：流水线的特点

- 流水过程由多个相关的子过程组成，这些子过程称为流水线的“级”或“段”。段的数目称为流水线的“深度”。
- 每个子过程由专用的功能段实现，各功能段的时间应基本相等，通常为1个时钟周期。
- 流水线需要经过一定的通过时间才能稳定。
- 流水技术适合于大量重复的时序过程。

流水线基础：流水线的分类

（1）单功能、多功能流水线

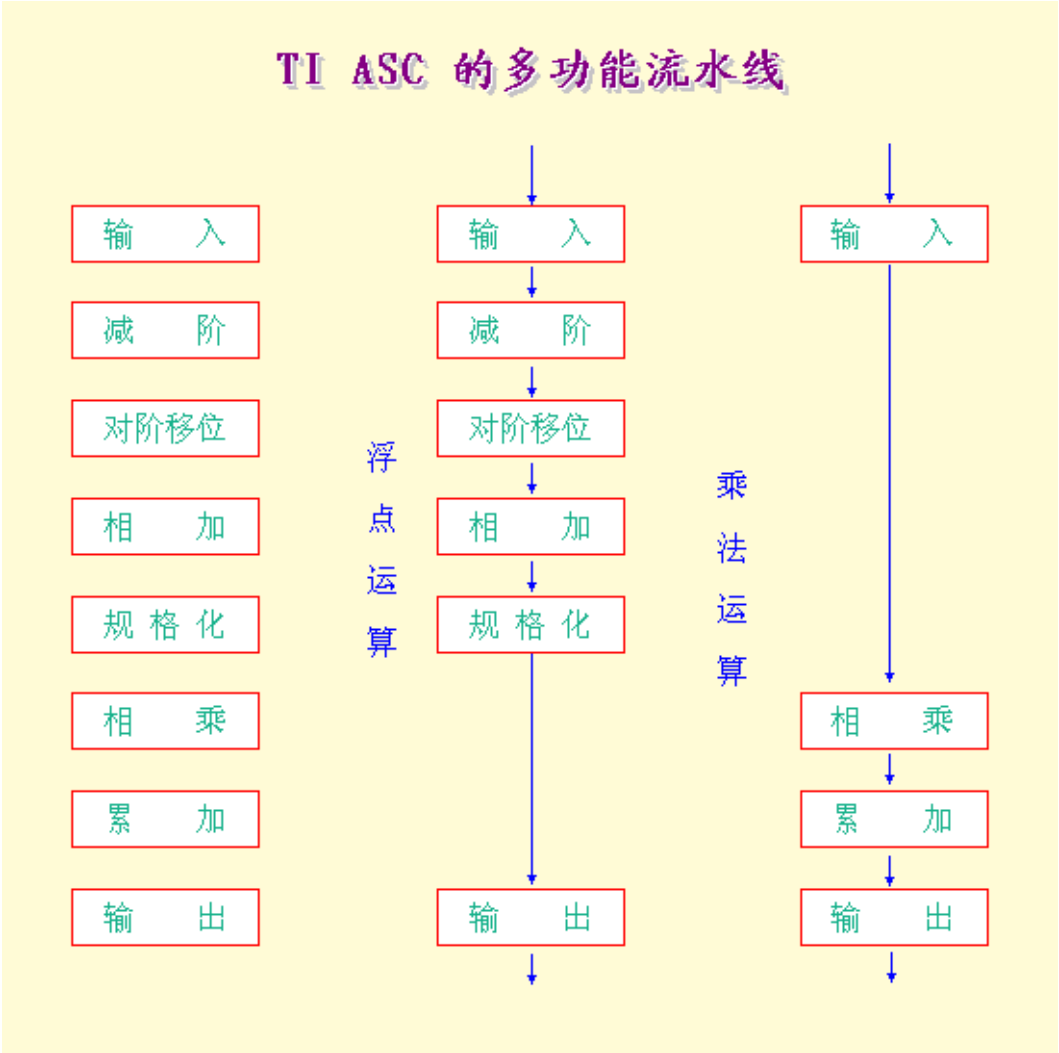
- 单功能流水线，是指只能完成一种固定功能的流水线。
  - 例如：功能单元流水线，浮点加法流水线等

流水线基础：流水线的分类

（1）单功能、多功能流水线

- 单功能流水线，是指只能完成一种固定功能的流水线。
  - 例如：功能单元流水线，浮点加法流水线等
- 多功能流水线，是指各段可以进行不同的连接，从而完成不同的功能。
  - 例如：TI ASC的多功能流水线

TI ASC多功能流水线



流水线基础：流水线的分类

（2）静态、动态流水线

- 静态流水线，是指在某段时间间隔内，流水线的各段只能按同一种功能的连接方式工作。
  - 例如：TI ASC的流水线，适合于处理一串相同的运算操作。

流水线基础：流水线的分类

（2）静态、动态流水线

- 静态流水线，是指在某段时间间隔内，流水线的各段只能按同一种功能的连接方式工作。
  - 例如：TI ASC的流水线，适合于处理一串相同的运算操作。
- 动态流水线，是指在某段时间内，当某些段正在实现某种运算时，另一些段却在实现另一种运算。
  - 能提高流水效率，但同时会使流水线的控制变得很复杂。

流水线基础：流水线的分类

（3）部件级、处理机级及处理机间流水线

- 部件级流水线，又叫运算操作流水线，是把处理机的算术逻辑部件分段，使得各种数据类型的操作能够进行流水。

流水线基础：流水线的分类

（3）部件级、处理机级及处理机间流水线

- 部件级流水线，又叫运算操作流水线，是把处理机的算术逻辑部件分段，使得各种数据类型的操作能够进行流水。
- 处理机级流水线，又叫指令流水线，是把解释指令的过程按照流水方式处理。

流水线基础：流水线的分类

（3）部件级、处理机级及处理机间流水线

- 部件级流水线，又叫运算操作流水线，是把处理机的算术逻辑部件分段，使得各种数据类型的操作能够进行流水。
- 处理机级流水线，又叫指令流水线，是把解释指令的过程按照流水方式处理。
- 处理机间流水线，又叫宏流水线，是由两个以上的处理机串行地对同一数据流进行处理，每个处理机完成一项任务。(如map-reduce)

流水线基础：流水线的分类

（3）部件级、处理机级及处理机间流水线

- 部件级流水线，又叫运算操作流水线，是把处理机的[illegible]各种数据类型的操作[illegible]
- 处[illegible]水线，是把解释指令的过程按照流水方式处理。
- 处理机间流水线，又叫宏流水线，是由两个以上的处理机串行地对同一数据流进行处理，每个处理机完成一项任务。(如map-reduce)

流水处理的粒度越来越大。

流水线基础：流水线的分类

（4）标量、向量流水处理机

- 标量流水处理机，是指处理机不具有向量数据表示，仅对标量数据进行流水处理。
  - 例如：IBM360/91，Amdahl 470V/6等

流水线基础：流水线的分类

（4）标量、向量流水处理机

- 标量流水处理机，是指处理机不具有向量数据表示，仅对标量数据进行流水处理。
  - 例如：IBM360/91，Amdahl 470V/6等
- 向量流水处理机，是指处理机具有向量数据表示，并通过向量指令对向量的各元素进行处理。
  - 例如：TI ASC、STAR-100、CRAY-1等

流水线基础：流水线的分类

（5）线性、非线性流水线

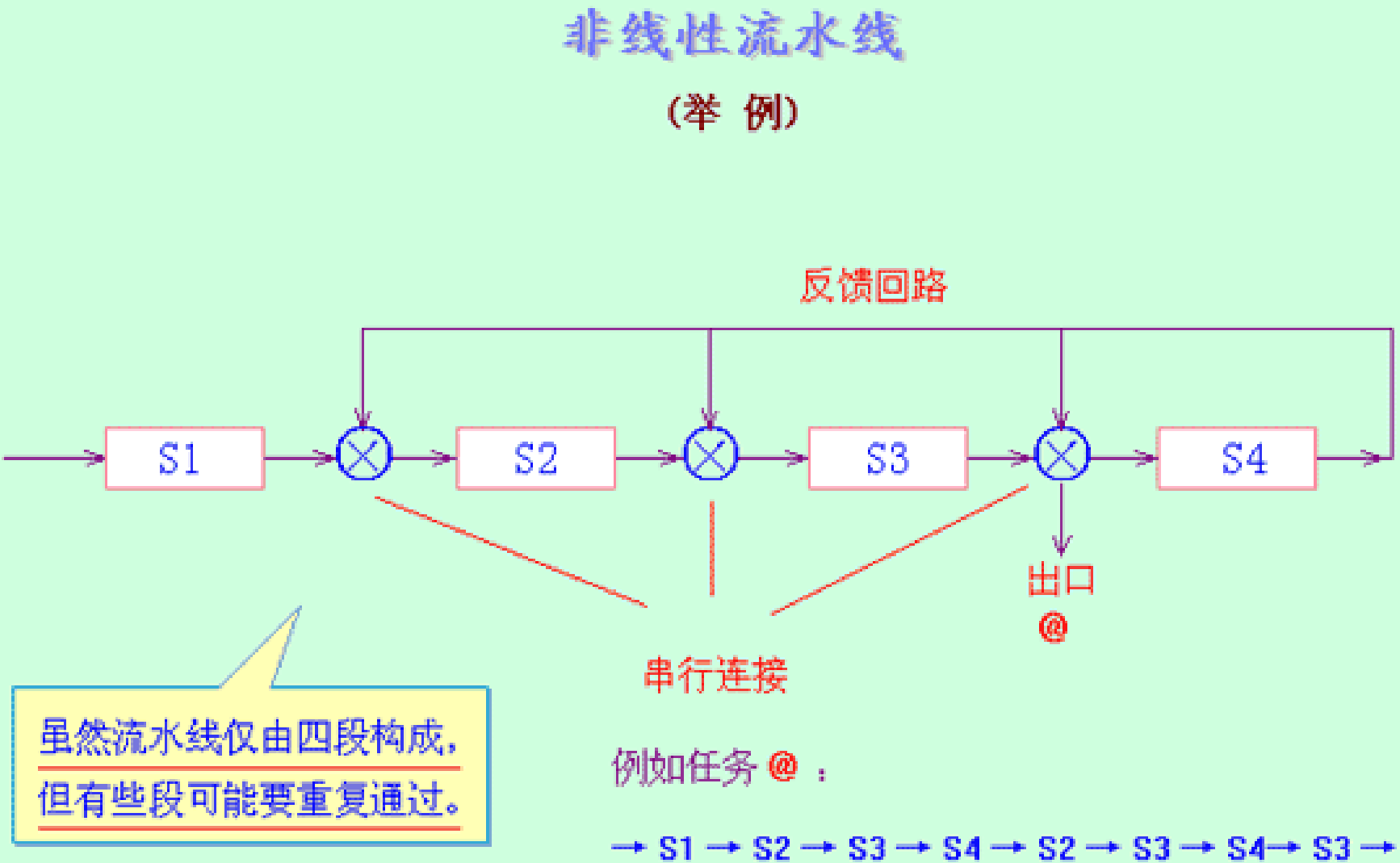
- 线性流水线是指流水线的各段串行连接，没有反馈回路。

流水线基础：流水线的分类

（5）线性、非线性流水线

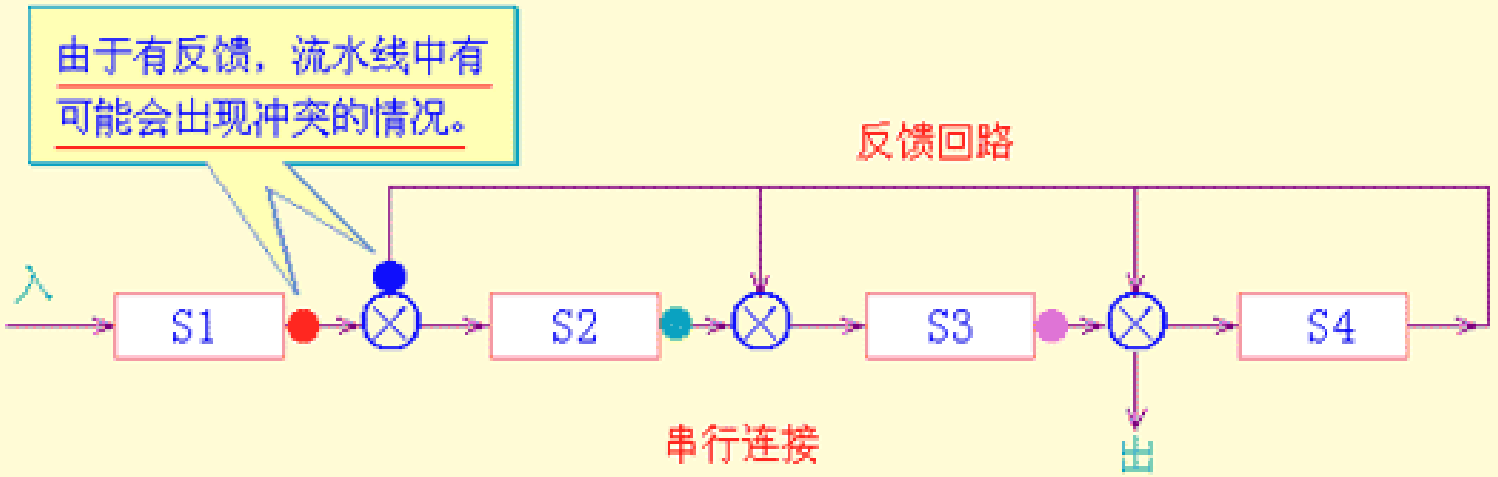
- 线性流水线是指流水线的各段串行连接，没有反馈回路。
- 非线性流水线是指流水线中除有串行连接的通路外，还有反馈回路。存在流水线调度问题：
  - 确定什么时候向流水线引进新的输入，从而使新输入的数据和先前操作的反馈数据在流水线中不产生冲突，此即所谓**流水线调度问题**。

非线性流水线示例


非线性流水线
(举 例)
反馈回路
S1
S2
S3
S4
出口
@
串行连接
虽然流水线仅由四段构成，
但有些段可能要重复通过。
例如任务 @ ：
→ S1 → S2 → S3 → S4 → S2 → S3 → S4→ S3 →

非线性流水线的调度问题

流水线的调度问题



所以：

在非线性流水线中，一个重要的问题是确定什么时候向流水线引进新的输入，从而使新输入的数据和先前操作的反馈数据在流水线中不产生冲突。这就是所谓的流水线调度问题。

流水线基础：流水线的分类

（6）顺序、乱序流动流水线

- 按照输出端任务流出顺序与输入端任务流入顺序是否相同划分
- 乱序流动流水线也可称为无序流水线、错序流水线

内容概要

- 流水线基础
  - 流水线概要
  - 时空图表示
  - 流水线分类
- **流水线的性能分析**
  - **流水线的吞吐率**
  - **流水线的加速比**
  - **流水线的效率**

吞吐率（throughput）

- 吞吐率是指单位时间内流水线所完成的任务数或输出结果的数量(指令数)。

吞吐率（throughput）

- 吞吐率是指单位时间内流水线所完成的任务数或输出结果的数量(指令数)。
- 最大吞吐率TP_{max}是指流水线在达到稳定状态后的吞吐率。

吞吐率（throughput）

- 吞吐率是指单位时间内流水线所完成的任务数或输出结果的数量(指令数)。
- 最大吞吐率TP_{max}是指流水线在达到稳定状态后的吞吐率。
- 设流水线由m段组成，完成n个任务的吞吐率称为实际吞吐率，记作TP。

最大吞吐率 （Max Throughput）

- 假设流水线各段的时间相等，均为$\triangle t_0$，则：

$$TP_{max} = 1/\triangle t_0$$

最大吞吐率 （Max Throughput）

- 假设流水线各段的时间相等，均为$\triangle t_0$，则：

$$TP_{max} = 1/\triangle t_0$$

- 假设流水线各段时间不等，第i段时间为$\triangle t_i$，则：

$$TP_{max} = 1/max\{\triangle t_i\}$$

最大吞吐率 （Max Throughput）

- 假设流水线各段的时间相等，均为$\triangle t_0$，则：

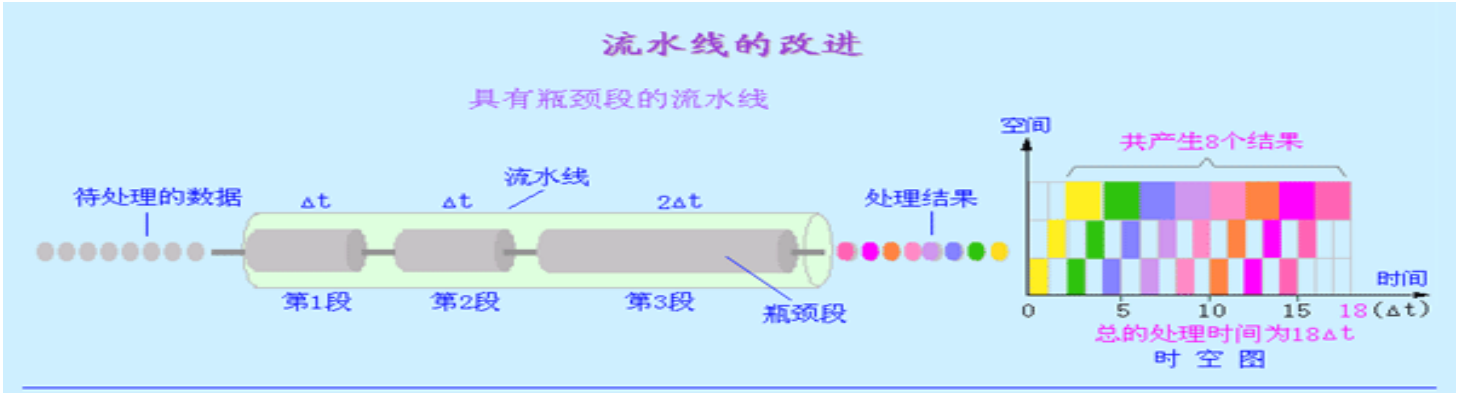
$$TP_{max} = 1/\triangle t_0$$

- 假设流水线各段时间不等，第i段时间为$\triangle t_i$，则：

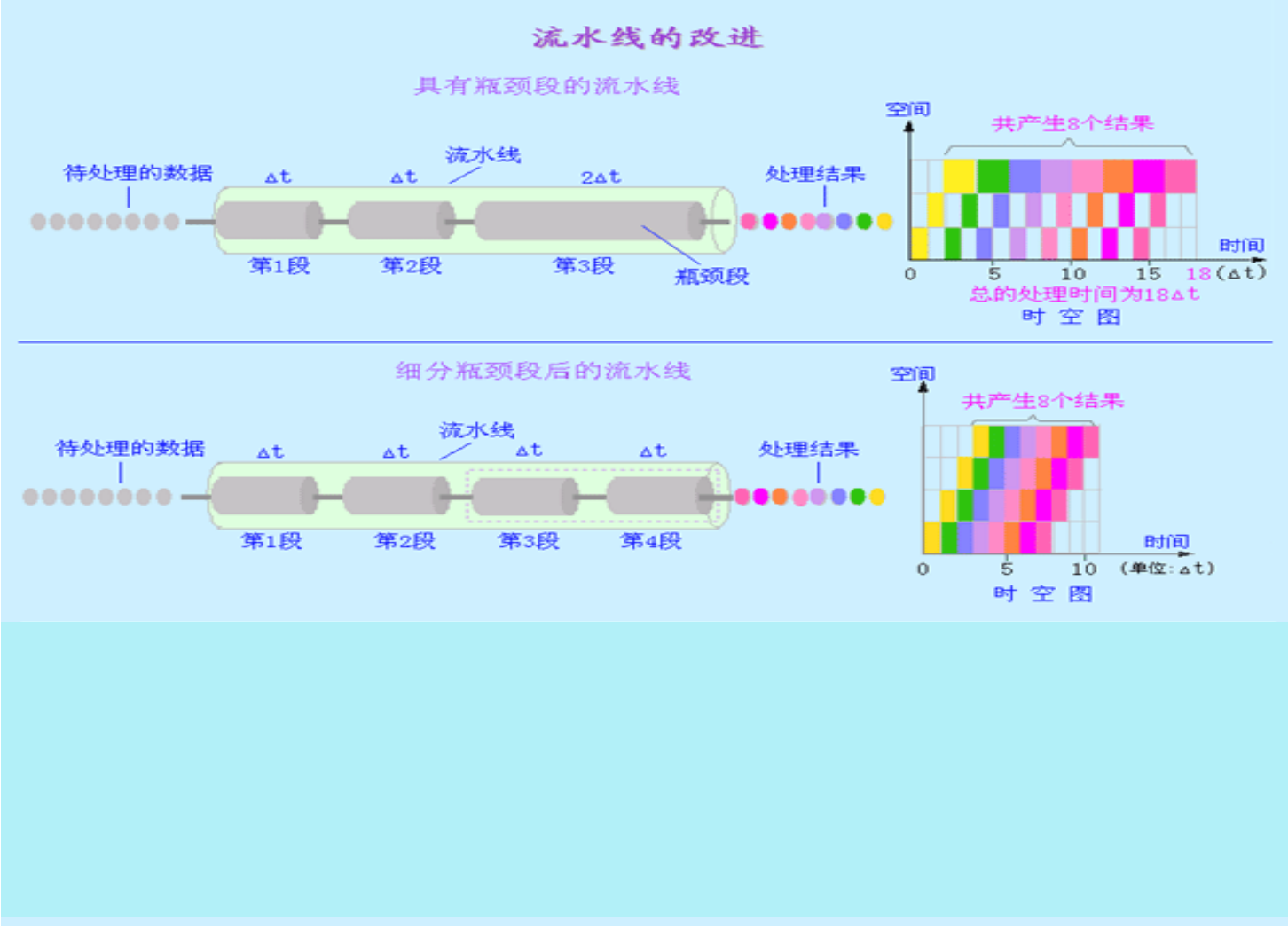
$$TP_{max} = 1/max\{\triangle t_i\}$$

- 最大吞吐率取决于流水线中最慢一段所需的时间，该段成为流水线的瓶颈。

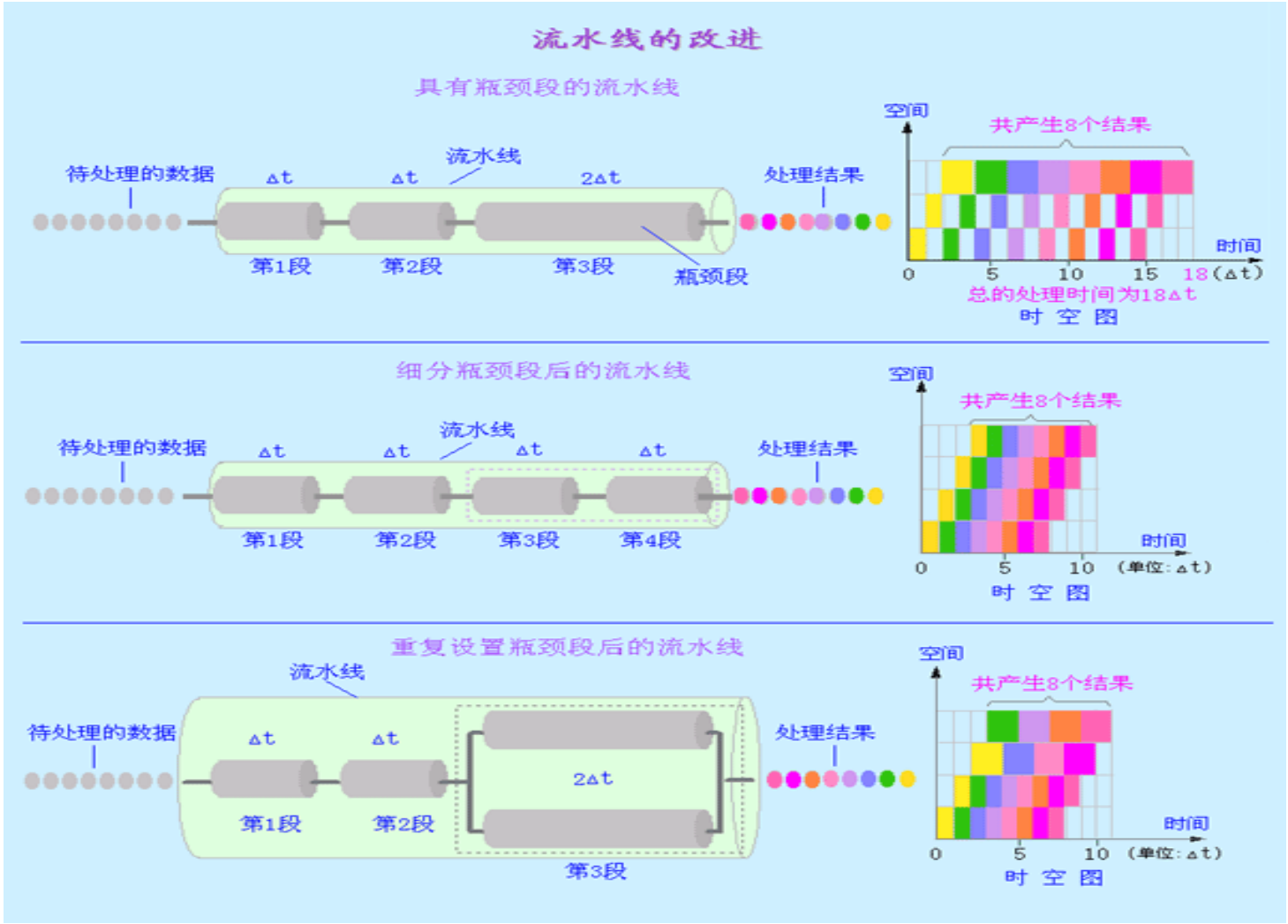
消除流水线瓶颈方法


流水线的改进
具有瓶颈段的流水线
待处理的数据
Δt
Δt
流水线
2Δt
处理结果
第1段
第2段
第3段
瓶颈段
空间
共产生8个结果
时间
0
5
10
15
18(Δt)
总的处理时间为18Δt
时 空 图

消除流水线瓶颈方法


流水线的改进
具有瓶颈段的流水线
待处理的数据
Δt
Δt
流水线
2Δt
处理结果
第1段
第2段
第3段
瓶颈段
空间
共产生8个结果
时间
0
5
10
15
18(Δt)
总的处理时间为18Δt
时 空 图
细分瓶颈段后的流水线
待处理的数据
Δt
Δt
流水线
Δt
Δt
处理结果
第1段
第2段
第3段
第4段
空间
共产生8个结果
时间
0
5
10
(单位:Δt)
时 空 图

消除流水线瓶颈方法


流水线的改进
具有瓶颈段的流水线
待处理的数据
Δt
Δt
流水线
2Δt
处理结果
第1段
第2段
第3段
瓶颈段
空间
共产生8个结果
时间
0
5
10
15
18(Δt)
总的处理时间为18Δt
时 空 图
细分瓶颈段后的流水线
待处理的数据
Δt
Δt
流水线
Δt
Δt
处理结果
第1段
第2段
第3段
第4段
空间
共产生8个结果
时间
0
5
10
(单位:Δt)
时 空 图
重复设置瓶颈段后的流水线
流水线
待处理的数据
Δt
Δt
2Δt
处理结果
第1段
第2段
第3段
空间
共产生8个结果
时间
0
5
10
(单位:Δt)
时 空 图

流水线的吞吐率（各段相等）

- 实际吞吐率TP：小于最大吞吐率。
  - 第一种情况：各段时间相等（设为$\triangle t_0$）

    假设流水线由 m 段组成，n 个任务；

流水线的吞吐率（各段相等）

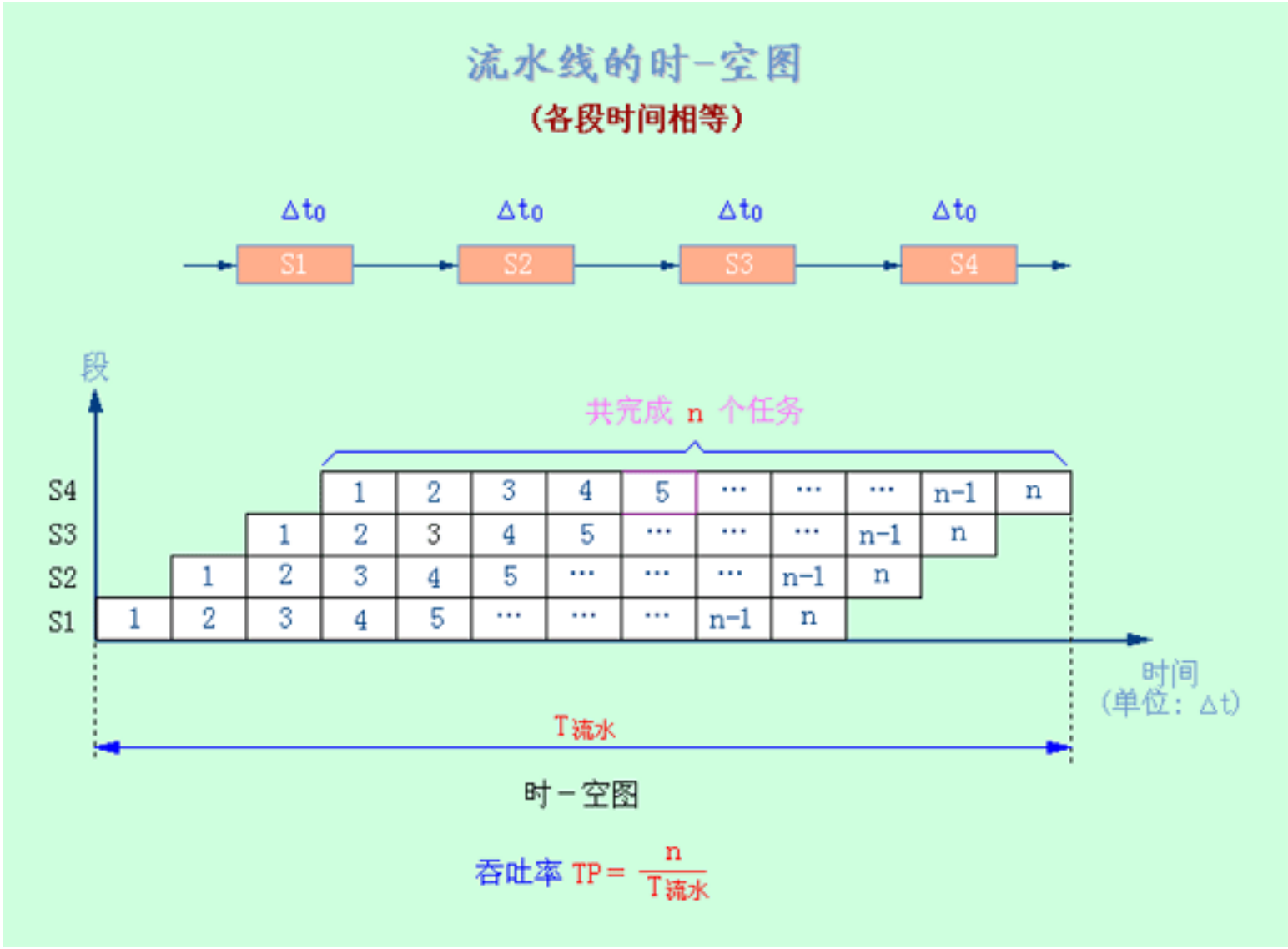
- 实际吞吐率TP：小于最大吞吐率。
  - 第一种情况：各段时间相等（设为$\triangle t_0$）
    假设流水线由 m 段组成，n 个任务；
  - 完成 n 个任务所需的时间：
    $T_{流水}=m\triangle t_0+(n-1)\triangle t_0$

流水线的吞吐率（各段相等）

- 实际吞吐率TP：小于最大吞吐率。
  - 第一种情况：各段时间相等（设为$\triangle t_0$）
    假设流水线由 m 段组成，n 个任务；
  - 完成 n 个任务所需的时间：
    $T_{流水}=m\triangle t_0+(n-1)\triangle t_0$

称为注入时间
(filling time)

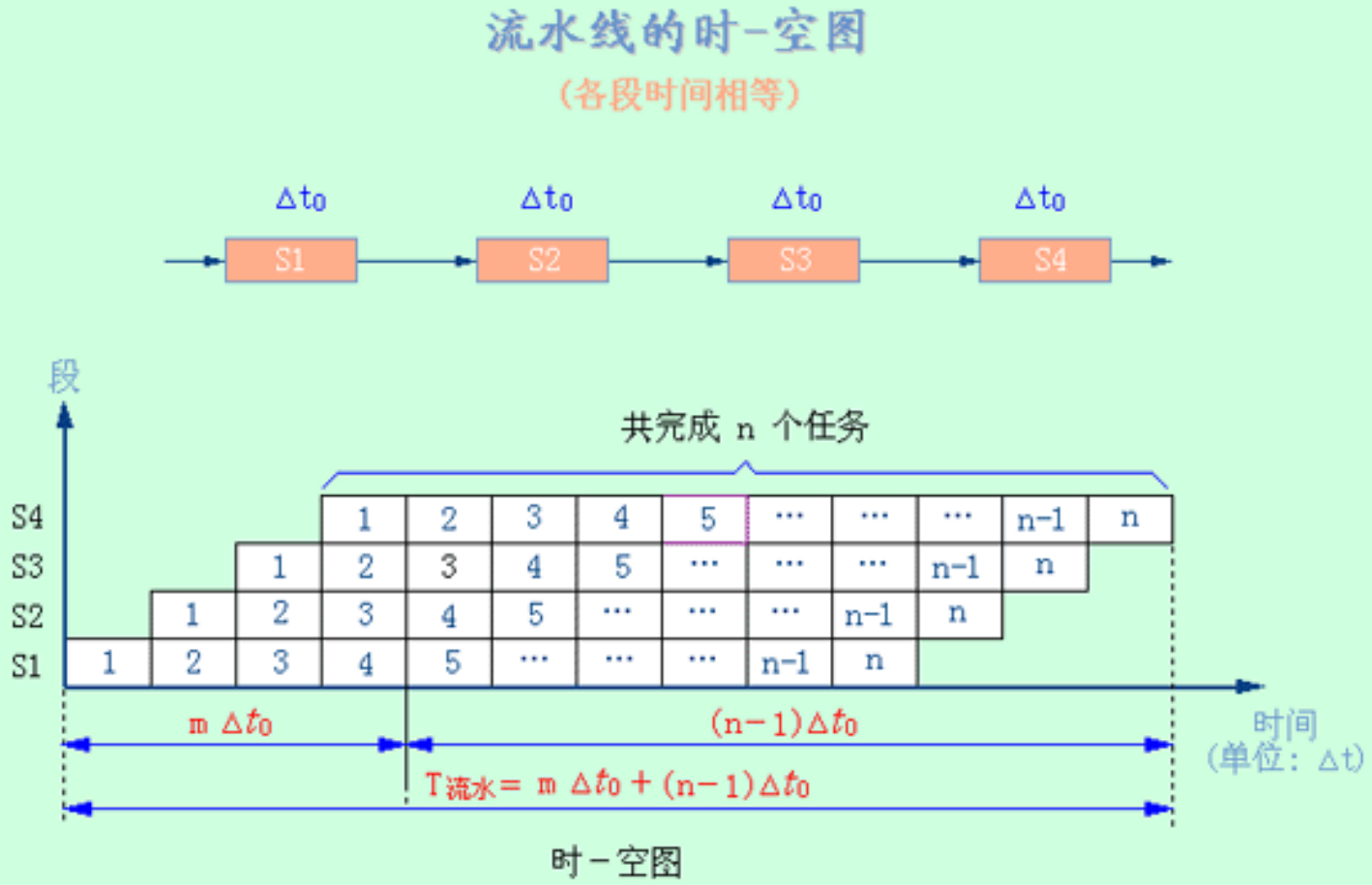
流水段相等时的时空图



时－空图

吞吐率 $TP = \frac{n}{T_{流水}}$

完成 n 个任务所需的时间



流水线的吞吐率（各段相等）

- 实际吞吐率

$$TP = \frac{n}{T_{流水}} = \frac{n}{m\Delta t_0 + (n-1)\Delta t_0}$$

$$= \frac{1}{(1+\frac{m-1}{n})\Delta t_0} = \frac{TP_{max}}{1+\frac{m-1}{n}}$$

流水线的吞吐率（各段相等）

- 实际吞吐率

$$TP = \frac{n}{T_{流水}} = \frac{n}{m\Delta t_0 + (n-1)\Delta t_0}$$

$$= \frac{1}{(1+\frac{m-1}{n})\Delta t_0} = \frac{TP_{max}}{1+\frac{m-1}{n}}$$

$TP < TP_{max}$

当$n >> m$时，$TP \approx TP_{max}$

流水线的吞吐率（各段相等）

- 实际吞吐率

$$TP = \frac{n}{T_{流水}} = \frac{n}{m\Delta t_0 + (n-1)\Delta t_0}$$

$$= \frac{1}{(1+\frac{m-1}{n})\Delta t_0} = \frac{TP_{max}}{1+\frac{m-1}{n}}$$

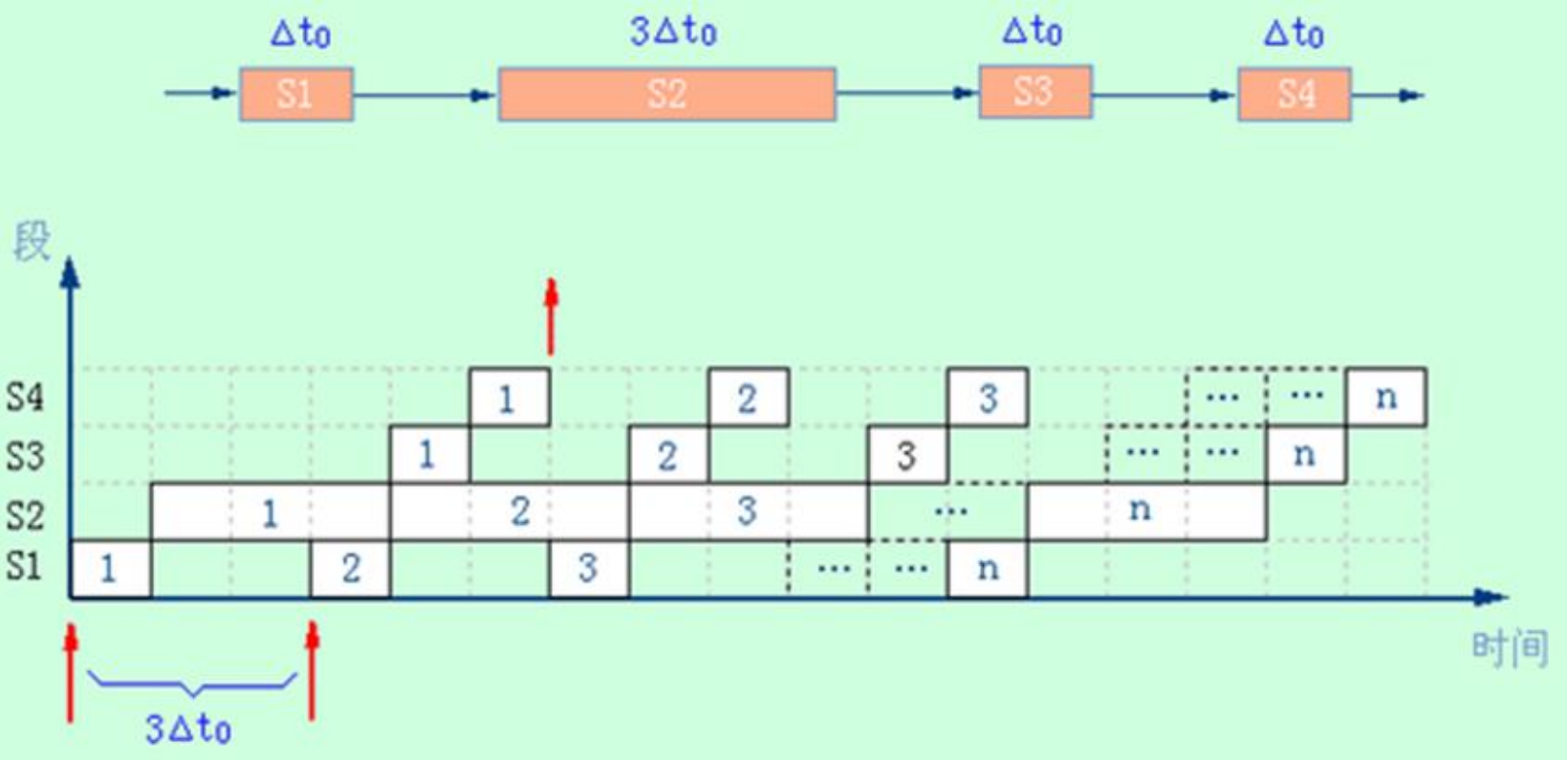
$TP < TP_{max}$

当$n >> m$时，$TP \approx TP_{max}$

这个结论说明什么？

流水线的吞吐率（各段不等）

- 实际吞吐率TP：小于最大吞吐率。



流水线的吞吐率（各段不等）

- 完成 n 个任务所需的时间

$$T_{流水} = \sum_{i=1}^{m} \Delta t_i + (n-1)\Delta t_j \quad \Delta t_j = \max\{\Delta t_i\}$$

流水线的吞吐率（各段不等）

- 完成 n 个任务所需的时间

$$T_{流水}=\sum_{i=1}^{m}\Delta t_i+(n-1)\Delta t_j \quad \Delta t_j=\max\{\Delta t_i\}$$

- 实际吞吐率

$$TP=\frac{n}{\sum_{i=1}^{m}\Delta t_i+(n-1)\Delta t_j}$$

$$TP < TP_{max}$$

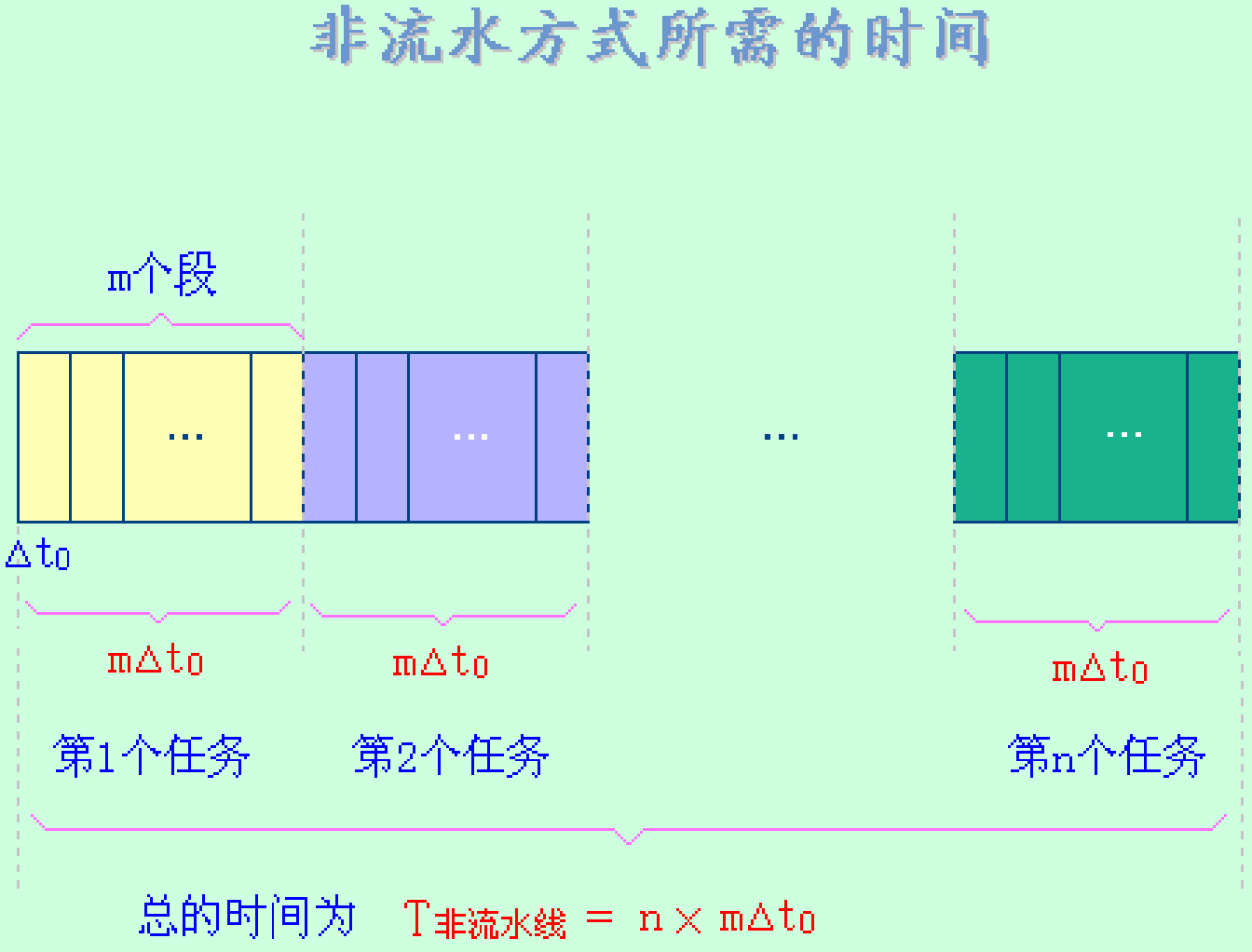
加速比（speedup）

- 加速比是指流水线的速度与等功能非流水线的速度之比。
  - $S = T_{非流水} / T_{流水}$
  - 其中$T_{流水}$和$T_{非流水}$分别为按流水和按非流水方式处理 n 个任务所需的时间

加速比（speedup）

- 加速比是指流水线的速度与等功能非流水线的速度之比。
  - $S = T_{非流水} / T_{流水}$
  - 其中$T_{流水}$和$T_{非流水}$分别为按流水和按非流水方式处理 n 个任务所需的时间
- 若流水线为 m 段，且各段时间相等，均为Δt_0，则：
  - $T_{非流水} = n\ m\Delta t_0$
  - $T_{流水} = m\Delta t_0 + (n-1)\Delta t_0$

非流水方式的任务所需时间



流水线的加速比

$$S = \frac{T_{非流水}}{T_{流水}} = \frac{nm\Delta t_0}{m\Delta t_0 + (n-1)\Delta t_0}$$

$$= \frac{mn}{m+n-1} = \frac{m}{1+\frac{m-1}{n}}$$

流水线的加速比

$$S = \frac{T_{非流水}}{T_{流水}} = \frac{nm\Delta t_0}{m\Delta t_0 + (n-1)\Delta t_0}$$

$$= \frac{mn}{m+n-1} = \frac{m}{1+\frac{m-1}{n}}$$

可以得出：当 $n >> m$ 时，$S \approx m$

流水线的效率（efficiency）

- 效率（E）指流水线的设备利用率。

流水线的效率（efficiency）

- 效率（E）指流水线的设备利用率。
- 由于流水线有通过时间和排空时间，所以流水线的各段并非一直满负荷工作，即：E<1。

流水线的效率（efficiency）

- 效率（E）指流水线的设备利用率。
- 由于流水线有通过时间和排空时间，所以流水线的各段并非一直满负荷工作，即：E<1。
- 若各段时间相等，则各段效率也相等，即$e_1 = e_2 = e_3 = \cdots = n\triangle t_0/T_{流水}$

流水线的效率（efficiency）

- 效率（E）指流水线的设备利用率。
- 由于流水线有通过时间和排空时间，所以流水线的各段并非一直满负荷工作，即：E<1。
- 若各段时间相等，则各段效率也相等，即$e_1 = e_2 = e_3 = \cdots = n\triangle t_0 / T_{流水}$
- 整个流水线效率：

$$E = \frac{n\triangle t_0}{T_{流水}} = \frac{n}{m+n-1} = \frac{1}{1+\frac{m-1}{n}}$$

当n>>m时，E≈1

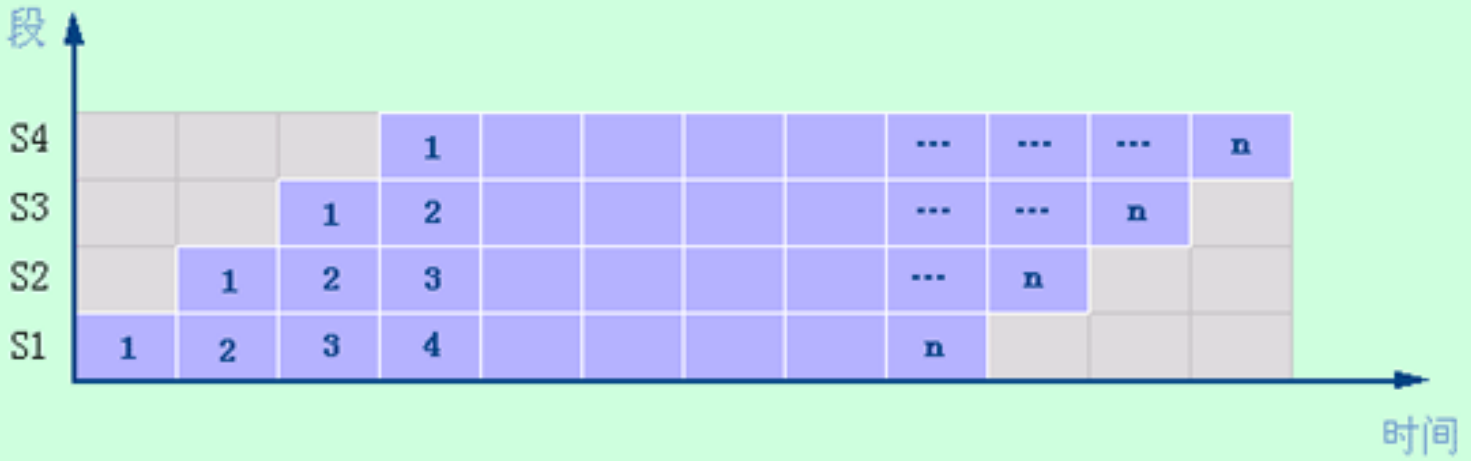
流水线的效率（efficiency）

- 从时空图上看，效率就是n个任务所占的时空区与m个段总的时空区之比。
- 根据这个定义，可以计算流水线各段时间不等时的流水线效率：

$$E = \frac{n \text{ 个任务占用的时空区}}{m \text{ 个段总的时空区}}$$

从时空图看流水线的效率

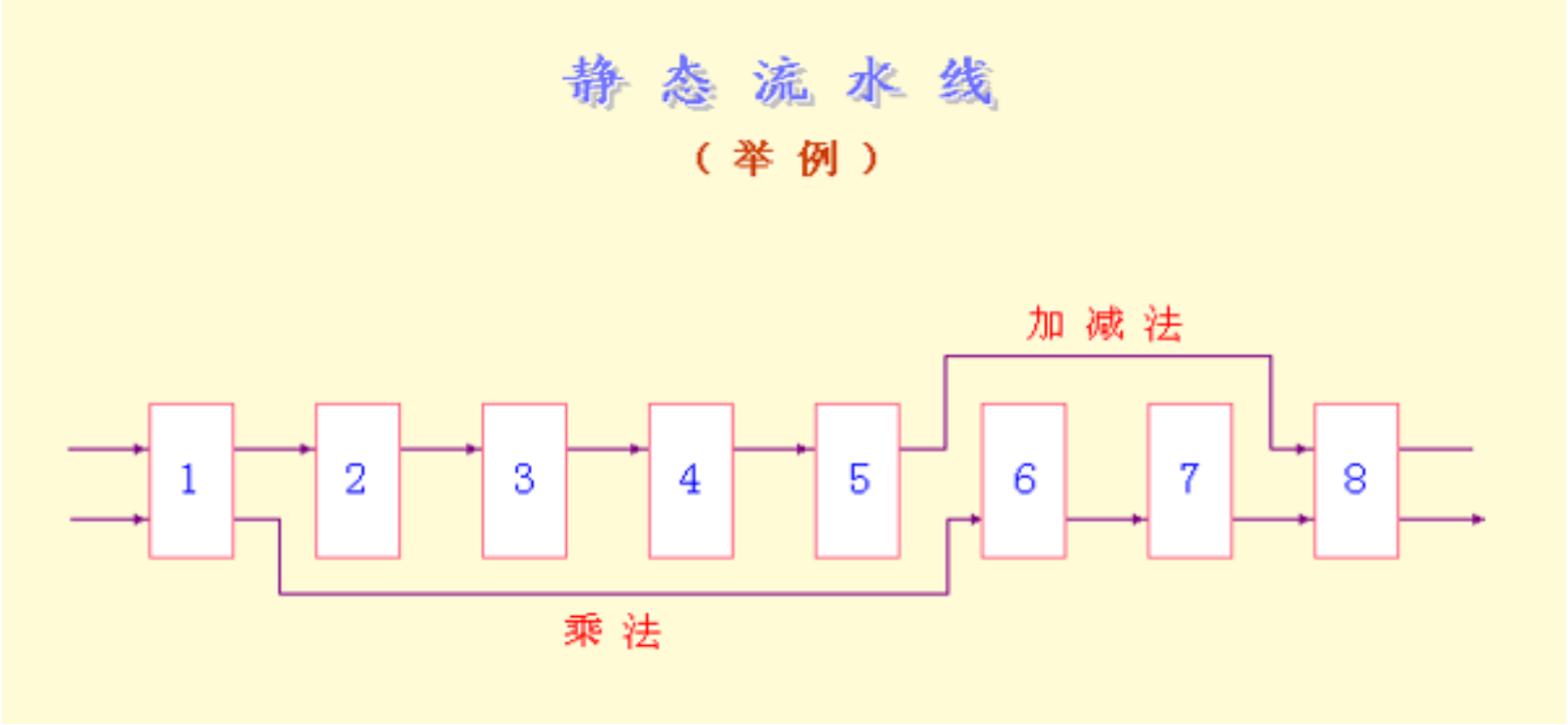
流水段的效率



所以 $e_2 = \frac{n\Delta t_0}{T_{流水}}$

流水线性能示例

例：（张晨曦教材） 在下面所示的静态流水线上计算：$\sum_{i=1}^{4} A_i B_i$，求：吞吐率、加速比、和效率。



流水线性能示例

- 此类问题的一般解决步骤：

  ① 根据目标公式确定计算过程

  ② 根据①中确定的计算过程画时空图

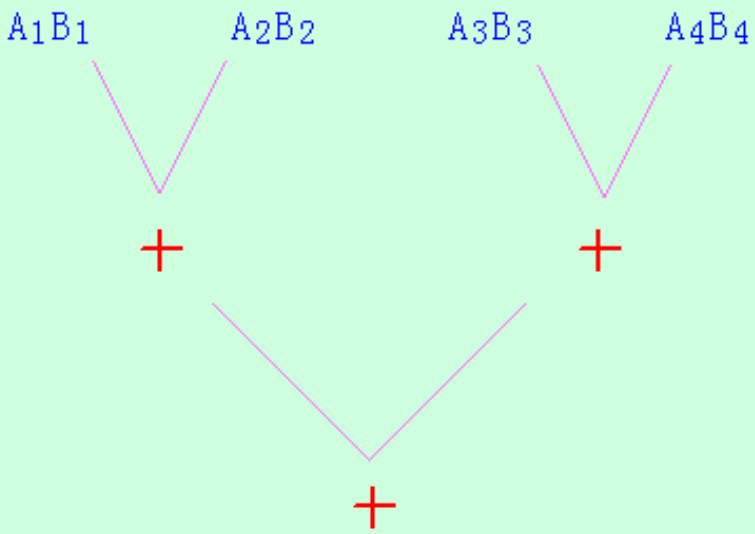
  ③ 根据②中给出的时空图计算性能

① 确定计算过程

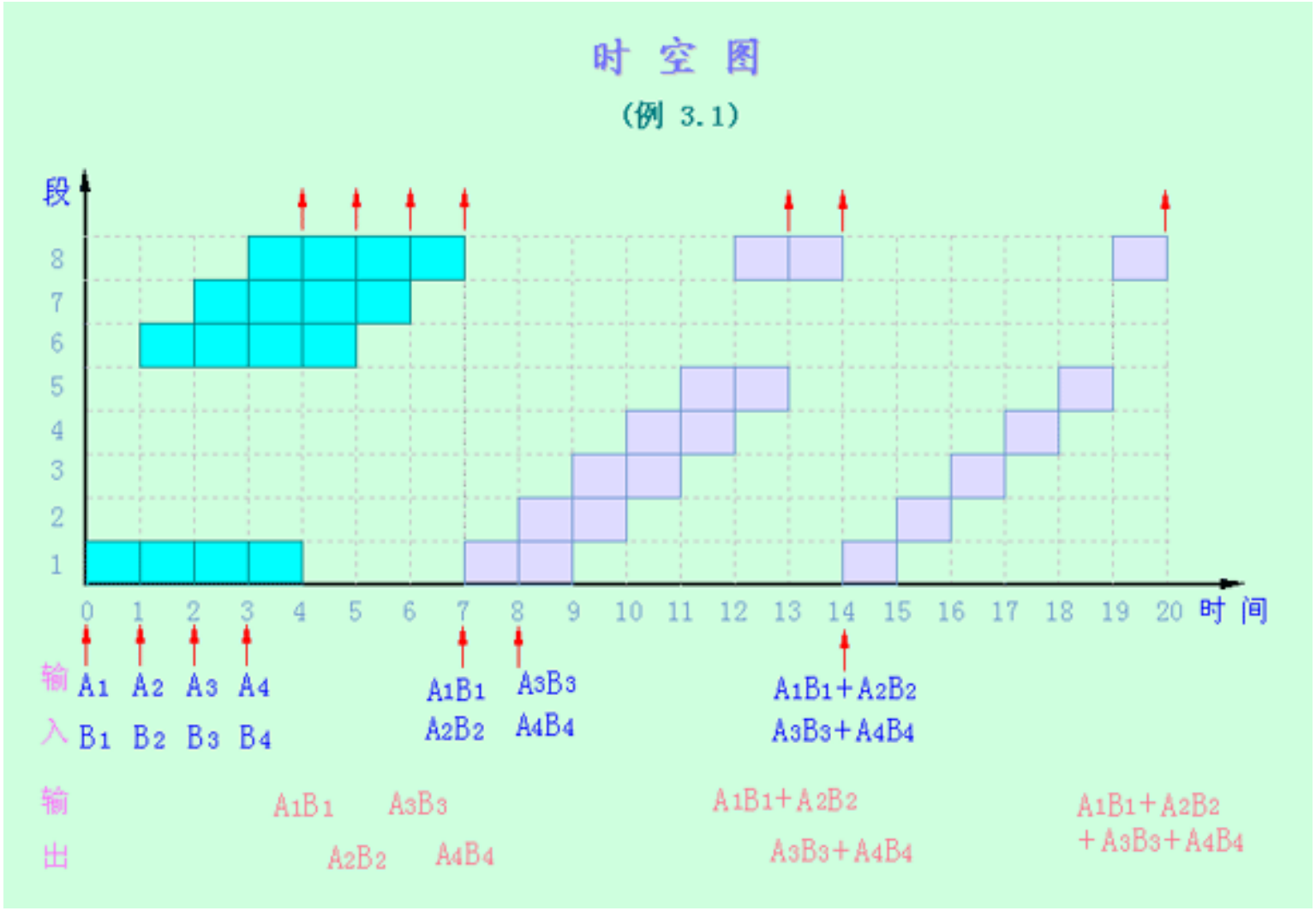
$\sum_{i=1}^{4} A_i B_i$ 的计算过程

（例 3.1）

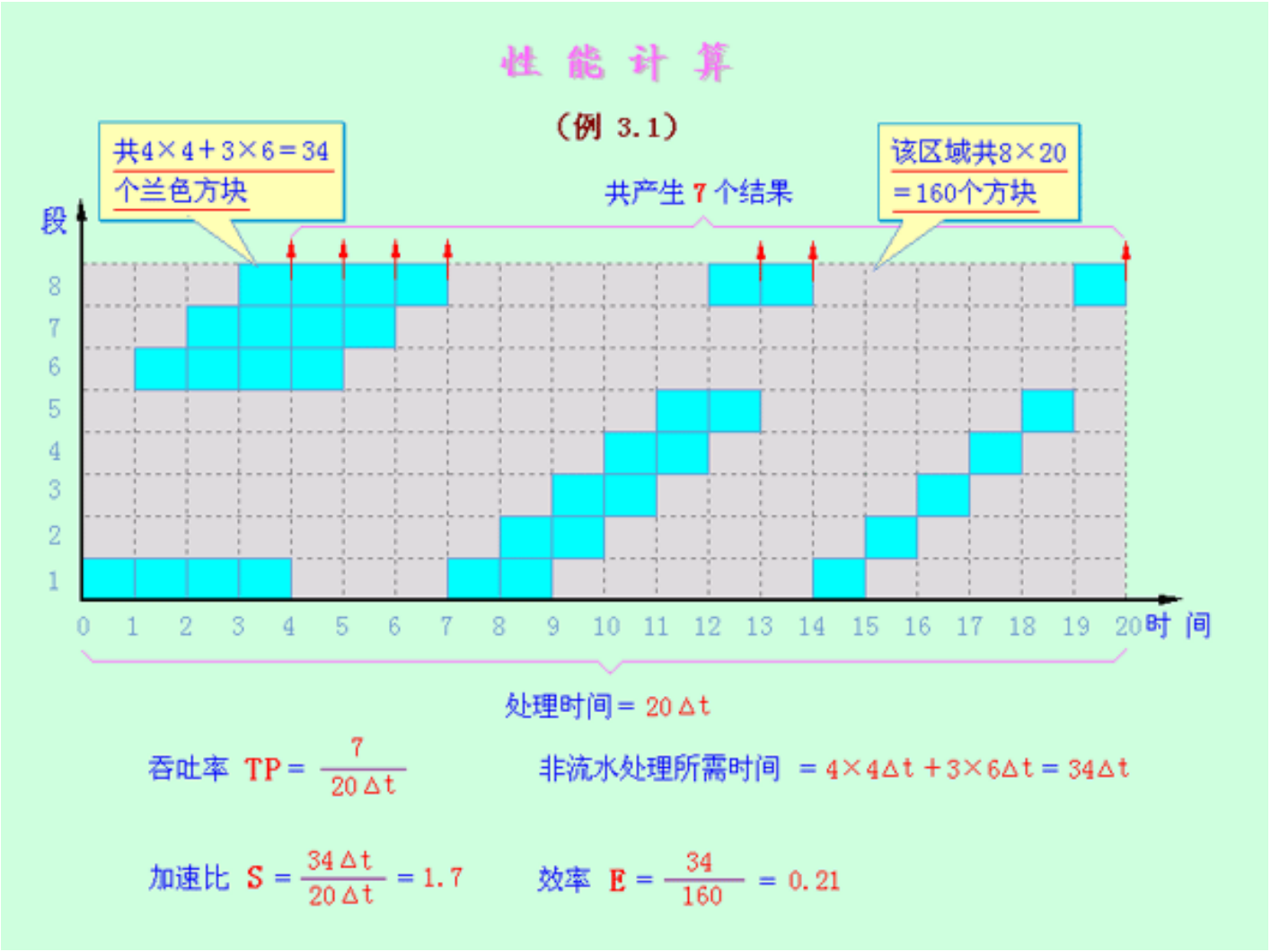
$$\sum_{i=1}^{4} A_i B_i = A_1B_1 + A_2B_2 + A_3B_3 + A_4B_4$$



② 画出时空图


时 空 图
(例 3.1)
段
8
7
6
5
4
3
2
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
时 间
输入
A1 A2 A3 A4
B1 B2 B3 B4
A1B1 A3B3
A2B2 A4B4
A1B1+A2B2
A3B3+A4B4
输出
A1B1 A3B3
A2B2 A4B4
A1B1+A2B2
A3B3+A4B4
A1B1+A2B2
+A3B3+A4B4

③ 根据时空图计算性能


性 能 计 算
(例 3.1)
共4×4+3×6=34
个兰色方块
共产生 7 个结果
该区域共8×20
=160个方块
段
8
7
6
5
4
3
2
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 时 间
处理时间 = 20 Δt
吞吐率 TP = 7 / 20 Δt
非流水处理所需时间 = 4×4Δt +3×6Δt = 34Δt
加速比 S = 34 Δt / 20 Δt = 1.7
效率 E = 34 / 160 = 0.21

流水线性能总结

- 流水线并不能减少（而且一般是增加）单条指令的执行时间，但能够提高吞吐率；
- 增加流水线的深度通常可以提高流水线性能；

流水线性能总结

- 流水线并不能减少（而且一般是增加）单条指令的执行时间，但能够提高吞吐率；
- 增加流水线的深度通常可以提高流水线性能；
- 流水线深度受限于流水线的延迟和额外开销；
- 流水线的额外开销包括：
  - *流水寄存器的延迟*
    - 建立时间：触发写操作的信号到达前寄存器输入保持稳定的时间；
    - 传输延迟：时钟信号到达后到寄存器输出可用的时间；
  - *时钟扭曲*
    - 时钟信号到达各流水寄存器的最大差值时间；