Optimal Control of Finite Capacity Batch Service Queues with General Holding Costs

Samuli Aalto
Helsinki University of Technology
Finland

Background

- Ph.D. Thesis (University of Helsinki)
  - S. Aalto (1998) “Studies in Queueing Theory”
- The part concerning the optimal control of batch service queueing systems based on two papers:
  - S. Aalto (1997) Reports of the Department of Mathematics 166, University of Helsinki
    - Poisson arrivals
  - S. Aalto (1998) to appear in Math Meth Oper Res
    - compound Poisson arrivals

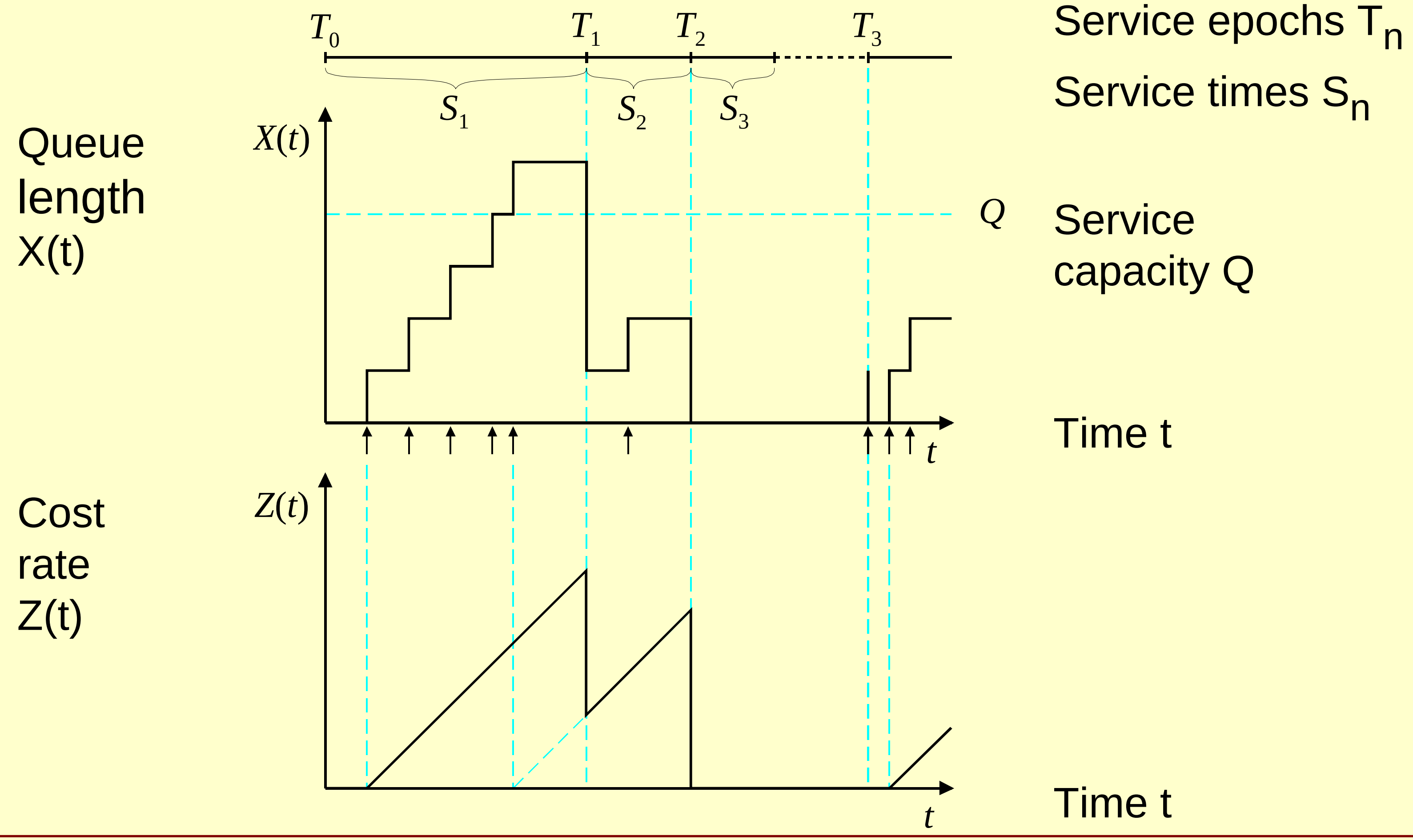
Contents

- Batch service queue
- Control problem
- Known results
- New results
- Open questions

Batch service queue

- In an ordinary queue
  - customers are served individually
- In a batch service queue
  - customers are served in batches of varying size
- Additional parameter needed:
  - Q = service capacity = the maximum number of customers served in a batch

Evolution

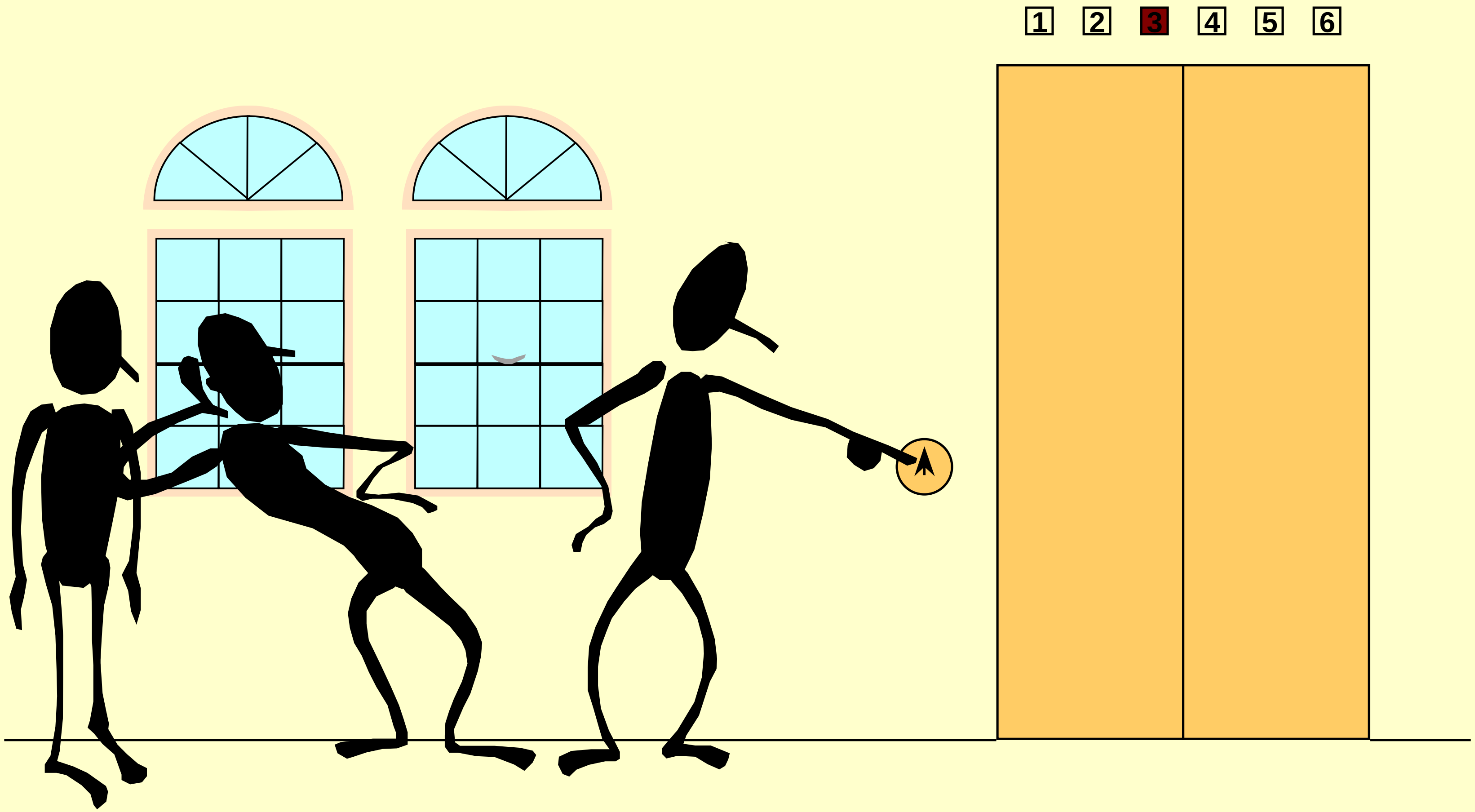


Queue length X(t)
Cost rate Z(t)
T0
T1
T2
T3
S1
S2
S3
X(t)
Q
Z(t)
t
t
Service epochs Tn
Service times Sn
Service capacity Q
Time t
Time t

Queueing models considered

- M/G(Q)/1
  - Poisson arrivals
  - generally distributed IID service times
  - single server with service capacity Q
- M^X/G(Q)/1
  - compound Poisson arrivals
  - generally distributed IID service times
  - single server with service capacity Q

An application



Contents

- Batch service queue
- Control problem
- Known results
- New results
- Open questions

Control problem

- Given
  - arrival process $A(t)$ and
  - service times S_n
- Determine
  - service epochs T_n
  - service batches B_n
- Operating policy $\pi = ((T_n),(B_n))$
  - should be **admissible**

Optimal control (1)

- Usual operating policy:
  - after a service completion, a new service is initiated **as soon as**

$$X(t) \geq 1$$

  - a service batch includes as many customers as possible
- This is certainly reasonable
- But is this optimal?

Optimal control (2)

- The answer depends on
  - how optimality is defined
- Thus, define first
  - the **cost structure**
- and then
  - the **objective function**

Cost structure

- **Holding costs**: Z(t)
  - described by the **cost rate** process Z(t)
  - cost rate depends on the number of waiting customers and the times they have been waiting
  - called **linear** if

$$Z(t) = h(X(t))$$

- **Serving costs**: $K + cB_n$
  - K per each service batch
  - c per each customer served

Objective function

- Minimize
  - the **long run average cost** ϕ^{π} or
  - the **discounted cost** V_{α}^{π}
- Among all the admissible operating policies π

Contents

- Batch service queue
- Control problem
- Known results
- New results
- Open questions

Known results

	Infinite service capacity $Q = \infty$	**Finite** service capacity $Q < \infty$
Linear holding costs $z = h(x)$	**Case A**: - Deb & Serfozo (1973) - Deb (1984)	**Case B**: - Deb & Serfozo (1973)
General holding costs $z = h(x,w)$	**Case C**: - Weiss (1979) - Weiss & Pliska (1982)	**Case D**

A	

Case A: linear holding costs & infinite service capacity

- Deb & Serfozo (1973) Adv Appl Prob
  - Poisson arrivals
  - average cost & discounted cost cases
- Deb (1984) Opsearch
  - compound Poisson arrivals
  - discounted cost case only
- **Result:**
  - h(x) is “uniformly increasing”
    => a **queue length threshold policy** is optimal

A	

Queue length threshold policies

- **Queue length threshold policy** π_x with threshold x:
  - after a service completion, a new service is initiated as soon as

$$X(t) \geq x$$

  - a service batch includes as many customers as possible
    - infinite capacity => all waiting customers
- Note: the usual operating policy = π_1

	B

Case B: linear holding costs & finite service capacity

- Deb & Serfozo (1973) Adv Appl Prob
  - Poisson arrivals
  - average cost & discounted cost cases
- **Result:**
  - h(x) is “uniformly increasing”
    => a **queue length threshold Q-policy** is optimal

B

Queue length threshold Q-policies

- Queue length threshold policy π_x with

$$x \leq Q$$

  is called a **queue length threshold Q-policy**
- Note:
  - finite capacity => service batch = min{X(t),Q}

A	B

Declaration for cases A and B

- Linear holding costs
  => cost rate remains constant between arrivals
  => no reason to start a new service until the next
  customer arrives
  => queue length threshold policies are optimal
- Semi-Markov decision technique can be applied:
  - system reviewed only when
    - a service has just been completed or
    - the server is free and a new customer arrives
- Sufficient to watch over the queue length process X(t)

Case C: general holding costs & infinite service capacity

C	

- Weiss (1979) Modeling and Simulation, Weiss & Pliska (1982) Opsearch
  - compound Poisson arrivals
  - average cost case only
- **Result:**
  - Z(t) is increasing (without limits when service is postponed forever)
    => a **cost rate threshold policy** is optimal

C

Cost rate threshold policies

- **Cost rate threshold policy** $\pi(z)$ with threshold z:
  - after a service completion, a new service is initiated as soon as

$$Z(t) \geq z$$

  - a service batch includes as many customers as possible
    - infinite capacity => all waiting customers

C	

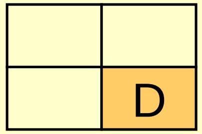
Declaration for case C

- Infinite capacity
  => queue can be emptied at every service epoch
  => no reason to watch over the queue length X(t)
- Cost rate Z(t) increasing (until the next service)
  => cost rate threshold policies are optimal
- Semi-Markov decision technique cannot be applied:
  - system needs to be reviewed continuously
- Each service starts a new regeneration cycle (as regards the **stationary** policies)
- Sufficient to watch over the cost rate process Z(t)

Contents

New results

	Infinite service capacity $Q = \infty$	**Finite** service capacity $Q < \infty$
Linear holding costs $z = h(x)$	**Case A**: - Deb & Serfozo (1973) - Deb (1984)	**Case B**: - Deb & Serfozo (1973) - Deb (1984)
General holding costs $z = h(x,w)$	**Case C**: - Weiss (1979) - Weiss & Pliska (1982)	**Case D**: - Aalto (1997) - Aalto (1998)



Case D:
General holding costs & finite service capacity (1)

- Aalto (1997) Univ of Helsinki
  - Poisson arrivals
  - average cost & discounted cost cases
- **Result:**
  - FIFO queueing discipline
  - **consistent** holding costs and
  - no serving costs included (K = c = 0)
    => a **cost rate threshold Q-policy** is optimal

D

Consistent holding costs

- Assume that

$$Z(t) = h(X(t), W(t))$$

where $W(t) = (W_1(t), W_2(t), \ldots)$ denotes

  – the vector of waiting times of the customers waiting at time t (in decreasing order)

- Holding costs are **consistent** if

$$x \le x' \ \text{ and } \ w \le w' \quad \Rightarrow \quad h(x,w) \le h(x',w')$$

- Examples: $h(x,w) = x$, $h(x,w) = w_1 + \ldots + w_x$

	D

Cost rate threshold Q-policies

- **Cost rate threshold Q-policy** $\pi_Q(z)$ with threshold z:
  - after a service completion, a new service is initiated as soon as

$$Z(t) \geq z \quad \text{or} \quad X(t) \geq Q$$

  - a service batch includes as many customers as possible
    - finite capacity => min{X(t),Q}

	D

Declaration

- Finite capacity
  => queue **cannot** be emptied at every service epoch
- First key observation:
  - To minimize holding costs, it is sufficient to consider the class of **Q-policies**
- Second key observation (due to single arrivals):
  - For each Q-policy, the queue becomes empty at every **non-trivial service epoch**
    => such an epoch starts a new regeneration cycle
    (as regards the **stationary** Q-policies)

D

Q-policies

- An operating policy π is called a **Q-policy** if
  - after a service completion, a new service is initiated **at latest** when

$$X(t) \geq Q$$

  - a service batch includes as many customers as possible
    - finite capacity => min{X(t),Q}

	D

Why just Q-policies?

- For each admissible policy π, it is possible to constract such a Q-policy π^Q that

$$X^{\pi^Q}(t) \le X^{\pi}(t) \qquad \forall t$$

- Due to FIFO principle and consistent holding costs, this implies that

$$Z^{\pi^Q}(t) \le Z^{\pi}(t) \qquad \forall t$$

D

Non-trivial service epochs

- **Idea**: Find such service completions that leave less than Q customers waiting
- For each Q-policy π, let

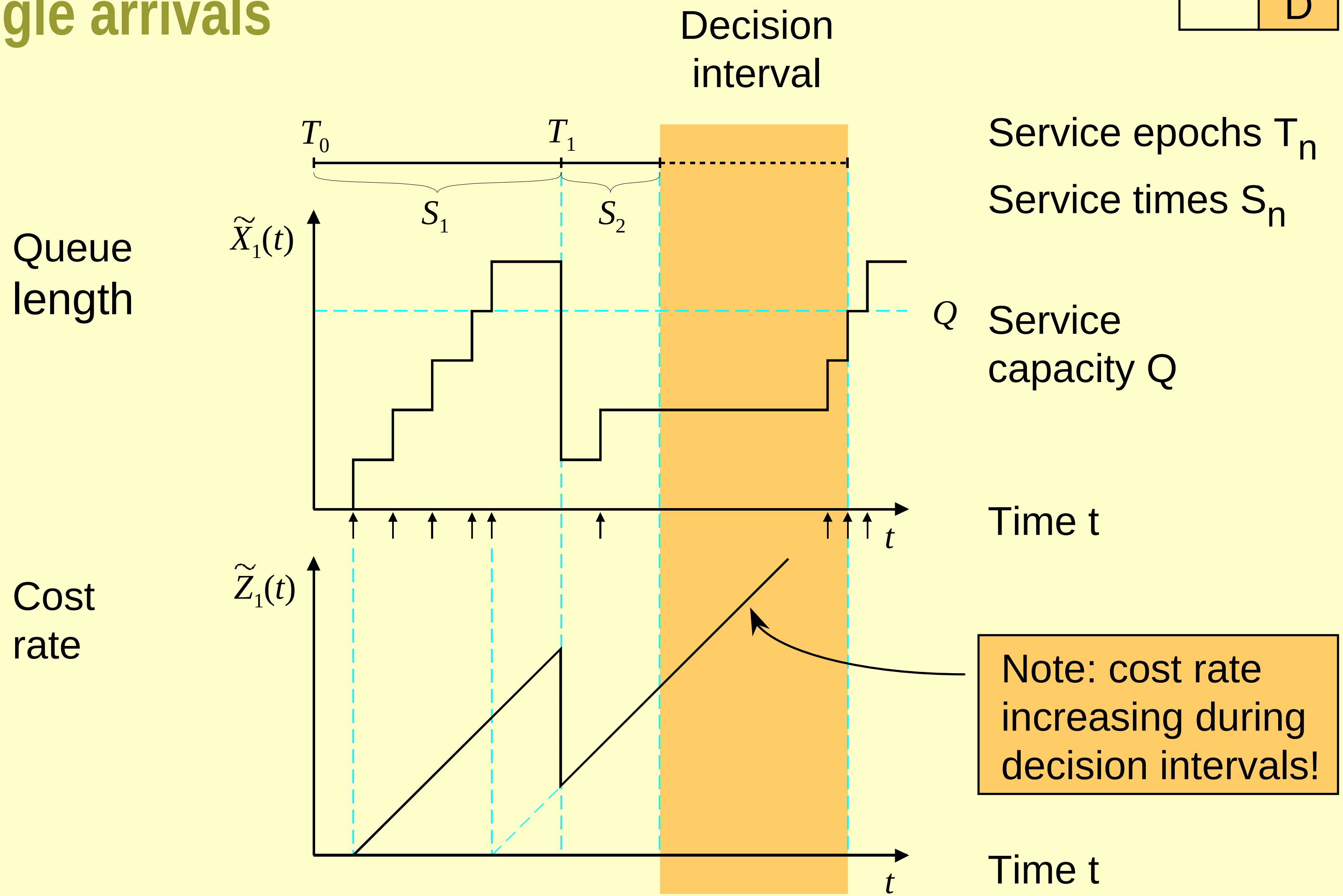
$$N_k^{\pi} = \min\left\{ n > N_{k-1}^{\pi} \,\middle|\, X^{\pi}(T_{n-1}^{\pi} + S_n) < Q \right\}$$

- Note: N_1 is the same for all Q-policies π
- **Non-trivial service epochs**:

$$\tilde{T}_k^{\pi} = T_{N_k^{\pi}}^{\pi}$$

D

Single arrivals

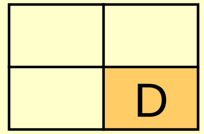


	D

Idea of the proof

- Let ϕ^{π} denote the aver. cost of a stationary policy π
- Then the cost rate threshold Q-policy $\pi_Q(\phi^{\pi})$ with threshold ϕ^{π} is better in the average cost sense

- Let V_{α}^{π} denote the disc. cost of a stationary policy π
- Then the cost rate threshold Q-policy $\pi_Q(\alpha V_{\alpha}^{\pi})$ with threshold αV_{α}^{π} is better in the discounted cost sense



Case D:
General holding costs & finite service capacity (2)

- Aalto (1998) Math Meth Oper Res
  - **compound** Poisson arrivals
  - discounted cost case only
- **Result:**
  - FIFO queueing discipline
  - **consistent** holding costs,
  - no serving costs included (K = c = 0) and
  - bounded arrival batches
    => a **general threshold Q-policy** is optimal

D

General threshold Q-policies

- **General threshold Q-policy** $\pi_Q(z,\zeta)$ with threshold z and value function ζ:
  - after a service completion, a new service is initiated as soon as

$$Z(t) + \zeta(X(t)) \geq z \quad \text{or} \quad X(t) \geq Q$$

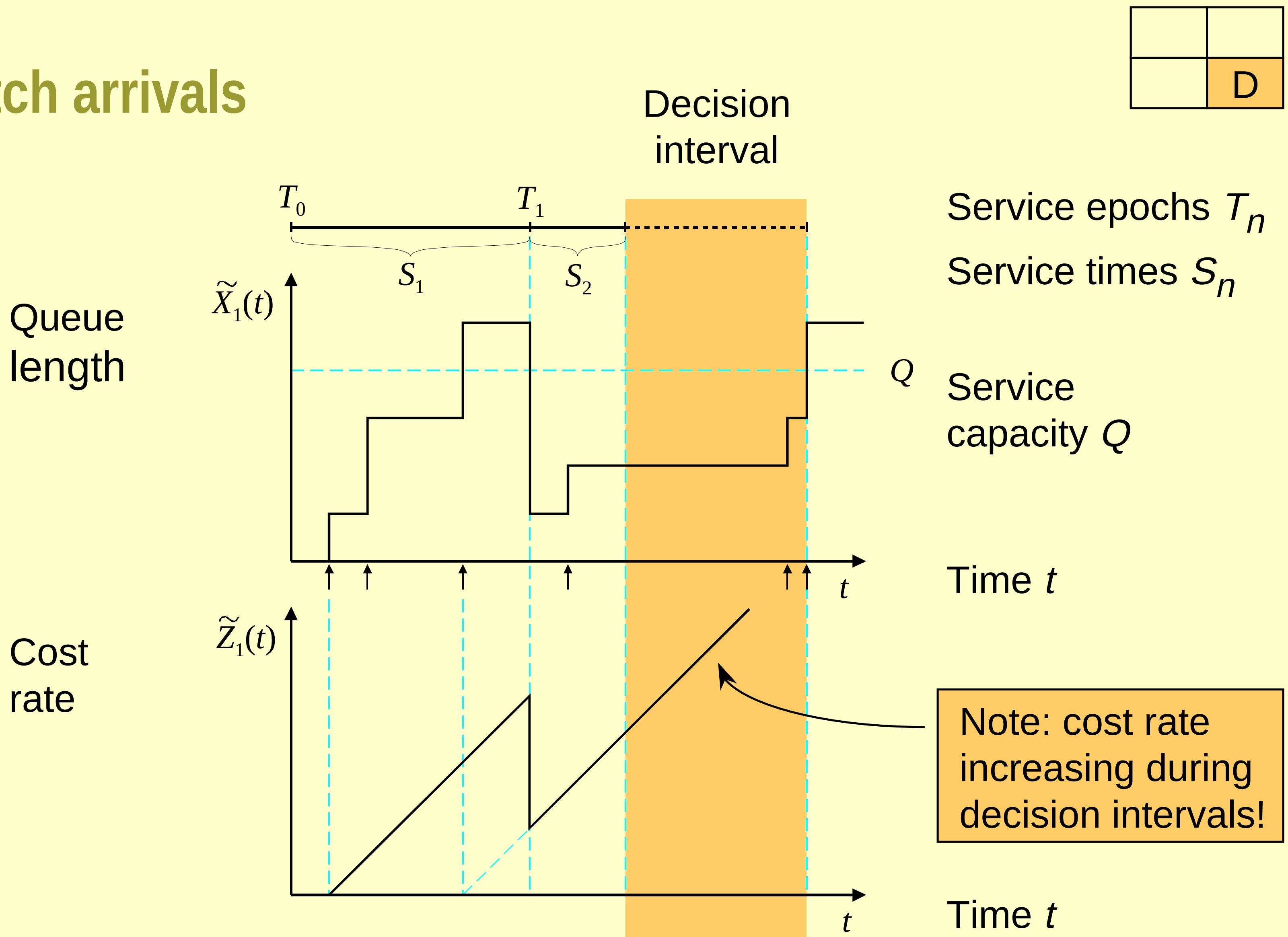
  - a service batch includes as many customers as possible
    - finite capacity => min{X(t),Q}

	D

Declaration

- First key observation:
  - To minimize holding costs, it is (still) sufficient to consider the class of **Q-policies**
- Second key observation (due to FIFO principle):
  - All those customers that remain waiting at a **non-trivial service epoch** arrived at that time
    => their waiting times are zero
    => such an epoch starts a new “semi-regeneration cycle” (as regards the **stationary** Q-policies)

Batch arrivals



Contents

- Batch service queue
- Control problem
- Known results
- New results
- Open questions

	D

Case D: General holding costs & finite service capacity

- How to get rid of the boundedness assumption concerning the arrival batches?
- If no serving costs are included ($K = 0$, $c = 0$),
  - Is it true that similar results are valid in the average cost case as in the discounted cost case?
- If serving costs are included ($K > 0$, $c > 0$),
  - What is the optimal policy in the average cost or discounted cost sense?

LOPPU