Service capacity and waiting times

Outline

Managerial motivation

Guided construction of the model

Worked example by hand

Case study and results

Sensitivity analysis: the price of a promise

Exercises

Learning objectives (1/2)

At the end of this chapter you will be able to:

- use the M/M/1 formulas inside an optimization problem;
- derive the optimum analytically and confirm it numerically;

Learning objectives (2/2)

At the end of this chapter you will be able to:

- put a price on a service promise;
- protect the sizing against uncertainty about demand.

Prerequisites

- From the module *Convex Optimization*, Chapter 3 «Convex and concave functions»: convex functions of one variable and their minimization
- From this module, Chapter 2 «Background: convexity, quadratic programming and KKT conditions»: convex quadratic programs, the KKT conditions and the multipliers as shadow prices
- From this module, Chapter 3 «The solver for nonlinear models»: functional constraints and the estimate of shadow prices by perturbation

Managerial motivation

Introduction: Managerial motivation

- In this chapter we decide **how much capacity** to assign to a service
- Problem class: **convex NLP** (M/M/1 queue)
- Queueing theory provides the quantitative relation between capacity and waiting
- Optimization turns it into a **decision**

Capacity versus waiting

- How much capacity should be assigned to a customer service, to a service desk, to a cloud service?
- More capacity costs money; too little capacity makes the **waiting times explode**
- Every service with random arrivals (call centres, emergency rooms, help desks, the API of a cloud service) lives with the same dilemma: capacity costs, and so does waiting
- Queueing theory gives the exact relation between the two; optimization chooses the right point

The central message

The message is counter-intuitive for anyone who reasons "by efficiency":

the optimal utilization is not 100%

It is by itself worth a discussion in a board meeting.

Summary: Managerial motivation

- Capacity and waiting are two conflicting costs
- Queueing theory links them with exact formulas
- The optimal utilization is **not** 100%

Guided construction of the model

Introduction: Guided construction of the model

- We describe the problem in words and fix the data
- We introduce the decision variable: the capacity μ
- We write the total cost model and solve it in **closed form**

The problem in words

- *We decide* how much service capacity to buy
- *The objective* is to minimize the total cost: the capacity that is bought plus the value of the time the customers spend in the system
- *The constraint* of a structural nature is only one: the stability of the queue ($\mu > \lambda$)
- In the variants the service promise $w \leq w_{\max}$ is added

The M/M/1 queue

- We use the simplest queueing model, in the standard notation **M/M/1**:
  - *Markovian* arrivals (a Poisson process)
  - *Markovian* service times (exponential)
  - *one* server
- For this system closed formulas hold
- What is new here: treating the capacity μ as an **optimization variable**

Data (input of the model)

Symbol	Type	Meaning
λ	$\in \mathbb{Q}_{>0}$	mean arrival rate of the customers (requests/hour)
c	$\in \mathbb{Q}_{>0}$	cost of one unit of capacity (€ per unit of μ per hour)
h	$\in \mathbb{Q}_{>0}$	value of one hour spent by a customer in the system (€)

Decision variable and derived quantities

- We introduce one continuous variable:

$$\mu = \text{service capacity (service rate, requests/hour)}$$

  with the stability requirement $\mu > \lambda$

- Mean time in the system (in the literature: W):

$$w(\mu) = \frac{1}{\mu - \lambda}$$

- Mean number of customers in the system (Little's law; in the literature: L):

$$l(\mu) = \frac{\lambda}{\mu - \lambda}$$

The model

Using this variable, the model for the problem is the following:

Total cost model: capacity + waiting

$$\min \quad c\,\mu + h\,\frac{\lambda}{\mu - \lambda} \tag{11.1a}$$

$$\text{subject to} \quad \mu > \lambda. \tag{11.1b}$$

Objective function and constraint

- The objective function is the sum of two terms:
  - $c\,\mu$ buys the capacity
  - $h \cdot l(\mu)$ monetizes the total time the customers spend in the system
- The objective is **convex** on $(\lambda, +\infty)$
- The constraint $\mu > \lambda$ defines the variable and guarantees the stability of the queue: with $\mu \leq \lambda$ the line grows without bound

Closed-form solution

Setting the derivative of the objective to zero gives:

$$C'(\mu) = c - \frac{h\,\lambda}{(\mu - \lambda)^2} = 0$$

$$\Longrightarrow \qquad \boxed{\mu^* = \lambda + \sqrt{h\,\lambda / c}}$$

Capacity = demand + safety stock

- The formula has the classical structure "capacity = demand + safety stock"
- The mean demand λ is covered and a margin $\sqrt{h\lambda/c}$ is added
- The margin grows with the value of the customers' time and decreases with the cost of capacity
- The margin grows like $\sqrt{\lambda}$: large systems enjoy economies of scale in the "buffer" (**pooling**)

Summary: Guided construction of the model

- A single continuous variable: the capacity $\mu > \lambda$
- Total cost $c\mu + h\lambda/(\mu - \lambda)$, convex
- Closed-form optimum: $\mu^* = \lambda + \sqrt{h\lambda/c}$
- The safety margin grows like $\sqrt{\lambda}$ (pooling)

Worked example by hand

Introduction: Worked example by hand

- We apply the formula for μ^* to a small service desk
- We compute utilization, mean time and cost at the optimum
- We see why "cutting capacity" is a bad idea

Example: a service desk with 8 customers per hour (1/2)

Data: $\lambda = 8$/h, $c = 2$ € per unit of capacity, $h = 4$ €/h.

- **Step 1.**

$$\mu^* = 8 + \sqrt{4 \cdot 8/2} = 8 + 4 = 12 \text{ customers/hour}$$

- **Step 2.** Utilization $\rho = 8/12 = 66.7\%$
- Mean time $w = 1/(12 - 8) = 15$ minutes
- Cost $C = 2 \cdot 12 + 4 \cdot 8/4 = 32$ €/h

Example: a service desk with 8 customers per hour (2/2)

- **Step 3 — why not cut the capacity?**
- With $\mu = 9$ ($\rho = 89\%$, "more efficient"):

$$C = 18 + 32/1 = 50 \text{ €/h}$$

  much worse

- The apparent waste (33% of unused capacity) is exactly what keeps the queues short

Summary: Worked example by hand

- At the optimum: $\mu^* = 12$, $\rho = 66.7\%$, $w = 15$ minutes, $C = 32$ €/h
- Raising the utilization to 89% pushes the cost up to 50 €/h
- Unused capacity is not waste: it keeps the queues short

Case study and results

Introduction: Case study and results

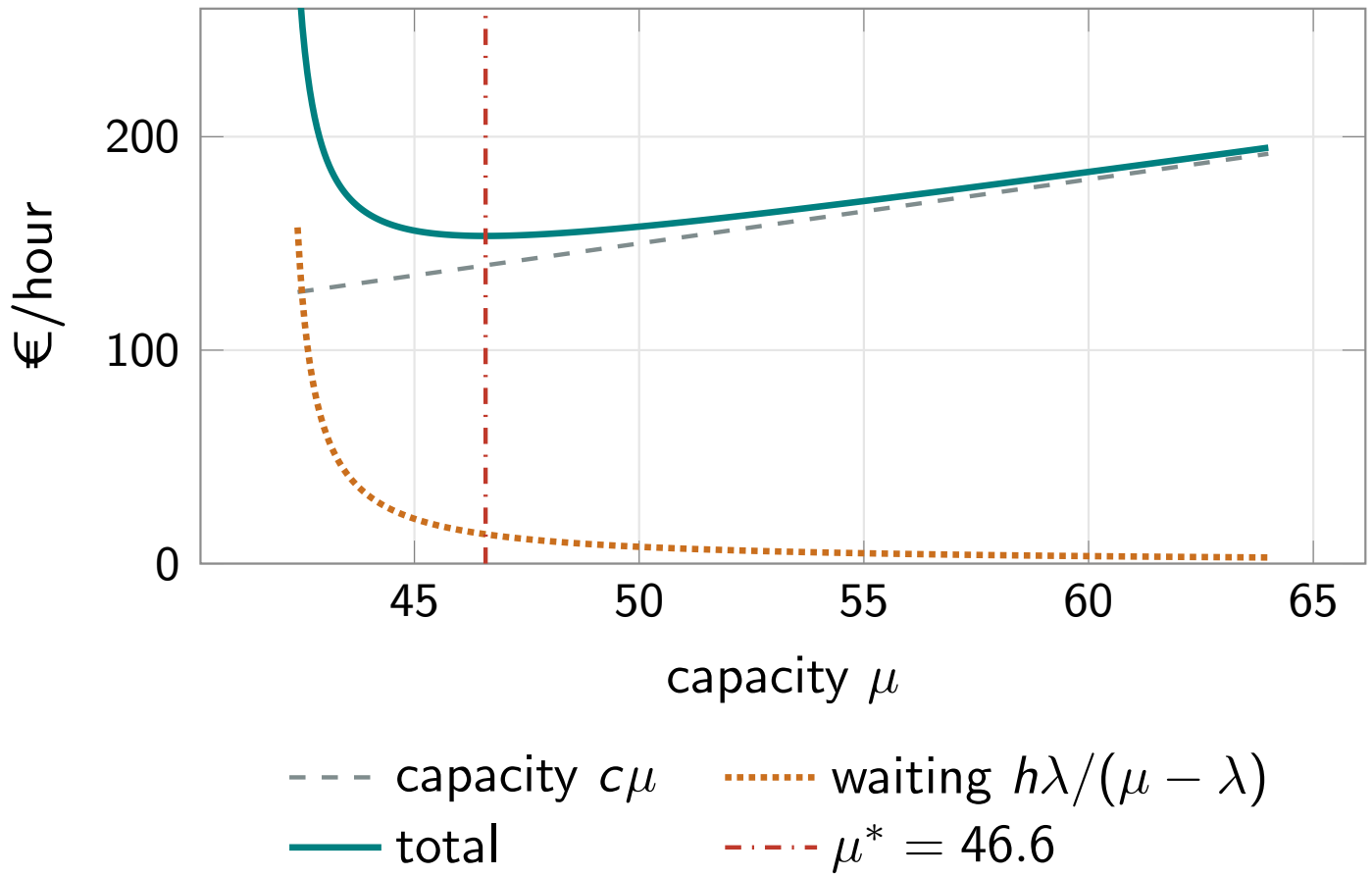
- A customer service: $\lambda = 42$ requests/hour, $c = 3$ €, $h = 1.5$ €
- We compare the analytical solution with the one of Gurobi
- We look at the two key plots: total cost and the "utilization wall"

Results

Customer service: $\lambda = 42$ requests/hour, $c = 3$ €, $h = 1.5$ €.

```
Analytical: mu* = 42 + sqrt(1.5*42/3) = 46.583
            cost 153.495 EUR/h
Gurobi    : mu* = 46.582 (bilinear w(mu-lam) = 1,
            NonConvex=2: matches)
At optimum: rho = 90.2%   w = 13.1 minutes
```

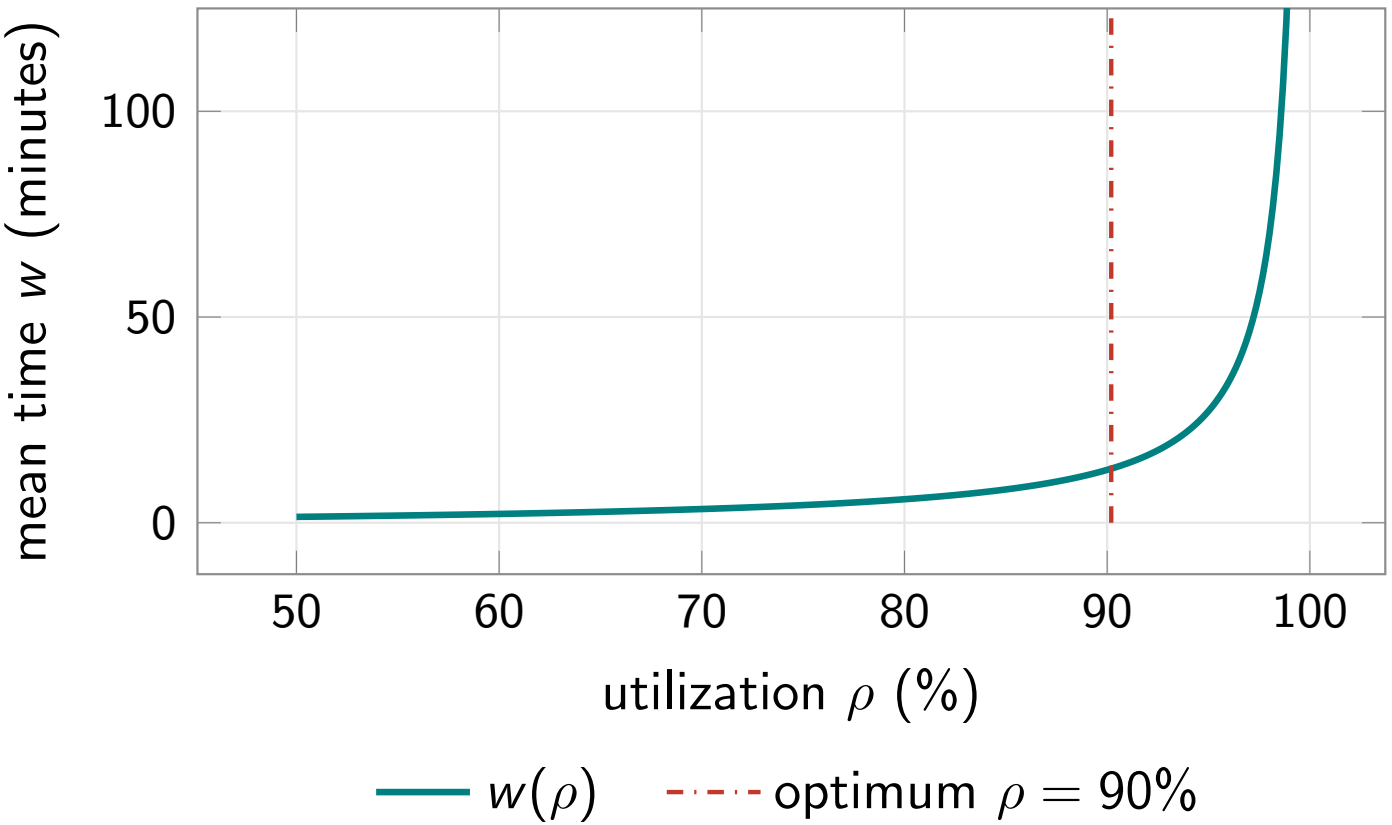
The total cost is convex in μ



Reading the plot

- The capacity cost $c\mu$ is increasing
- The waiting cost $h\lambda/(\mu - \lambda)$ is decreasing
- Their sum, the total cost, is convex
- The minimum is at $\mu^* = 46.6$

The utilization wall



Beyond 90% the waiting time explodes hyperbolically.

Managerial insight

- The utilization wall is perhaps the most important plot to show to a manager
- Between $\rho = 90\%$ and $\rho = 99\%$ the waiting time is multiplied by ten
- "Saturating the resources" and "giving good service" are objectives in **mathematical**, not organizational, conflict
- No reorganization removes the hyperbola $w = 1/(\mu - \lambda)$

Summary: Case study and results

- Optimum: $\mu^* = 46.58$, cost 153.50 €/h; Gurobi confirms it
- At the optimum $\rho = 90.2\%$ and $w = 13.1$ minutes
- Close to 100% utilization the waiting time explodes: saturating resources and serving well are in conflict

Sensitivity analysis: the price of a promise

Introduction: the price of a promise

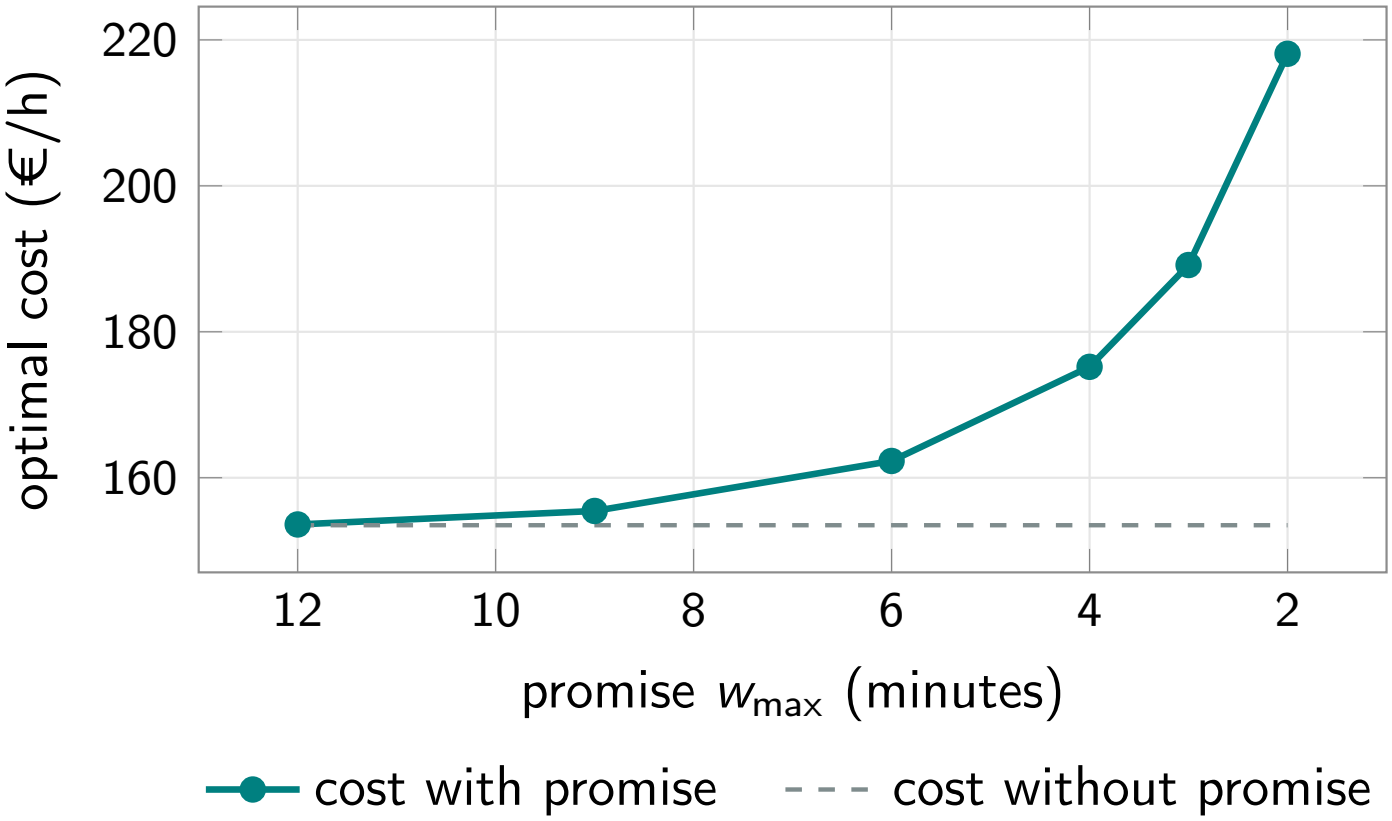
- Marketing promises a mean time below a threshold $w_{\max}$
- How much does keeping this promise cost?
- We compute the marginal cost of the promise (shadow price)
- Finally we protect the sizing against uncertainty on λ (robustness)

The service promise

- For a "mean time below $w_{\max}$" we need $\mu \geq \lambda + 1/w_{\max}$
- The constraint becomes active when it is tighter than the economic optimum (13.1 minutes)

```
w_max = 12 min: mu = 47.00  cost 153.60   (almost free)
w_max =  9 min: mu = 48.67  cost 155.45
w_max =  6 min: mu = 52.00  cost 162.30
w_max =  4 min: mu = 57.00  cost 175.20
w_max =  3 min: mu = 62.00  cost 189.15
w_max =  2 min: mu = 72.00  cost 218.10   (+42% on the base cost)
```

How much a more ambitious service promise costs



Reading the plot

- The axis is reversed: towards the right the promise becomes more ambitious
- The dashed line is the cost without any promise
- Up to 12 minutes the promise is almost free
- Below 4 minutes every additional minute cut costs explosively

The marginal cost of the promise (1/2)

- The shadow price of the constraint grows explosively:
  - going from 12 to 9 minutes costs 1.85 €/h
  - from 3 to 2 minutes costs 28.95 €/h
- When the constraint is active, $\mu = \lambda + 1/w_{\max}$ and

$$C = c\lambda + c/w_{\max} + h\lambda\, w_{\max}$$

- From which (with $w_{\max}$ in hours):

$$\frac{dC}{dw_{\max}} = -\frac{c}{w_{\max}^2} + h\lambda$$

The marginal cost of the promise (2/2)

- At $w_{\max} = 6$ minutes it is equal to

$$-300 + 63 = -237 \text{ €/h per hour of promise}$$

  exactly the finite-difference estimate of the script
- This is the number to give to marketing *before* it signs the SLA
- SLA = *Service Level Agreement*, the contractual commitment on the service level

Robustness

- If λ is uncertain in $[36, 48]$, sizing on the nominal optimum $\mu^* = 46.58$ is **unstable**
- With $\lambda = 48 > \mu^*$ the queue diverges (infinite cost)
- Minimizing the cost of the worst case gives $\mu_{\text{rob}} = 52.90$, with a worst-case cost of 173.39 €/h
- An insurance premium of $\approx 13\%$ is paid in order to be protected from a disaster
- Capacity decisions must always be checked against the *maximum* plausible demand, not only against the mean

Summary: the price of a promise

- A promise $w \leq w_{\max}$ imposes $\mu \geq \lambda + 1/w_{\max}$
- The marginal cost $dC/dw_{\max} = -c/w_{\max}^2 + h\lambda$ explodes for ambitious promises
- Sizing on the mean demand can make the queue unstable
- The robust solution costs about 13% more and protects from disaster

Exercises

Introduction: Exercises

- Four exercises to consolidate the concepts of the chapter
- Checking the marginal cost, quadratic penalty, pooling, capacity-service frontier
- They are solved with the formulas seen and with the lab script

Exercise: check

- Derive on your own the formula

$$\frac{dC}{dw_{\max}} = -\frac{c}{w_{\max}^2} + h\lambda$$

- Verify it numerically at $w_{\max} = 4$ and 9 minutes by finite differences

Exercise: quadratic penalty

- Replace the linear waiting cost with

$$h \left[\frac{1}{\mu - \lambda} \right]^2$$

- Derive μ^* analytically
- Compare the optimal utilization with the linear case

Exercise: two queues vs one

- Two separate service desks each receive $\lambda = 21$
- Alternatively, a single system receives $\lambda = 42$ with aggregate capacity
- At equal capacity cost, which configuration gives shorter waiting times?
- This is the quantitative argument for **pooling**

Exercise: capacity-service frontier

- Plot the frontier (cost, w) by varying $w_{\max}$ continuously
- Identify the point beyond which every additional promised minute cut costs more than 10 €/h

Summary (1/2)

Model:

- M/M/1 queue, capacity μ as the variable, stability constraint $\mu > \lambda$
- Total cost $c\mu + h\lambda/(\mu - \lambda)$, convex
- Optimum $\mu^* = \lambda + \sqrt{h\lambda/c}$: demand + safety stock

Managerial message:

- The optimal utilization is not 100%
- Close to saturation the waiting time explodes (utilization wall)

Summary (2/2)

Service promises:

- $w \leq w_{\max}$ becomes $\mu \geq \lambda + 1/w_{\max}$
- The shadow price explodes for ambitious promises: compute it before signing the SLA

Robustness:

- Size on the maximum plausible demand, not only on the mean
- The insurance premium ($\approx 13\%$) prevents the queue from diverging