Queueing models for appointment-driven systems

Stefan Creemers and Marc Lambrecht

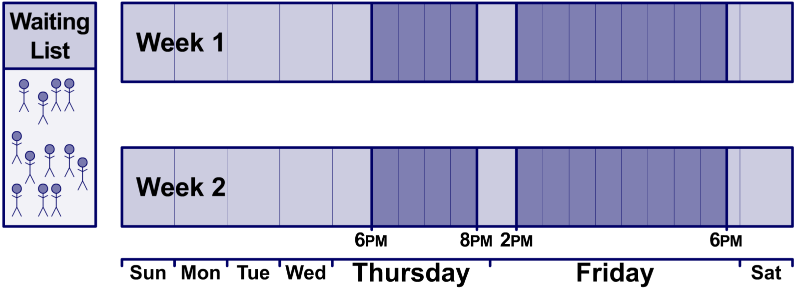
Research Center for Operations Management

Catholic University Leuven

July 28, 2008

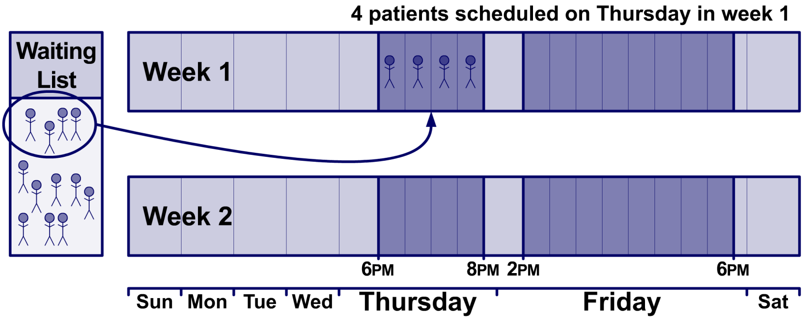
Problem setting: example doctor's office

- Opening hours on Thursday from 6PM until 8PM and on Friday from 2PM until 6PM
- On Thursday a maximum of 4 patients is served, on Friday up to 8 patients receive service



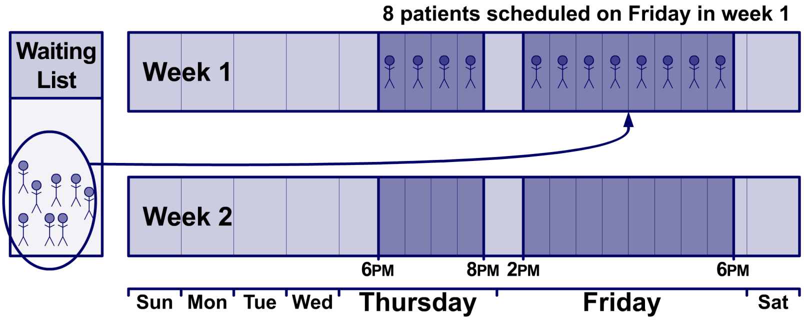
Problem setting: example doctor's office

- Opening hours on Thursday from 6PM until 8PM and on Friday from 2PM until 6PM
- On Thursday a maximum of 4 patients is served, on Friday up to 8 patients receive service



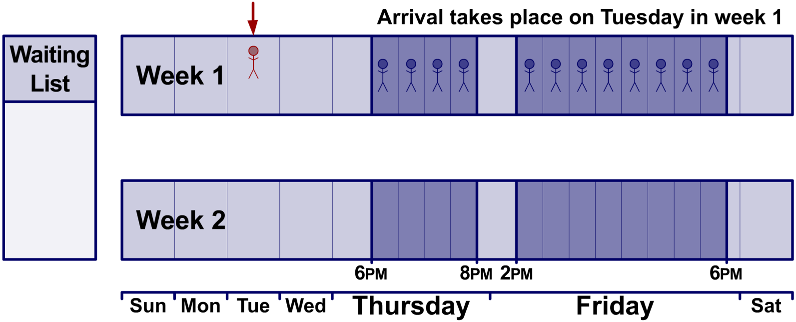
Problem setting: example doctor's office

- Opening hours on Thursday from 6PM until 8PM and on Friday from 2PM until 6PM
- On Thursday a maximum of 4 patients is served, on Friday up to 8 patients receive service



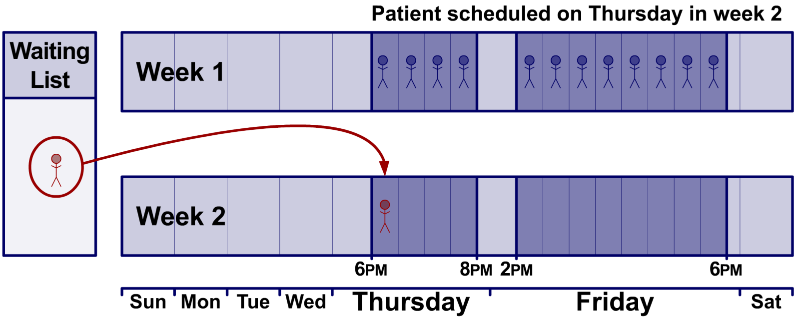
Problem setting: example doctor's office

- Opening hours on Thursday from 6PM until 8PM and on Friday from 2PM until 6PM
- On Thursday a maximum of 4 patients is served, on Friday up to 8 patients receive service

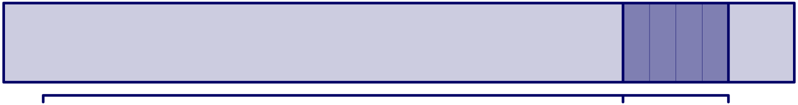


Problem setting: example doctor's office

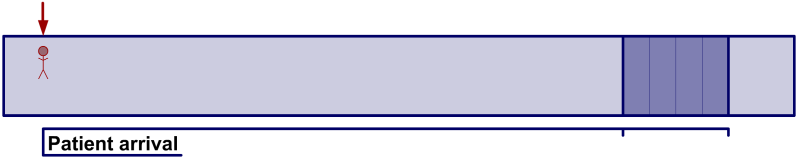
- Opening hours on Thursday from 6PM until 8PM and on Friday from 2PM until 6PM
- On Thursday a maximum of 4 patients is served, on Friday up to 8 patients receive service



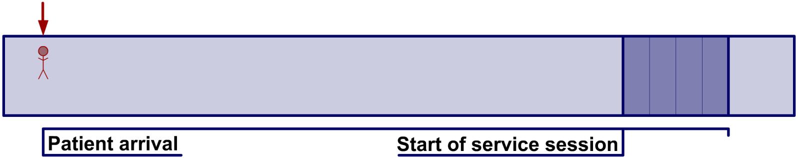
Problem setting: patient point of view



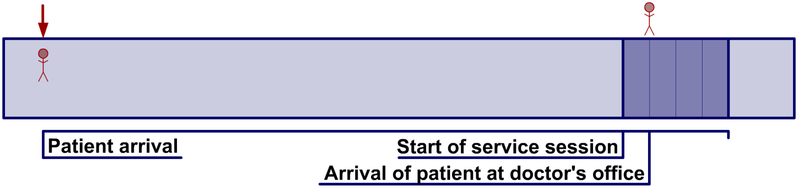
Problem setting: patient point of view



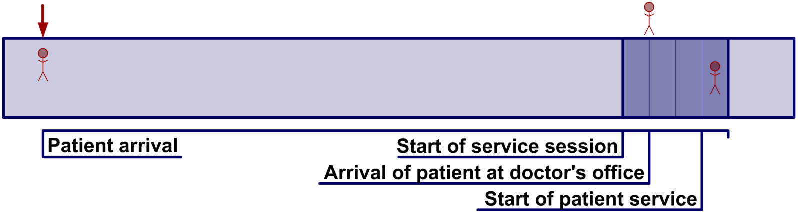
Problem setting: patient point of view



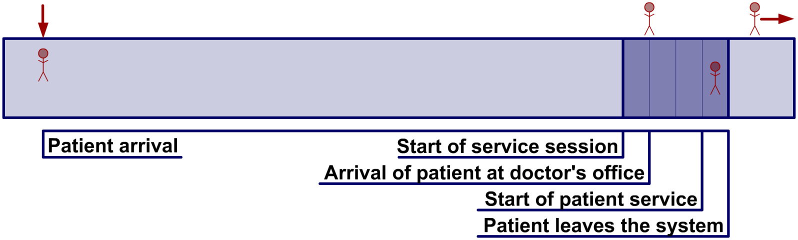
Problem setting: patient point of view



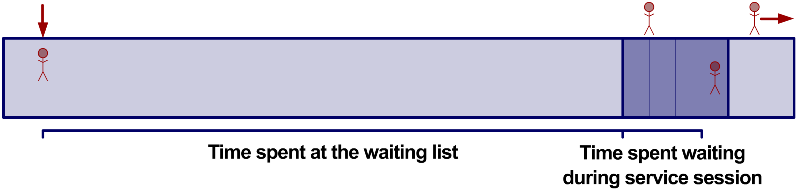
Problem setting: patient point of view



Problem setting: patient point of view



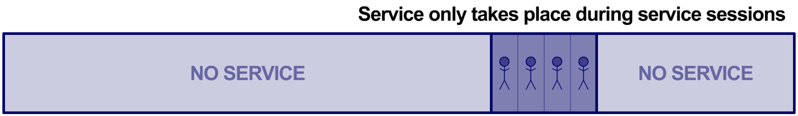
Problem setting: patient point of view



Main measures of interest:

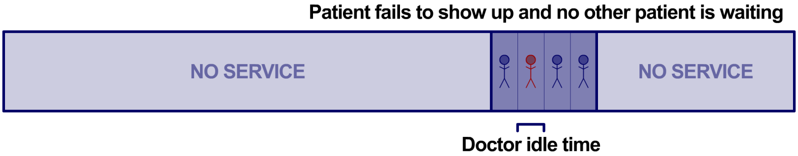
- Patient waiting time at the waiting list
- Patient waiting time during a service session

Problem setting: server point of view



Main measures of interest:

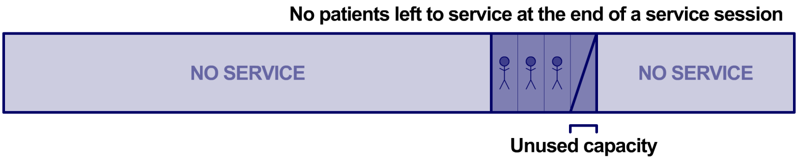
Problem setting: server point of view



Main measures of interest:

- Server idle time

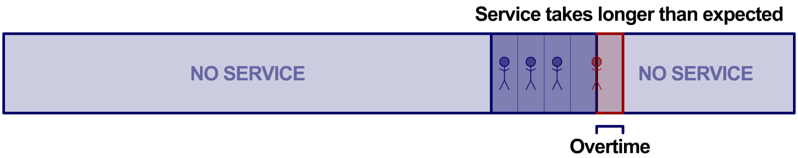
Problem setting: server point of view



Main measures of interest:

- Server idle time
- Unused server capacity at the end of a service session

Problem setting: server point of view



Main measures of interest:

- Server idle time
- Unused server capacity at the end of a service session
- Server overtime

Problem setting: performance measures

- Measures of interest:
    - Patient-related measures:
        - Patient waiting time at the waiting list
        - Patient waiting time at the doctor's office
    - Server-related measures:
        - Unused server capacity
        - Server idle time
        - Server overtime
- These performance measures can be used in an optimization procedure to aid strategic decision-making:
    - Optimal location in space and time of service sessions
    - Optimal number of patients to be treated during each service session
- Currently no models are available that provide all of these performance measures

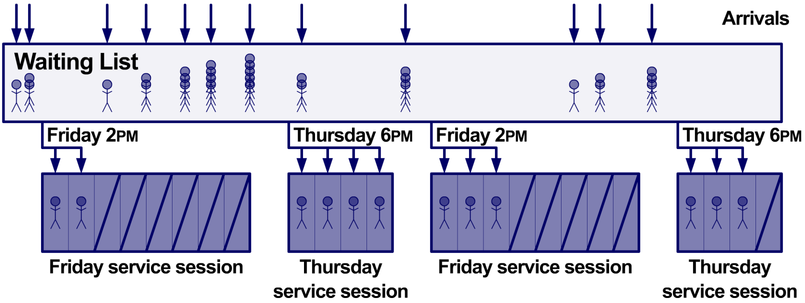
Two queues, two models

- In appointment-driven systems, patients join two queues:
  - The waiting list
  - The queue at the service facility
- Both queues are modeled using distinct models
  - A vacation model (AMQ), that regulates the assignment of patients to service sessions
  - An appointment system (AS), that models everything that happens during the service session itself
- Combining results of both models allows the assessment of performance measures at an appointment-driven system as a whole

Appointment system (AS)

- During a service session, how should patients be scheduled in order to optimize:
  - Patient waiting time at the service facility
  - Staff performance (overtime, idle time, unused capacity)
- Two approaches:
  - Procedures to obtain (optimal) interarrival times of patients
  - Appointment scheduling rules (ASR)

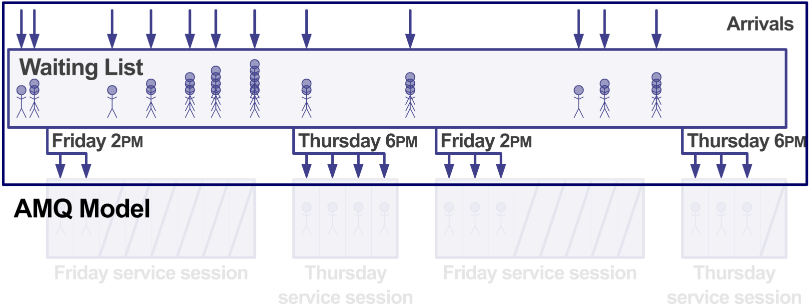
Patient assignment system (AMQ)



We obtain:

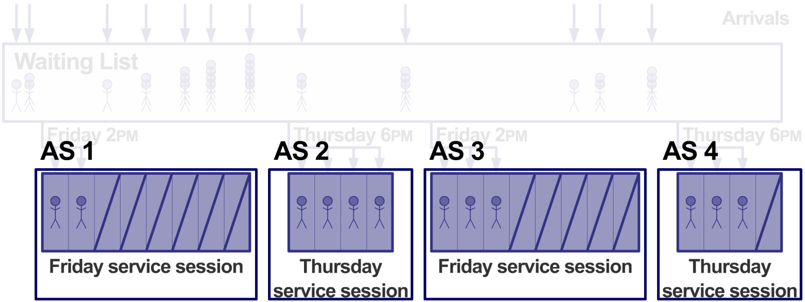
- Waiting list performance measures
- Distribution of the number of patients present at the start of a service session

Myopic separately, Global combined



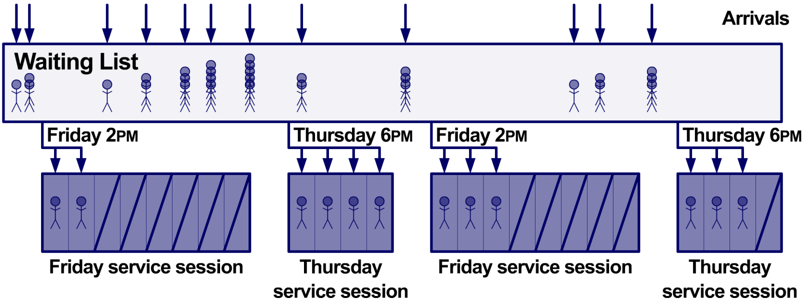
- Separately only a myopic view is offered:
  - The AMQ only observes the waiting list
  - The AS only observes a realization of a single service session
- Together they describe the appointment-driven system

Myopic separately, Global combined



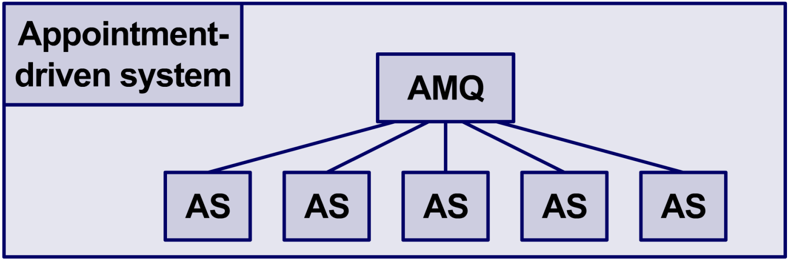
- Separately only a myopic view is offered:
  - The AMQ only observes the waiting list
  - The AS only observes a realization of a single service session
- Together they describe the appointment-driven system

Myopic separately, Global combined



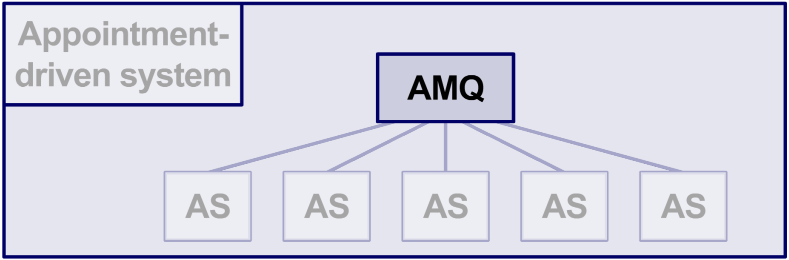
- Separately only a myopic view is offered:
  - The AMQ only observes the waiting list
  - The AS only observes a realization of a single service session
- Together they describe the appointment-driven system

Overview research



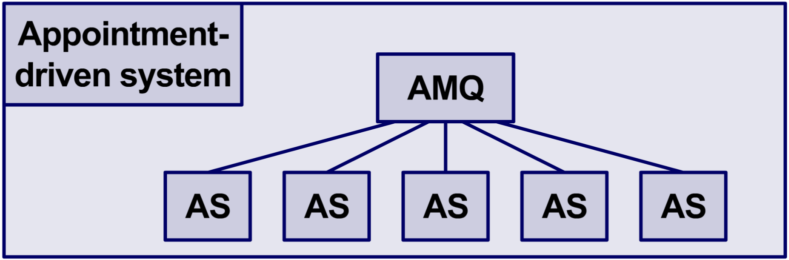
Queueing models for appointment-driven systems
Creemers S. and Lambrecht M.R. (2008)
First round of revision in Annals of OR

Overview research



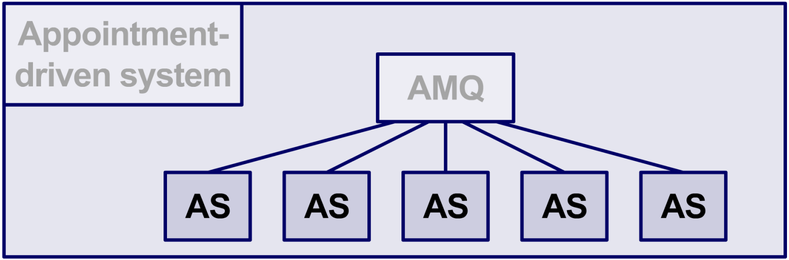
Advanced queueing models for appointment-driven systems
Creemers S. and Lambrecht M.R. (2008)

First model overview



Queueing models for appointment-driven systems
Creemers S. and Lambrecht M.R. (2008)
Accepted for publication in Annals of OR

First model overview



Queueing models for appointment-driven systems
Creemers S. and Lambrecht M.R. (2008)
Accepted for publication in Annals of OR

AS: Problem setting

- Problem setting: during a single service session, schedule a set of patients as to optimize some some performance measures
- Appointment Scheduling Rule (ASR): block appointment rule
- Performance measures:
  - Patient waiting time at the service facility
  - Server overtime
  - Server idle time
  - Unused server capacity
- Methodology: closed form results due to the use of the gamma distribution to model the service process

AS: Assumptions

- Block appointment rule (i.e. all patients arrive at the start of the service session)
- Patients all show up and arrive on time
- All patients in the service session are served
- Patient service time distribution is assumed to be i.i.d.
- Patient service time distribution is approximated using a gamma distribution of parameters α and θ (which matches the first two moments)
- Service takes place in an uninterrupted fashion

AS: Model

- Patient waiting time is maximized, staff overtime is minimized (there is no idle time)
- The uninterrupted duration of servicing n patients follows a gamma distribution of parameters $n\alpha$ and θ
- We obtain closed form results for average overtime performed ($E[\phi_n]$) and average waiting time ($E[W_{AS}]$):

$$E[\phi_n] = \frac{[-O\gamma(n\alpha, {}^{O}\!/_{\theta})] + \left[O^{n\alpha}\left(\frac{O}{\theta}\right)^{-n\alpha}\theta^{1-n\alpha}\gamma(1+n\alpha, {}^{O}\!/_{\theta})\right]}{\Gamma(n\alpha)},$$

$$E[W_{AS}] = \frac{n-1}{2\mu},$$

where:

- O indicates when the server works overtime
- μ is the service rate of a single patient

AS: Example

- Input parameters:
  - A maximum of 4 patients are to be served in a duration of 120 minutes (i.e. $n = 4$ and $O = 120$)
  - Service of an individual patient takes on average half an hour (i.e. $\mu = 1/30$)
  - SCV of service times equals $2/3$
- Results (validated through simulation):

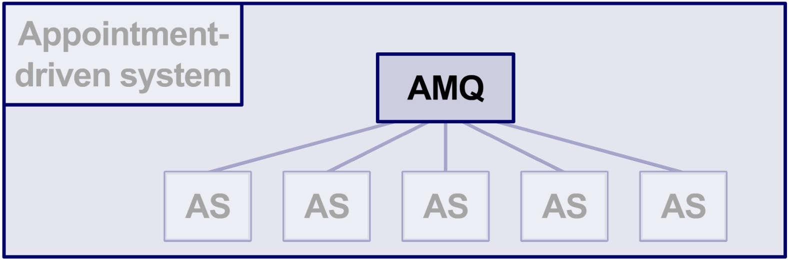
n	0	1	2	3	4
$P(n)$	0.1	0.2	0.2	0.2	0.3
$E[\phi_n]$	0	0.1577	1.6360	7.1308	19.275
$E[W_{AS}]$	0	0	15	30	45

  - Mean overtime over all AS: 7.57 minutes
  - Mean patient waiting time over all AS: 22.5 minutes

AS: Output

- For a given number of patients served at a given service session we obtain:
  - The average waiting time of a patient at the service facility
  - The average amount of overtime performed
  - The average amount of idle time (which always equals zero in our AS)
- In order to obtain general results at the appointment-driven system, these data need to be aggregated using the probability distribution obtained at the AMQ

First model overview



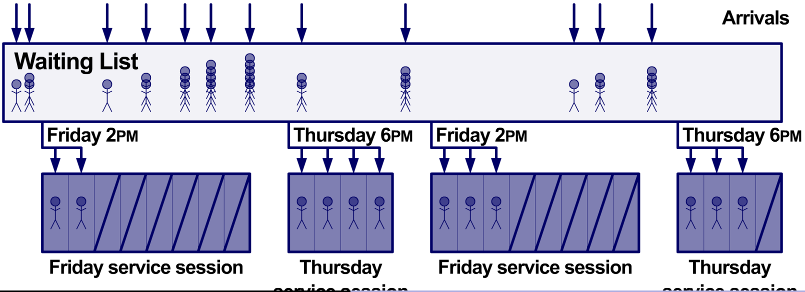
Queueing models for appointment-driven systems
Creemers S. and Lambrecht M.R. (2008)
Accepted for publication in Annals of OR

AMQ: Problem setting

- Problem setting: assigning patients to the first available service session
- Performance measures:
  - Average number of patients in the waiting list
  - Distribution of the number of patients present in queue at the start of a particular service session (i.e. the probability of having an AS with n patients to be served)
- Methodology: vacation model and matrix analytical techniques to obtain the stationary distribution of the corresponding CTMC (Continuous Time Markov Chain)

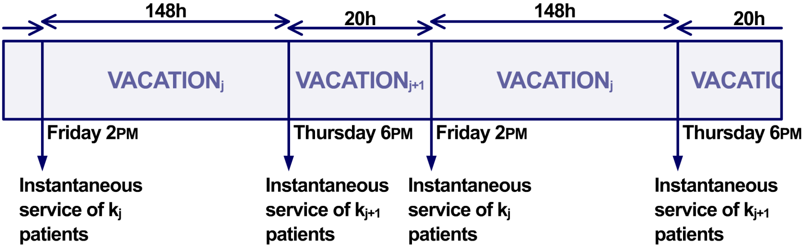
AMQ: Assumptions

- Time-independent, Poisson arrivals that are allowed to occur at any time
- Gated *k*-limited service discipline
- Bulk service queue with instantaneous service
- State-dependent, deterministic, cyclic vacations



AMQ: Assumptions

- Time-independent, Poisson arrivals that are allowed to occur at any time
- Gated *k*-limited service discipline
- Bulk service queue with instantaneous service
- State-dependent, deterministic, cyclic vacations



AMQ: basic idea of the model

i/j	0	1	2	3	4	5	6	7	...
0	μ	λ	0	0	0	0	0	0	...
1	μ	0	λ	0	0	0	0	0	...
2	μ	0	0	λ	0	0	0	0	...
3	μ	0	0	0	λ	0	0	0	...
4	μ	0	0	0	0	λ	0	0	...
5	0	μ	0	0	0	0	λ	0	...
6	0	0	μ	0	0	0	0	λ	...
7	0	0	0	μ	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...

With:

- μ the rate until the next vacation
- λ the arrival rate

AMQ: model

- Use of Erlang distribution of sufficient phases to model the deterministic vacation durations
- Threedimensional stochastic process $X = \{X(t) : t \geq 0\}$
- The statespace may be represented by triplets (Q, j, v) where:
  - Q is the queue size,
  - j is the type of vacation (e.g. Thursday or Friday)
  - v is the phase of the vacation process
- Possible transitions:
  - Arrival of a new patient: $(Q, j, v) \longrightarrow (Q + 1, j, v)$
  - End of a vacation: $(Q, j, V + 1) \longrightarrow (max((Q - k_j), 0), j + 1, 1)$

AMQ: model

- Infinitesimal generator $\mathbf{Q}$ is endowed with a QBD (Quasi-Birth-Death) structure:

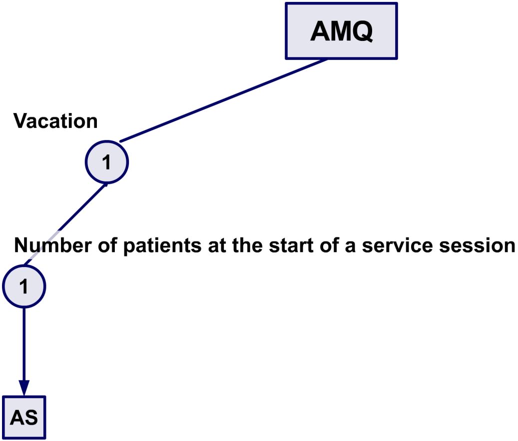
$$\mathbf{Q} = \begin{bmatrix} \hat{\mathbf{L}} & \mathbf{F} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{B} & \mathbf{L} & \mathbf{F} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B} & \mathbf{L} & \mathbf{F} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B} & \mathbf{L} & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \ddots \end{bmatrix},$$

  where $\mathbf{0}$ is a matrix of appropriate size containing only zeros and where $\hat{\mathbf{L}}$, $\mathbf{L}$, $\mathbf{F}$ and $\mathbf{B}$ are the respective "local", "forward" and "backward" transition rate matrices
- The stationary distribution of such structured CTMC is efficiently obtained using matrix analytical techniques

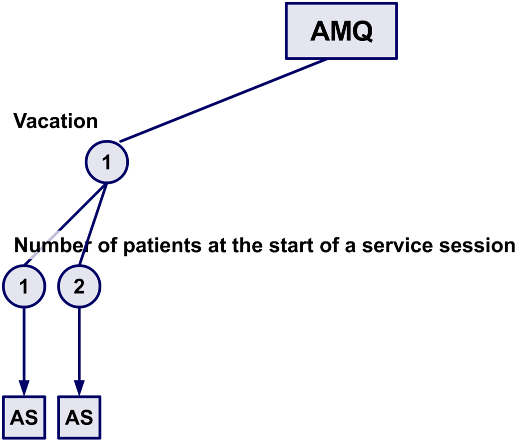
AMQ: Output

- The average patient waiting time at the waiting list
- The distribution of the number of patients present at the start of a particular service session (e.g. on Thursday or on Friday)
- Combined with the analysis of the AS that correspond to those service session realizations (with non-zero probability), general results for an appointment-driven system are obtained.

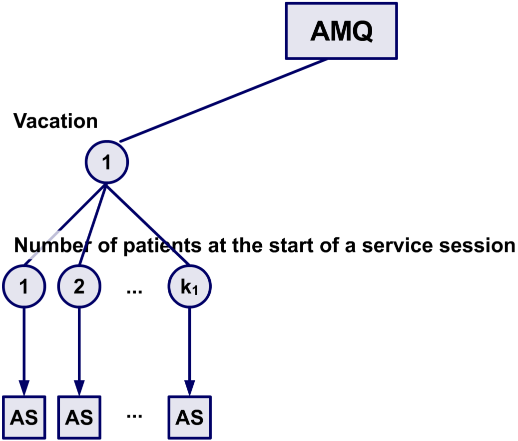
Appointment-driven system: General results



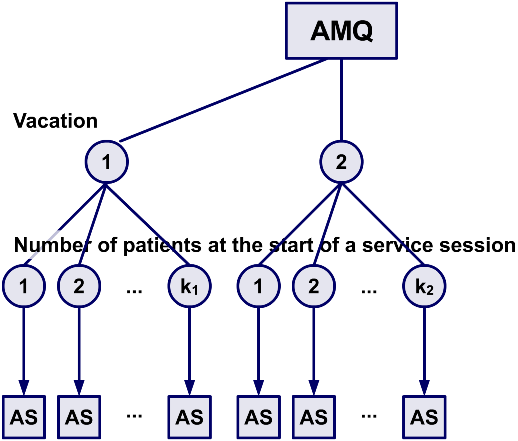
Appointment-driven system: General results



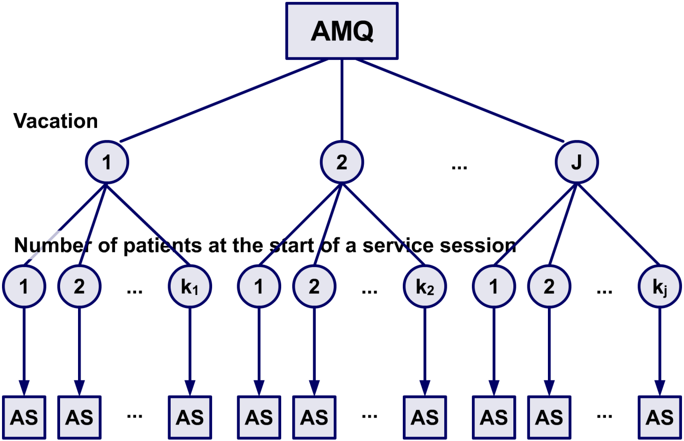
Appointment-driven system: General results



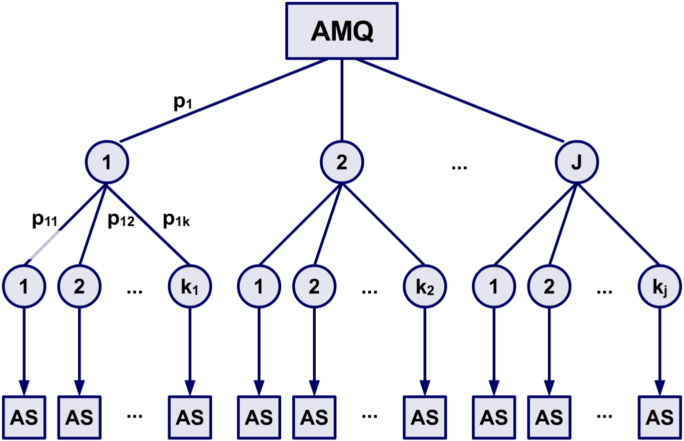
Appointment-driven system: General results



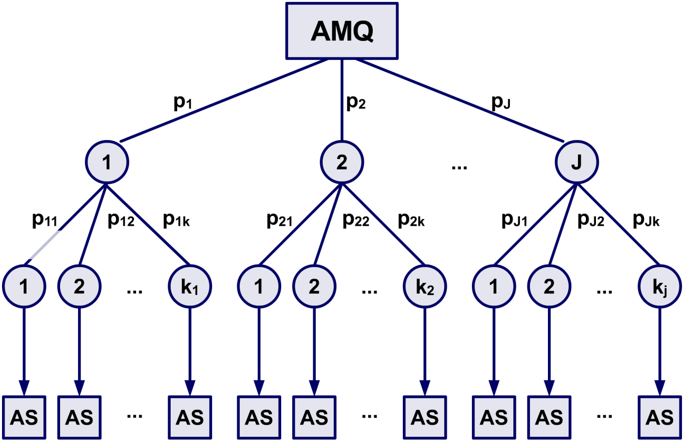
Appointment-driven system: General results



Appointment-driven system: General results



Appointment-driven system: General results



Appointment-driven system: Example

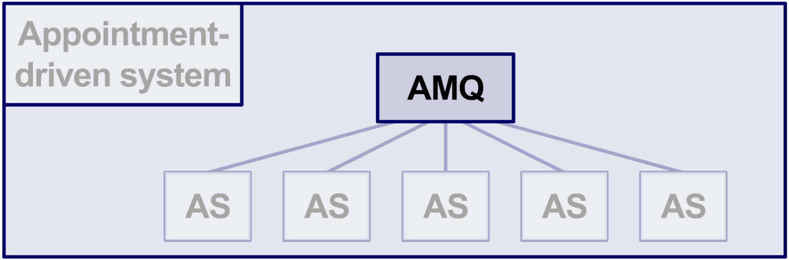
- Doctor's office with opening hours on Thursday from 6PM until 8PM and on Friday from 2PM until 6PM
- On Thursay a maximum of 4 patients receives service, on Friday a maximum of 8 patients is served
- On average 8 patients arrive each week (arrivals follow a Poisson distribution)
- AS remains unchanged (i.e. a block appointment rule with mean service time of 30 minutes and a SCV of 2/3)

V	$E[W_{AMQ}]$	$E[W_{AS}]$	p_ϕ	$E[\phi]$
10	5,126.9400	105.9466	0.1979	10.0794
50	4,440.3660	106.5436	0.1882	9.5996
100	4,360.2300	106.6709	0.1868	9.5329
200	4,320.7920	106.7410	0.1861	9.4991
∞	4,281.3099	106.8222	0.1852	9.4721

Conclusions

- Contribution:
  - One of the few models that are able to assess:
    - Patient waiting time at the waiting list
    - Patient waitign time at the service facility
    - Staff overtime (and overtime probability)
    - Staff idle time
    - Unused staff capacity
  - Resulting performance measures are of sufficient accuracy
  - Performance measures can be used to address strategical issues
- Model limitations:
  - Rather simple AS (block appointment rule)
  - Time-independent, Poisson arrival process

Second model overview



Advanced queueing models for appointment-driven systems
Creemers S. and Lambrecht M.R. (2008)

Comparison with previous AMQ model

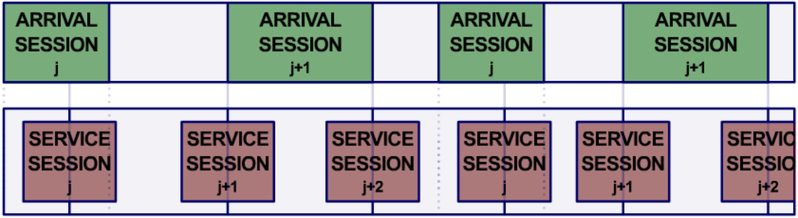
- Relaxed assumptions:
  - Arrivals are no longer assumed to follow a Poisson process, the first two moments of the interarrival time distribution are matched using a phase type distribution
  - Patients are only allowed to arrive during arrival sessions (e.g. during office hours on weekdays).
  - Different arrival session are allowed to feature different interarrival time distributions (i.e. time-dependent arrivals)
- Performance: increased accuracy and compuational performance
- Methodology: set of DTMC (Discrete Time Markov Chain) analyzed using matrix analytical methods

Vacation classes

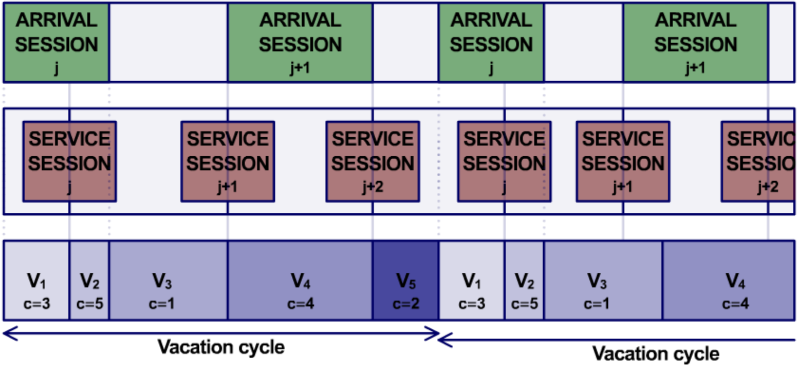
- Due to the incorporation of arrival sessions, we make a distinction between 5 different classes of arrivals
- A new vacation is initiated whenever:
  - A service session starts
  - An arrival session starts
  - An arrival session ends
- Each vacation class requires a distinct modeling approach

Vacation classes

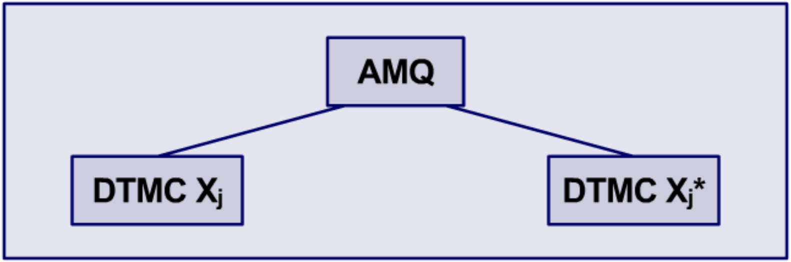
Vacation classes



Vacation classes

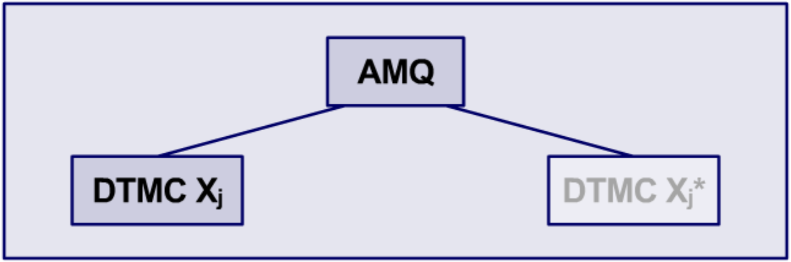


Division into two sets of DTMC



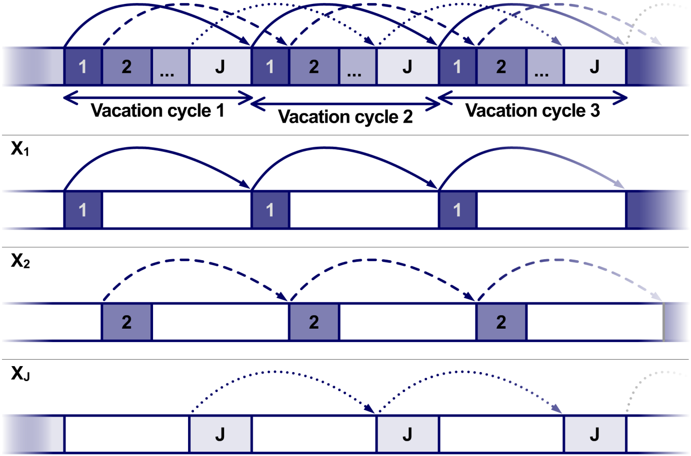
- Reduces complexity of the DTMC to be analyzed (i.e. it is more efficient to analyze a number of smaller DTMC as compared to analyzing one large DTMC)
- Increased model flexibility and accuracy

The DTMC X_j: Introduction



- DTMC X_j observes the status of the queue only at the start of vacations of type j
- Analysis of the DTMC X_j yields the stationary distribution of the number of patients in queue at the start of a vacation of type j

The DTMC X_j: Illustration



The DTMC X_j: Modeling issue

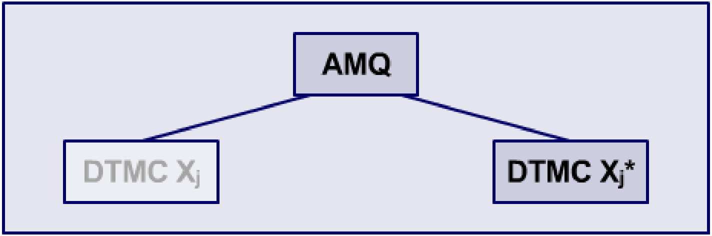
- Origin: a DTMC X_j observes the queue only at the start of vacations of type j
- Issue: Any alterations of the queue in between observation moments are left unobserved (i.e. we need to take these alterations into account)
- Solution: Determine the probability to move from a given state at the start of a vacation type j towards a state at the next vacation of type j
- Solution procedure:
  - A counting process allows the exact computation of the probability of having a number of arrivals during a given vacation i of deterministic duration T_i
  - These probabilities are the input of an efficient algorithm that determines the required probabilities

The DTMC X_j: Model

- Probabilities obtained from the algorithm are the only input of the DTMC X_j
- The DTMC X_j is twodimensional and its statespace may be represented by pairs (Q, a) where:
  - Q indicates the queue size
  - a indicates the phase of the arrival process (phase type)
- The transition matrix $\mathbf{Q}_j$ is of upper block Hessenberg form:

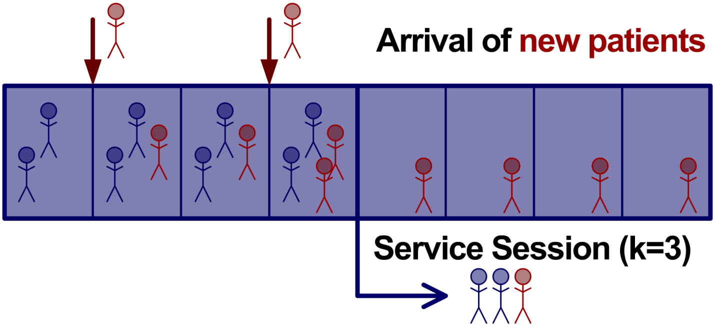
$$\mathbf{Q}_j = \begin{bmatrix} \mathbf{L}_j & \mathbf{F}_j^{(1)} & \mathbf{F}_j^{(2)} & \mathbf{F}_j^{(3)} & \mathbf{F}_j^{(4)} & \cdots \\ \mathbf{B}_j & \mathbf{L}_j & \mathbf{F}_j^{(1)} & \mathbf{F}_j^{(2)} & \mathbf{F}_j^{(3)} & \cdots \\ \mathbf{0} & \mathbf{B}_j & \mathbf{L}_j & \mathbf{F}_j^{(1)} & \mathbf{F}_j^{(2)} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_j & \mathbf{L}_j & \mathbf{F}_j^{(1)} & \cdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_j & \mathbf{L}_j & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \ddots \end{bmatrix}$$

The DTMC X_j^*: Introduction



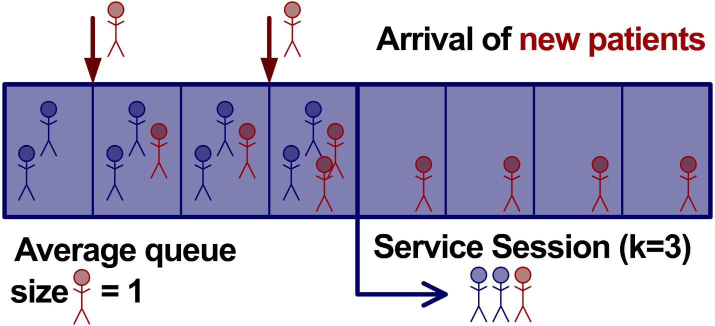
- DTMC X_j^* observes the queueing behavior of patients arriving at a vacation of type j
- From X_j^* we obtain the stationary distribution of the patients in queue that were not already present in queue at the start of the vacation of type j (those patients are already accounted for in DTMC X_j)

The DTMC X_i^*: Illustration



- Measure of interest: average queue size of only those patients that were not present at the start of the vacation
- Methodology: a resetting DTMC that can be analysed using matrix analytical techniques

The DTMC X_i^*: Illustration



- Measure of interest: average queue size of only those patients that were not present at the start of the vacation
- Methodology: a resetting DTMC that can be analysed using matrix analytical techniques

The DTMC X_j^*: Model

- The DTMC X_j^* is threedimensional and its statespace may be represented by pairs (Q, a, v) where:
  - Q indicates the queue size
  - a indicates the phase of the arrival process (phase type)
  - v indicates the phase of the vacation process (Erlang)
- The transition matrix $\mathbf{Q}_j^*$ is of lower block Hessenberg form:

$$\mathbf{Q}_j^* = \begin{bmatrix} \hat{\mathbf{L}}_j^* & \mathbf{F}_j^* & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{B}_j^* & \mathbf{L}_j^* & \mathbf{F}_j^* & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{B}_j^* & \mathbf{0} & \mathbf{L}_j^* & \mathbf{F}_j^* & \mathbf{0} & \dots \\ \mathbf{B}_j^* & \mathbf{0} & \mathbf{0} & \mathbf{L}_j^* & \mathbf{F}_j^* & \dots \\ \mathbf{B}_j^* & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{L}_j^* & \dots \\ \dots & \dots & \dots & \dots & \dots & \ddots \end{bmatrix}$$

Combining both sets of DTMC

- Output:
  - DTMC X_j provides us with the stationary distribution of the queue size at the start of a vacation of type j
  - DTMC X_j^* provides us with the stationary queue size of those patients that arrived during a vacation of type j (i.e. those patients that were not already in queue at the start of the vacation)
- Adding the average queue size at the start of a vacation of type j ($\overline{X}_j$) and the average queue size of arriving patients ($\overline{X}_j^*$) yields the average queue size at a vacation of type j.
- Aggregating over all vacation types j allows the assessment of performance measures at the AMQ as a whole

Numerical example

- Service sessions take place on Thursday at 12AM (), and on Friday at 7AM and 12AM.
- Arrival sessions are installed on Thursday from 6AM until 6 PM and on Friday from 7AM until 12AM:
  - On Thursday interarrivals are highly variable with a mean hourly rate of $1/3$ and a SCV equal to 2
  - On Friday interarrivals are much less variable with a mean hourly rate of $1/5$ and a SCV equal to 0.68
- The resulting vacation cycle consists of 5 vacations which can be characterized as follows:

j	c_j	T_j	k_j
1	3	6	0
2	5	6	3
3	1	13	0
4	4	5	1
5	2	138	2

Numerical example

- Individual vacations:

j	Q_j Model $V = 200$	Simulation
1	6.0362	6.0266
2	5.5505	5.5648
3	6.5528	6.5848
4	5.6633	5.6661
5	4.7401	4.7428

- The AMQ as a whole:
  - Average queue size: 4.9941 (simulation: 4.9989)
  - Average waiting time (weeks): 0.9419 (simulation: 0.9453)

Conclusions

- Renders the previous AMQ model virtually obsolete:
    - Computationally less burdensome
    - Increased model accuracy
    - Relaxed assumptions
    - Increased model flexibility
- Limitations:
    - The time-dependent character of the arrival process might be modelled in a more proficient fashion
    - There probably exists a more efficient way to obtain the required performance measures at the DTMC X_j^*

Conclusions and future research

Contributions and limitations

- Contributions:
  - Development of new queueing models that allow the detailed study of appointment-driven systems:
    - Strategic important performance measures: patient waiting time (internal as well as external), staff performance (overtime, idle time, unused capacity)
    - Input for an optimization procedure
- Limitations:
  - Use of rather simple AS
  - No optimization yet
  - Time-dependent arrival process may be made more general

Future research

- Determining a good AS
  - Based on the Welch-Bailey appointment rule (which has been found to perform very well among different ASR)
  - More than just a myopic analysis of an AS
- Development of an optimization procedure to answer strategic questions such as:
  - How many service sessions should be installed
  - What should be the size of these service sessions
  - How many patients should be allowed in each service session
  - Where in time should service and arrival sessions be installed
  - ...

Time for questions