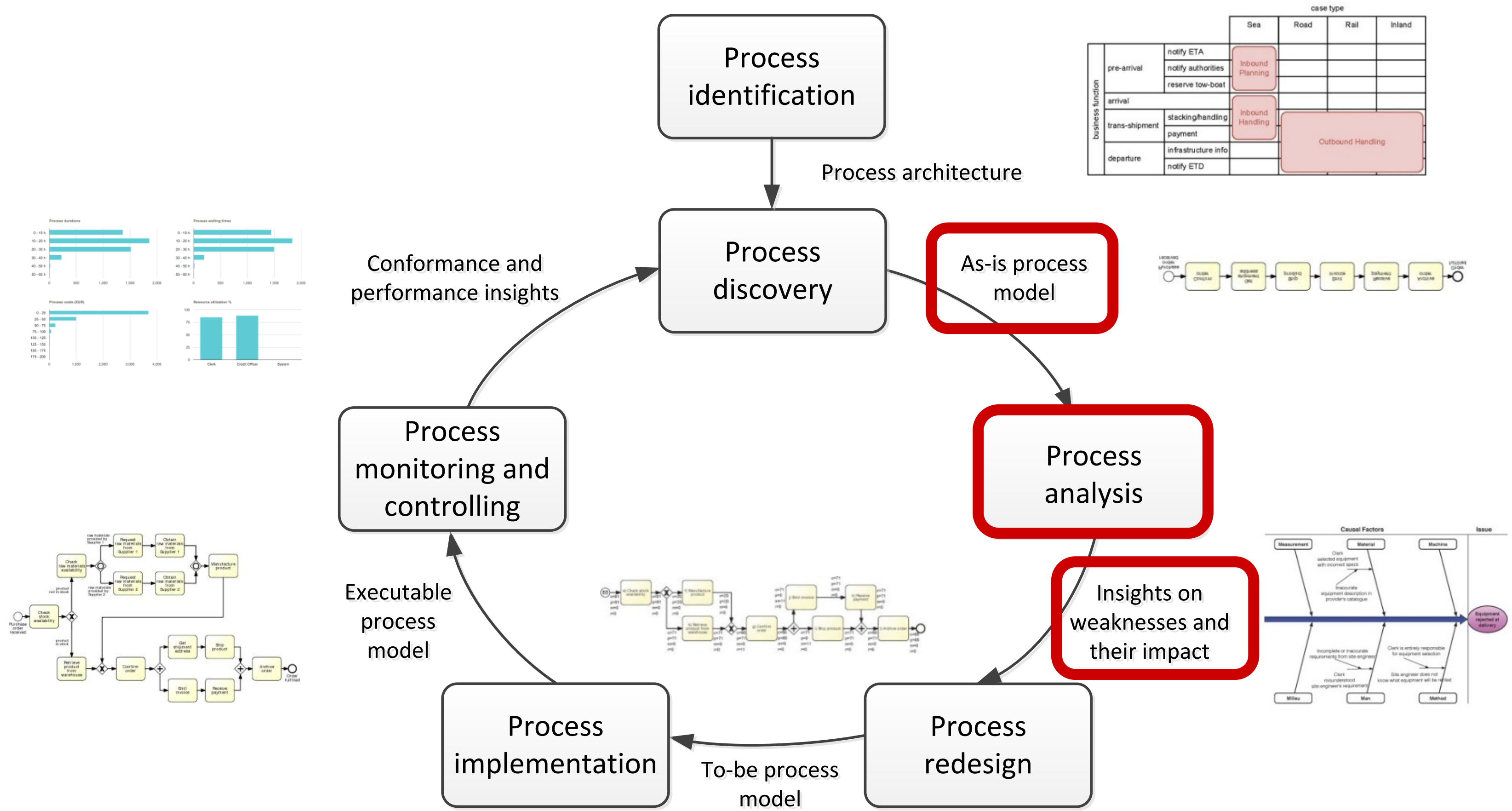
INFO-H420
Management of Data Science and Business Workflows
Part I
5. Quantitative Process Analysis

Dimitris SACHARIDIS

2023-2024

BPM Lifecycle



Process Analysis Techniques

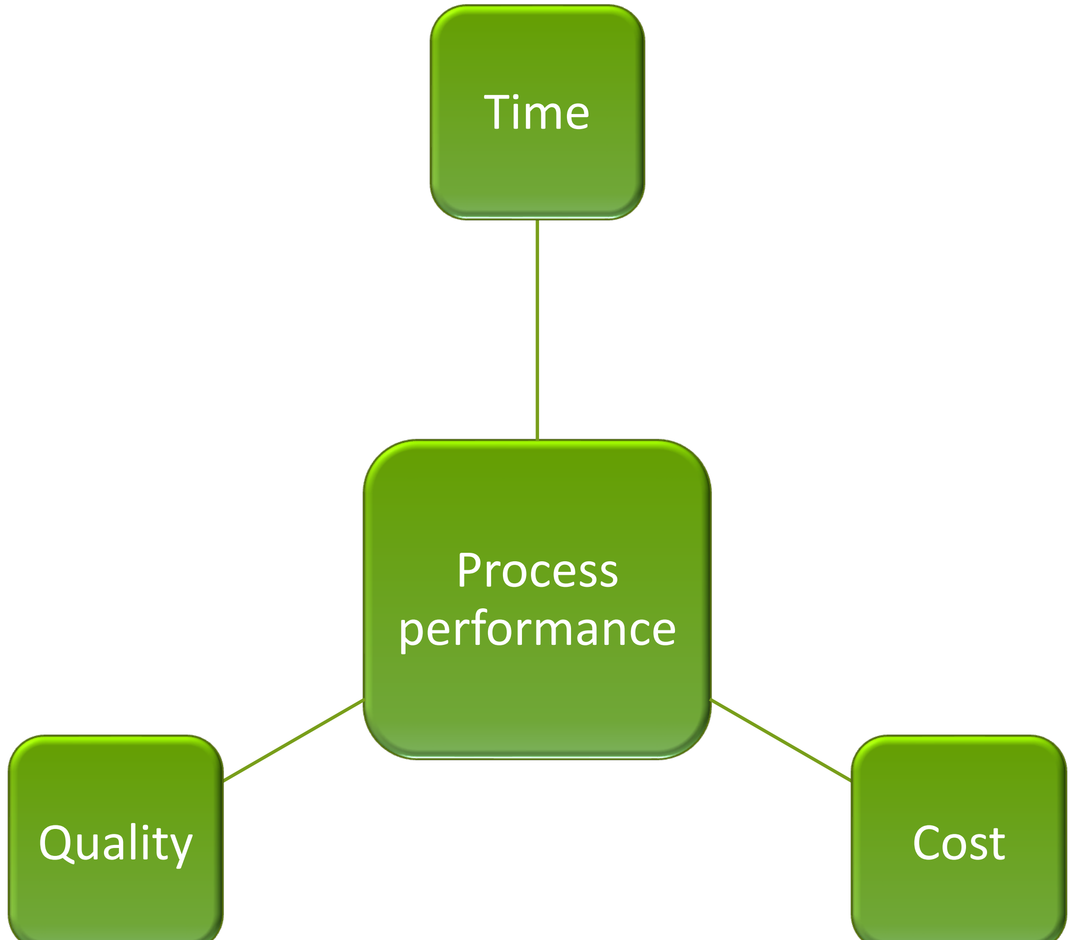
Qualitative analysis

- Value-Added & Waste Analysis
- Root-Cause Analysis
- Pareto Analysis
- Issue Register

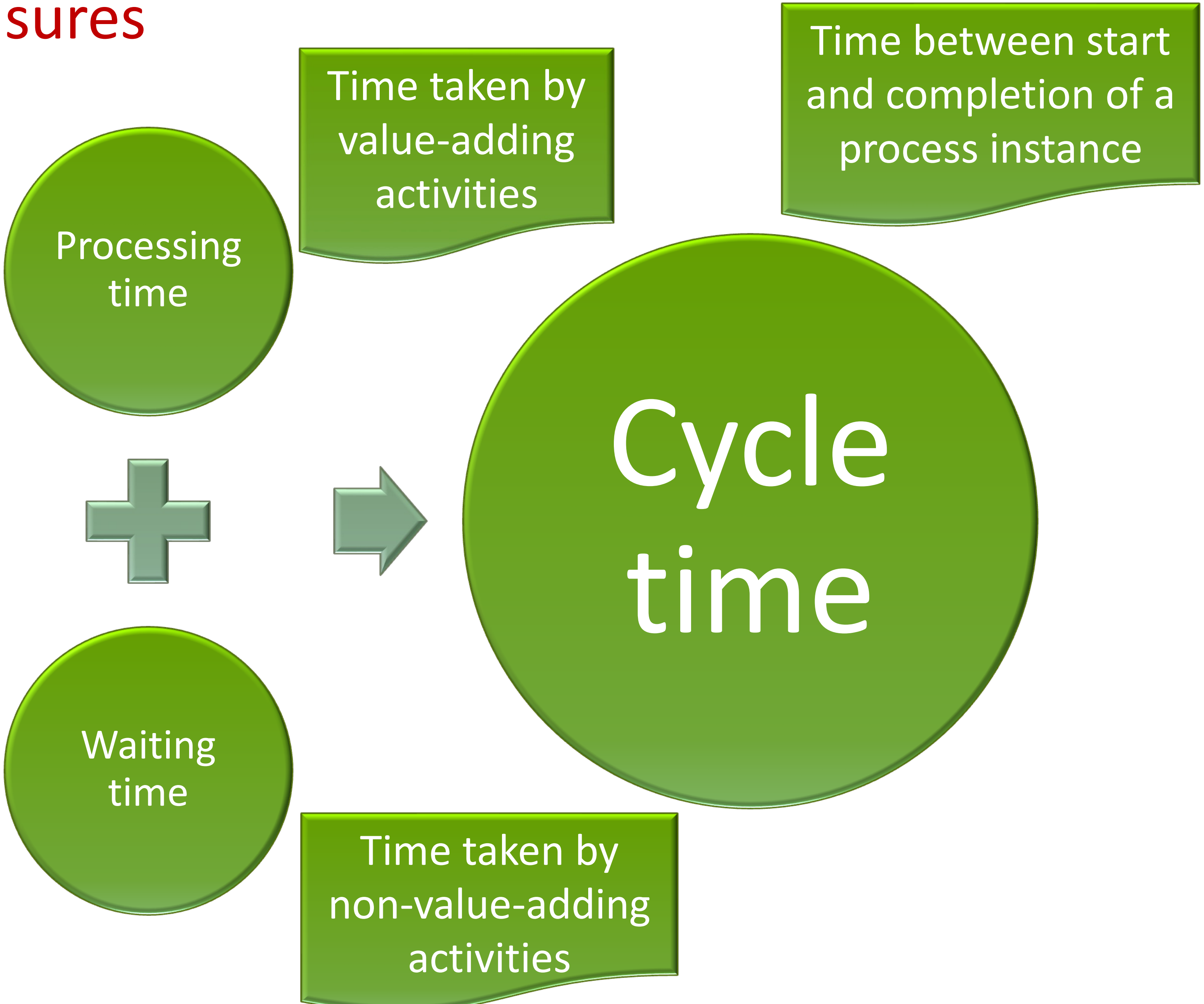
Quantitative Analysis

- Flow analysis
- Queuing analysis
- Simulation

Process performance



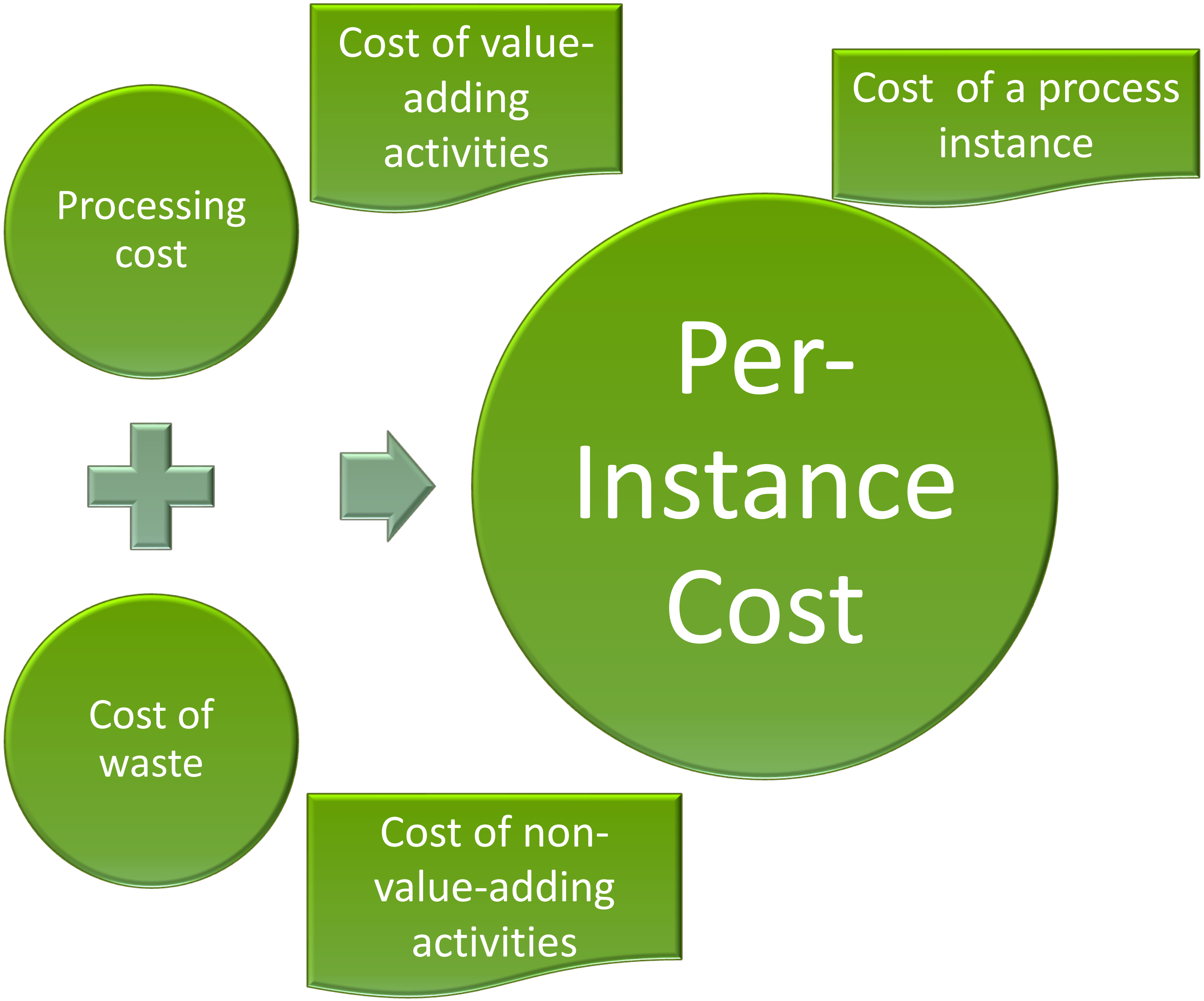
Time measures



Cycle time efficiency

Processing Time ÷ Cycle Time = Cycle Time Efficiency

Cost measures



Typical components of cost

Material cost

- Cost of tangible or intangible resources used per process instance

Resource cost

- Cost of person-hours employed per process instance

Resource utilization

Time spent per resource on process work ÷ Time available per resource for process work = Resource utilization

Resource utilization = 60%

➔ on average resources are idle 40% of their allocated time

Resource utilization vs. waiting time

Typically, when resource utilization > 90%
➔ Waiting time increases steeply

Quality

Product quality

- Defect rate

Delivery quality

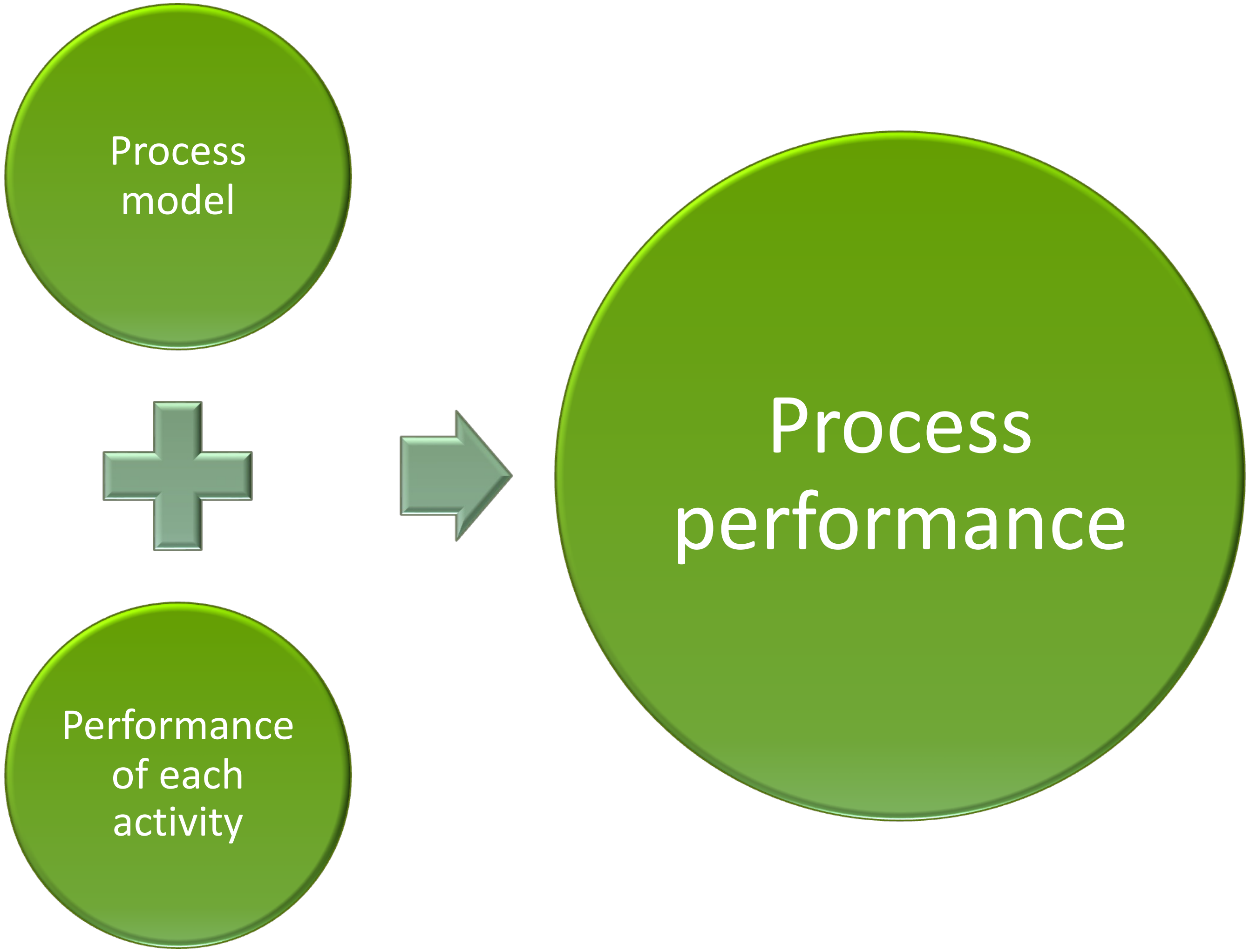
- On-time delivery rate
- Cycle time variance

Customer satisfaction

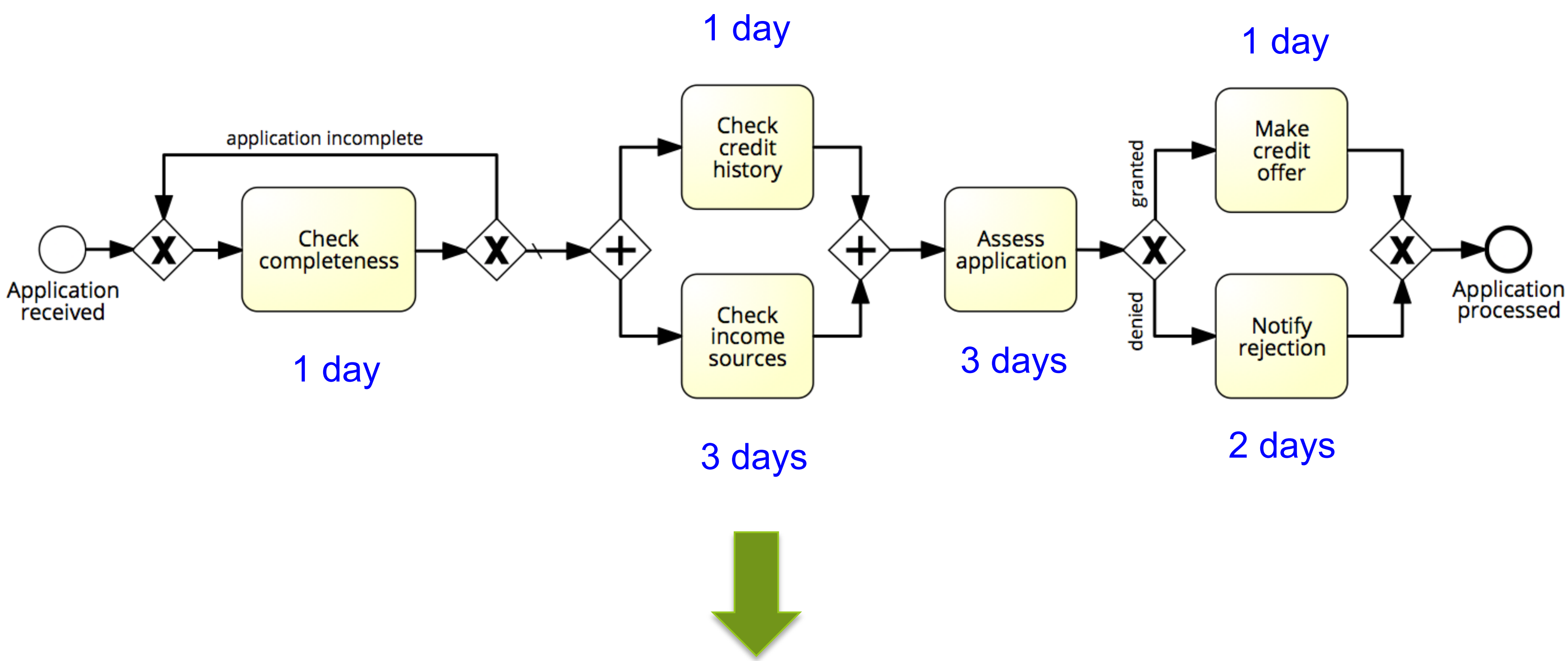
- Customer feedback score

FLOW ANALYSIS

Flow analysis



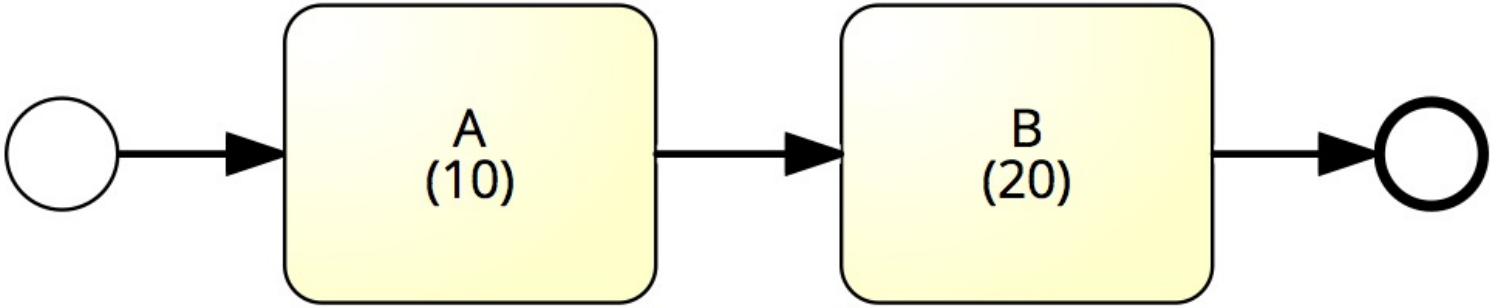
Flow analysis of cycle time



Cycle time = X days

Sequence – Example

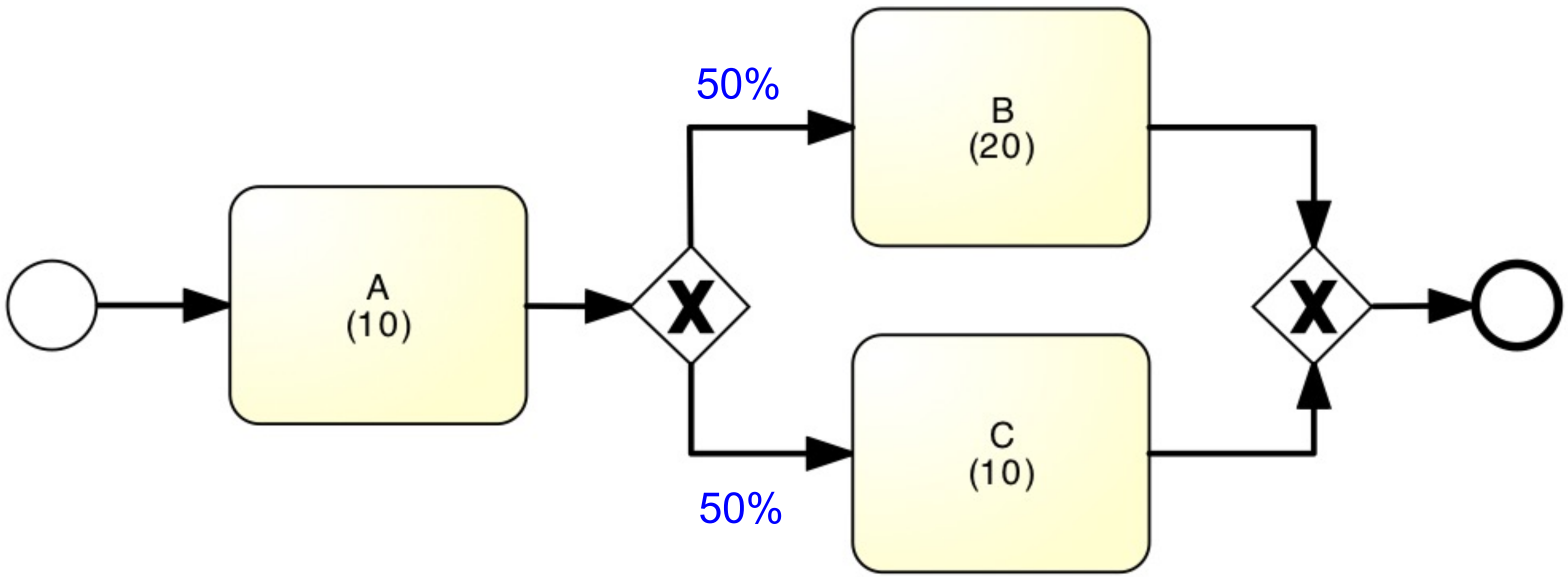
- What is the average cycle time?



Cycle time = 10 + 20 = 30

Example: Alternative Paths

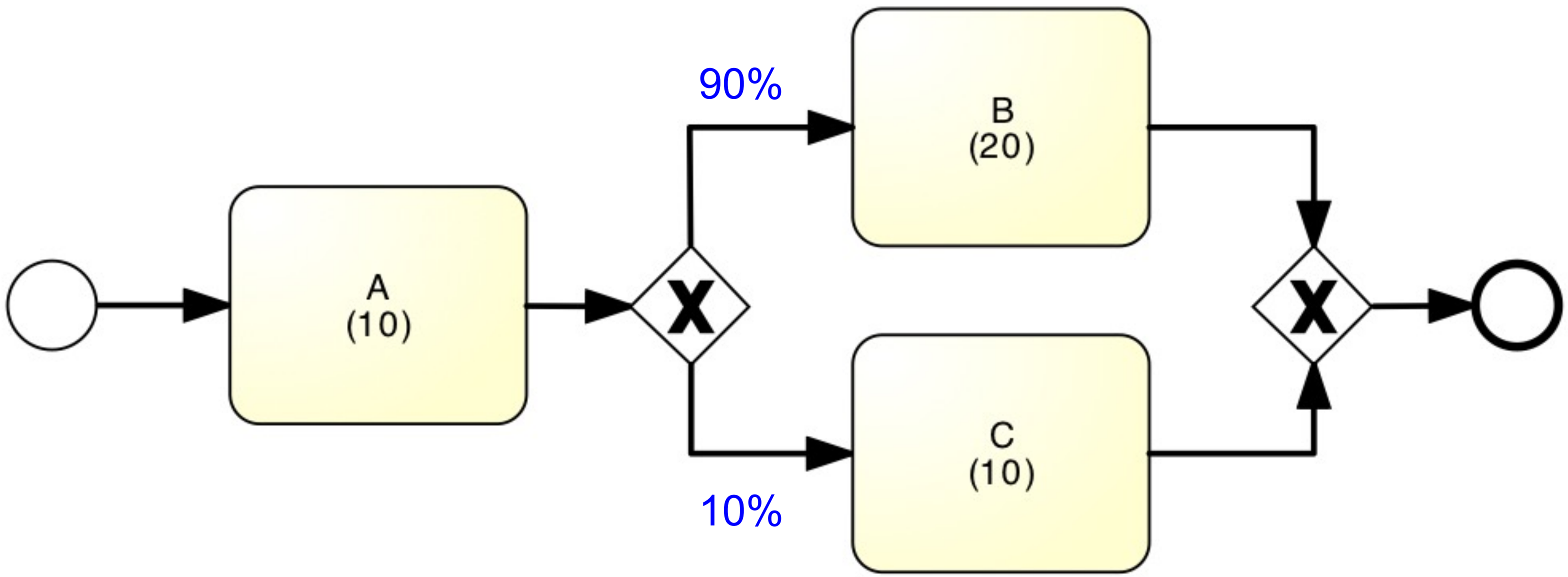
- What is the average cycle time?



Cycle time = 10 + (20+10)/2 = 25

Example: Alternative Paths

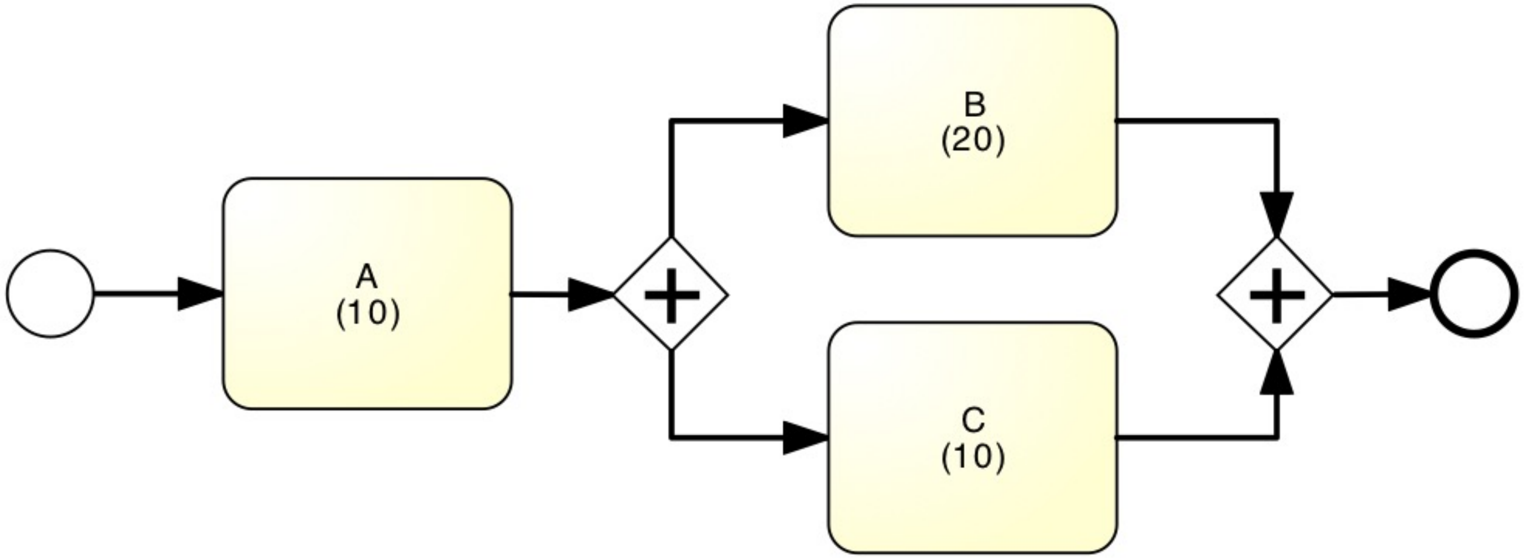
- What is the average cycle time?



Cycle time = 10 + 0.9*20+0.1*10 = 29

Example: Parallel paths

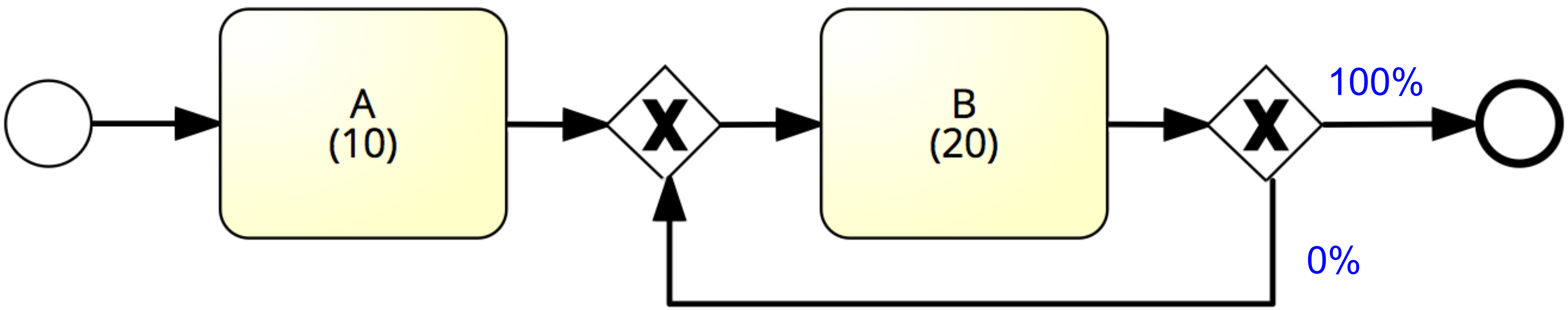
- What is the average cycle time?



Cycle time = 10 + 20 = 30

Example: Rework loop

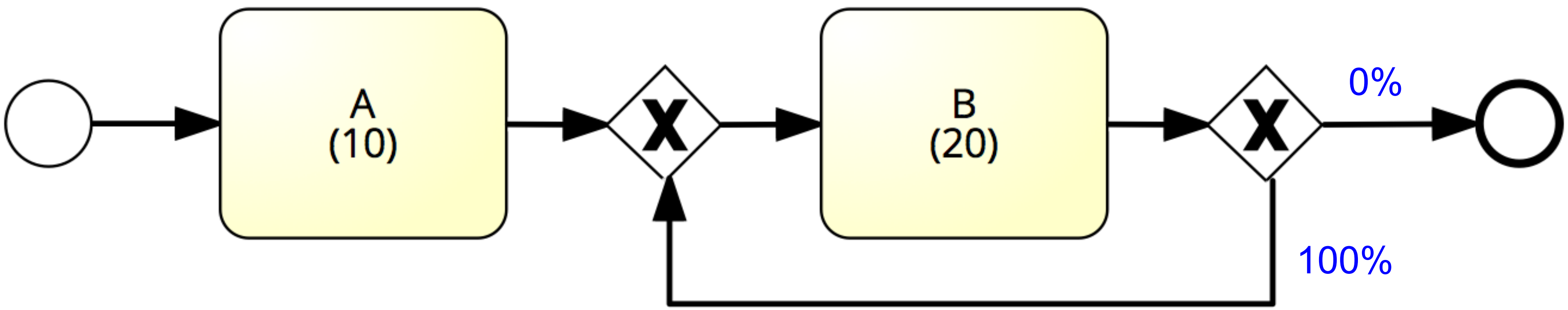
- What is the average cycle time?



Cycle time = 10 + 20 = 30

Example: Rework loop

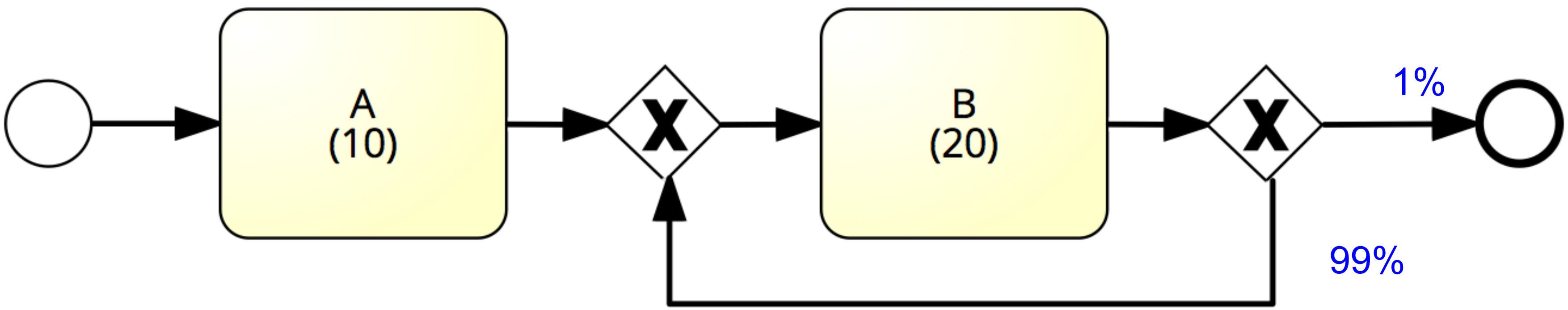
- What is the average cycle time?



Cycle time = ∞

Example: Rework loop

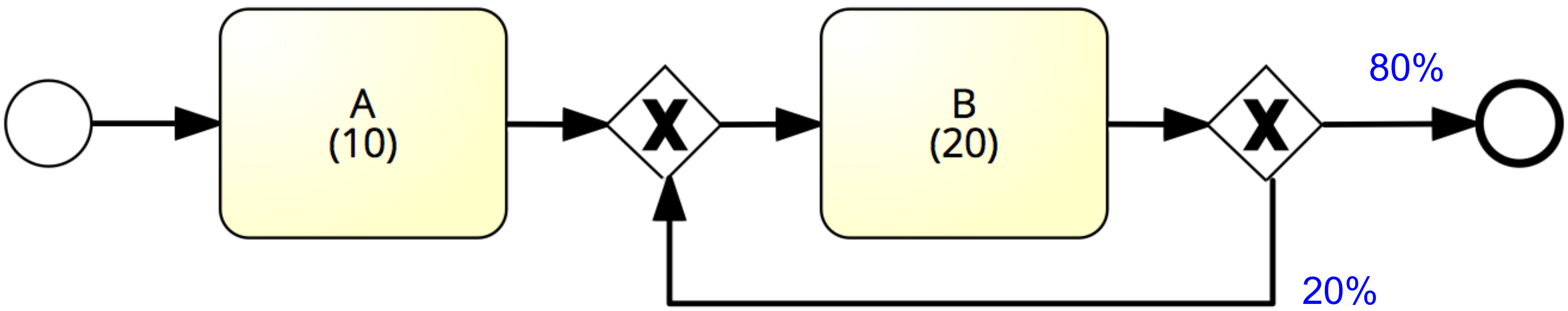
- What is the average cycle time?



Cycle time = $10 + 20*(1+0.99+0.99^2+0.99^3+\ldots) =$
$= 10 + 20/(1-0.99) = 2010$

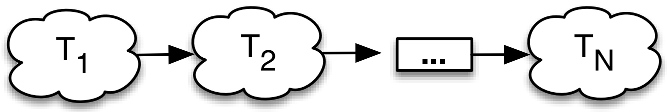
Example: Rework loop

- What is the average cycle time?

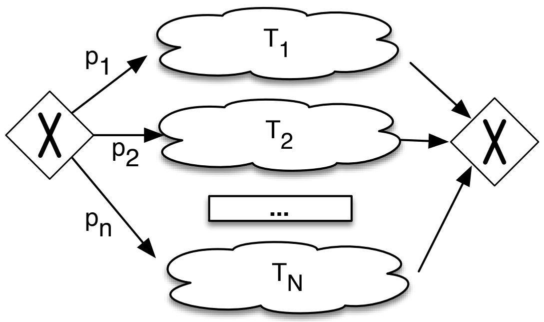


Cycle time = 10 + 20/0.8 = 35

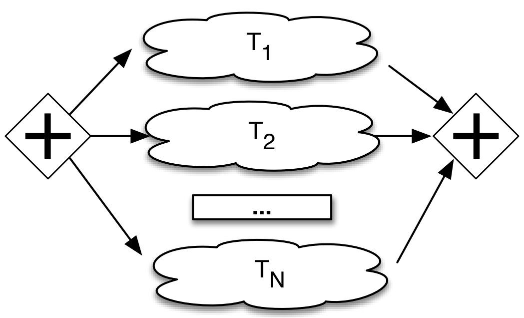
Flow analysis equations for cycle time



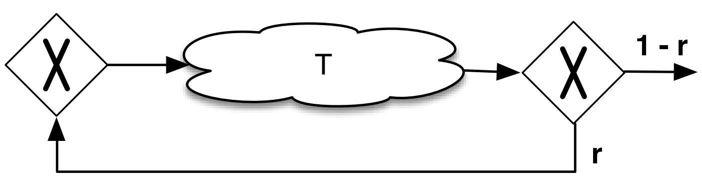
$$CT = T_1 + T_2 + \ldots + T_N$$



$$CT = p_1 * T_1 + p_2 * T_2 + \ldots + p_n * T_N$$

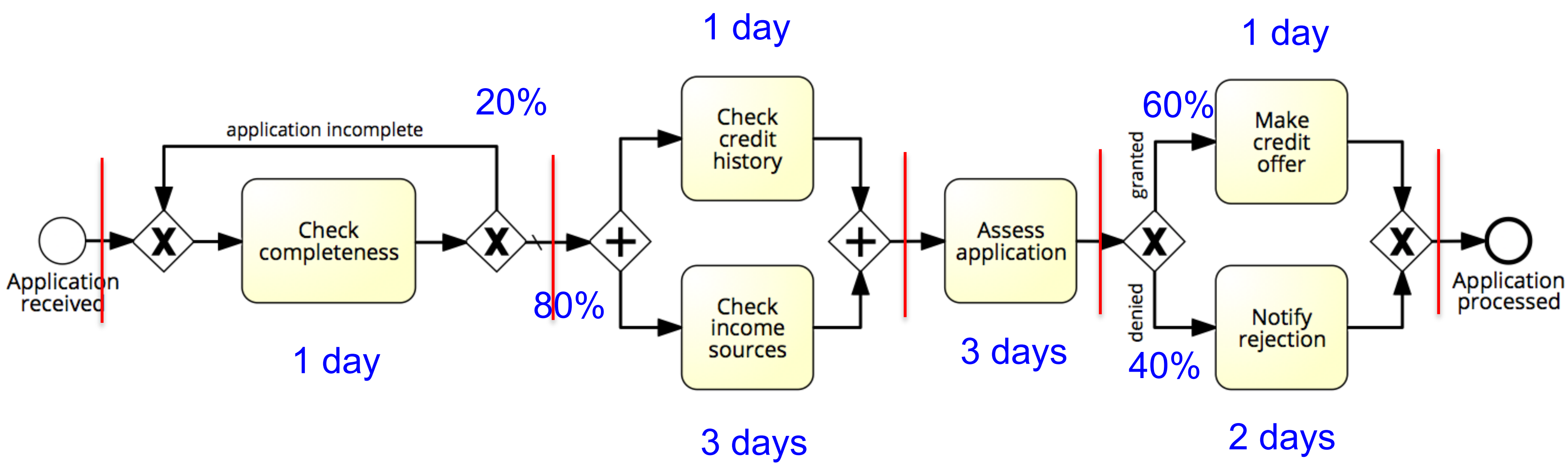


$$CT = \max(T_1, T_2, \ldots, T_N)$$



$$CT = T / (1-r)$$

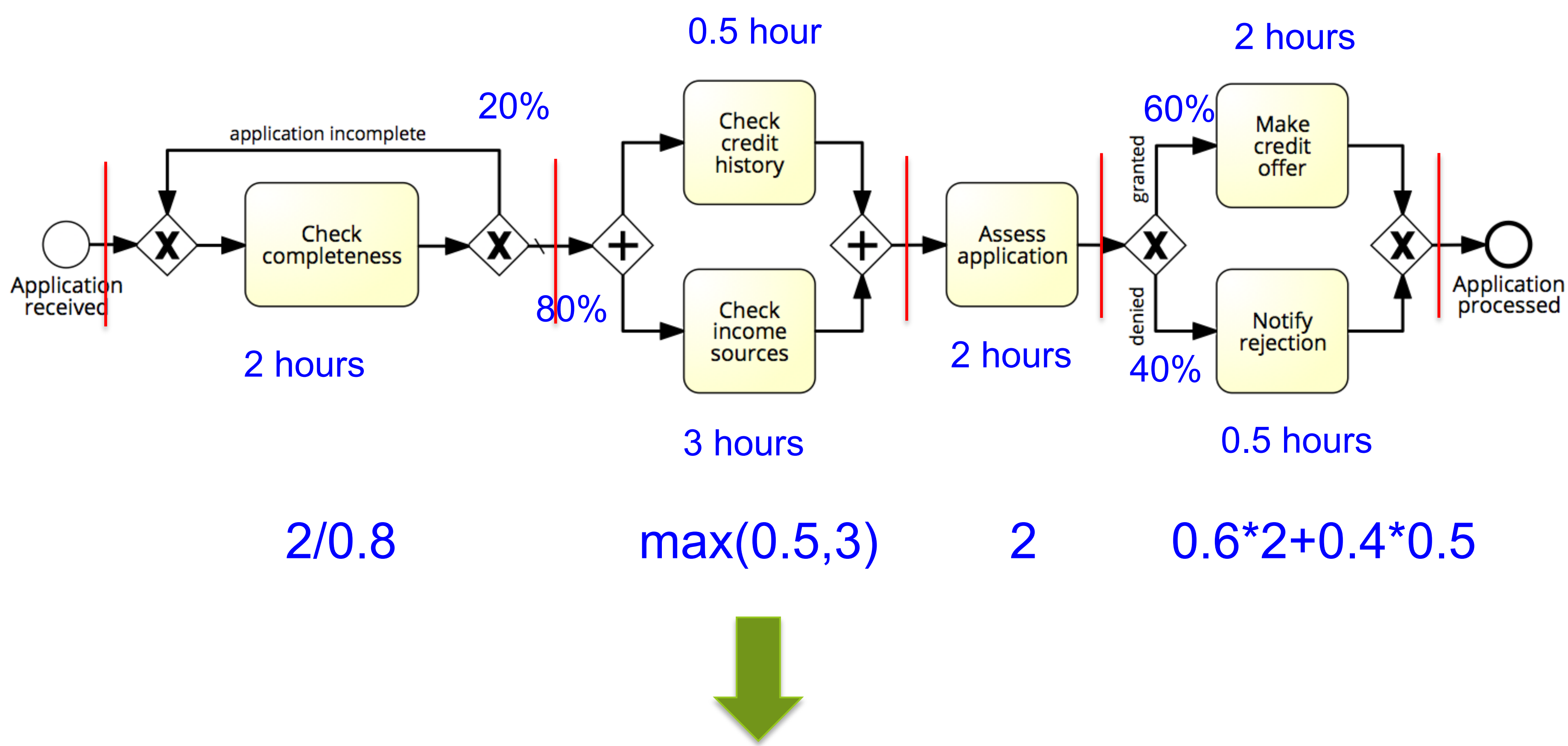
Flow analysis of cycle time



1/0.8 max(1,3) 3 0.6*1+0.4*2

Cycle time = 1.25 + 3 + 3 + 1.4 = 8.65 days

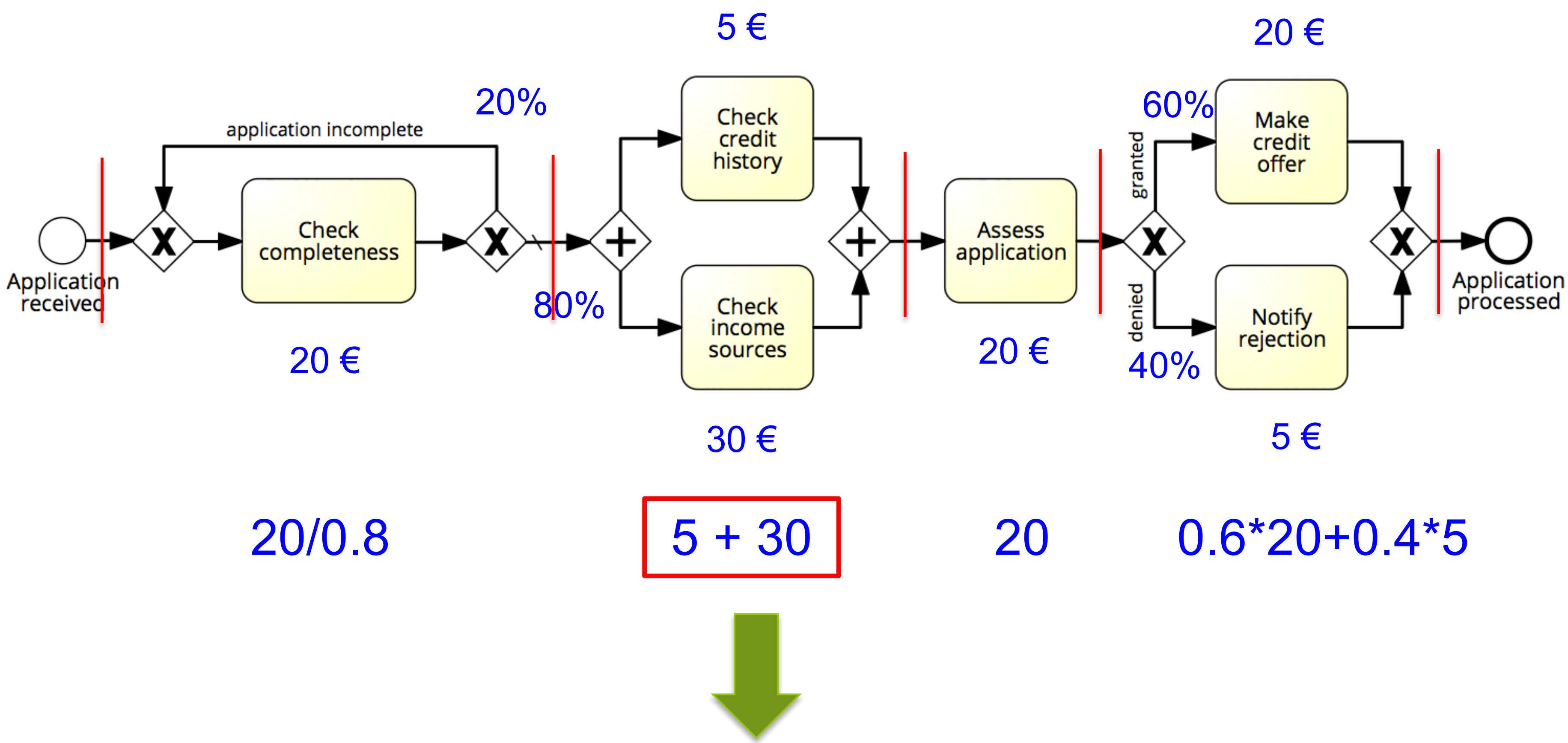
Flow analysis of processing time



Processing time = 2.5 + 3 + 2 + 1.4 = 8.9 hours

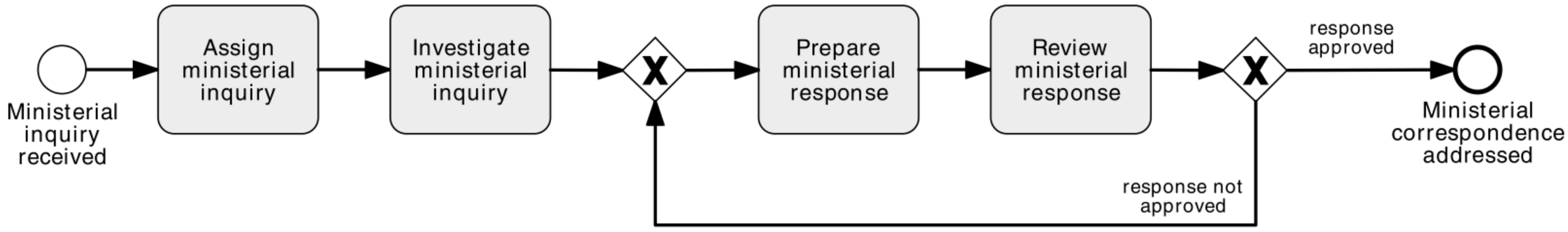
Cycle time efficiency = 8.9 hours / 8.65 days = 12.9%

Flow analysis of cost



Cost = 25 + 35 + 20 + 14 = 94 €

Exercise: Calculate the Cycle Time Efficiency

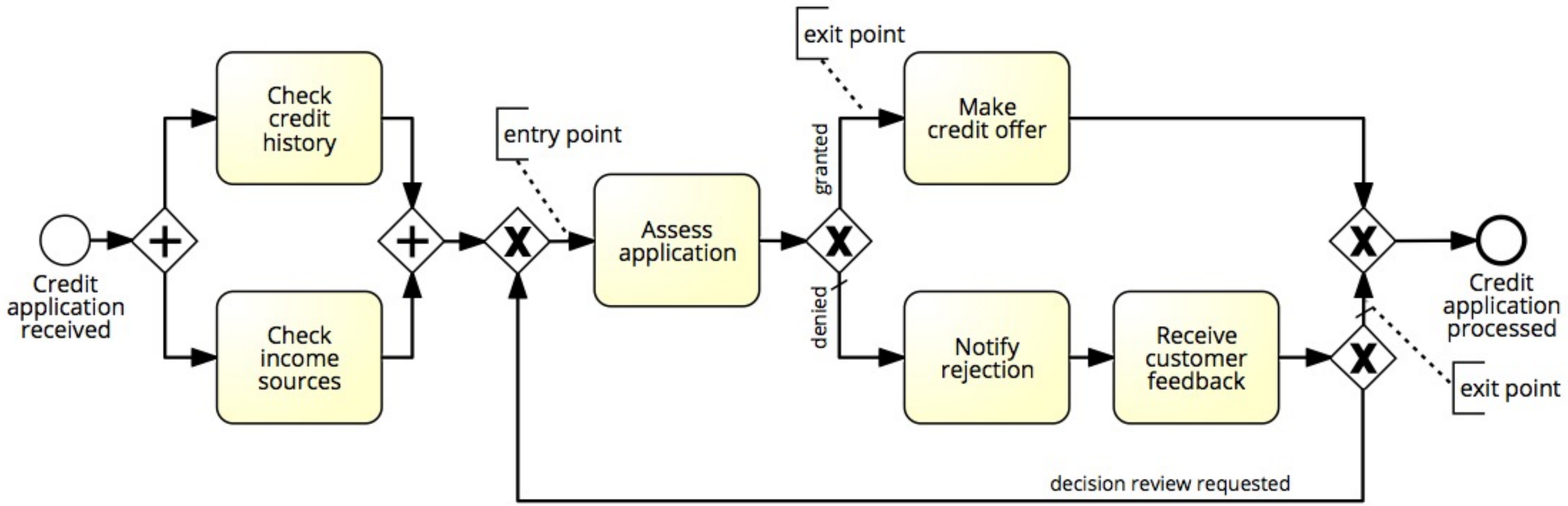


Activity	Cycle time	Processing time
Register ministerial enquiry	2 days	30 mins
Investigate ministerial enquiry	8 days	12 hours
Prepare ministerial response	4 days	4 hours
Review ministerial response	4 days	2 hour

Flow analysis: scope and limitations

- Flow analysis for cycle time calculation
- Other applications:
  - Calculating cost-per-process-instance
  - Calculating error rates at the process level
  - Estimating capacity requirements
- But it has its limitations…

Limitation 1: Not all Models are Structured



Limitation 2: Fixed arrival rate capacity

- Cycle time analysis does not consider:
  - The rate at which new process instances are created (arrival rate)
  - The number of available resources
- Higher arrival rate at fixed resource capacity
  ➔ high resource contention
  ➔ higher activity waiting times (longer queues)
  ➔ higher activity cycle time
  ➔ higher overall cycle time
- The slower you are, the more people have to queue up…
  - and vice-versa

Process Analysis Techniques

Qualitative analysis

- Value-Added & Waste Analysis
- Root-Cause Analysis
- Pareto Analysis
- Issue Register

Quantitative Analysis

- Flow analysis
- Queuing analysis
- Simulation

Why flow analysis is not enough?

Flow analysis does not consider waiting times due to resource contention

Queuing analysis and simulation address these limitations and have a broader applicability

Exercise

A fast-food restaurant receives on average 1200 customers per day (between 10:00 and 22:00). During peak times (12:00-15:00 and 18:00-21:00), the restaurant receives around 900 customers in total, and 90 customers can be found in the restaurant (on average) at a given point in time. At non-peak times, the restaurant receives 300 customers in total, and 30 customers can be found in the restaurant (on average) at a given point in time.

1. What is the average time that a customer spends in the restaurant during <u>peak</u> times?
2. What is the average time that a customer spends in the restaurant during <u>non-peak</u> times?
3. The restaurant plans to launch a marketing campaign to attract more customers. However, the restaurant's capacity is limited and becomes too full during peak times. What can the restaurant do to address this issue without investing in extending its building?

Cycle Time & Work-In-Progress

- WIP = (average) Work-In-Process
  - Number of cases that are running – in other words: cases that have started but not yet completed.
  - Example: the number of active orders in an order-to-cash process.
- WIP is a form of waste

- **Little's Formula:** $L = \lambda \cdot W$
  - L = WIP
  - λ = arrival rate (number of new cases per time unit)
  - W = cycle time

Queuing Analysis

- Capacity problems are common and a key driver of process redesign
  - Need to balance the cost of increased capacity against the gains of increased productivity and service
- Queuing and waiting time analysis is particularly important in service systems
  - Large costs of waiting and/or lost sales due to waiting

Prototype Example – ER at a Hospital

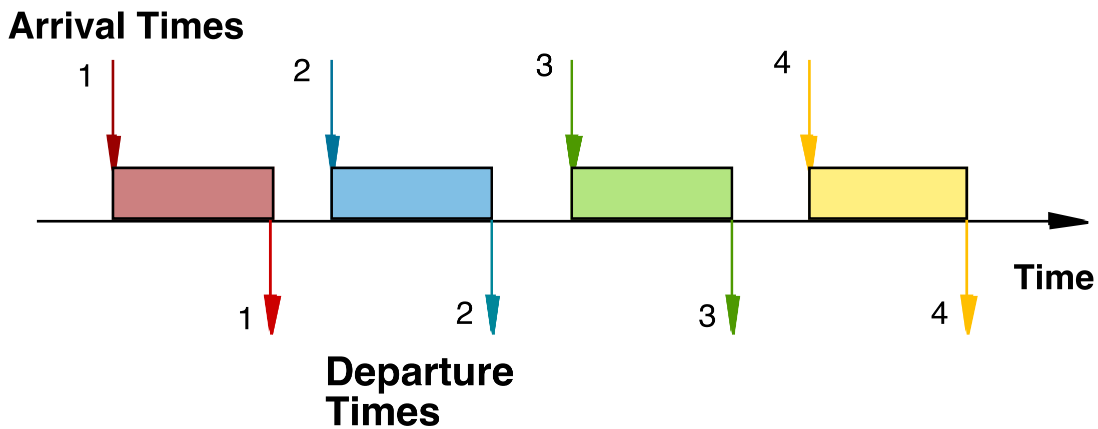
- Patients arrive by ambulance or by their own accord
- One doctor is always on duty
- More patients seeks help $\Rightarrow$ longer waiting times

➢ ***Question***: ***Should another MD position be instated?***

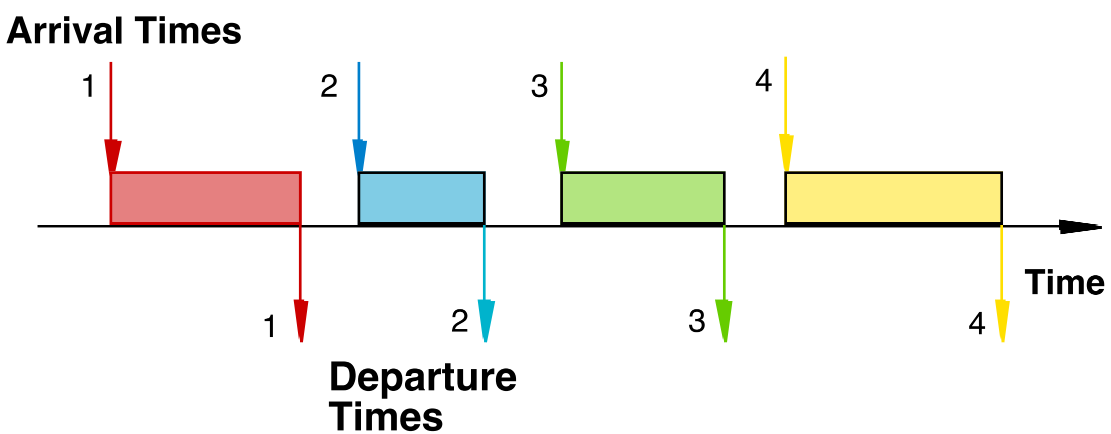
© Laguna & Marklund

Delay is Caused by Job Interference

If arrivals are regular or sufficiently spaced apart, no queuing delay occurs



Deterministic traffic

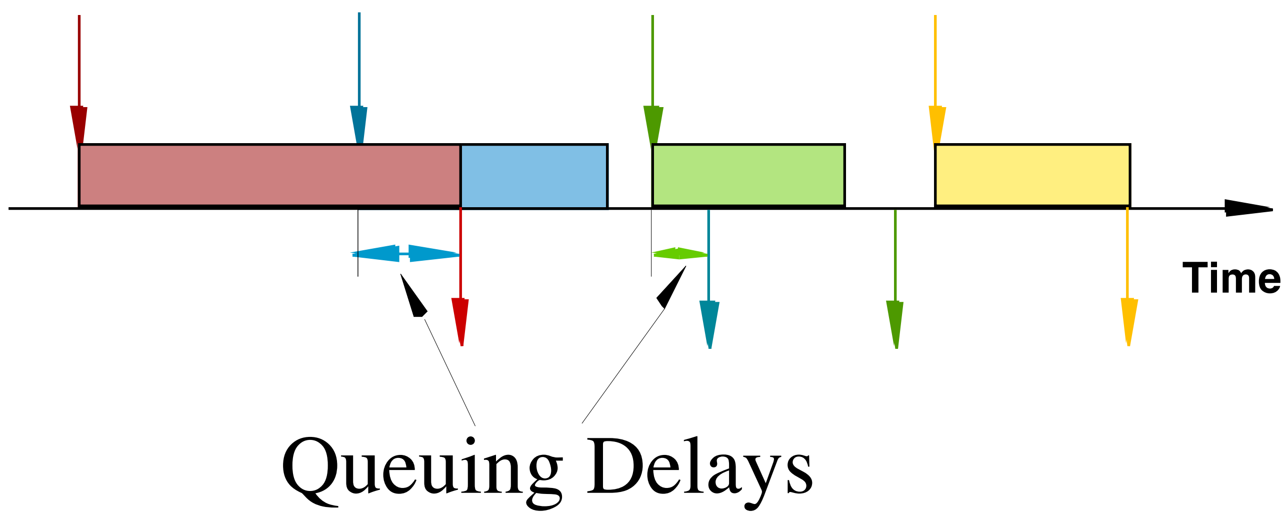


Variable but spaced apart traffic

© Dimitri P. Bertsekas

Job Size Variation Causes Interference

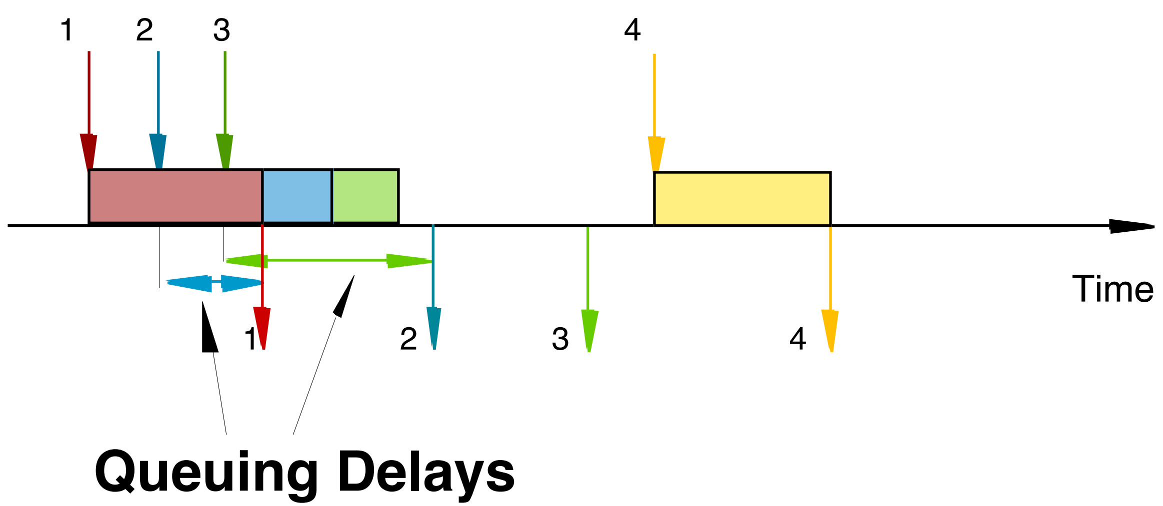
- Deterministic arrivals, variable job sizes



© Dimitri P. Bertsekas

Burstiness Causes Interference

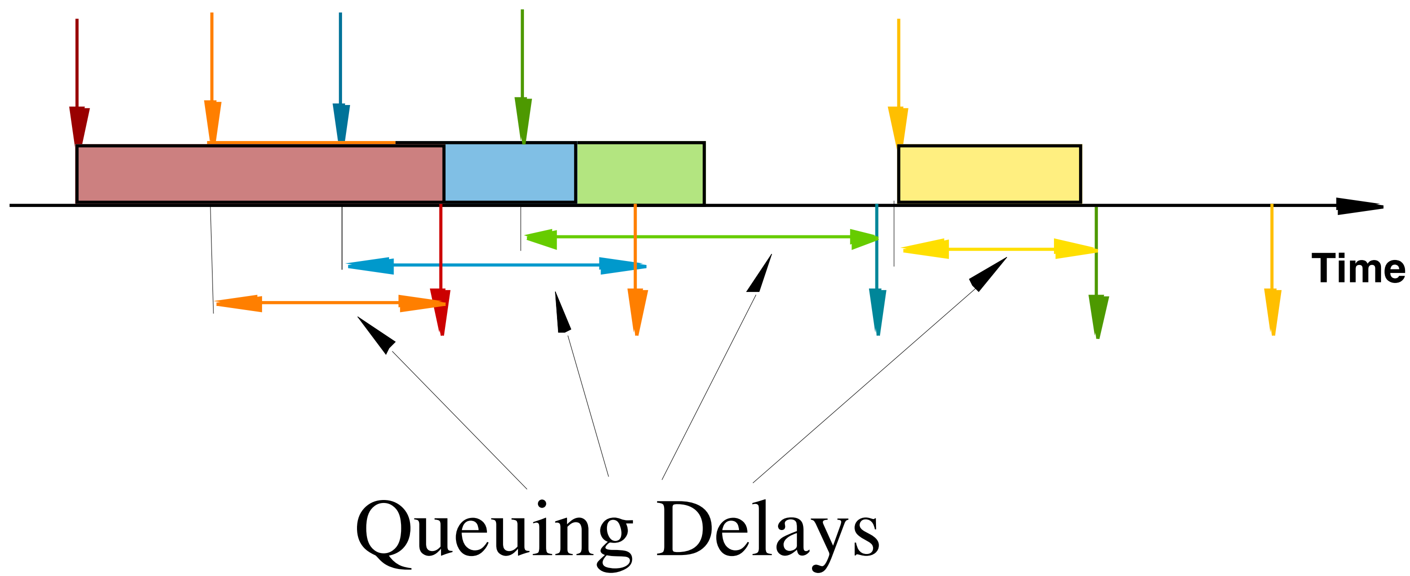
- Deterministic job size, variable arrivals



Bursty Traffic

High Utilization Exacerbates Interference

- The queuing probability increases as the load increases
- Utilization close to 100% is unsustainable → too long queuing times



© Dimitri P. Bertsekas

The Poisson Process

- A Poisson Process is a model for a series of discrete events where the *average time* between events is known, but the exact timing of events is random.
- The arrival of an event is independent of the event before

Example Poisson Process with average time between events of 60 days

https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459

The Poisson Process

- Common arrival assumption in many queuing and simulation models

- has a single parameter:
- λ is the **arrival rate** (number of events per time unit)

- Expected number of events in a time interval of length T is λT

- if the average time between events is 60 minutes, what is λ?
  - $\lambda = 1/60$

The Poisson Process

Key property: **no memory**

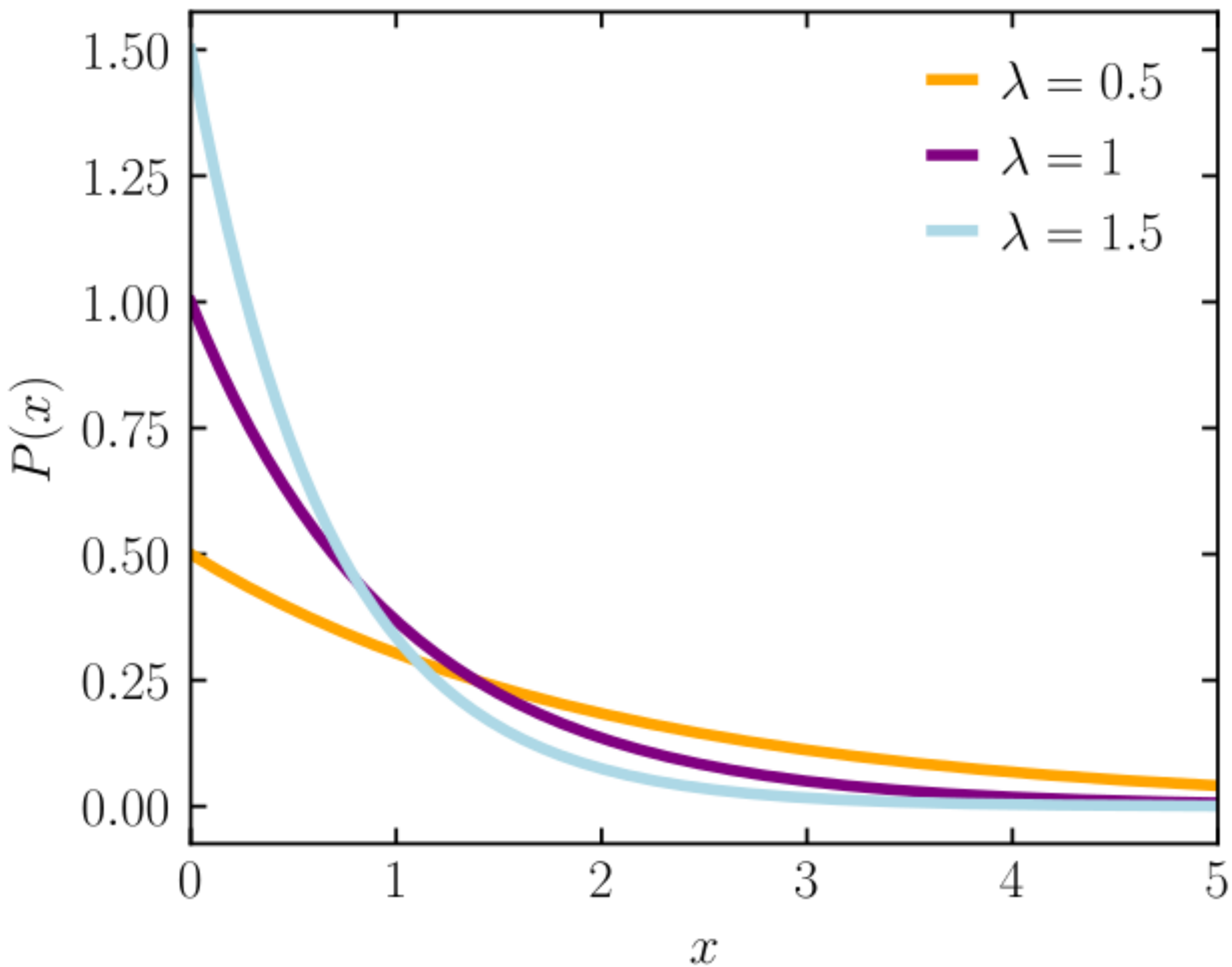
- The fact that a certain event has not happened tells us nothing about how long it will take before it happens
- Let T represent the random variable denoting the time elapsed since the last event (=inter-arrival time)

$$P(\ T>t+s \mid T>s\ ) = P(T>t)$$

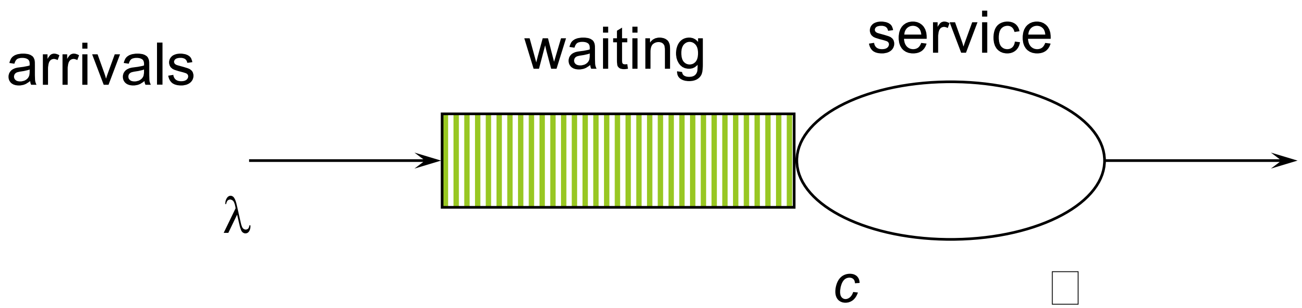
  - If we've already waited s minutes without seeing an event, the probability that an event won't occur in the next t minutes is the same as if we hadn't already waited s minutes

- The times between arrivals are independent, identically distributed and follow the **exponential distribution**
  - $P\ (T > t) = e^{-\lambda t}$

Probability Density Function of Exponential Distribution



Queuing theory: basic concepts

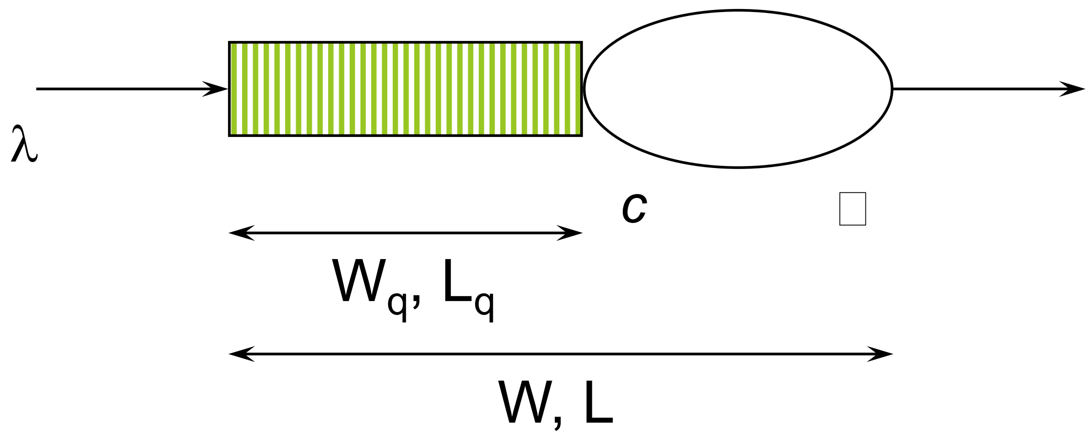


Basic characteristics:

- λ (mean arrival rate) = average number of arrivals per time unit
- μ (mean service rate) = average number of jobs that can be handled by one server per time unit:
- *c* = number of servers

© Wil van der Aalst

Queuing theory concepts (cont.)

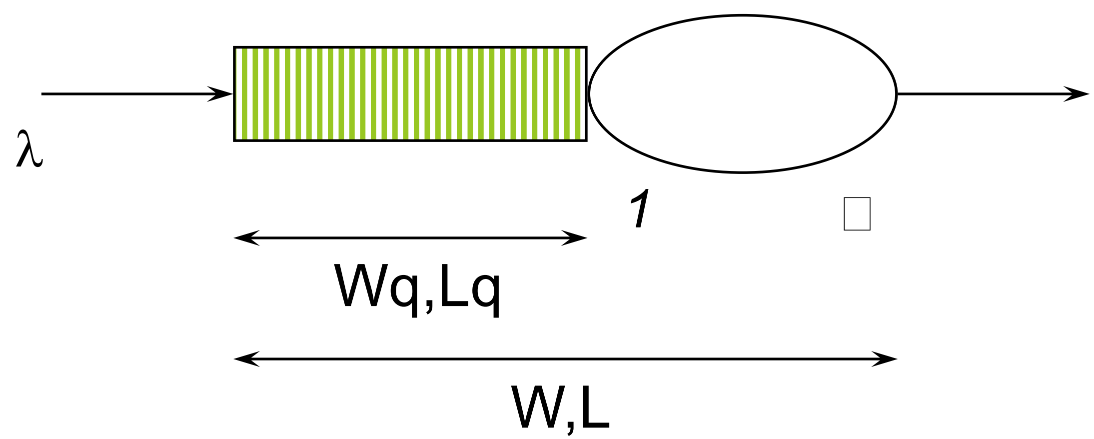


Given λ, μ, and c, we can calculate :

- ρ = resource utilization
- W_q = average time a job spends in queue (i.e., waiting time)
- W = average time in the "system" (i.e., *cycle time*)
- L_q = average number of jobs in queue (i.e., length of queue)
- L = average number of jobs in system (i.e., *Work-in-Progress*)

© Wil van der Aalst

M/M/1 queue



Assumptions:

- time between arrivals and processing time follow an exponential distribution
- 1 server (c = 1)
- FIFO

$$\rho = \frac{\text{Capacity Demand}}{\text{Available Capacity}} = \frac{\lambda}{\mu}$$

$L=\rho/(1-\rho)$

\mu

$W=L/\lambda=1/(\mu-\lambda)$

$L_q=\rho^2/(1-\rho) = L-\rho$

\mu \mu

$W_q=L_q/\lambda=\lambda/(\mu(\mu-\lambda))$

© Laguna & Marklund

M/M/c queue

- Now there are c servers in parallel, so the expected capacity per time unit is then c*□ \mu

$$\rho = \frac{\text{Capacity Demand}}{\text{Available Capacity}} = \frac{\lambda}{c * \mu}$$

Little's Formula $\Rightarrow$ $W_q = L_q/\lambda$

$W = W_q + (1/\mu)$

Little's Formula $\Rightarrow$ $L = \lambda W$

Tool Support

- For M/M/c systems, the exact computation of L_q is rather complex, but…

- **these calculations can be done by a queuing theory calculator:**
  - http://www.supositorio.com/rcalc/rcalclite.htm
  - https://qsa.inf.unideb.hu/prod/frontend/schemes/

Example – ER at County Hospital

- **Situation**
  - Patients arrive according to a Poisson process with intensity λ ($\Leftrightarrow$ the time between arrivals is exp(λ) distributed)
  - The service time (the doctor's examination and treatment time of a patient) follows an exponential distribution with mean $1/\mu$ (=exp(μ) distributed)

  $\Rightarrow$ *The ER can be modeled as an M/M/c system where c = the number of doctors*

- **Data gathering**

  $\Rightarrow$ λ = 2 patients per hour

  $\Rightarrow$ μ = 3 patients per hour

- **Question**

– Should the capacity be increased from 1 to 2 doctors?

© Laguna & Marklund

Queuing Analysis – Hospital Scenario

- Interpretation
  - To be in the queue = to be in the waiting room
  - To be in the system = to be in the ER (waiting or under treatment)

Characteristic	One doctor (c=1)	Two Doctors (c=2)
ρ	2/3	1/3
L_q	4/3 patients	1/12 patients
L	2 patients	3/4 patients
W_q	2/3 h = 40 minutes	1/24 h = 2.5 minutes
W	1 h	3/8 h = 22.5 minutes

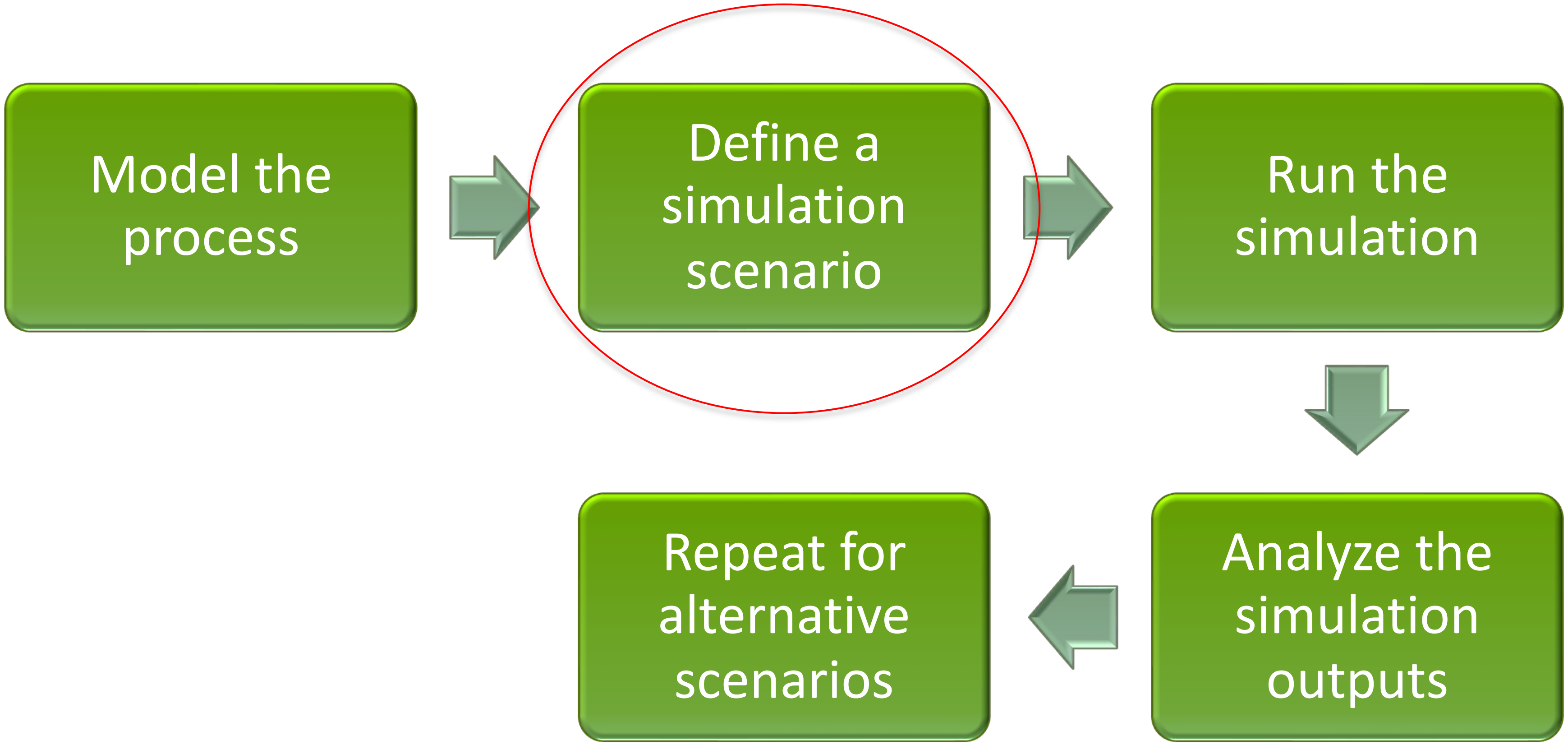
- Is it warranted to hire a second doctor ?

© Laguna & Marklund

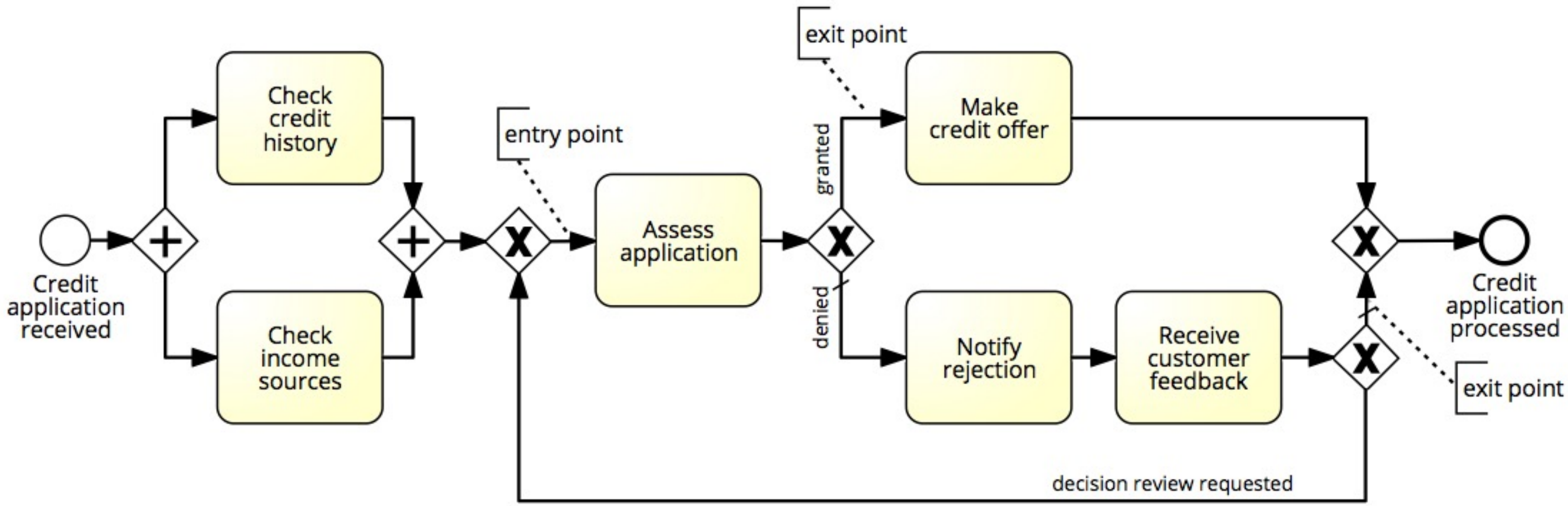
Process Simulation

- Versatile quantitative analysis method for
  - As-is analysis
  - What-if analysis
- In a nutshell:
  - Run a large number of process instances
  - Gather performance data (cost, time, resource usage)
  - Calculate statistics from the collected data

Process Simulation



Example



Check credit history
Check income sources
Credit application received
entry point
Assess application
exit point
granted
denied
Make credit offer
Notify rejection
Receive customer feedback
Credit application processed
exit point
decision review requested

Elements of a simulation scenario

1. Processing times of activities
   - Fixed value
   - Probability distribution

ULB

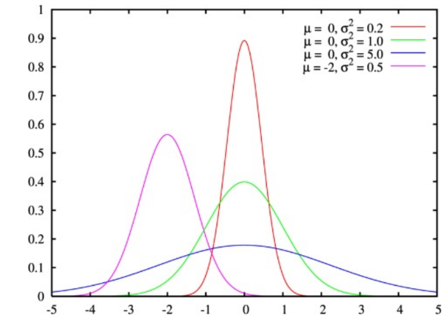
Choice of probability distribution

Fixed

- Can be used to approximate cases where the activity processing time varies very little
- Example: a task performed by a software application

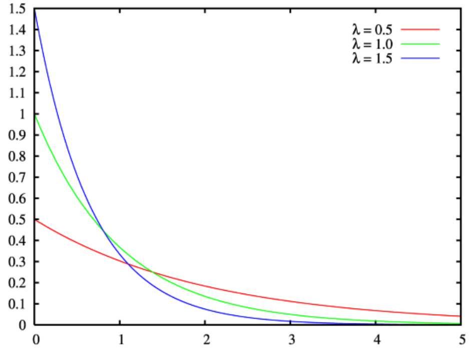
Normal

- Repetitive activities
- Example: "Check completeness of an application"
- Requires us to specify the <u>mean and the std deviation</u>

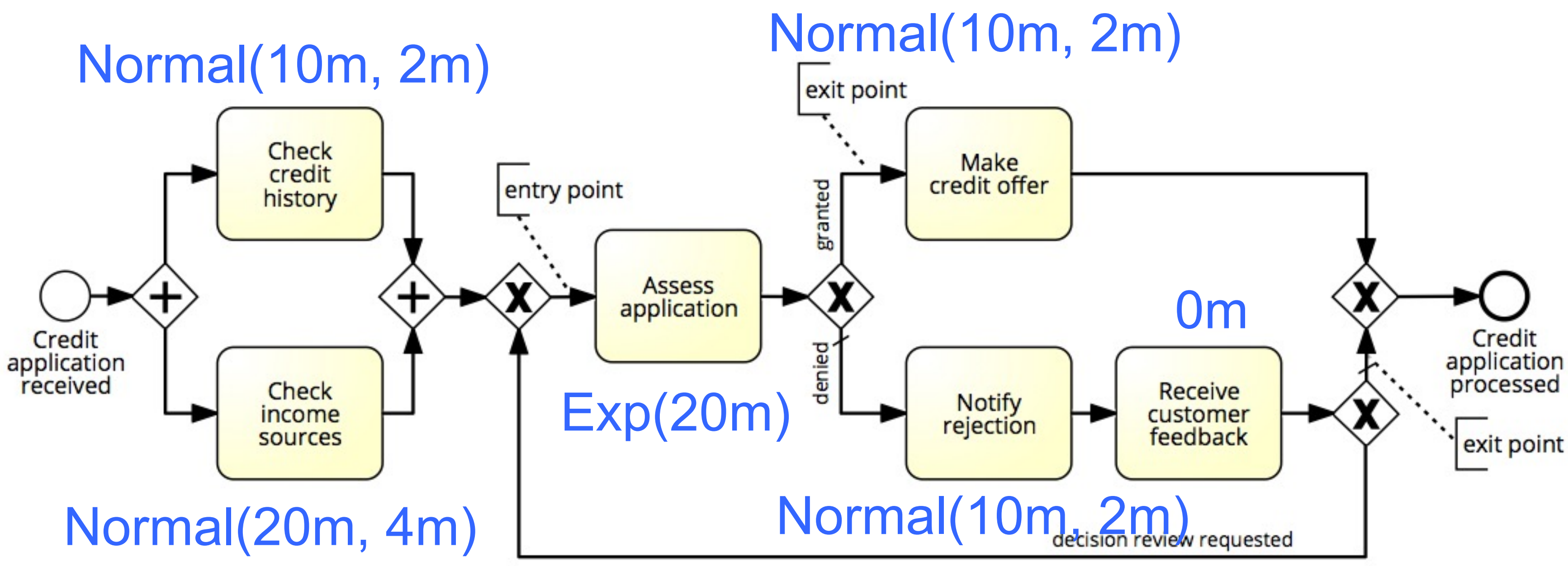


Exponential

- Complex activities that may involve detailed analyses or decisions
- Example: "Assess an application"
- Requires us to specify the <u>mean</u> only



Simulation Example

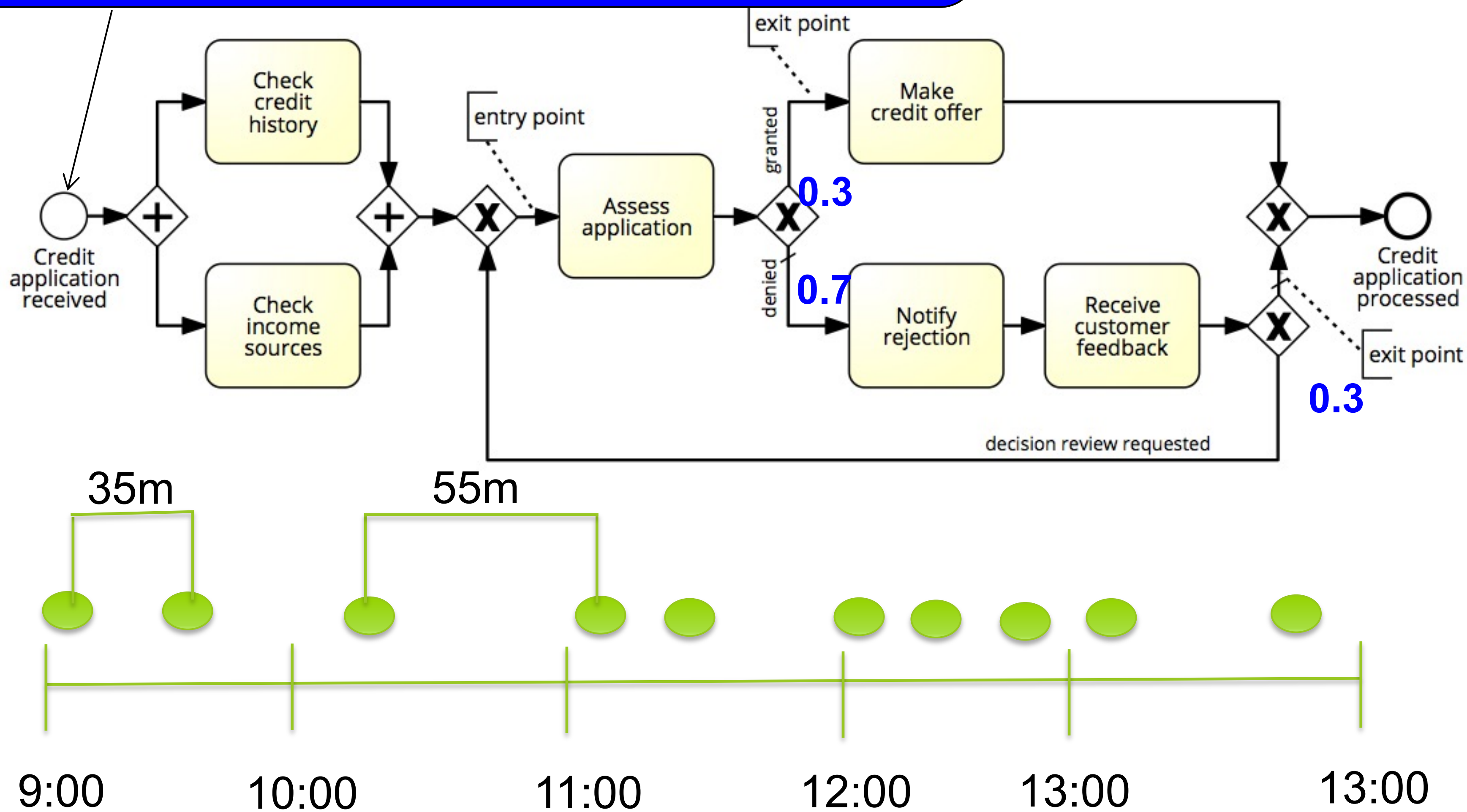


Elements of a simulation model

1. Processing times of activities
   - Fixed value
   - Probability distribution
2. Conditional branching probabilities
3. Arrival rate of process instances and probability distribution
   - Typically, exponential distribution with a given mean inter-arrival time
   - Arrival calendar, e.g., Monday-Friday, 9am-5pm, or 24/7

Branching probability and arrival rate

Arrival rate = 2 applications per hour
Inter-arrival time = 0.5 hour
Exponential distribution
From Monday-Friday, 9am-5pm



Elements of a simulation model

1. Processing times of activities
   - Fixed value
   - Probability distribution
2. Conditional branching probabilities
3. Arrival rate of process instances and probability distribution
   - Typically, exponential distribution with a given mean inter-arrival time
   - Arrival calendar, e.g., Monday-Friday, 9am-5pm, or 24/7
4. Resource pools

Resource pools

- Name
- Size of the resource pool
- Cost per time unit of a resource in the pool
- Availability of the pool (working calendar)
- Examples:

<u>Clerk</u>

€ 25 per hour

Mon-Fri, 9am-5pm

<u>Credit Officer</u>

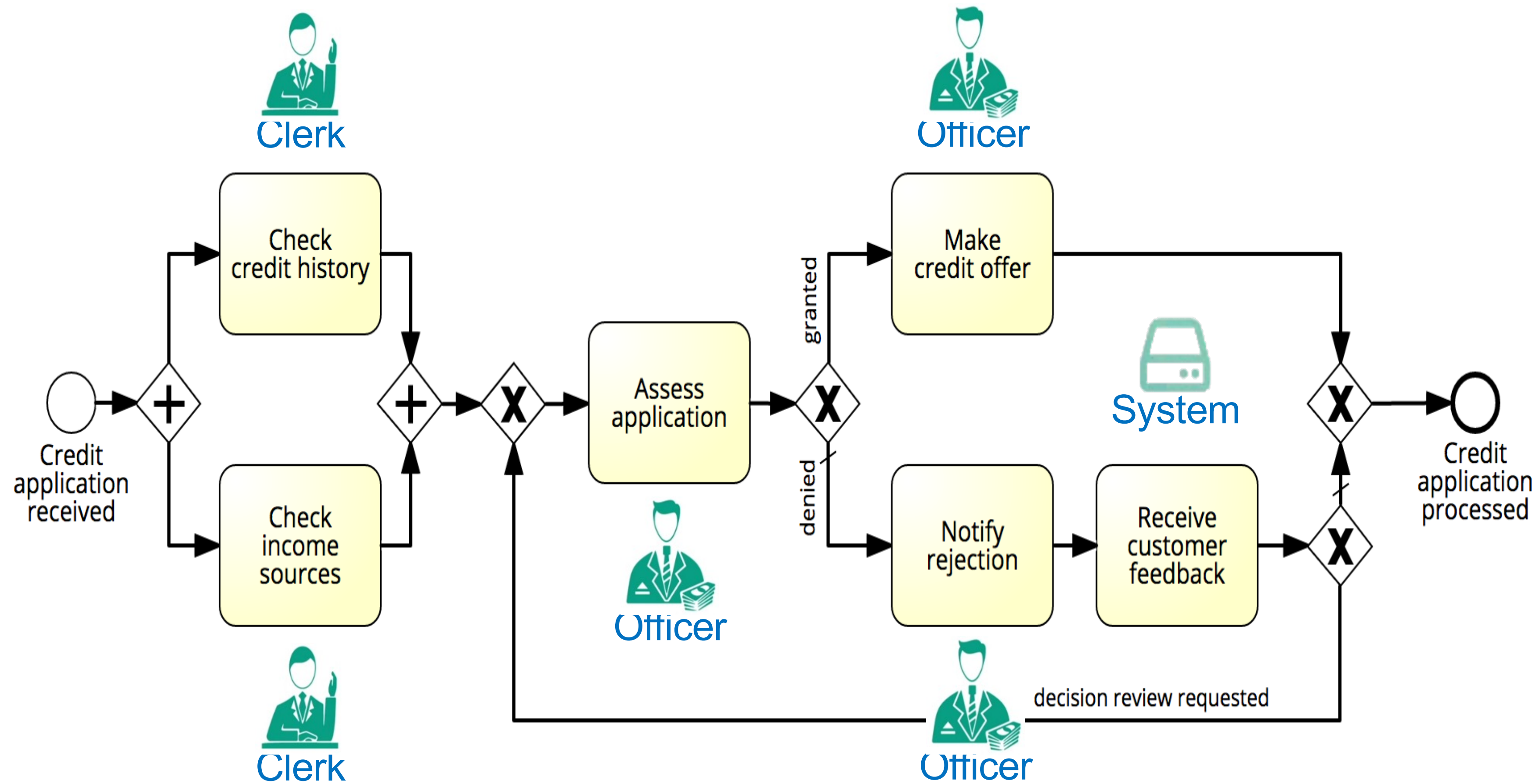
€ 35 per hour

Mon-Fri, 9am-4pm

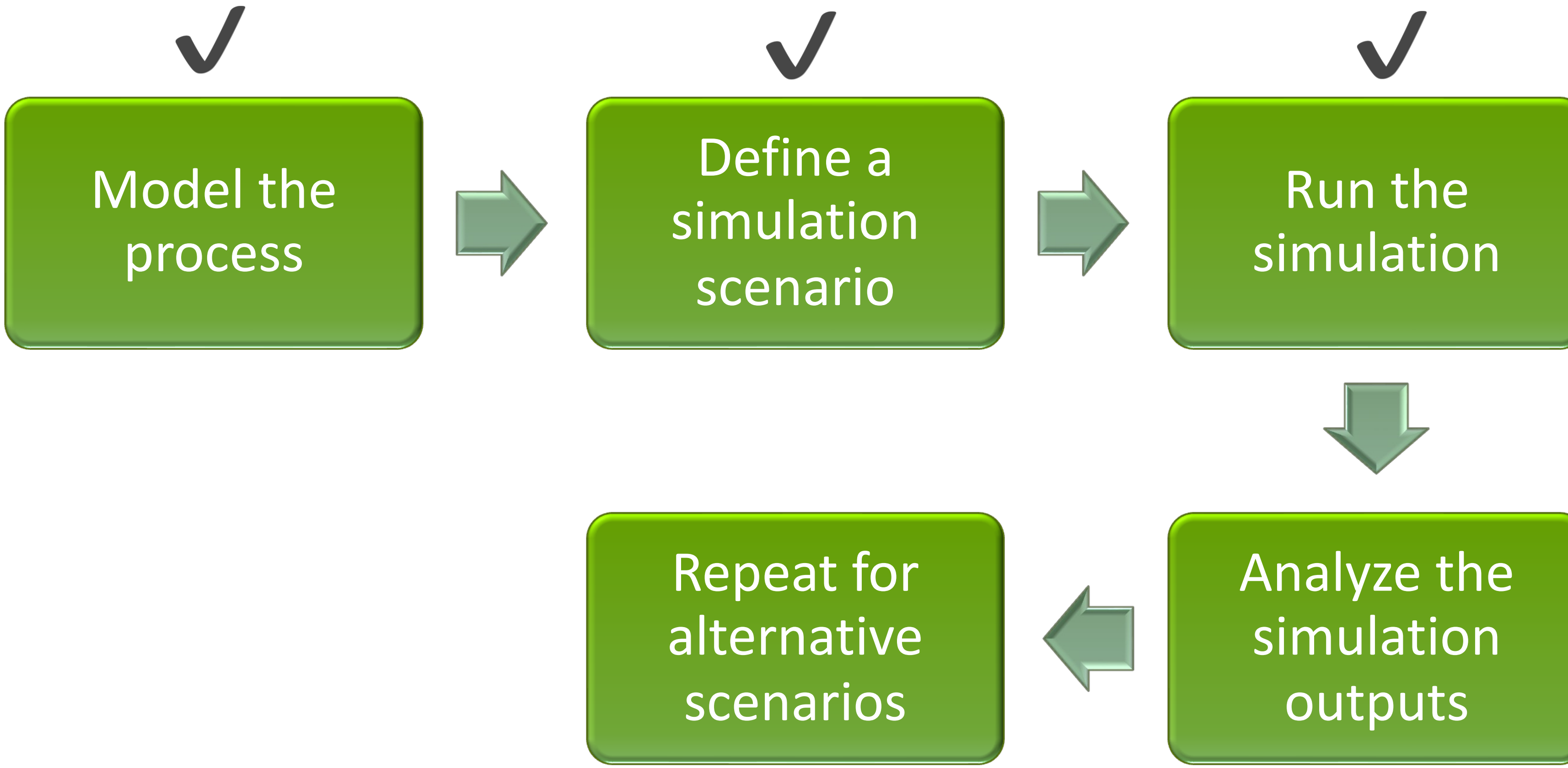
Elements of a simulation model

1. Processing times of activities
   - Fixed value
   - Probability distribution
2. Conditional branching probabilities
3. Arrival rate of process instances and probability distribution
   - Typically, exponential distribution with a given mean inter-arrival time
   - Arrival calendar, e.g., Monday-Friday, 9am-5pm, or 24/7
4. Resource pools
5. Assignment of tasks to resource pools

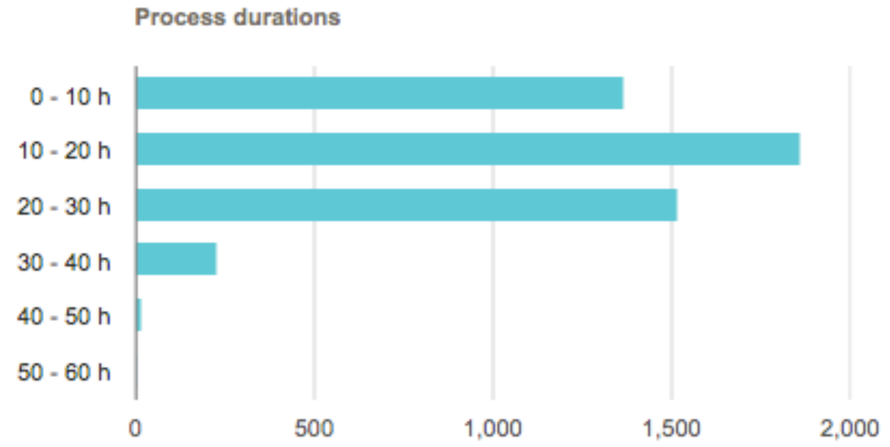
Resource pool assignment

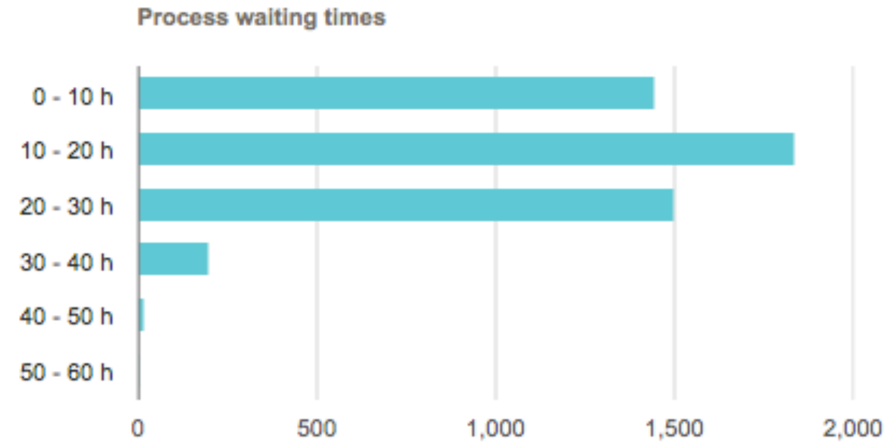


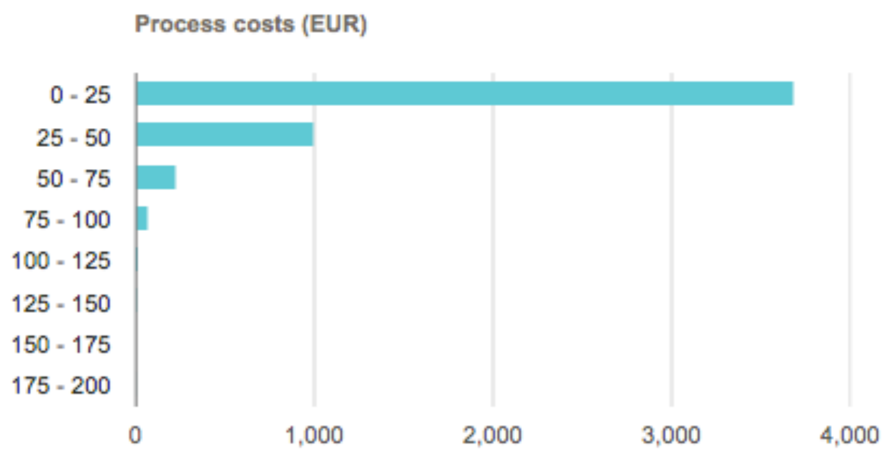
Process Simulation

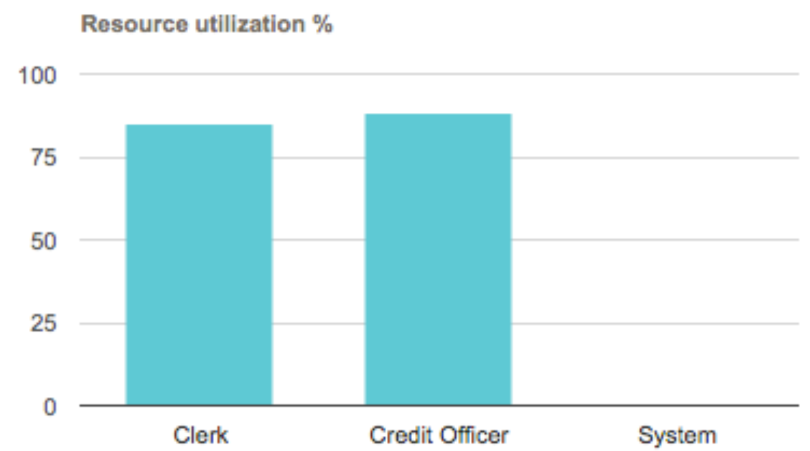


Output: Performance measures & histograms

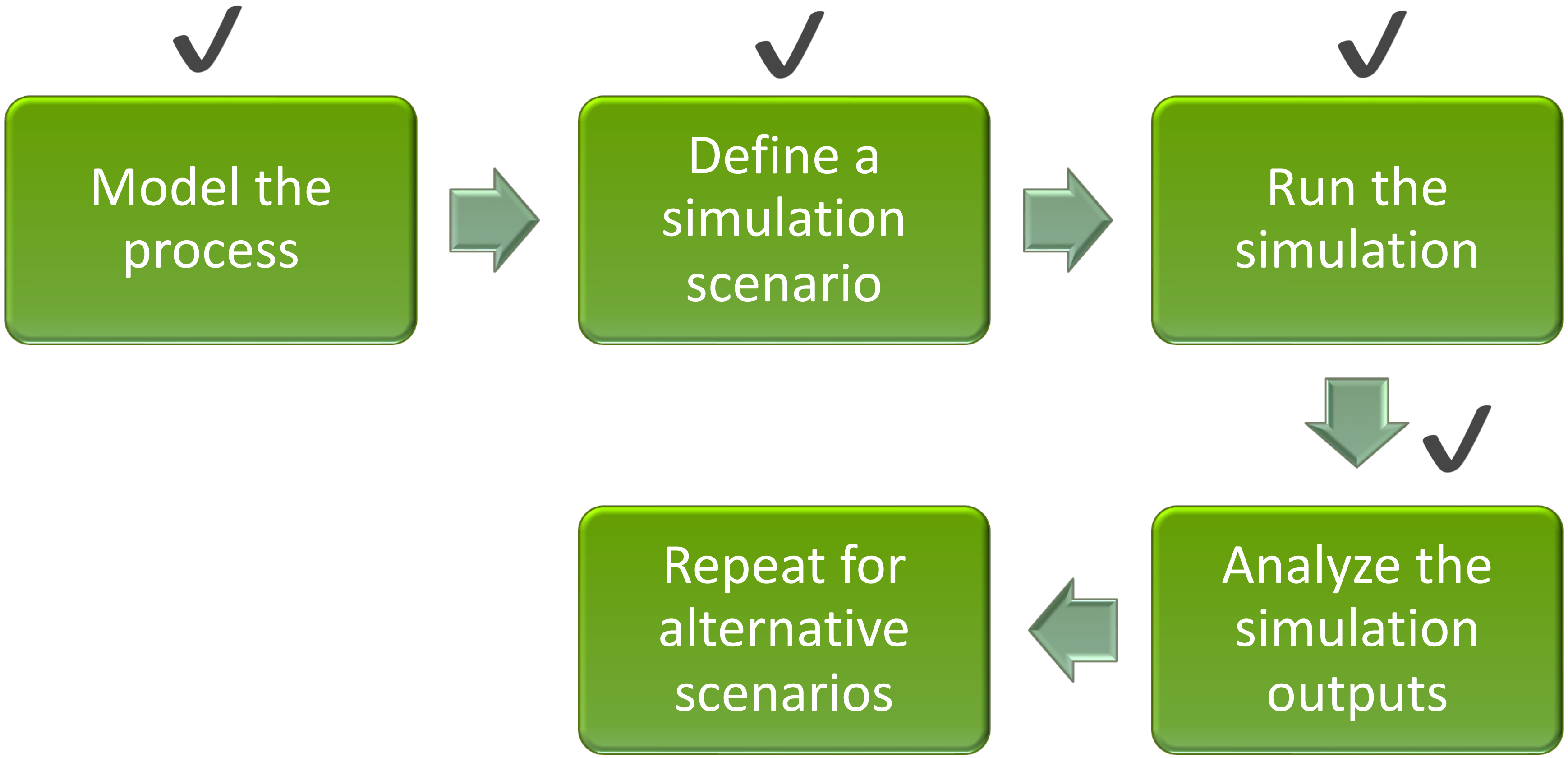




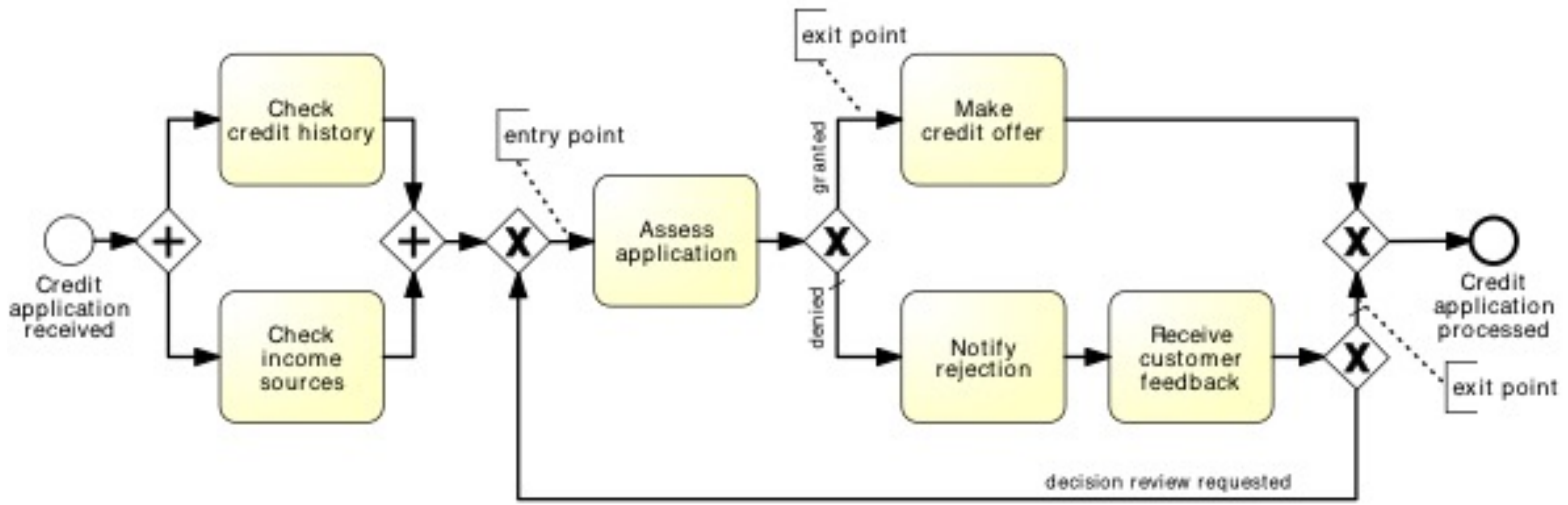




Process Simulation



Demo: Simulation in BIMP



https://bimp.cs.ut.ee/simulator/trial?sample=credit_card_application

Pitfalls of simulation

- Stochasticity
- Data quality
- Simplifying assumptions

Stochasticity

- Problem
  - Simulation results may differ from one run to another
- Solutions
  1. Make the simulation timeframe long enough to cover weekly and seasonal variability, where applicable
  2. Use multiple simulation runs, average results of multiple runs, compute confidence intervals

Data quality

- Problem
  - Simulation results are only as trustworthy as the input data
- Solutions:
  1. Rely as little as possible on “guesstimates”. Use input analysis where possible:
     - Derive simulation scenario parameters from numbers in the scenario
     - Use statistical tools to check fit the probability distributions
  2. Simulate the “as is” scenario and cross-check results against actual observations

Simulation simplifying assumptions

- That the process model is always followed to the letter
  - No deviations
  - No workarounds
- That a resource only works on one task
  - No multitasking
- That if a resource becomes available and a work item (task) is enabled, the resource will start it right away
  - No batching
- That resources work constantly (no interruptions)
  - Every day is the same!
  - No tiredness effects
  - No distractions beyond “stochastic” ones

Acknowledgements

- All material comes from Marlon Dumas, Marcello La Rosa, Jan Mendling, Hajo A. Reijers, authors of the “Fundamentals of Business Process Management” book.