Methods for the specification and verification of business processes MPB (6 cfu, 295AA)

Roberto Bruni http://www.di.unipi.it/~bruni

22 - Quantitative Analysis



We overview some techniques for the quantitative analysis of business processes

Ch.7 of Fundamental of Business Process Management. M. Dumas et al. (inspired by slides available at https://courses.cs.ut.ee/2014/bpm/)

Performance Analysis

Validation

is concerned with the relation between the model and reality Verification

is typically used to answer **qualitative** questions Is there a deadlock possible? It is possible to successfully handle a specific case? Will all cases terminate eventually? It is possible to execute two tasks in any order?

Performance analysis

is typically used to answer **quantitative** questions How many cases can be handled in one hour? What is the average flow time? How many extra resources are required? How many cases are handled within 2 days?

Performance dimensions

Any company would like to make its processes

faster, (time)

cheaper, (finance)

and better. (quality)

KPI

To estimate the performance along any dimension we need to fix something that can be measured

A process performance measure is a quantity that can be unambiguously determined for a given business process

> Time, finance, quality can be refined to a number of **Key Performance Indicators (KPI)**

1. Time Cycle time:

the time needed to handle one case from start to end. One can aim to reduce the *average* cycle time or to reduce the *maximal* cycle time or to meet a cycle time negotiated with the customer

Processing time (also service time):

the time that resources spend on actually handling the case

Waiting time:

the time that a case spends in *idle* mode (e.g., it includes **queueing time** due to unavailability of resources to handle the case)

2. Finance

Cost, turnover, yield or revenue are all concerned with finance related performance dimensions.

A yield increase may have the same effect as a cost decrease w.r.t. the organization profit.

Business process redesign is typically concerned with cost. There are several types of cost: cost of production, of delivery, of human resources, ... Each type can be refined into performance measures by selecting an **aggregation function** such as: count, average, variance, minimum, maximum, ... Example: average delivery cost per item

Cost

We can distinguish between:

Fixed cost: overhead costs not affected by the intensity of processing (e.g., use of infrastructure, maintenance costs).

Variable cost: positively correlated with some variable quantity (e.g. the level of sales, the number of purchased goods, the number of new hires)

Operational cost: closer to productivity, often directly related to the output of a business process (e.g. labor cost in producing a good or delivering a service)

Operational cost

Process redesign is often aimed to reduce operational cost, particularly labor cost

Although task automation may reduce labor cost, it may cause incidental cost involved with developing the respective application and fixed maintenance cost for it

3. Quality

External quality: from the viewpoint of the client (e.g. client satisfaction with the delivered product or with the way the process has been executed) Important factors: amount, relevance, quality and timeliness of the information a client receives as process progresses

Internal quality: from the viewpoint of process participants Important factors: the level of control of the work performed, of variation experienced, of challenges faced

Quality vs time

External process quality is often measured in terms of time (e.g. the average cycle time or the percentage of cases where deadlines are missed)

In the following we assume that any performance measure where time is involved is classified under the time dimension, even if it is related to quality

Deriving performance measures

One possible method for deriving performance measures for a given process is the following:

1. Formulate performance objectives of the process at the high level, in the form of a desirable state that the process should ideally reach

2. For each performance objective, identify the relevant performance dimensions and aggregation functions and derive one or more KPI for the objective

3. Define a target objective for each KPI

A restaurant has recently lost many customers due to **poor customer service**.

The management team has decided to address this issue first of all by **focussing on the delivery of meals**.

The team gathered data by asking customers about how quickly they liked to receive their meals and what they considered as an **acceptable wait**.

The data suggested that half of the customers would prefer their meals to be served in 15 minutes or less.

All customers agreed that a waiting time of **30 minutes or more** is unacceptable.

1. Formulate performance objectives of the process at the high level, in the form of a desirable state that the process should ideally reach (e.g., customers should be served in less than 30 minutes)

2. For each performance objective, identify the relevant performance dimensions and aggregation functions and derive one or more KPI for the objective (e.g., time dimension, ST₃₀ be the percentage of customers served in less than 30 minutes)

3. Define a target objective for each KPI (e.g., $ST_{30} \ge 97\%$)

1. Formulate performance objectives of the process at the high level, in the form of a desirable state that the process should ideally reach (e.g., customers should be served in about 15 minutes)

2. For each performance objective, identify the relevant performance dimensions and aggregation functions and derive one or more KPI for the objective (e.g., time dimension, ST_{15} be the percentage of customers served in less than 15 minutes)

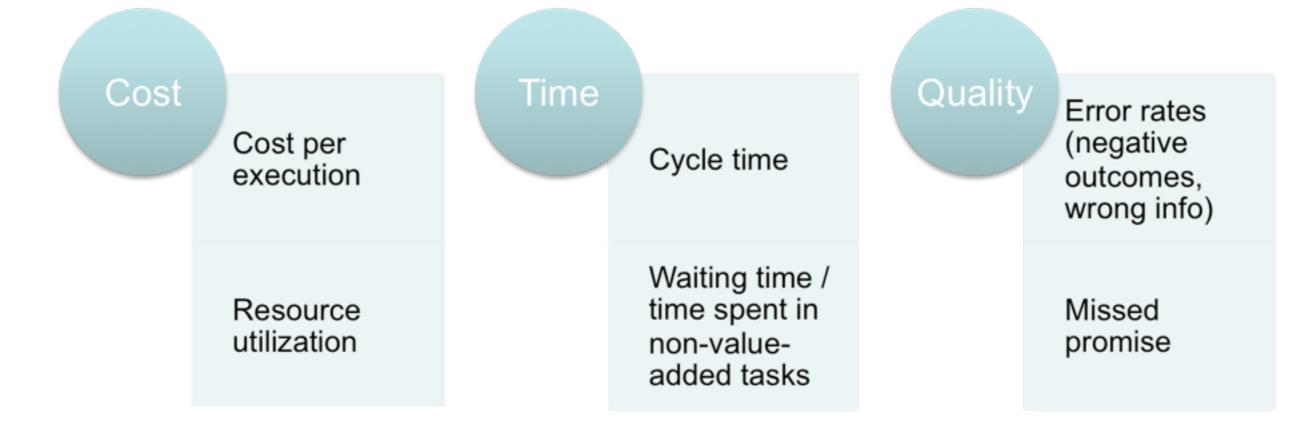
3. Define a target objective for each KPI (e.g., $ST_{15} \ge 85\%$)

1. Formulate performance objectives of the process at the high level, in the form of a desirable state that the process should ideally reach (e.g., customers should be served in about 15 minutes)

2. For each performance objective, identify the relevant performance dimensions and aggregation functions and derive one or more KPI for the objective (e.g., time dimension, AMDT be the average meal delivery time)

3. Define a target objective for each KPI (e.g., AMDT \leq 15')

Typical process performance measures



Flow analysis

Flow analysis is a family of techniques that allow us to estimate the overall performance of a process given some knowledge about the performance of its activities

Examples:

we can calculate the min/max/average **cycle time** of an entire process if we know the min/max/average cycle time of each activity involved in the process.

We can compute the average **cost** of a process instance knowing the cost-per-execution of each activity. We can calculate the **error rate** of a process given the error rate of each activity.

Cycle time analysis

Cycle time analysis

Cycle time = difference between the start time (ready to be executed) and the end time (completion) of a case

Cycle time analysis = the task of calculating the average cycle time of an entire process or some process fragment

Assumption: average activity times are available for all the activities involved in the process

Activity time = waiting time + processing time

Flow patterns

The simplest case is that of a single activity, but then we can take into account different structure patterns that frequently occur:

paths composed in sequence

alternative paths (XOR split and join)

parallel paths (AND split and join)

rework (1-or-more cycles, 0-or-more cycles)

Notation

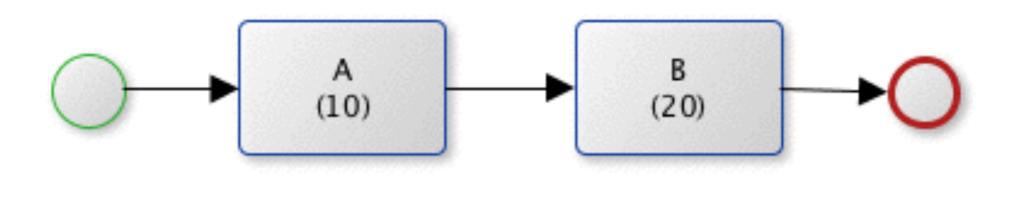
We denote the average cycle time by **CT** and call it simply *cycle time*

When several (sub)processes P₁, P₂, ..., P_n are involved we refer to their cycle times by CT₁, CT₂, ..., CT_n

Similarly, if activities A, B, ... are involved we refer to their cycle times by CT_A , CT_B , ...

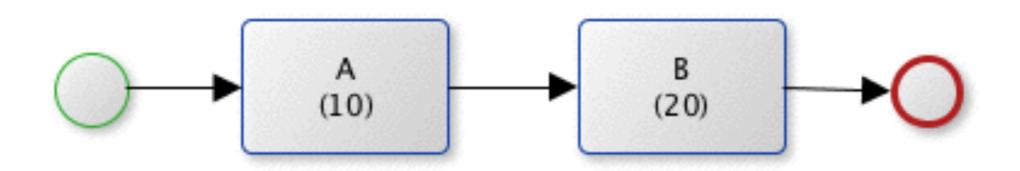
In diagrams, we will often write activities cycle time within parentheses





CT = ?

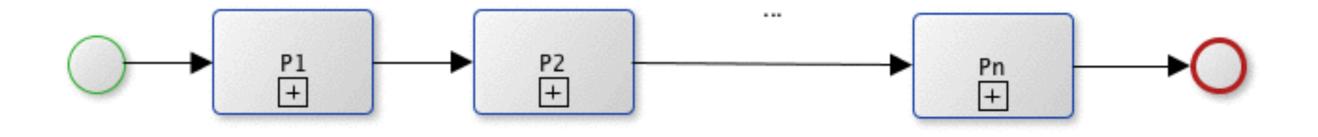




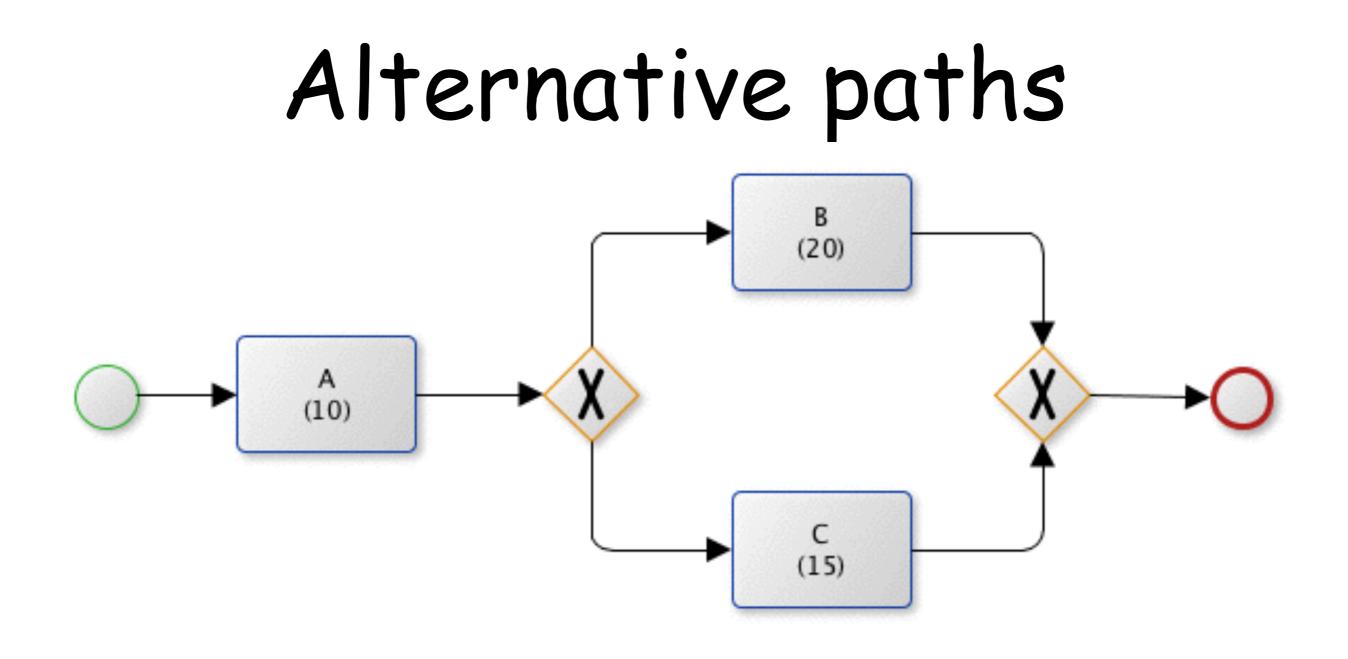
 $CT = CT_A + CT_B = 10 + 20 = 30$

Sequence

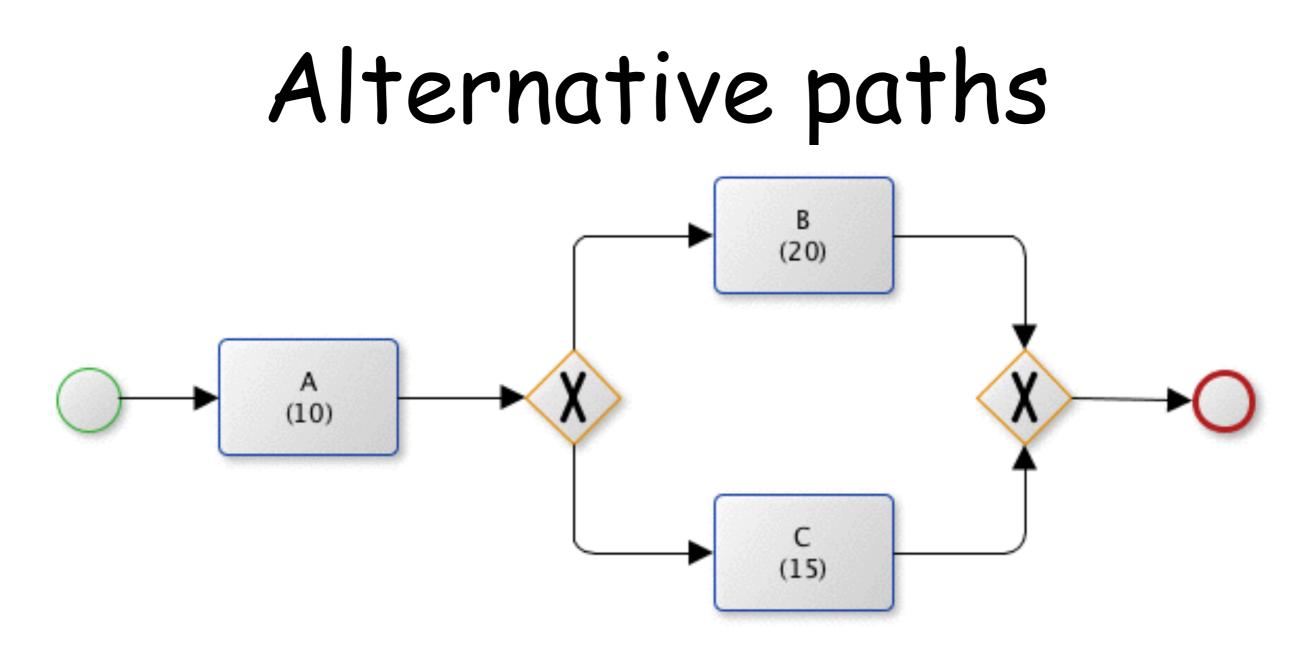
The cycle time of a purely sequential fragment of a process is the sum of the cycle times of the activities in the fragment



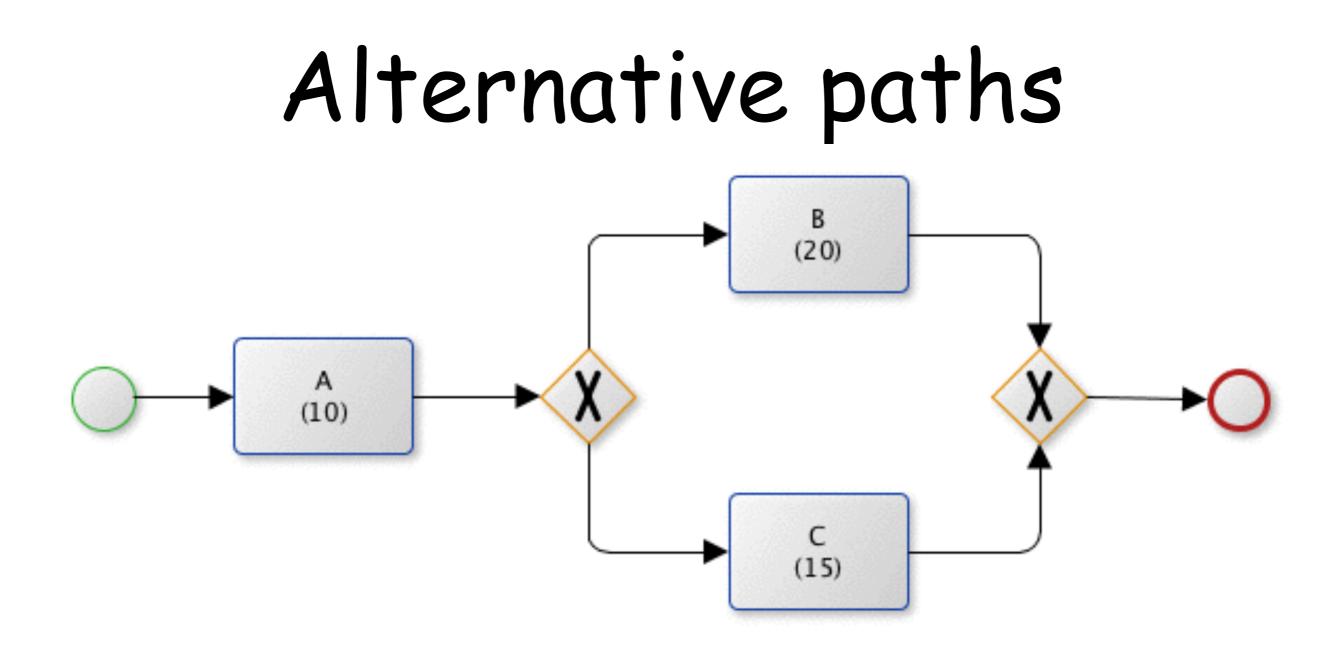
 $CT = \sum_{i=1}^{n} CT_i$



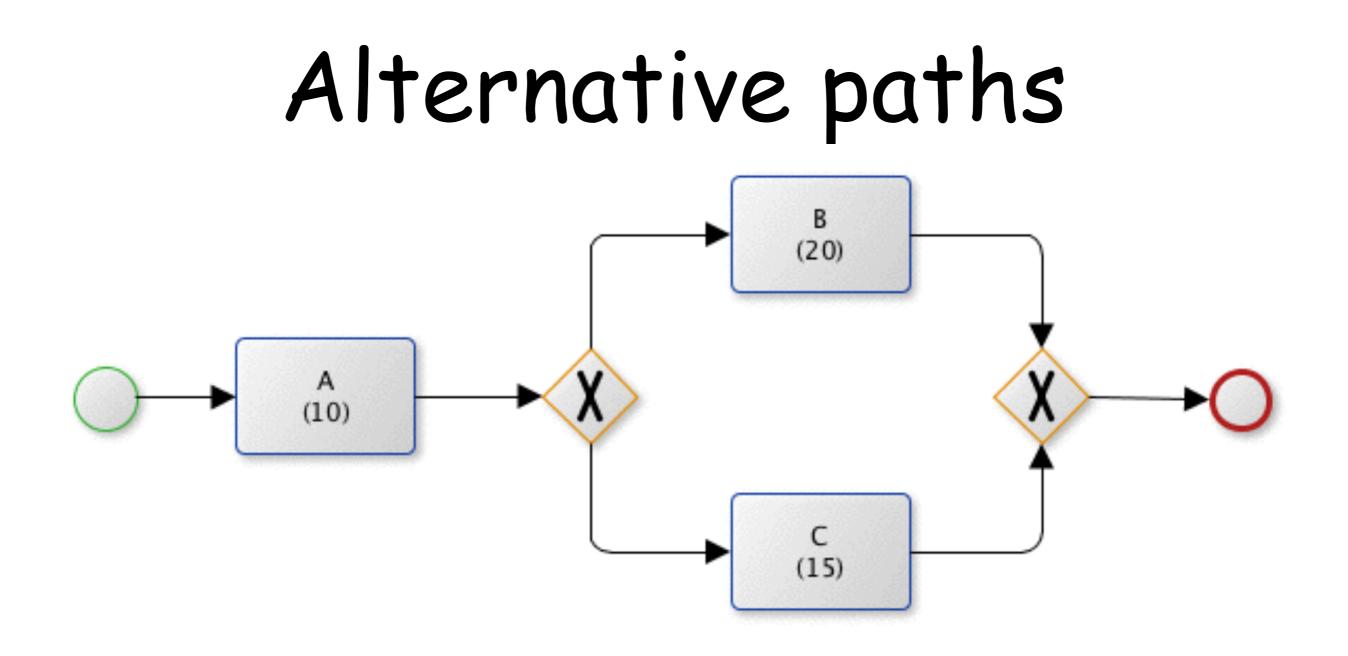
CT = ?



in some cases $CT = CT_A + CT_B = 10 + 20 = 30$ in other cases $CT = CT_A + CT_C = 10 + 15 = 25$ whether the average is closer to 25 or to 30 depends on **how frequently** each branch is taken

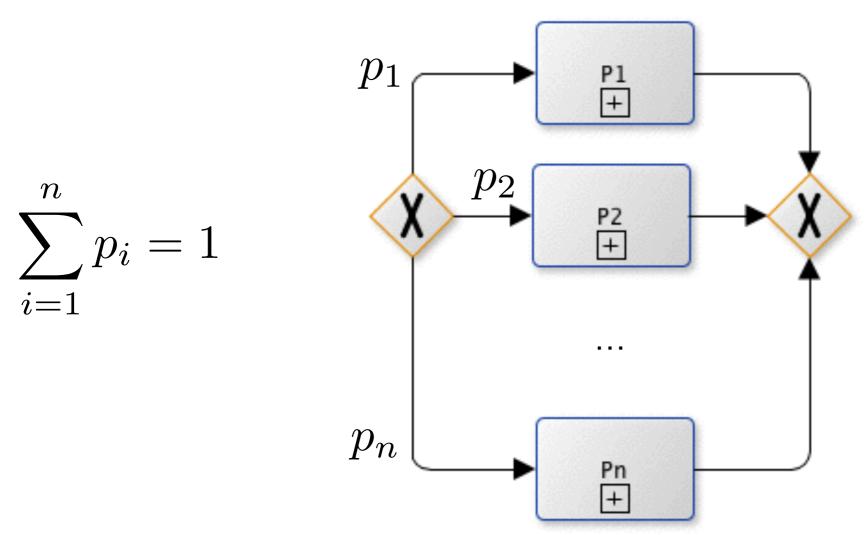


if B and C are taken in 50% of the cases each, then the average will sit in the middle between 25 and 30

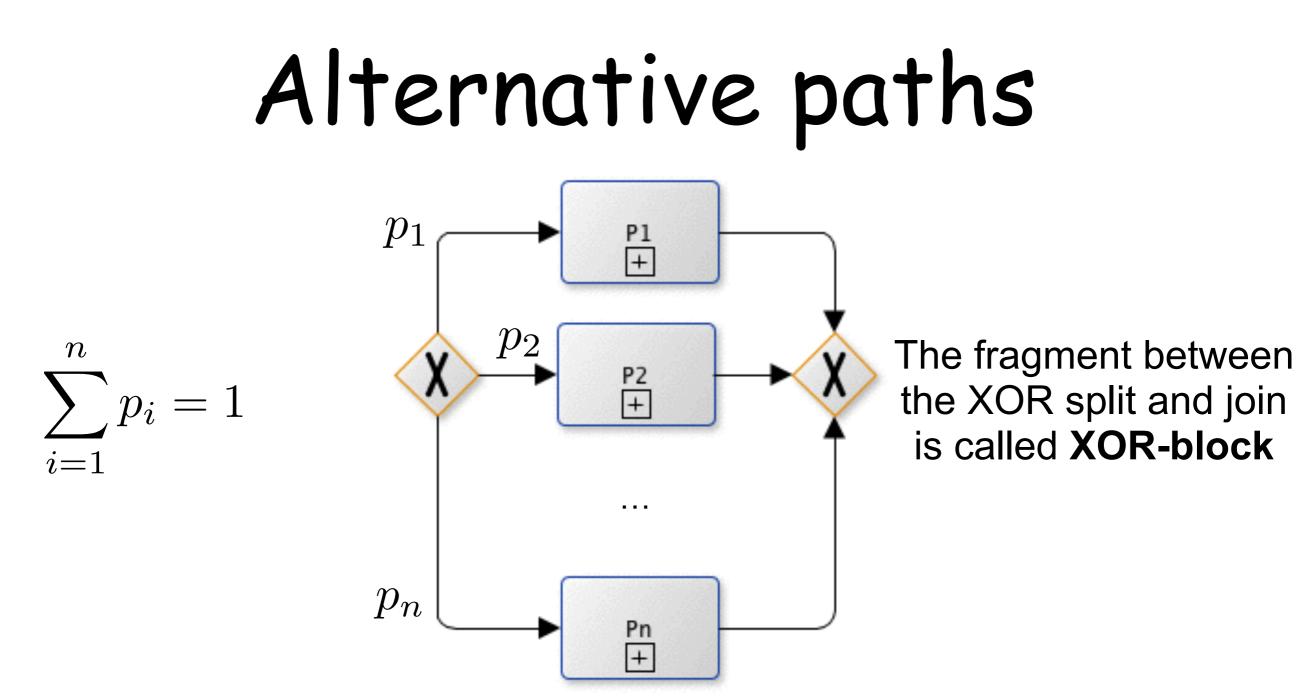


if B is taken in 90% of the cases, then the average will be closer to 30 than 25

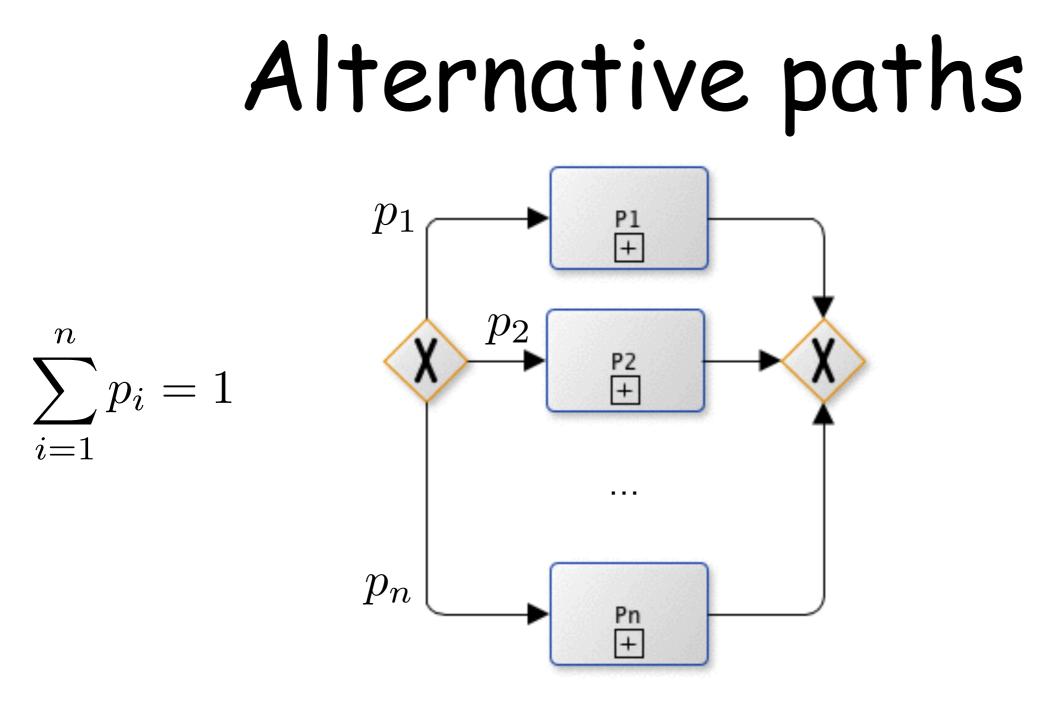


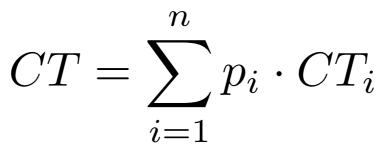


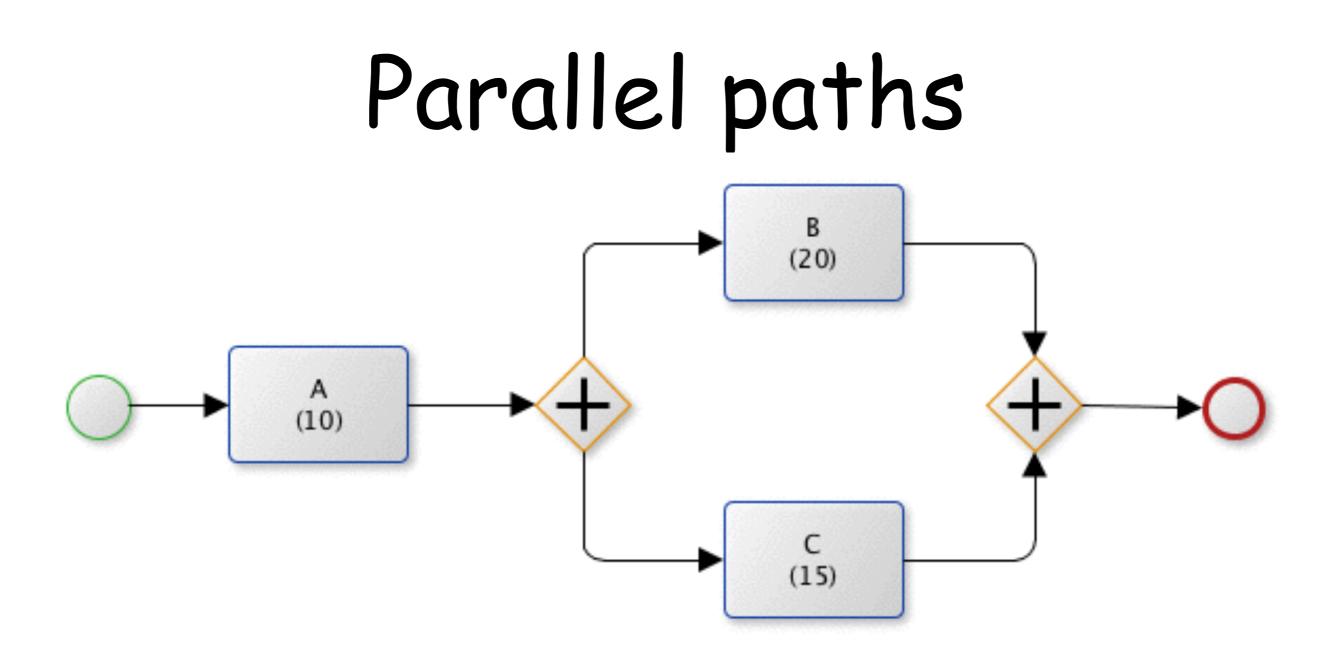
Branching probability p_i: is the frequency with which a given branch of a decision gateway is taken



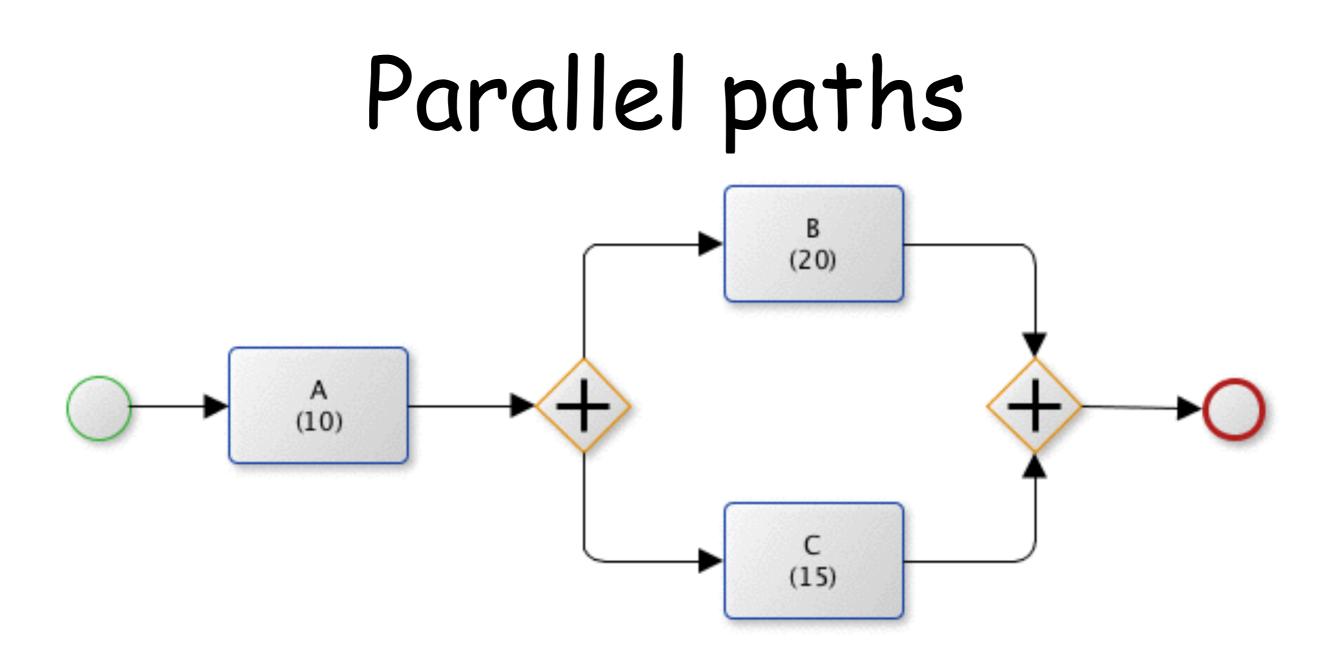
The cycle time of a XOR-block fragment is the **weighted average** of the cycle times of the branches



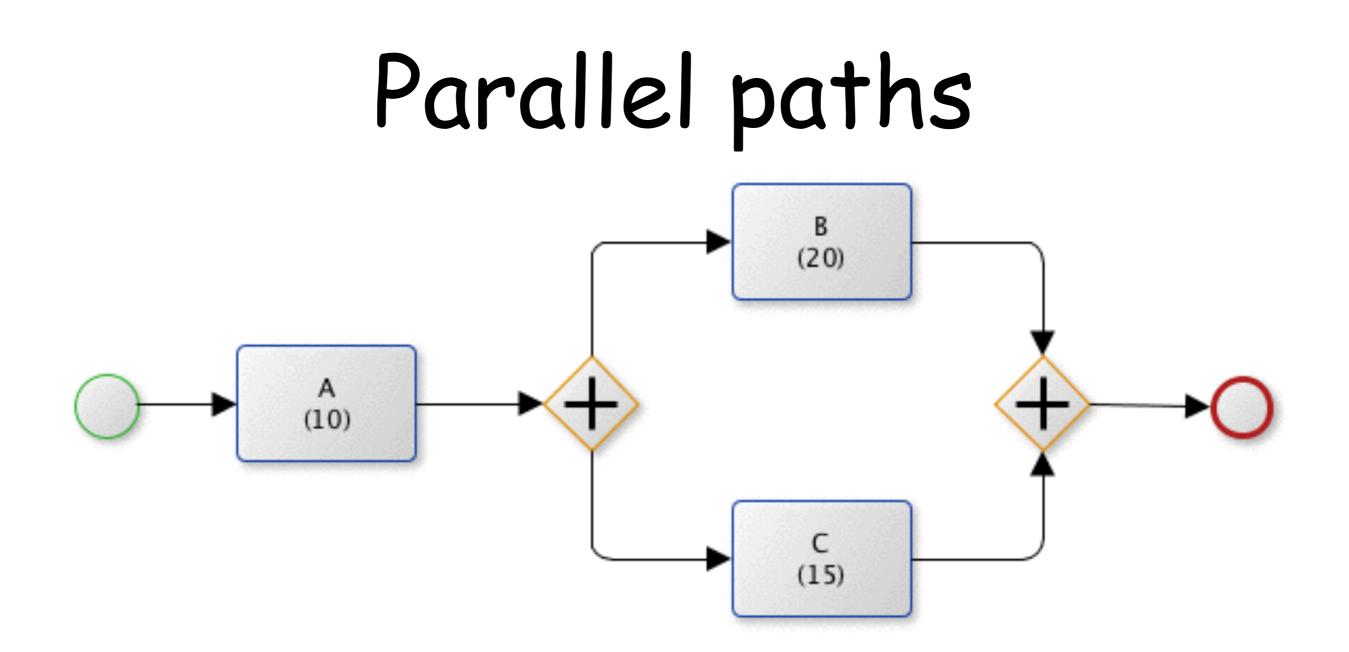




 $CT = CT_A + CT_B + CT_C = 10 + 20 + 15 = 45$?

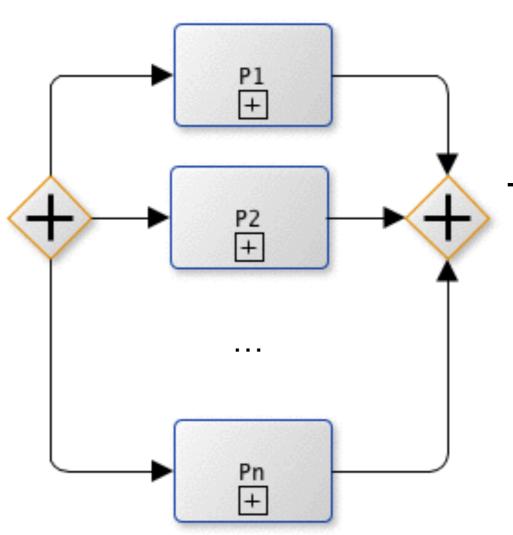


 $CT = CT_A + CT_B + CT_C = 10 + 20 + 15 = 45$? (but while B is executed, also C is executed, and B takes longer than C)



 $CT = 10 + max \{20, 15\} = 10 + 20 = 30$

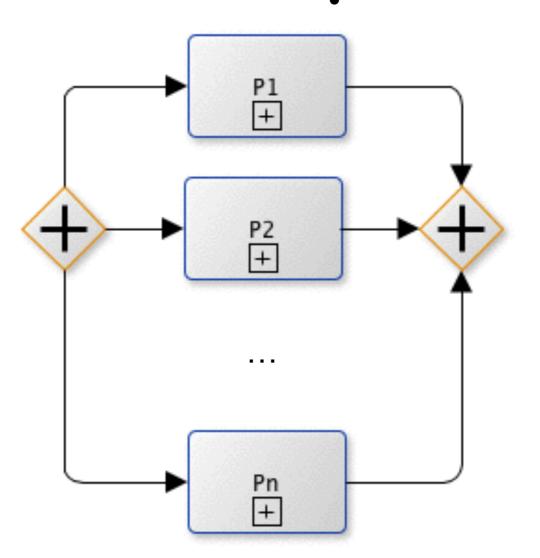
Parallel paths



The fragment between the AND split and join is called **AND-block**

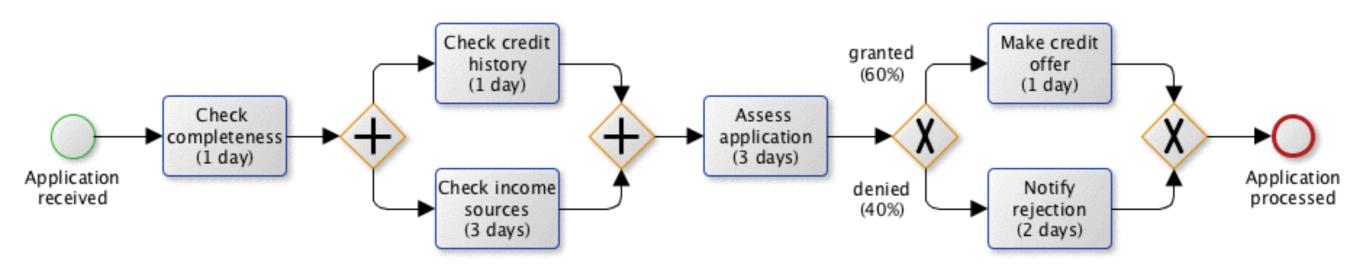
The cycle time of an AND-block fragment is the cycle time of the **slowest** branch

Parallel paths



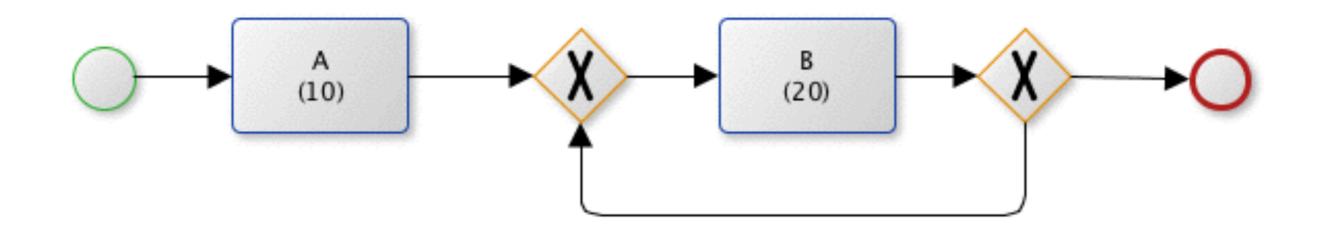
$$CT = \max_{i} \{CT_i\} = \max\{CT_1, CT_2, ..., CT_n\}$$

Question time

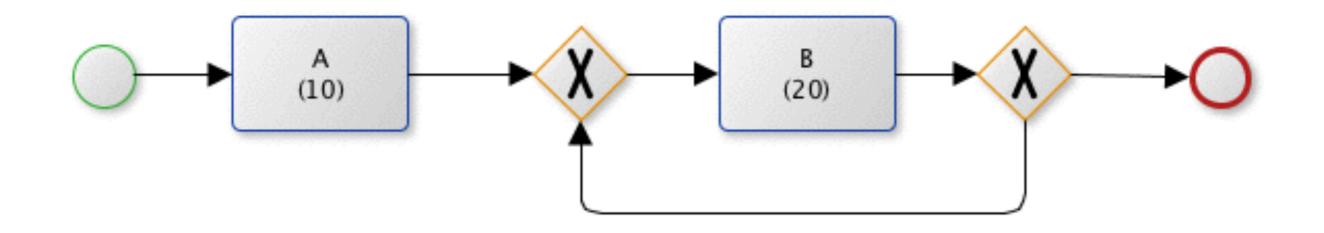


CT = ?

Rework loop (1 or more times)

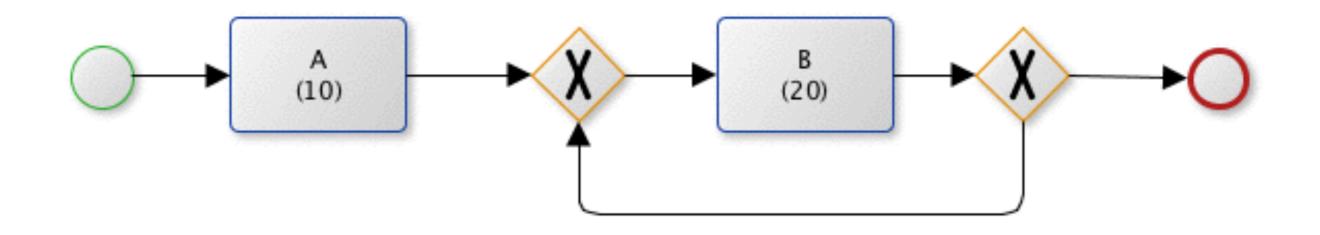


Rework loop (1 or more times)

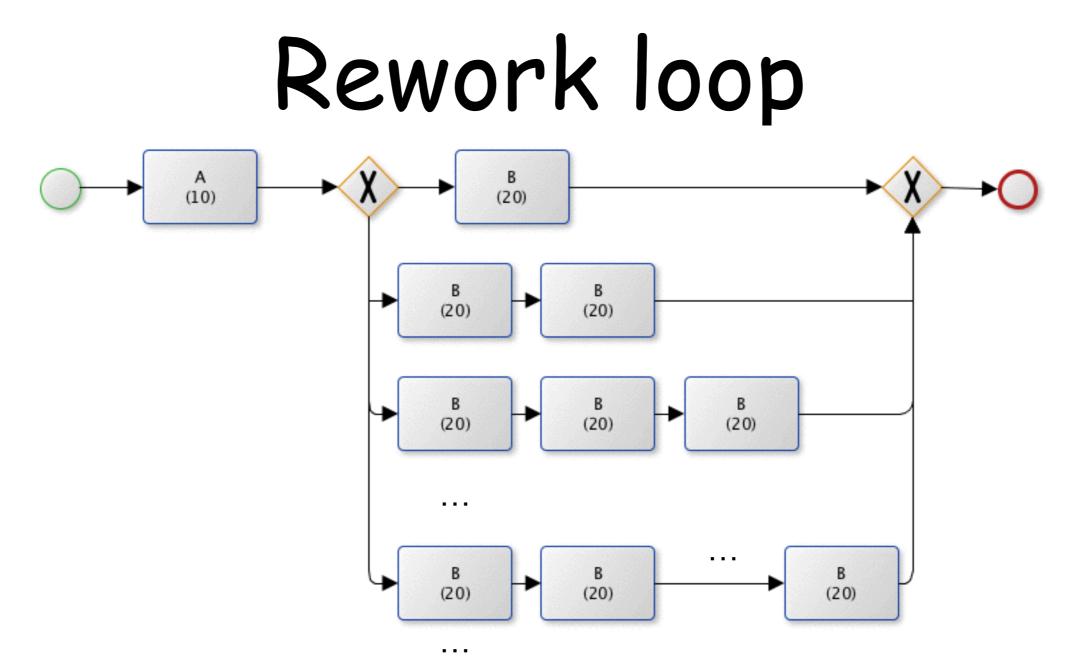


CT = 10 + 20 + 20 + 20 + ... ?

Rework loop (1 or more times)

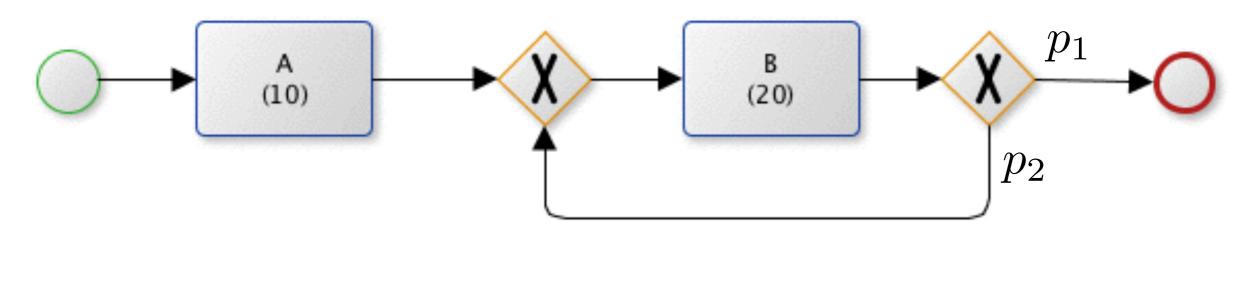


For sure we can say that B will be executed once. (CT ≥ 10 + 20 = 30) Then, depending on a choice, B can be executed twice. Then, a third time, and so on ...



For sure we can say that B will be executed once. Then, depending on a choice, B can be executed twice. Then, a third time, and so on ...

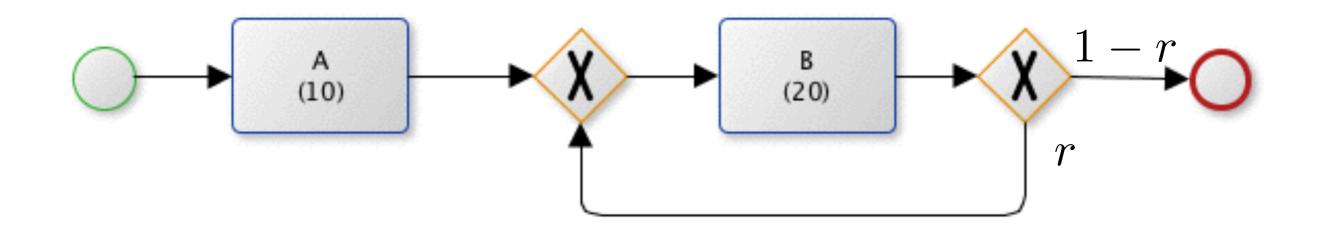
Branching probability, again...



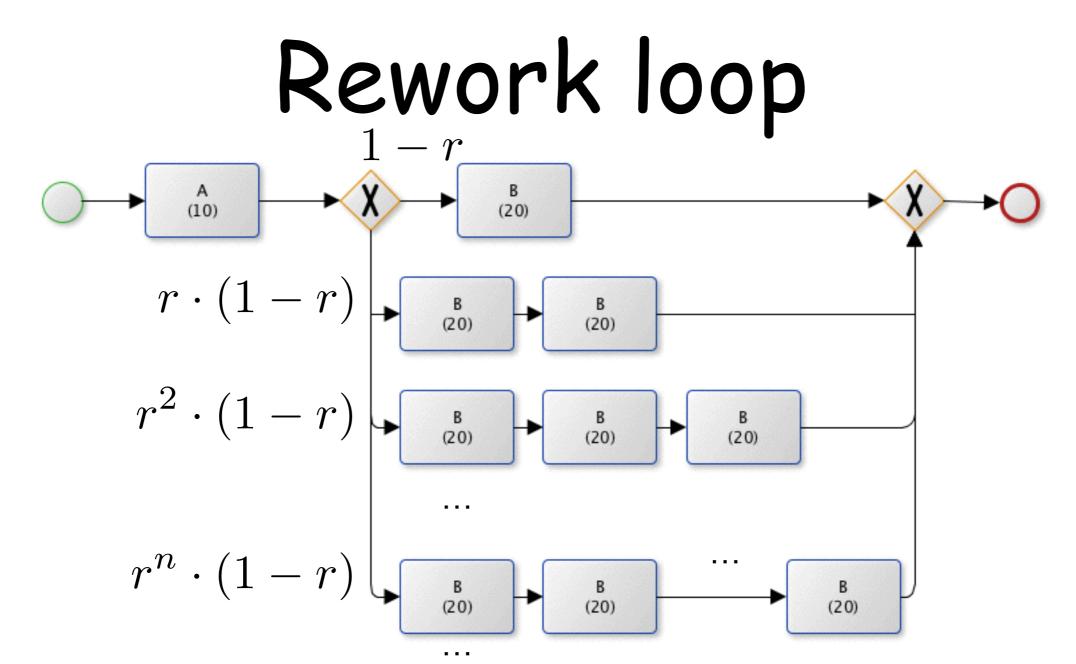
 $p_1 + p_2 = 1$

Branching probability p_i : is the frequency with which a given branch of a decision gateway is taken

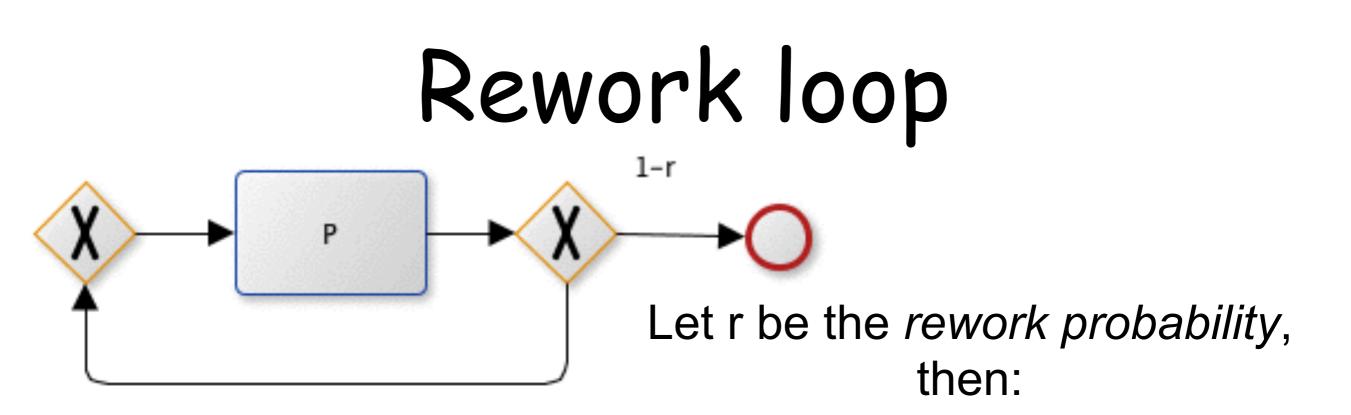
Rework probability



Rework probability r: is the frequency with which the task is reworked



For sure we can say that B will be executed once. Then, depending on a choice, B can be executed twice. Then, a third time, and so on ... **but always with less and less probability**

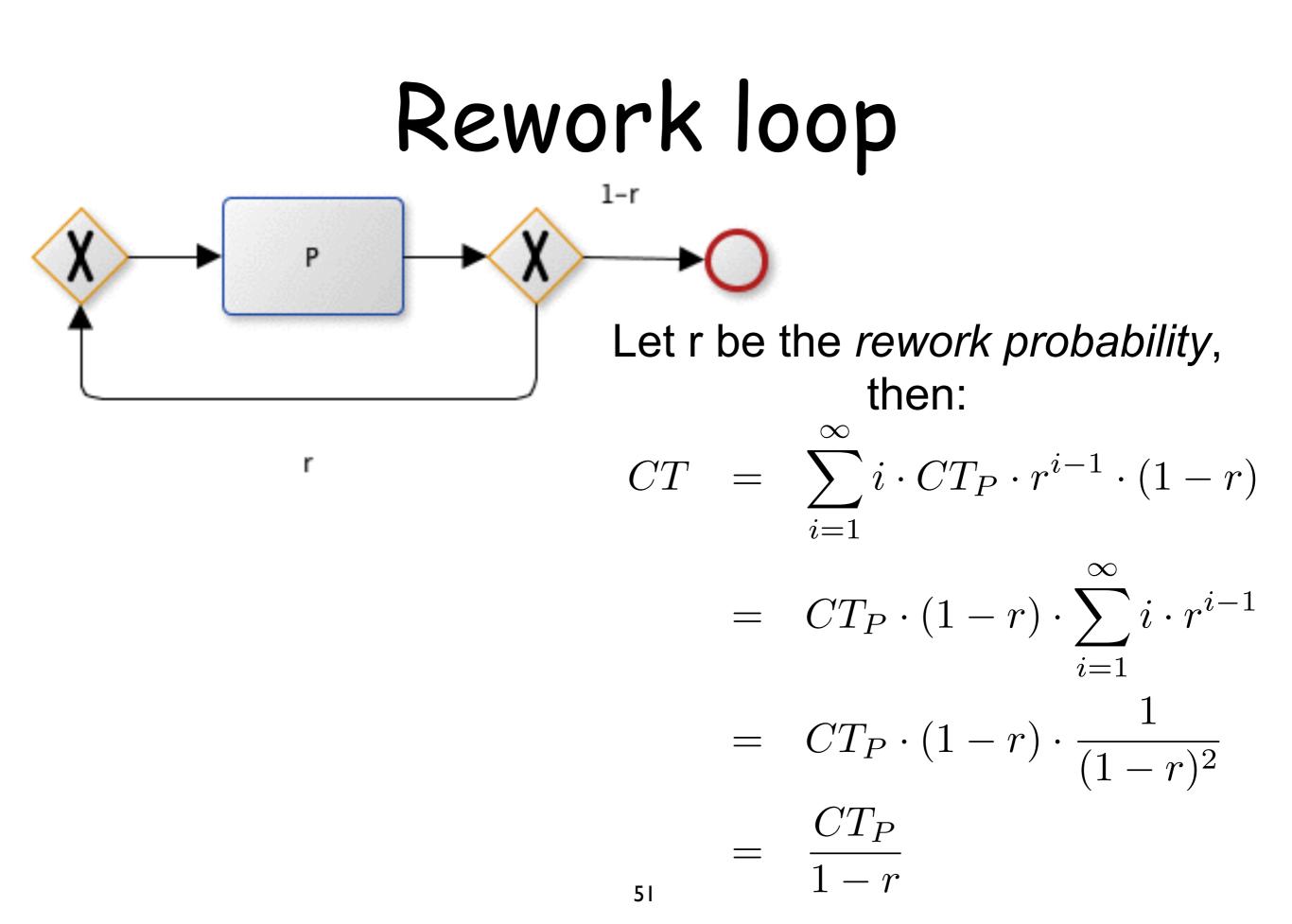


r

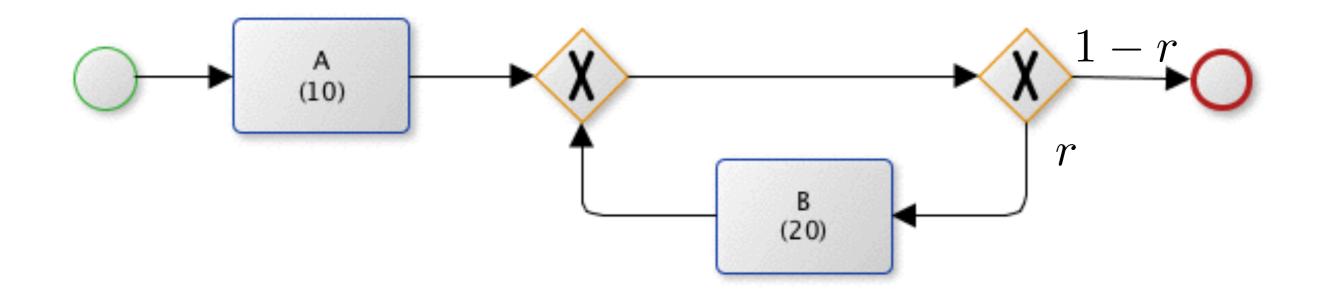
$$CT = 1 \cdot CT_P \cdot r^0 \cdot (1 - r) + 2 \cdot CT_P \cdot r^1 \cdot (1 - r) + \dots + n \cdot CT_P \cdot r^{n-1} \cdot (1 - r) + \dots = \sum_{i=1}^{\infty} i \cdot CT_P \cdot r^{i-1} \cdot (1 - r)$$

i=1

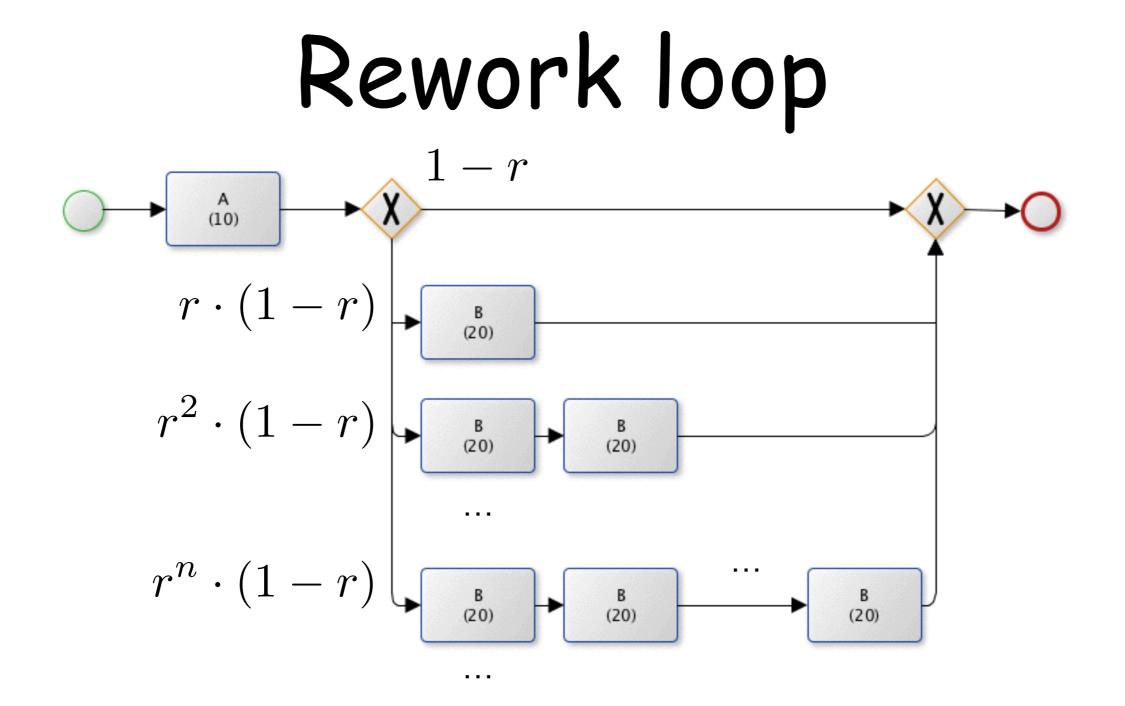
-r

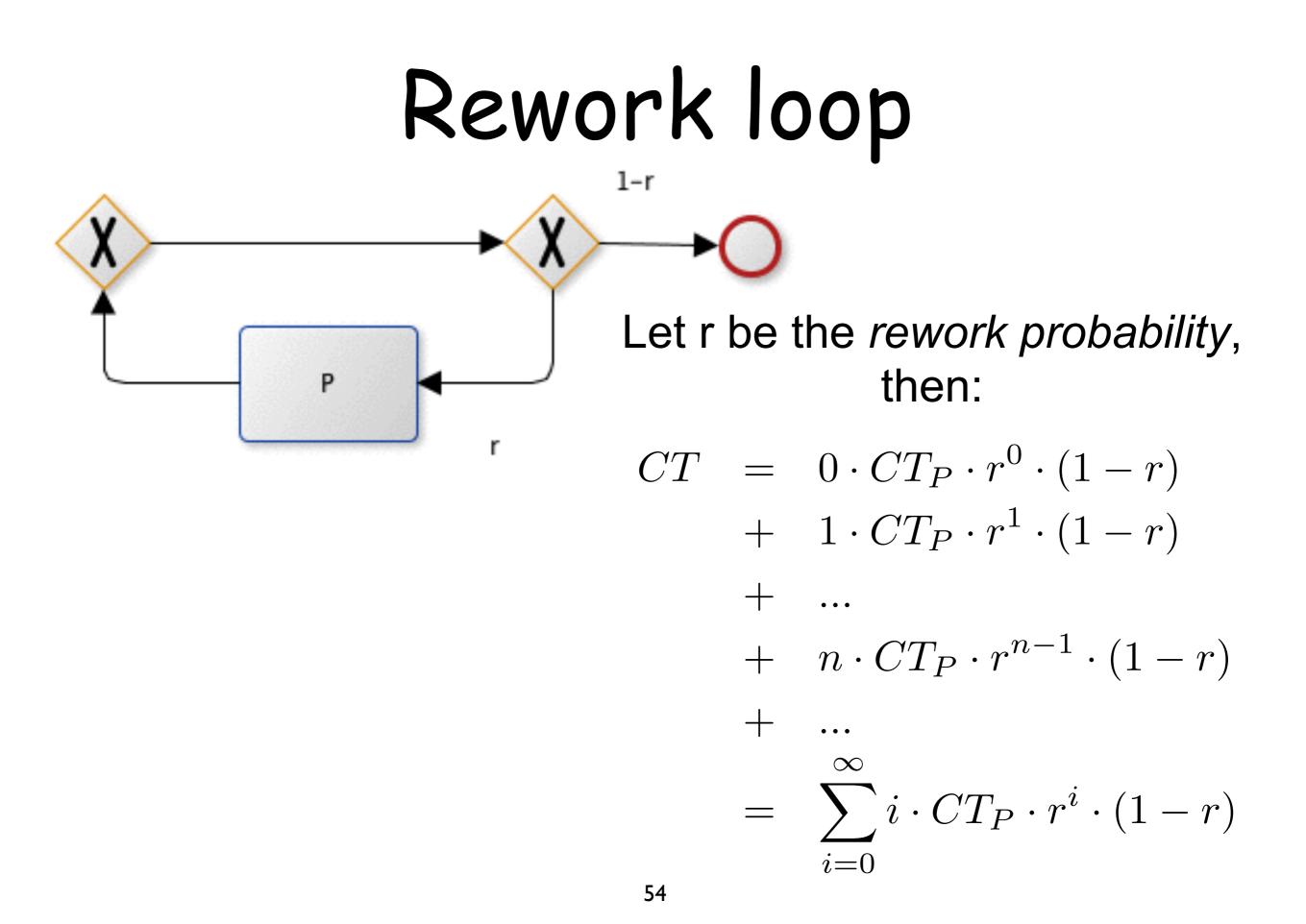


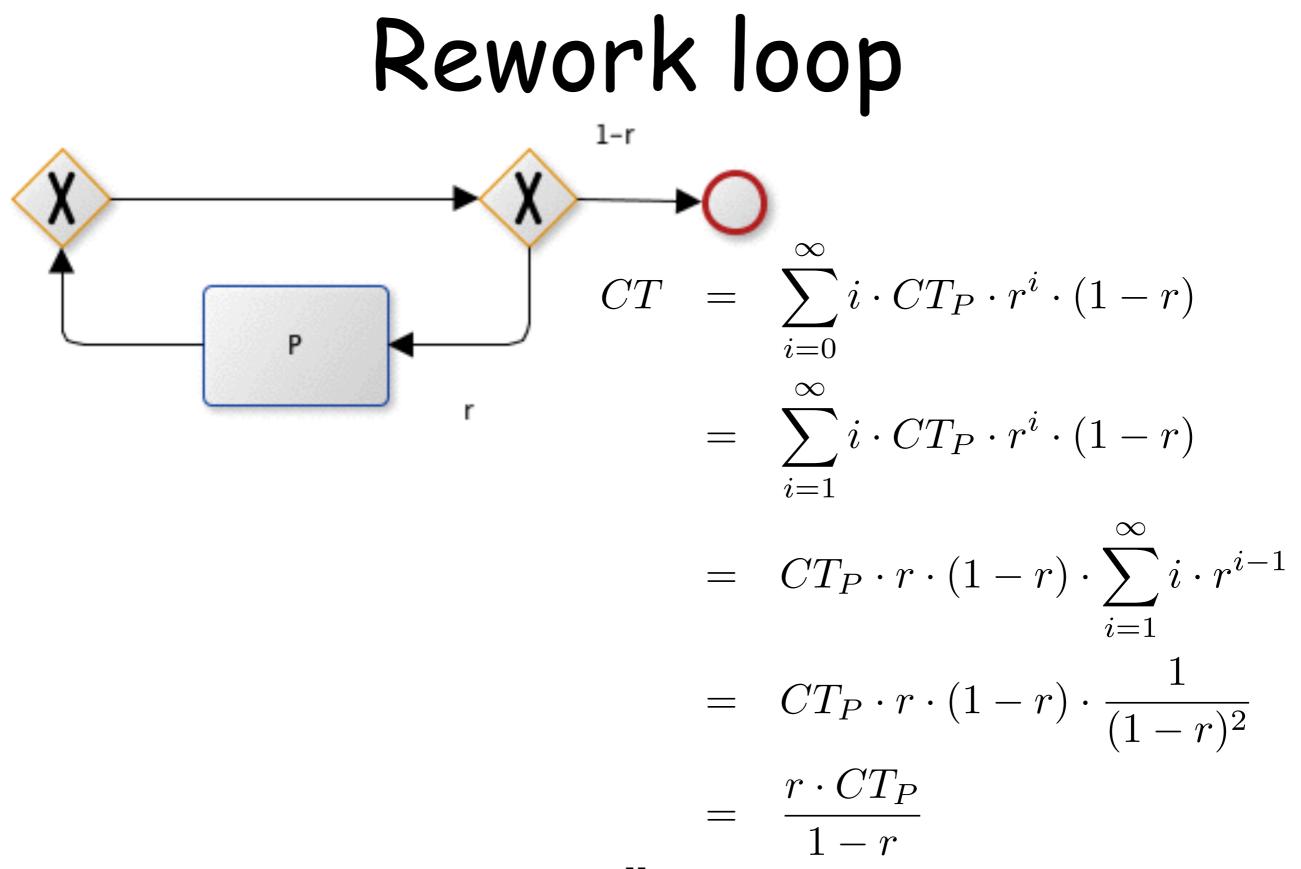
Rework loop (O or more times)

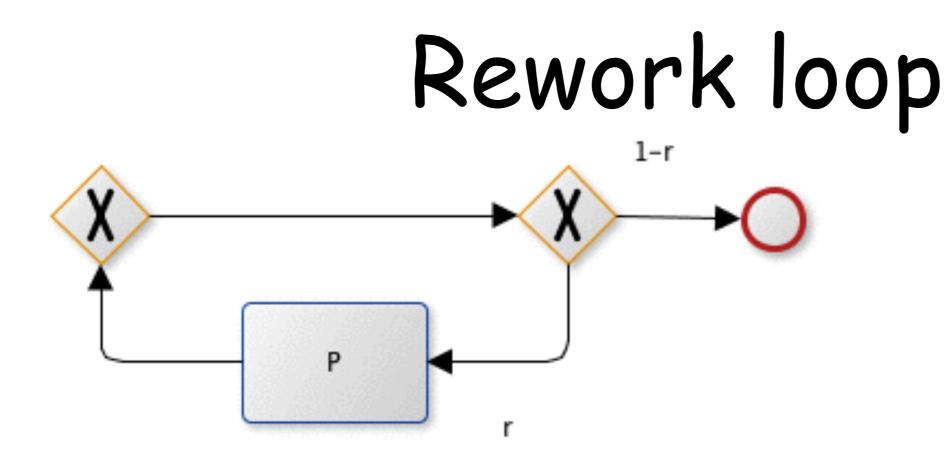


CT = ?







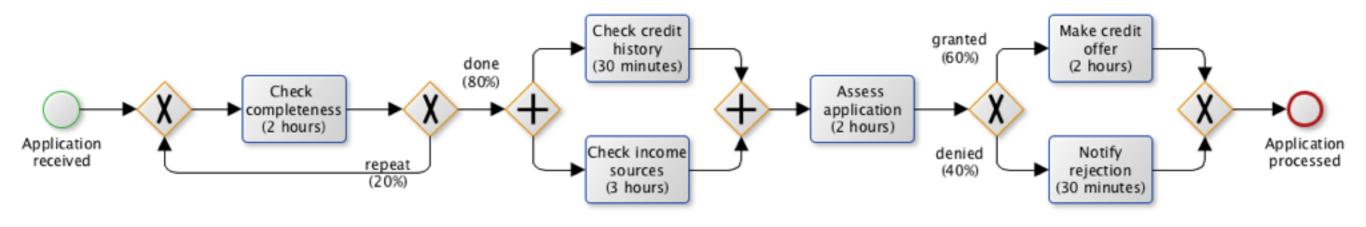


Intuitively,

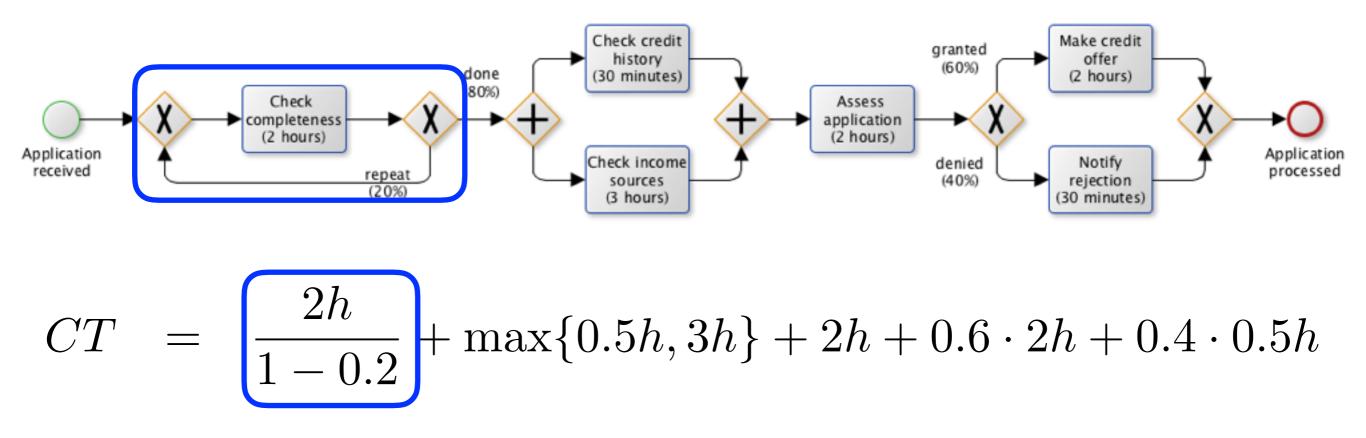
if $\frac{C' T_P}{1-r}$ is the average cycle time for reworking P one or more times,

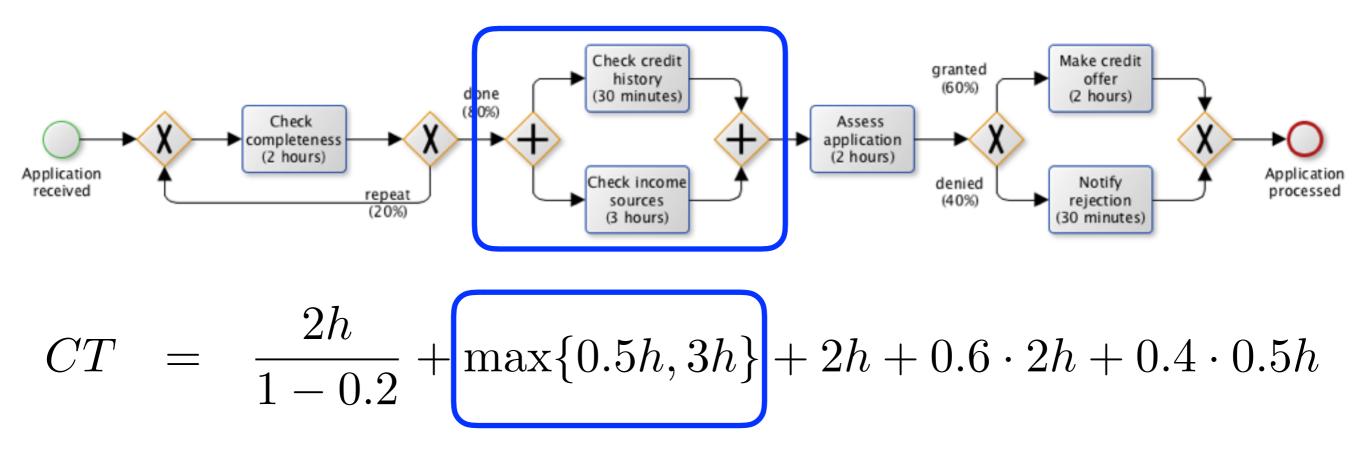
then $\frac{CT_P}{1-r} - CT_P = \frac{(1-(1-r)) \cdot CT_P}{1-r} = \frac{r \cdot CT_P}{1-r}$ is the average cycle time for reworking P zero or more times

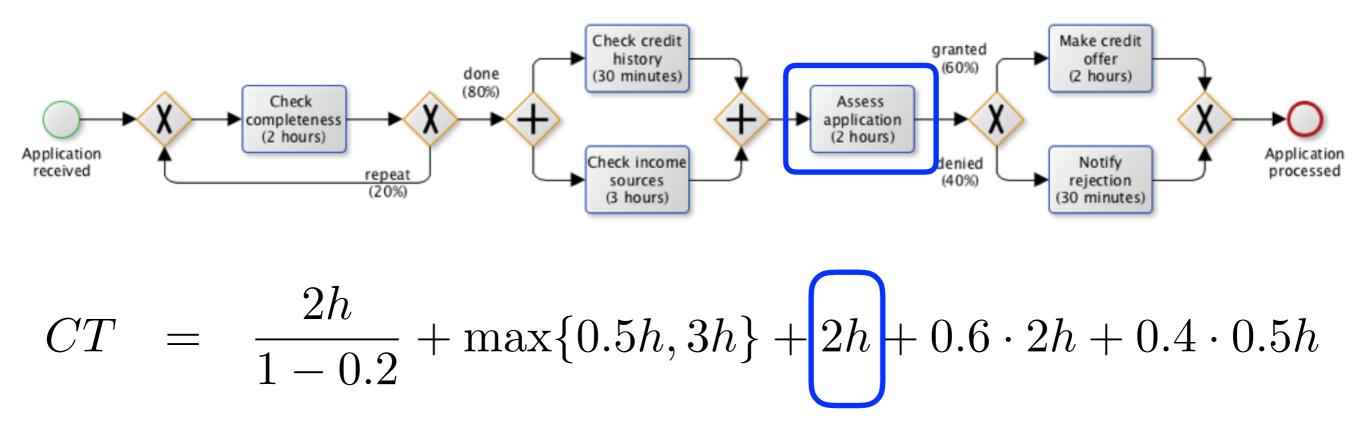
Compute the average cycle time CT of the process below

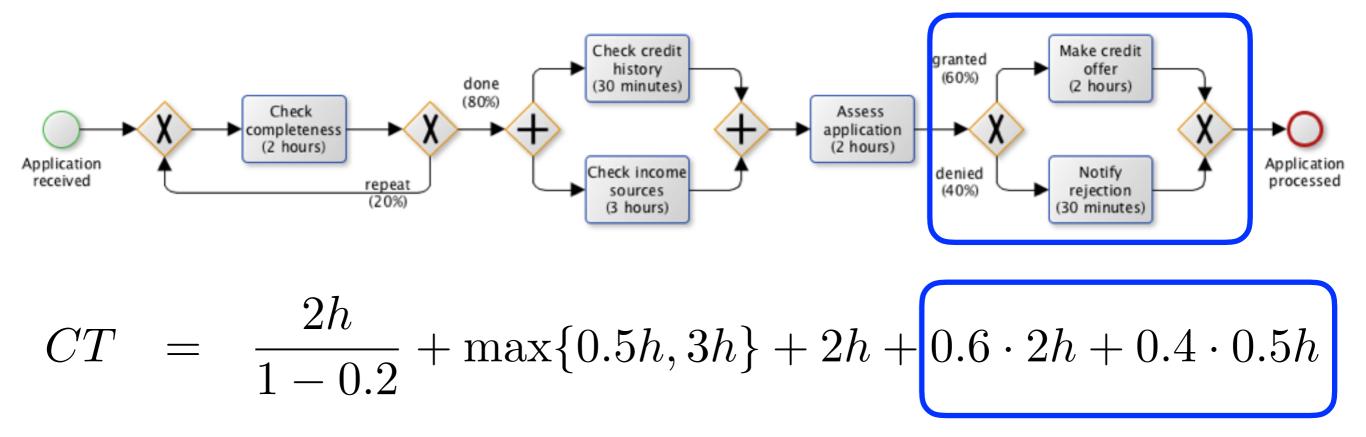


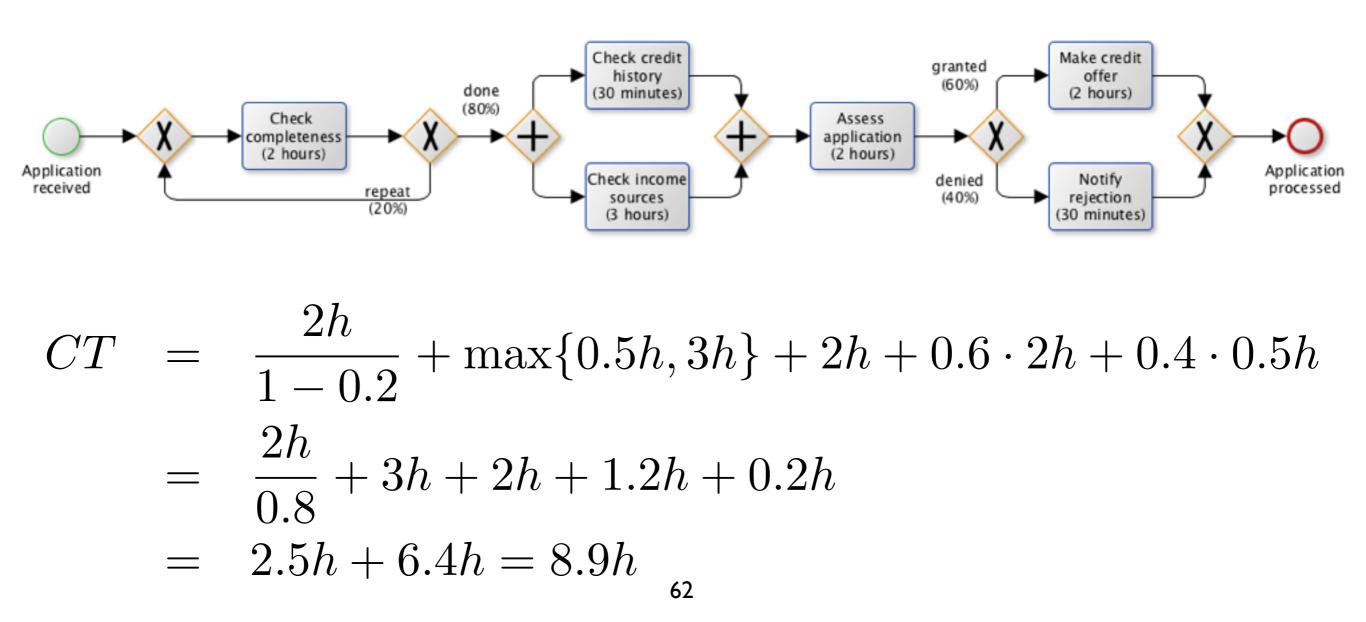
 $CT = \frac{2h}{1 - 0.2} + \max\{0.5h, 3h\} + 2h + 0.6 \cdot 2h + 0.4 \cdot 0.5h$



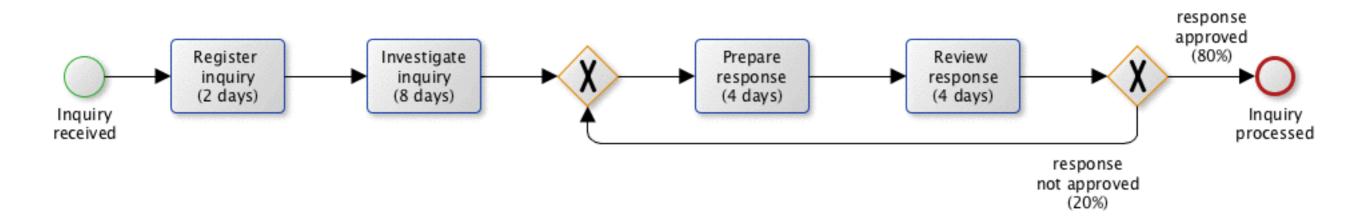








Exercise



Waiting vs processing

As mentioned at the beginning, the cycle time of an activity or a process can be divided into *waiting time* and *processing time*

Waiting time:

is the portion of the cycle time where no work is being done to advance the process (e.g. time spent in transferring documents or waiting for an actor to perform the work)

Processing time:

is the time that actors spend doing actual work

Waiting vs processing

In most processes,

the waiting time is a considerable portion of the cycle time!

For example,

in many situations cases are processed in batches (e.g. applications, surveys)

and in many other cases actors are just not ready (e.g. supervisor approval, medical prescription)

Cycle time efficiency

Cycle time efficiency (CTE): is the ratio of overall processing time relative to the overall cycle time

A ratio close to 1 indicates that there is little room for improving the cycle time (unless radical changes in the process)

A ratio close to zero indicates that there a significant amount of room for improving cycle time (by reducing the waiting time)

Theoretical cycle time

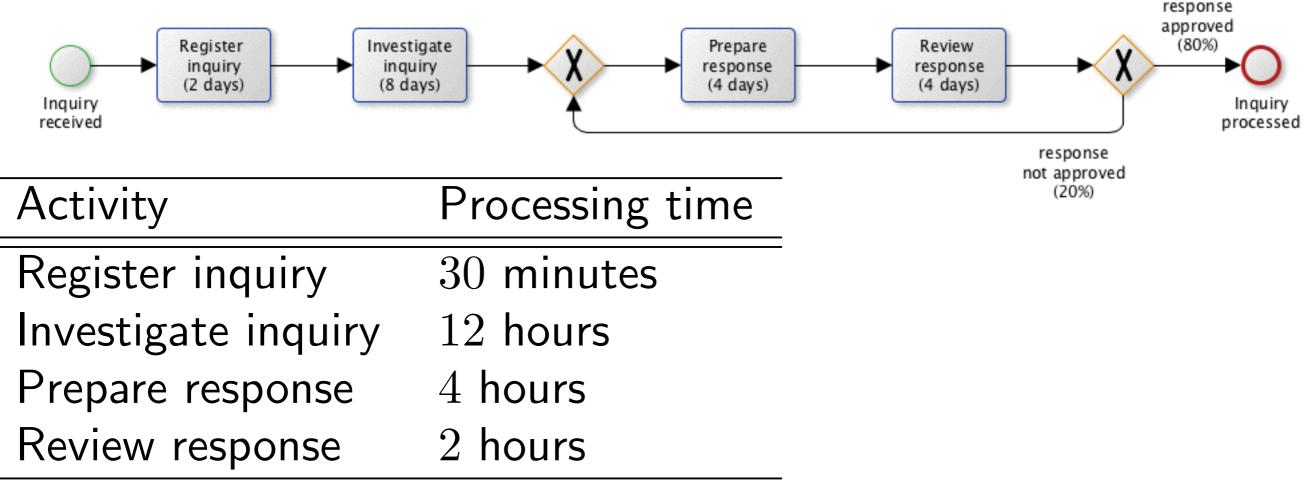
Assume that for each activity of the process both the processing time and the cycle time are known

Let **TCT** denote the **theoretical cycle time** of the process: this is computed in the same ways as CT, but using the processing time of activities (it is the amount of time a process would take on average if no waiting time was necessary)

$$CTE = \frac{TCT}{CT}$$

Exercise

Compute the TCT and CTE of the process below, given the processing times reported in the table (assume 1 day = 8 working hours)



Little's law

Arrival rate and Work-In-Process

Cycle time is directly related to two other important measures:

Arrival rate λ of a process: is the average number of new instances of the process (i.e. cases) that are created per time unit

Work-In-Process (WIP):

is the average number of instances of a process (i.e. cases) that are active (i.e. not yet completed) at a given point in time

Little's law

In a paper from 1954, operation research professor John Little assumed the following formula holds true (without giving a proof):

The long-term average number of customers in a **stable** system (WIP) is equal to the long-term average effective arrival rate (λ) multiplied by the average time a customer spends in the system (CT) algebraically: **WIP = \lambda \cdot CT**

> (``stable" means that the number of customers in the system is not increasing infinitely)

$WIP = \lambda \cdot CT$

Little's law tell us that:

WIP increases if the cycle time (CT) increases or if the arrival rate (λ) increases (if the process slow down there will be more active cases and the faster new cases are created the higher will be the number of active instances)

If the arrival rate (λ) increases and we want to keep WIP constant, then we must decrease the cycle time (CT) (i.e., we must work faster)

A note on Little's law

The law is classically stated using different symbols $\mathbf{L} = \mathbf{\lambda} \cdot \mathbf{W}$

In a subsequent paper from 1961, John Little proved the equality later followed by simpler proofs in 1967, 1969, 1972

Since we can estimate WIP and λ by observing the system, we can use Little's law as an alternative way to calculate the average cycle time CT:

$$CT = \frac{WIP}{\lambda}$$

Assume there are 250 business days per year.

If the total number of applications received over the last year is 2500 we can infer that the average number of *applications per day* is 10 (i.e. λ =10).

By sampling (e.g. checking every week), we observed that on average there were 200 *applications concurrently active* (i.e. WIP=200).

$$CT = \frac{WIP}{\lambda} = \frac{200}{10} = 20 \,\mathrm{days}$$

Exercise

A restaurant receives on average 1200 customers per day (from 10am to 10pm).

During peak times (12pm to 3pm, and 6pm to 9pm) the restaurant receives around 900 customers and, on average, 90 customers can be found in the restaurant at a given time.

At non-peak times, the restaurant receives 300 customers in total and, on average, 30 customers can be found in the restaurant at a given time.

What are the average times that a customer spends in the restaurant during peak/non-peak times?

Exercise (continued)

The maximum capacity of the restaurant is sometimes reached during peak times.

The restaurant manager expects that the number of customers during peak times will increase slightly in the coming months.

What action can be taken to address this issue without investing in extending the building?

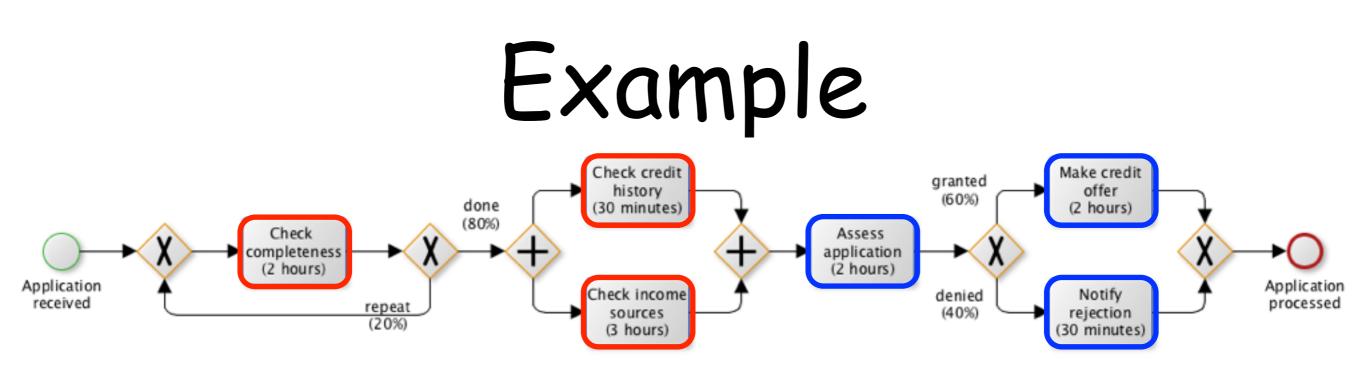
Cost analysis

Cost analysis

Analogously to the case of cycle time computation, flow analysis can be used to calculate other performance measures.

If we know the average cost of each activity, then we can calculate the average cost of the process more or less as we have just seen.

In fact the formulas for sequences, XOR-blocks and reworks are the same, but for AND-blocks we need to take the sum (instead of max)

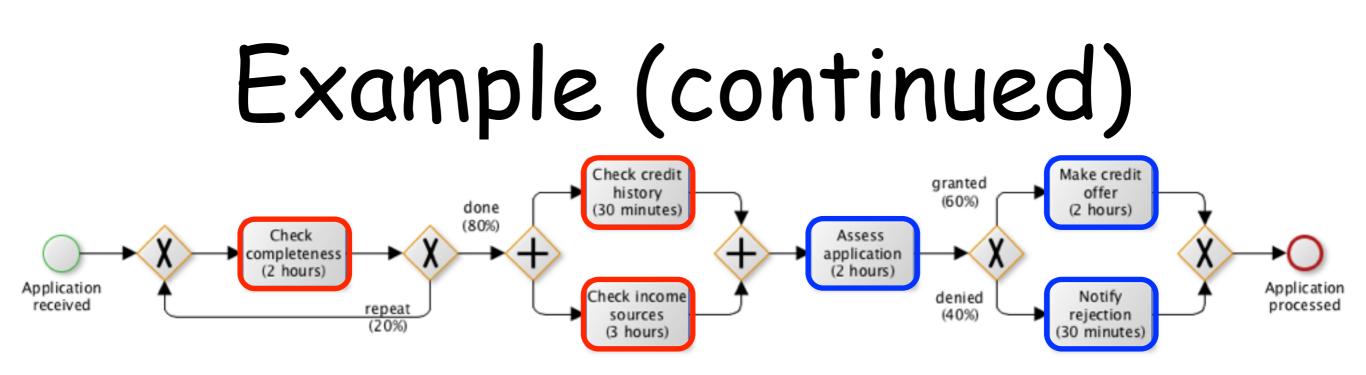


Assume

activities are annotated with processing time, "red" activities are performed by a clerk (hourly cost 25€), while "blue" activities by a credit officer (hourly cost 50€). Assume also that

the bank is charged **1**€ for each ``credit history check".

What is the average cost of the process?

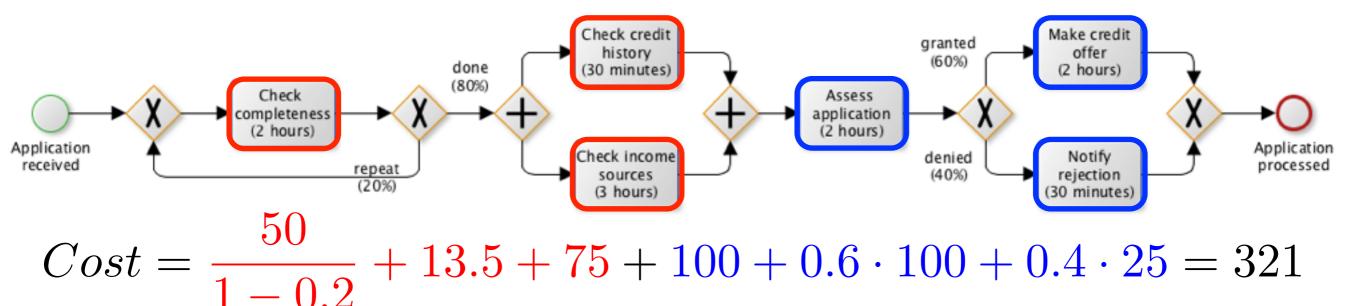


We can distinguish two kinds of costs:

human resource costs: can be calculated as the product of the (hourly) cost and the processing time of the task

other costs: fixed costs that are incurred by an execution of a task (not related to the time spent by human resources)

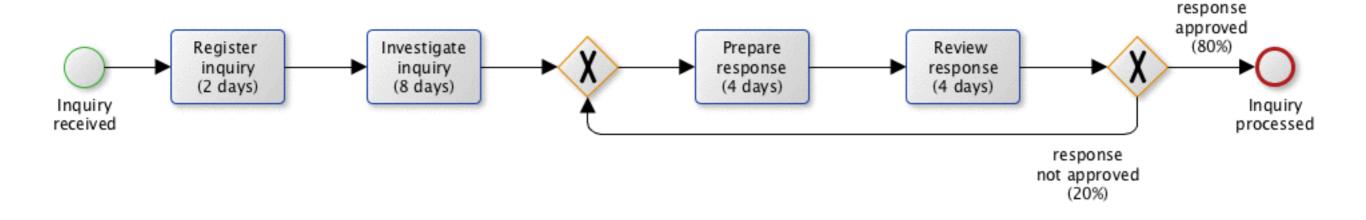
Example (continued)



Activity	Resource cost	Other cost	Total cost
Check completeness	$2 \cdot 25 = 50$		50
Check credit history	$0.5 \cdot 25 = 12.5$	1	13.5
Check income resources	$3 \cdot 25 = 75$		75
Assess application	$2 \cdot 50 = 100$		100
Make credit offer	$2 \cdot 50 = 100$		100
Notify rejection	$0.5 \cdot 50 = 25$		25

Exercise

Compute the cycle time efficiency and total cost of the process below assuming a working day last 8 hours



Activity	Resource	Resource hourly cost	Processing time
Register inquiry	clerk	$25 \mathrm{euros}$	30 minutes
Investigate inquiry	advisor	$50 { m euros}$	$12 {\rm hours}$
Prepare response	senior advisor	$75 { m euros}$	4 hours
Review response	counselor	100 euros	2 hours

Resource allocation

Resource management

In a process we can indicate: which tasks need to be performed, the order in which they must be carried out, who should do it

The way in which work items are allocated to resources (people, machines) is very important to the efficiency and effectiveness of the workflow

Resources

The basic characteristic of a resource is that it is able to carry out particular tasks

We assume each resource is uniquely identifiable and has capacity one, i.e., each resource may be working on no more than one activity at any given time

Resource classification

A resource is permitted to carry out a number of tasks (e.g., in a bank, a teller is not allowed to grant a mortgage)

A task can be performed only by certain resources

Which resources are able to carry out which task? It is impracticable to indicate them one by one in the process (staff can change)

We classify them using **resource classes** 1) classification based on functional properties (roles) 2) classification based on the position in the organization (groups or organizational units)

Roles (skill, competence, qualification): counter-staff, travel-agent, assessor, printer, administrator, chief-executive, senior-doctor

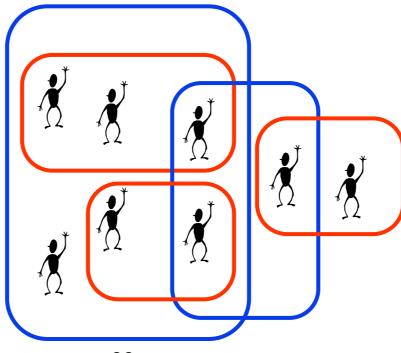
> Groups (department, office, team): sales-department, purchasing-department, development-team, Avana-branch

The same resource can belong to more roles and groups

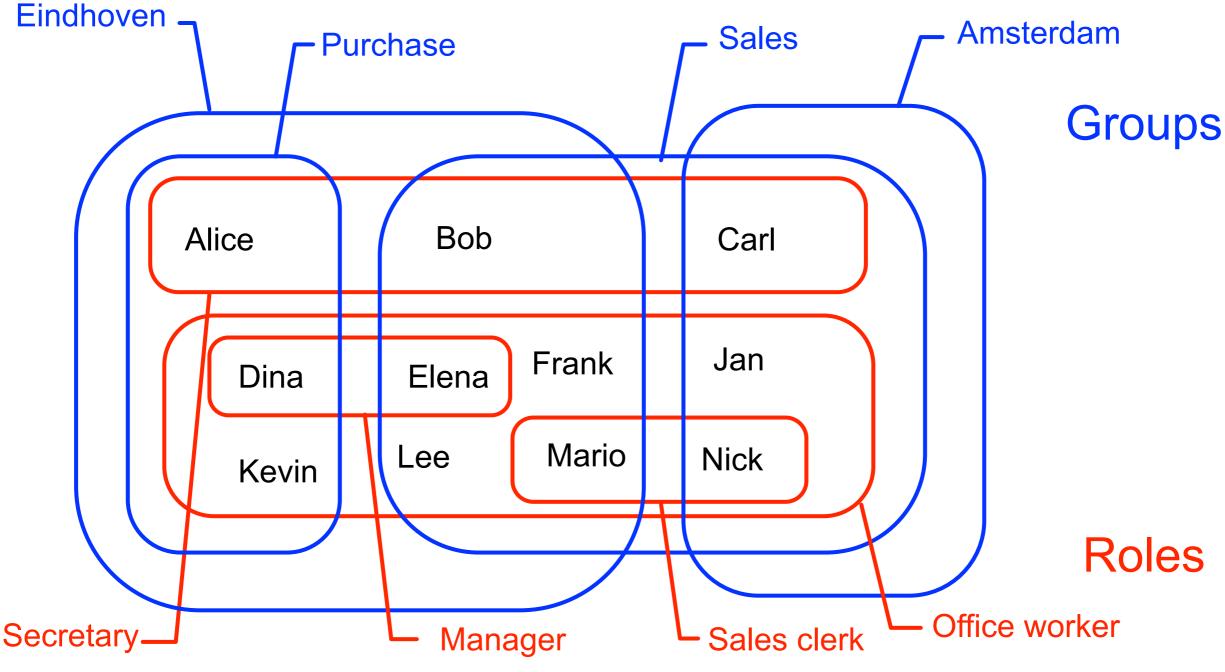
Classification diagram

A classification diagram shows which resources belong to each class and group

We can list resource ids and enclose them in vertical blue boxes that represent groups and in horizontal red boxes that represent roles



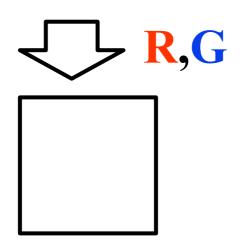
Classification diagram: example

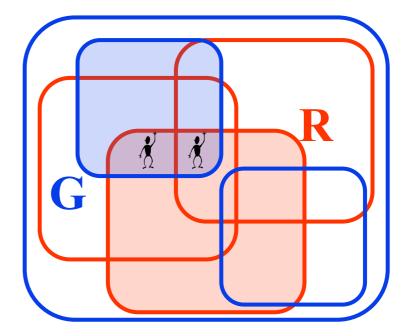


Resource management rules

Resource management rules: tell how to map work onto resources

Each task executed by a resource is labelled with one role and one group (allowed resources must lie in the intersection)





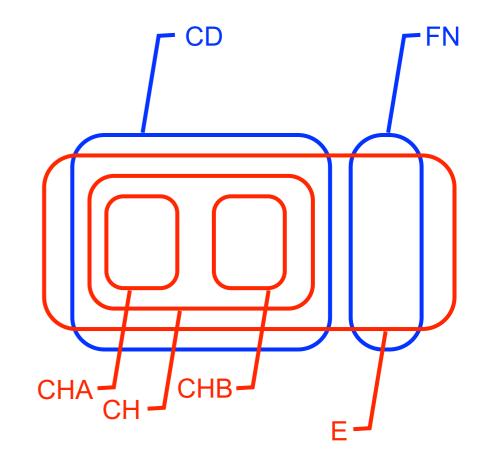
- An insurance company uses the following procedure for the processing of the claims
- Every claim, reported by a customer, is registered by an employee of department CD
- After the registration, the claim is classified by a <u>claim handler (CH)</u> of <u>rank A or B</u> within <u>CD</u>.
- There are two categories: **simple** and **complex** claims.
 - For simple claims two tasks need to be executed: check insurance and phone garage. These tasks are *independent* of each other.
 - The complex claims require three tasks: check insurance, check damage history and phone garage. These tasks need to be executed sequentially in the order specified.
- Both for the simple and complex claims, the tasks are done by <u>employees</u> of <u>department CD</u>. After executing the two/three tasks a **decision** is made by a <u>claim</u> <u>handler of rank A</u> and has two possible outcomes: **OK** (positive) or **NOK** (negative).
- If the decision is positive, then insurance company will pay.
 An <u>employee</u> of the <u>finance department</u> handles the payment.
- In any event, the insurance company sends a letter to the customer. An <u>employee</u> of the <u>department CD</u> writes this letter.

The following roles are identified:

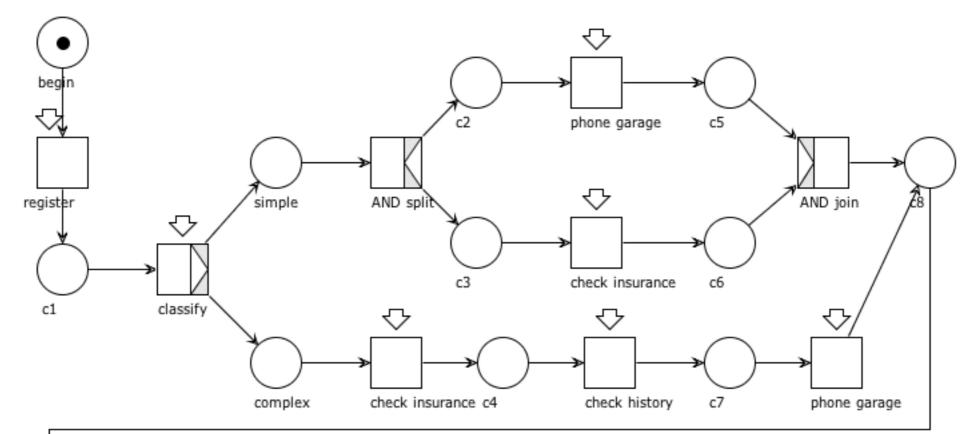
- *Employee* (E)
- Claim handler (CH)
- Claim handler A (CHA)
- Claim handler B (CHB)

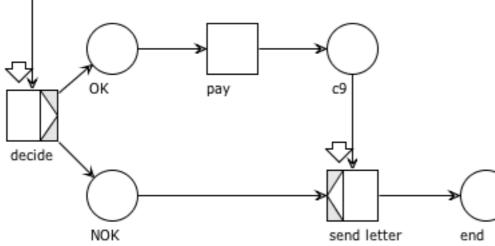
The following groups are identified:

- Car Damages Department (CD)
- Finance Department



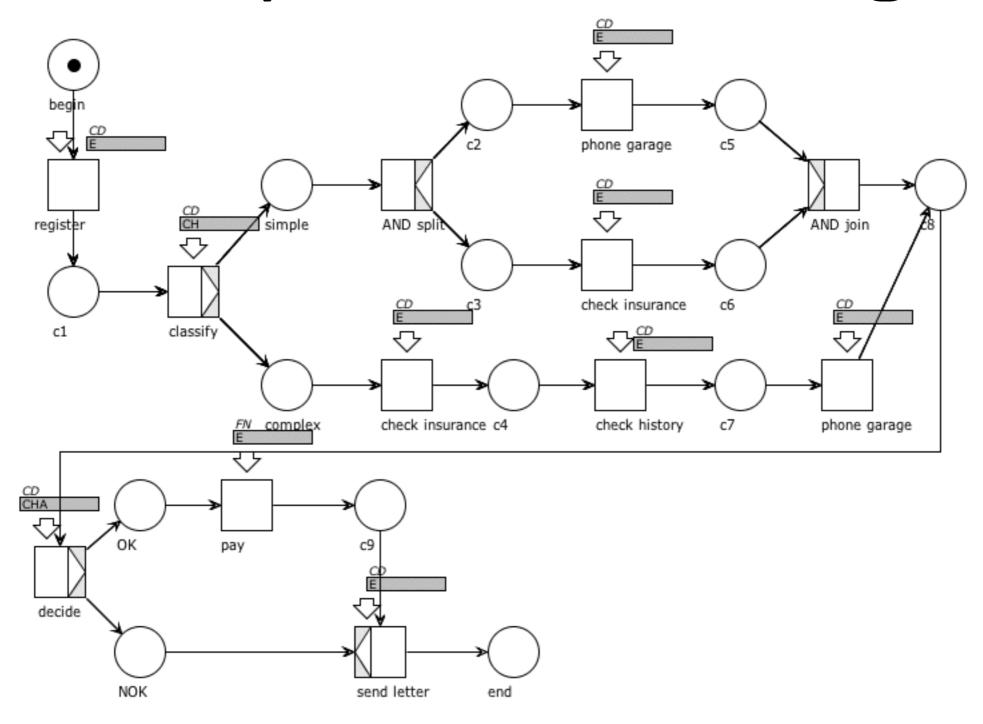
(FN)





	00	Untitled – 0
(Process Resources BPEL preview
	Objects Unassigned Assigned	Roles Roles </th

Zoom: 100%



Capacity plan

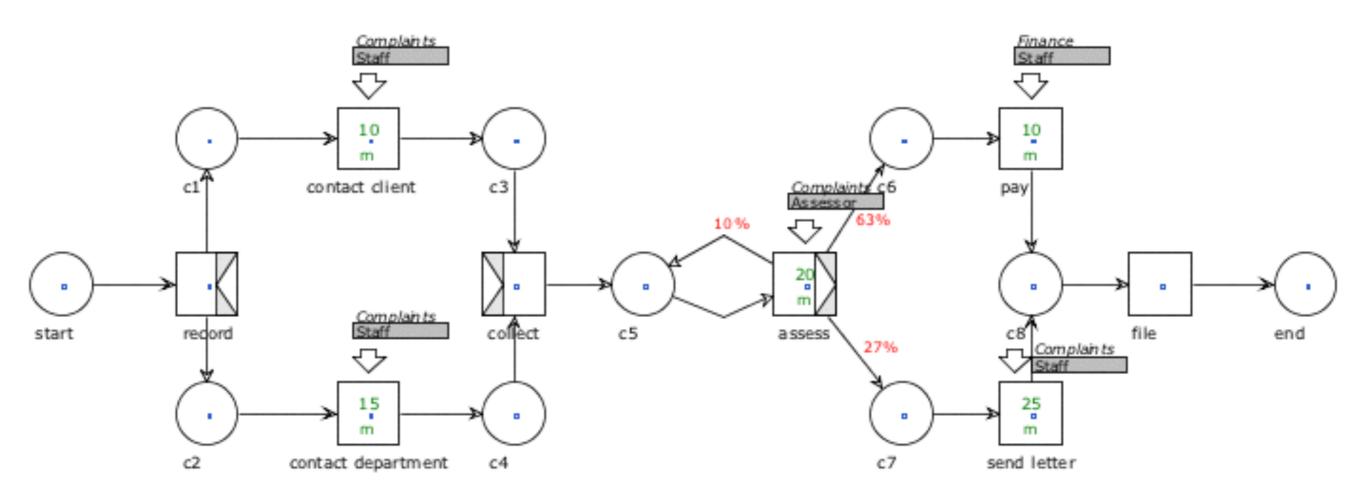
The staff may vary according to seasonal factors (vacation, illness, leaving the company)

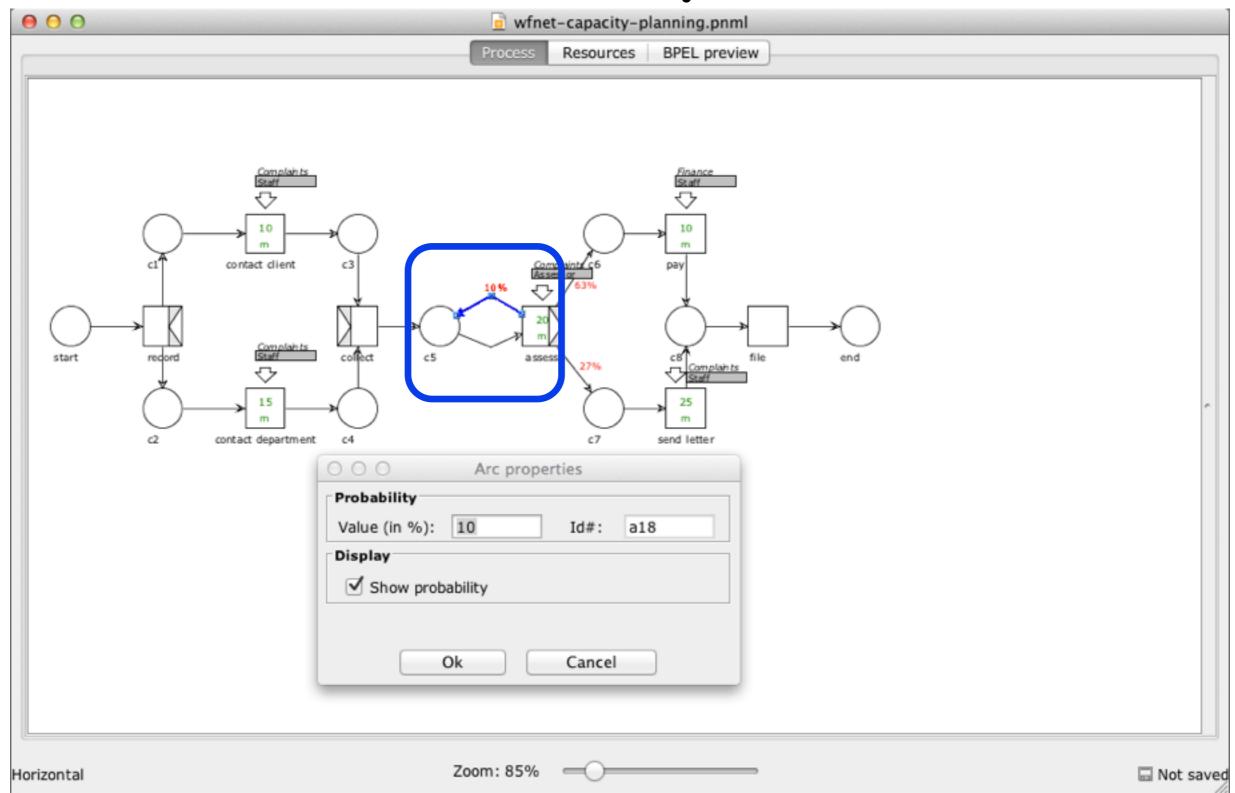
The capacity plan shows what resources are needed for a period of time

If we have a forecast of the supply of new cases, then the capacity requirement can be estimated

(wfnet-capacity-planning.pnml
1		Process Resources BPEL preview
	Choose view Choose view Graphical view Objects Unassigned Assigned Choose view Choose vie	
		Compound roles Compound groups Staff

Zoom: 85% -





00	🖻 wfnet-capacity-	000	Transition p	properties		
	Process Resources		General	BPEL		
		Identification				_
		Name:	assess	Id#:	t5	
		Branching				
Complain ts Staff		O None	AND-split		O AND-join	
	\frown		 XOR-split 		XOR-join	
$() \longrightarrow \mathbb{I}_{m}^{10} \longrightarrow \mathbb{I}_{m}^{10}$			🔘 XOR-join-split	80	 AND-join-split 	
c1 contact client	c3		O AND-join-XOR-split	88	🔘 XOR-join-AND-split	88
		Trigger				
		 Automatic 	Resource	,	₽	
start redord Staff	collect c5 assess 27%	Message	🖂 🗌 🖂 Time	(Ð	
	\wedge	Orientation				
$() \longrightarrow []{15} \\ m \\ m \\ m \\ \end{pmatrix} \longrightarrow ()$	\odot		() N	orth		
c2 contact department	c4 c7		🔾 West 😽	€) ⊙ Ea	ist	
			⊖ so	buth		
		Service time				
		Average:	20	n	ninute(s)	\$
		Resource mappir	ng			
		Role:	Assessor ‡ G	Group:	Complaints	\$
		No. of assigned r	resource objects:		0	
					Ľ	
			Ok	Cance	2	
rizontal	Zoom: 85%				🗔 Not saved	

000		Cap	acity planning			
Parameters Observation period: Arrivals per period (λ Loop termination thr	reshold (ε): 0.001	hour(s)	\$		_	Compute
Capacity requirements	per task	Tir	ne unit: 1.0 minute	(s) Floating poin	t precision:	1.00
Task	Service time	Items/case	Time/case	Items/period	Time/period	Group / Role
contact client (t2)	10.00	1.00	10.00	50.00		nplaints/Staff
contact departmen	15.00	1.00	15.00	50.00		nplaints/Staff
assess (t5)	20.00	1.11	22.22	55.56	1111.10 Con	nplaints/Assessor
pay (t6)	10.00	0.70	7.00	35.00	350.00 Fina	ance/Staff
send letter (t7)	25.00	0.30	7.50	15.00	375.00 Con	nplaints/Staff
Whole process			61.72		3086.09	
Capacity requirements	per resource class			Average resource	utilization:	80 🕨
Resource class	per resource class		Aggre	gate time		resource objects
Resource class Employee	per resource class		Aggre	gate time 0.00		resource objects 0.00
Resource class Employee Finance	per resource class		Aggre	gate time 0.00 350.00		resource objects 0.00 0.91
Resource class Employee Finance Staff	per resource class		Aggre	gate time 0.00 350.00 1974.99		resource objects 0.00 0.91 5.14
Resource class Employee Finance	per resource class		Aggre	gate time 0.00 350.00		resource objects

Limitation of flow analysis

Pitfalls and limitation

The equations we have presented deals with *structured* process only: we cannot calculate the cycle time for any processes.

The equations exploit the average cycle time of activities: we need to estimate such values (interviewing stakeholders, inspecting logs)

Flow analysis does not account for the fact that a process can behave differently depending on the load: when the load goes up and the resources are constant, the waiting time increases.

Resource contention

This last phenomenon is called **resource contention**: there is more work to be done than resources available to perform the work

Flow analysis does not take into account the effects of increased resource contention. The estimates obtained from flow analysis are only applicable if the level of resource contention remains relatively stable over the long-run