MTAT.03.231
Business Process Management (BPM)
(for Masters of IT)

Lecture 4: Quantitative Process Analysis

Marlon Dumas

marlon.dumas ät ut . ee
Once I’ve got a model, what’s next?

Analyze:
- Cycle time
- Capacity, resource utilization
- Cost (Activity-Based Costing)
- QoS
- Risk
- ...

Business Process Analysis Techniques

• Qualitative analysis
  – Step-by-Step Animation
  – Cause-Effect-Analysis

• Quantitative Analysis
  – Cycle Time Analysis
  – Capacity Analysis
  – Queuing Theory
  – Process Simulation
  – Markovian Analysis
Process Throughput

• Inflow and Outflow rates typically vary over time
  – $\text{IN}(t) =$ Arrival/Inflow rate of jobs at time $t$
  – $\text{OUT}(t) =$ Departure/Outflow rate of finished jobs at time $t$
  – $\text{IN} =$ Average inflow rate per time unit
  – $\text{OUT} =$ Average outflow rate per time unit

• A stable system must have $\text{IN}=\text{OUT}=\lambda$
  – $\lambda =$ the process flow rate
  – $\lambda =$ process throughput
Work-In-Process

• Jobs that have entered the process but not yet left it
• A long lasting trend in manufacturing has been to lower WIP by reducing batch sizes
  – The JIT philosophy
  – Forces reduction in set up times and set up costs
• WIP = Average work in process over time
Cycle Time

- Cycle time: Difference between a job’s start time and end time
- Little’s Formula: General relationship between the average WIP, the throughput ($\lambda$) and Cycle time (CT)

$$\text{Little's Formula: WIP} = \lambda \cdot \text{CT}$$

- Implications, everything else equal
  - Shorter cycle time $\iff$ lower WIP
  - If $\lambda$ increases $\Rightarrow$ to keep WIP at current levels CT must be reduced
Exercise 1

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 peak times?
2. What is the average time that a customer spends in the restaurant during non-peak 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 Analysis

- *Cycle time analysis*: the task of calculating the *average* cycle time for an entire process or process segment
  - Assumes that the average activity times for all involved activities are available (activity time = waiting time + processing time)
- In the simplest case a process consists of a sequence of activities on a single path
  - The average cycle time is the sum of the average activity times
- … but in general we must be able to account for
  - Rework
  - Multiple paths
  - Parallel activities
Rework

- Many processes include control or inspection points where if the job does not meet certain standard, it is sent back for rework
  - The rework will affect the average cycle time
- Definitions
  - $T = \text{sum of activity times in the rework loop}$
  - $r = \text{percentage of jobs requiring rework (rejection rate)}$
- Assuming a job is never reworked more than once
  \[ CT = (1+r)T \]
- Assuming a reworked job is no different than a regular job
  \[ CT = \frac{T}{1-r} \]
Example – Rework effects on the average cycle time

• Consider a process consisting of
  – Three activities, A, B & C taking on average 10 min. each
  – One inspection activity (I) taking 4 minutes to complete.
  – X% of the jobs are rejected at inspection and sent for rework

What is the average cycle time?

a) If no jobs are rejected and sent for rework.

b) If 25% of the jobs need rework but never more than once.

c) If 25% of the jobs need rework but reworked jobs are no different in quality than ordinary jobs.
It is common that there are alternative routes through the process
  - For example: jobs can be split in “fast track” and normal jobs

Assume that \( m \) different paths originate from a decision point
  - \( p_i \) = The probability that a job is routed to path \( i \)
  - \( T_i \) = The time to go down path \( i \)

\[
CT = p_1T_1 + p_2T_2 + \ldots + p_mT_m = \sum_{i=1}^{m} p_i T_i
\]
Example – Processes with Multiple Paths

- Consider a process segment consisting of 3 activities A, B & C with activity times 10, 15 & 20 minutes respectively.
- On average 20% of the jobs are routed via B and 80% go straight to activity C.

What is the average cycle time?
Processes with Parallel Activities

• If two activities related to the same job are done in parallel the contribution to the cycle time for the job is the maximum of the two activity times.

• Assuming
  – M process segments in parallel
  – $T_i =$ Average process time for process segment i to be completed

\[ CT_{\text{parallel}} = \text{Max}\{T_1, T_2, \ldots, T_M\} \]
Consider a process segment with 5 activities A, B, C, D & E with average activity times: 12, 14, 20, 18 & 15 minutes.

What is the average cycle time for the process segment?
**Cycle Time Efficiency**

- Measured as the percentage of the total cycle time spent on value adding activities.

\[
\text{Cycle Time Efficiency} = \frac{\text{Theoretical Cycle Time}}{\text{CT}}
\]

- Theoretical Cycle Time = the cycle time which we would have if only value adding activities were performed
  - That is if the activity times, which include waiting times, are replaced by the processing times
Cycle time Reduction

- Cycle time analysis provides valuable information about process performance
  - Helps to quantify efficiency problems and bottlenecks
  - Useful for assessing the effect of design changes
- Ways of reducing cycle times through process redesign
  1. Eliminate activities
  2. Reduce waiting time
  3. Eliminate or reduce rework
  4. Perform activities in parallel
  5. Move processing time to activities not on the critical path
Example – Critical Activity Reduction

• Consider a process with three sequences or paths

<table>
<thead>
<tr>
<th>Sequence (Path)</th>
<th>Time required (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. A→B→E</td>
<td>12+14+15 = 41</td>
</tr>
<tr>
<td>2. A→C→E</td>
<td>12+20+15 = 47 = CT</td>
</tr>
<tr>
<td>3. A→D→E</td>
<td>12+18+15 = 45</td>
</tr>
</tbody>
</table>

⇒ By moving 2 minutes of activity time from path 2 to path 1 the cycle time is reduced by 2 minutes to CT=45 minutes
Generalization: Time, Cost, and Reliability Analysis of Structured Models

\[
\left\langle \sum_i T_{SC_i}, \sum_i C_{SC_i}, \prod_i R_{SC_i} \right\rangle
\]

\[
\left\langle \sum_i p_i \cdot T_{SC_i}, \sum_i p_i \cdot C_{SC_i}, \sum_i p_i \cdot R_{SC_i} \right\rangle
\]

\[
\left\langle \max\{T_{SC_i}\}, \sum_i C_{SC_i}, \prod_i R_{SC_i} \right\rangle
\]

\[
\left\langle (1 - p)^{-1} \cdot T_{SC}, (1 - p)^{-1} \cdot C_{SC}, R_{SC}^{(1-p)^{-1}} \right\rangle
\]
Limitation: Not all Models are Structured
Capacity Analysis

• Focus on assessing the capacity needs and resource utilization in the process
  1. Determine the number of jobs flowing through different tasks (*flow rate*)
  2. Determine capacity requirements and utilization based on the flow rates obtained in 1.
  3. Determine bottlenecks

• Complements cycle time analysis…
Computing Flow Rate (1)

Exclusive choice and flow rates

- The flow rate along a certain path depends on
  - The number of jobs entering the process as a whole \((n)\)
  - The probability for a job to go along a certain path

- Defining
  - \(N_i\) = number of jobs taking path \(i\)
  - \(p_i\) = Probability that a job goes along path \(i\)

\[ N_i = n \cdot p_i \]

Parallel Activities and flow rates

- All jobs still have to go through all activities
  - if they are in parallel or sequential does not affect the number of jobs flowing through a particular activity
Computing Flow Rate (2)

- A rework loop implies an increase of the flow rate for that process segment
- Definitions
  - $N =$ Number of jobs flowing through the rework loop
  - $n =$ Number of jobs arriving to the rework loop from other parts of the process
  - $r =$ Probability that a job needs rework
- Assuming a job is never reworked more than once

\[ N = (1+r)n \]

- Assuming a reworked job is no different than a regular job

\[ N = \frac{n}{1-r} \]
Example – Flow Rate of Rework Loop

\[ N = (1 + r)n = (1 + 0.25)100 = 125 \]
Analyzing Capacity Needs and Utilization

Need to know

- Processing times for all activities
- The type of resource required to perform the activity
- The number of jobs flowing through each activity (flow rate)
- The number of available resources of each type

Step 1 – Calculate unit load for each resource

- The total resource time required to process one job
  - $N_i =$ Number of jobs flowing through activity $i$ for every new job entering the process
  - $T_i =$ The processing time for activity $i$ in the current resource
  - $M =$ Total number of activities using the resource

Unit load for resource $j = \sum_{i=1}^{M} N_i \cdot T_i$
Analyzing Capacity Needs and Utilization

Step 2 – Calculate the “unit capacity”
- The number of processed jobs per time unit

Unit capacity for resource j = 1/Unit load for resource j

Step 3 – Determine the resource pool capacity
- A resource pool is a set of identical resources available for use
- Pool capacity is the number of jobs per time unit that can be processed
  - Let M = Number of resources in the pool

Pool capacity = M·Unit capacity = M/unit load
Process Capacity and Capacity Utilization

• The process capacity is determined by the bottleneck
  – The bottleneck is the resource or resource pool with the smallest capacity (the slowest resource in terms of jobs/time unit)
  – The slowest resource will limit the throughput

Capacity Utilization

• The theoretical process capacity is obtained by focusing on processing times as opposed to activity times
  – Delays and waiting times are disregarded
  ⇒ The actual throughput ≤ The theoretical capacity!

\[
\text{Capacity Utilization} = \frac{\text{Actual Throughput}}{\text{Theoretical Process Capacity}}
\]
Limitations of Cycle Time/Capacity Analysis

- Cycle time analysis and capacity do not consider waiting times due to resource contention
- Queuing analysis and simulation address these limitations and have a broader applicability
Queuing Theory: Notation

- State of the system = number of customers in the system
- Queue length = (state of the system) – (number of customers being served)

\[ \lambda = \text{Average arrival intensity (= # arrivals per time unit)} \]

\[ \mu = \text{Average service intensity for the system} \]

\[ \rho = \text{Utilization factor = The expected fraction of time that the service facility is being used} \]
Why is Queuing Analysis Important?

• Capacity problems are very common in industry and one of the main drivers 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 of 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 ⇒ longer waiting times

➤ Question: Should another MD position be instated?
Probability Distributions: Uniform
Probability Distributions: Normal

\[ \mu = 0, \sigma^2 = 0.2 \]
\[ \mu = 0, \sigma^2 = 1.0 \]
\[ \mu = 0, \sigma^2 = 5.0 \]
\[ \mu = -2, \sigma^2 = 0.5 \]
Probability Distributions: 
Negative Exponential
Queuing theory: basic concepts

Basic characteristics:

- average number of arrivals per time unit: \( \lambda \) (mean arrival rate)
- average number that can be handled by one server per time unit: \( \mu \) (mean service rate)
- number of servers: \( c \)
Queuing theory concepts (cont.)

Given $\lambda$, $\mu$ and $c$, we can calculate:

- occupation rate: $\rho$
- $W_q =$ average time in queue
- $W =$ average system in system (i.e. cycle time)
- $L_q =$ average number in queue (i.e. length of queue)
- $L =$ average number in system average (i.e. Work-in-Progress)
M/M/1 queue

Assumptions:
- time between arrivals and service time follow a negative exponential distribution
- 1 server (c = 1)
- FIFO

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

\[ L = \frac{\rho}{1 - \rho} \]
\[ W = L/\lambda = 1/(\mu - \lambda) \]
\[ L_q = \frac{\rho^2}{(1 - \rho)} = L - \rho \]
\[ W_q = L_q/\lambda = \frac{\lambda}{(\mu - \lambda)} \]
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*  \( \Rightarrow \ W_q = \frac{L_q}{\lambda} \)

\[
W = W_q + \frac{1}{\mu}
\]

*Little’s Formula*  \( \Rightarrow \ L = \lambda W = \lambda (W_q + 1/\mu) = L_q + \frac{\lambda}{\mu} \)
Example – ER at County Hospital

- **Situation**
  - Patients arrive according to a Poisson process with intensity $\lambda$ ($\Leftrightarrow$ the time between arrivals is $\exp(\lambda)$ distributed).
  - The service time (the doctor’s examination and treatment time of a patient) follows an exponential distribution with mean $1/\mu$ (= $\exp(\mu)$ distributed)
  - **The ER can be modeled as an M/M/c system where c=the number of doctors**

- **Data gathering**
  - $\lambda = 2$ patients per hour
  - $\mu = 3$ patients per hour

- **Questions**
  - Should the capacity be increased from 1 to 2 doctors?
  - How are the characteristics of the system ($\rho$, $W_q$, $W$, $L_q$ and $L$) affected by an increase in service capacity?
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)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>One doctor (c=1)</th>
<th>Two Doctors (c=2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>2/3</td>
<td>1/3</td>
</tr>
<tr>
<td>( L_q )</td>
<td>4/3 patients</td>
<td>1/12 patients</td>
</tr>
<tr>
<td>( L )</td>
<td>2 patients</td>
<td>3/4 patients</td>
</tr>
<tr>
<td>( W_q )</td>
<td>2/3 h = 40 minutes</td>
<td>1/24 h = 2.5 minutes</td>
</tr>
<tr>
<td>( W )</td>
<td>1 h</td>
<td>3/8 h = 22.5 minutes</td>
</tr>
</tbody>
</table>

- **Is it warranted to hire a second doctor?**
Process Simulation

- **Drawbacks of queuing theory:**
  - Generally not applicable when system includes parallel activities
  - Requires case-by-case mathematical analysis
  - Assumes “steady-state” (valid only for “long-term” analysis)

- **Process simulation is more versatile (also more popular)**

- **Process simulation = run a large number of process instances, gather data (cost, duration, resource usage) and calculate statistics from the output**

- **Simulation ≠ animation**
  - Simulation is a batch process, animation is interactive
  - Some tools allow one to animate while simulating, but in practice this is too slow!
Process Simulation

• Basic steps in evaluating a process model with simulation
  1. Building the simulation model
  2. Running the simulation
  3. Analyzing the simulation results (performance measure)
  4. Evaluation of alternative scenarios
Elements of a simulation model

- The process model including:
  - Activities, control-flow relations (flows, gateways)
  - Resources and resource pools (i.e. roles)
- Resource requirements: mapping between activities and resource pools
- Processing times (per activity, or per activity-resource pair)
- Costs (per activity, or per activity-resource pair)
- Arrival rate (also called: token creation)
- Conditional branching probabilities (XOR gateways)
Simulation Example – BPMN model
Resource Pools (Roles)

• Two options to define resource pools
  – Define individual resources of type clerk
  – Or assign a number of “anonymous” resources all with the same cost
• E.g.
  – 3 anonymous clerks with cost of € 10 per hour, 8 hours per day
  – 2 individually named clerks
    • Jim: € 12, 4 hours per day
    • Mike: € 14, 8 hours per day
  – 1 manager John at € 20 per hour, 8 hours per day
## Resource pools and execution times

<table>
<thead>
<tr>
<th>Task</th>
<th>Role</th>
<th>Execution Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Normal distribution: mean and std deviation</td>
</tr>
<tr>
<td>Receive application</td>
<td>system</td>
<td>0</td>
</tr>
<tr>
<td>Check completeness</td>
<td>Clerk</td>
<td>30 mins</td>
</tr>
<tr>
<td>Perform checks</td>
<td>Clerk</td>
<td>2 hours</td>
</tr>
<tr>
<td>Request info</td>
<td>system</td>
<td>1 min</td>
</tr>
<tr>
<td>Receive info (Event)</td>
<td>system</td>
<td>48 hours</td>
</tr>
<tr>
<td>Make decision</td>
<td>Manager</td>
<td>1 hour</td>
</tr>
<tr>
<td>Notify rejection</td>
<td>system</td>
<td>1 min</td>
</tr>
<tr>
<td>Time out (Time)</td>
<td>system</td>
<td>72 hours</td>
</tr>
<tr>
<td>Receive review request (Event)</td>
<td>system</td>
<td>48 hours</td>
</tr>
<tr>
<td>Notify acceptance</td>
<td>system</td>
<td>1 min</td>
</tr>
<tr>
<td>Deliver Credit card</td>
<td>system</td>
<td>1 hour</td>
</tr>
</tbody>
</table>

**Alternative: assign execution times to the tasks only (like in cycle time analysis)**
Arrival rate and branching probabilities

10 applications per hour (one at a time)
Poisson arrival process (negative exponential)

Alternative: instead of branching probabilities one can assign “conditional expressions” to the branches based on input data
Simulation output: KPIs

Resource Cost

$4,260.95

Cycle Time - Histogram

Days

# PIs

0 10 20 30 40 50 60

0 2 4 6 8 10 12

0 500.00 1,000.00 1,500.00 2,000.00 2,500.00 3,000.00 3,500.00 4,000.00 4,500.00
## Simulation output: detailed logs

<table>
<thead>
<tr>
<th>Process Instance</th>
<th># Activities</th>
<th>Start</th>
<th>End</th>
<th>Cycle Time</th>
<th>Cycle Time (s)</th>
<th>Total Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>5</td>
<td>4/06/2007 13:00</td>
<td>4/06/2007 16:26</td>
<td>03:26:44</td>
<td>12403.586</td>
<td>03:26:44</td>
</tr>
</tbody>
</table>

### Activity Details

<table>
<thead>
<tr>
<th>Process Instance</th>
<th>Activity ID</th>
<th>Activity Name</th>
<th>Activity Type</th>
<th>Resource</th>
<th>Start</th>
<th>End</th>
<th>Cycle Time</th>
<th>Cycle Time (s)</th>
<th>Total Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>7aed54717-f044-4da1-b543-82a960809ecb</td>
<td>Check for completeness</td>
<td>Task</td>
<td>Manager</td>
<td>4/06/2007 14:00</td>
<td>4/06/2007 14:31</td>
<td>18:14:56</td>
<td>65695.612</td>
<td>18:14:56</td>
</tr>
</tbody>
</table>