5th Seminar – Performance modelling

Dimensional analysis. When solving performance modelling exercises it's recommended to perform dimensional analysis[1] known from physics. Let's look at the kinematics equation $s = v_0 t + \frac{a}{2} t^2$

With dimensions: $s[\mathrm{m}] = v_0 \left[\frac{\mathrm{m}}{\mathrm{s}}\right] t[\mathrm{s}] + \frac{a}{2} \left[\frac{\mathrm{m}}{\mathrm{s}^2}\right] t^2[\mathrm{s}^2] = v_0 t[\mathrm{m}] + \frac{a}{2} t^2[\mathrm{m}]$

The main motivation for dimensional analysis is to check whether the dimensions of certain values match or not. If not, then the used formula is probably incorrect.[2] Dimensional analysis often helps us to choose the correct formula. It's important to note that units like "piece", "request", etc. don't have their own dimensions, thus dimensions $\frac{\text{request}}{\mathrm{s}}$ and $\frac{1}{\mathrm{s}}$ are considered the same.

Basic formulas. Little's law: $N = X \cdot T$, $N\ [1] = X \left[\frac{1}{s}\right] \cdot T\ [s]$

Utilization derived intuitively from Little's law in case of *one exclusive resource instance*:
$U = \frac{X}{X^{\max}} = \frac{T_{\text{busy}}}{T_{\text{measured}}} = N = X \cdot T$

Maximum throughput (when the utilization is 100%) derived from processing time in case of *one exclusive resource instance*: $X = \frac{U}{T} \Rightarrow X^{\max} = \frac{1}{T}$

1 Disk's performance

A disk serves 50 requests per second. Each request takes 0,005 seconds to serve. There is no overlap in the system.

a. What is the maximal workload (arrival rate) that can be served?
b. What is the utilization of the system?

2 Inspecting midterm exam evaluation

During the inspection of the exam students have the opportunity to complain about accidental grading mistakes. The final exam result may be modified in case of a justifiable complaint. An exam grader can review 10 exercises in an hour in case of the first exercise (E1) or 20 for the second exercise (E2). Each exercise has its own dedicated exam grader who graded that specific exercise. For each of the following subtasks create a process model describing the given scenario and calculate the maximum throughput (reviewed students per hour) of the inspection process!

a. Students complain about the first, then the second exercise to the respective graders.
b. The students become resourceful and they give the two exercises to the two graders at the same time, since they were written on different sheets of paper. What are the effects of this parallelization?
c. Due to the long queue the students complain only about one of the exercises based on which one of the graders is free at the moment.
d. Word got around that the grader of the second exercise is less strict, thus 80% of students form a queue in front of the second grader. The remaining 20% of students goes to the grader of the first exercise.
e. 10% of students only need 1 or 2 points to get the better grade after reviewing their exam so they keep repeating the complaining process. Assume that the complaining process is the same as in subtask a.
f. What would be the difference, if both graders were ready to review both exercises (reviewing the individual exercises takes the same time as before), and the students were planning to complain first about the first, then the second exercise to a grader.

[1] Dimensional analysis (Wikipedia), `https://en.wikipedia.org/wiki/Dimensional_analysis`
[2] Recommended reading: what if? – Droppings, `http://what-if.xkcd.com/11/`

3 2-tier architecture

We have a webserver (WS) and a cluster of two database servers (DB1, DB2). We chose between the database servers using a weighted round robin load balancer with ratio 1:2. We use both kinds of resources to serve each request. In the peak period we monitor the system for 30 minutes, during which it serves 9000 requests. The measured busy times are the following: WS – 1350 s CPU time, DB1 – 810 s, DB2 – 1320 s disk IO time.

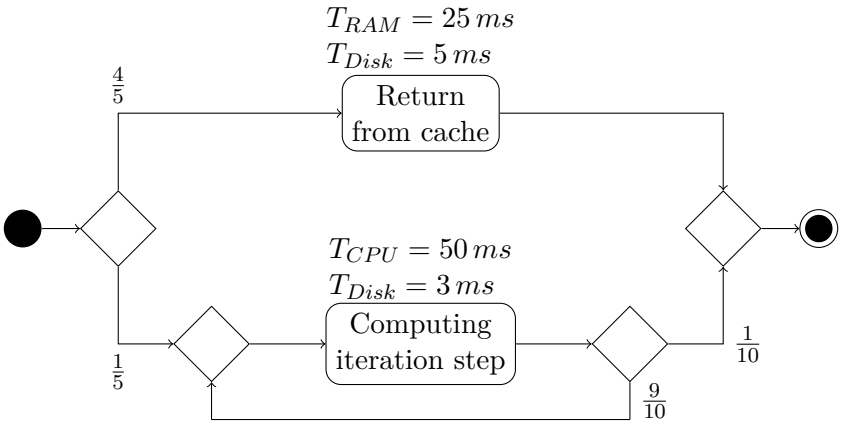
a. Create a process model based on the above text that models the processing of a request!
b. What is the current throughput of the servers?
c. How much time does it take for the servers to serve one request?
d. What is the maximum throughput of the system?
e. Why did we use different types of busy times for the two types of resources?
f. What kind of simplifications does the model make?

4 Microservice (∗ previous exam exercise)

Our microservice implements an approximation algorithm which other services can utilize over the network in order to achieve their goals. Many of the requests are identical, so our service caches their answers, enabling the possibility of returning previously calculated results substantially faster than requests yet unseen. The computation of the approximation is a resource-heavy iterative process, which must be repeated until the result is accurate enough.

The model below summarizes the experience gained during experimental runs. We observed that 80% of the requests can be served from the cache, and that after an iterative computational step the result's accuracy is satisfactory with a 10% chance. We also measured the average busy times of the resources for both activities. This is displayed above the activities on the figure below (only for the resources that were busy for a non-negligible amount of time).

We used the same server for the measurements as final product will. This server is equipped with 2 CPU-s, 1 RAM-module and 1 Disk.



a. If we consider only the resource used for the longest amount of time for each activity, then what is the maximum throughput of the system?
b. If we consider only the one resource (Disk) that was used substantially by both activities, then what is maximum throughput of the system?
c. Based on this, what is the actual maximum throughput of the system? Which resource could be scaled up to further increase it?
d. If the utilization of the system is 50% and typically 100ms elapse between the arrival of a request and the return of its answer (counting time spent waiting), then how many yet unprocessed requests are in the system on average?

5 Island's traffic network (∗ previous exam exercise)

The habitants of an island cross the lake around the island every day when they go to work. They can go north on a bridge or south by car ferry. The (in each direction) single-lane bridge is 200 meters long, the speed limit is 60 $\frac{km}{h}$, and the safety distance (tailgating) is 30 meters from tail light to tail light. There are four ferryboats. Each boat takes the island-land-island trip in 15 minutes, and thus together they take at most 800 cars to the land per hour.

a. What is the maximum throughput of the bridge (northwards)?

b. How many cars can a ferryboat take?
c. What is the combined maximum throughput of the two routes leaving the island in the morning rush hours?
d. If the highway on the land gets closed due to an accident, and the traffic is diverted through the island (over the bridge, then by ferry), what is the maximum throughput of the diverted path?
e. One morning 900 cars left the island by ferry between 7:00 and 8:30. What was the throughput and the utilization of the ferries in this period?
f. In the above scenario how many cars on average were waiting on the coast at the same time, if the cars arrived to the port in a well-timed manner, on average half minutes before getting in?

6 Knowledge base (∗)

Our company's public professional knowledge base offers articles that may reference each other. On average the server takes 60 ms serving an article request. After reading an article, the reader only leaves the page 30% of the time; most of the cases, they click on a reference to another article.

a. How much server time is needed on average to satisfy a reader's total thirst for knowledge?
b. Assume that the requests can't be parallelized. How many unique users can the server serve per hour?