

17-423/723 Homework 2: Design for Scalability

Spring 2025

Released: April 14, 2025

Due: April 25, 2025 11:59 pm (on Gradescope)

Estimated Time Effort: 3 to 4 hours

Learning Objectives:

- Specify scalability requirements for a software system.
- Apply design patterns and techniques to improve the scalability of a system.
- Evaluate design alternatives and their trade-offs between scalability and other quality attributes.
- Apply an appropriate abstraction to describe design decisions for scalability.

NOTE: Although the system under study is the scheduling system that you are working on for the course project, this is an **individual homework assignment**. You are free (and encouraged) to discuss the assignment with your team members or classmates, but the final submitted solution must be your own.

Tasks

For the below questions, consider the scheduling system that you and your classmates have been developing throughout the semester. The system under consideration includes not only the scheduling application, but all of the shared services (i.e. public information service, central DB, health administrator service, and policy maker service).

(Optional) Q1. Identify scalability bottlenecks (1 page max): Use a load testing tool like locust.io ([quick start guide](#)) to identify performance bottleneck(s) on your scheduling app or shared service. In the 1-page description, provide a brief description of how you performed the load testing, what metrics you used, and the results of the load testing (i.e., the identified bottleneck). If available, include a link to any script that you used for the testing.

Note: Please coordinate with the rest of your team to avoid multiple people performing load testing on the same endpoint simultaneously; otherwise, the bottleneck that you've identified might not be accurate.

Q2. Specify scalability requirements (0.5 page max): Specify **two** scalability requirements: One for the scheduling application that interacts with end users (i.e., patients) and another one for one of the shared services in the system.

Recall that a scalability requirement includes a description of **workload** and an acceptable level of **performance**. To describe both the workload and performance, use **quantitative** metrics that you consider to be reasonable for the system. Briefly justify why those metrics are reasonable; you may refer to an online source as part of the justification.

If you've done the load testing in Q1, describe how your scalability requirements are informed by the identified bottleneck(s).

Q3. Evaluate alternative design decisions (2 page max): Suppose that you are tasked with redesigning the existing system to satisfy the scalability requirements specified above. For this new design, please consider the following list of design decisions that are relevant for designing scalable systems:

- What data do we need to store for system operations?
- What data storage model (relational vs. document) should be used for which data?
- What type of scaling (vertical, horizontal, or both) do we apply?
- How do we distribute data (replication, partitioning, or both)?
- If replication is used, synchronous vs. asynchronous replication?
- Which data do we cache? Where is the cache stored?

Provide your decisions for **all** of the above questions. For **three** of the above questions, also include a discussion of an alternative design decision that was considered and justification for why you settled on your final decision. When justifying your decision, we encourage you to consider tradeoffs with other quality attributes that are affected by the decision.

Q4. Communicate design decisions (1 page max): Construct a component diagram that illustrates the new design of your scheduling system. The diagram must be annotated with additional text that describes the responsibilities of the components, including those that are responsible for achieving the scalability requirements of your system. Please ensure that your design description communicates the essential aspects of the system with respect to scalability.

Submission & Grading

Compile your answers into a single PDF file and upload it through Gradescope by the indicated deadline. **Please correctly map the pages in the PDF to the corresponding questions on Gradescope.**

This assignment is out of **100** points (110 points with the bonus). For full points, we expect:

- **Q1 (10 pt bonus):** A report on load testing on the scheduling app or the shared service
- **Q2 (20 pt):** A specification of two scalability quality attribute requirements
- **Q3 (50 pt):** A description of the scalability-relevant design decisions, including (1) responses to all of the questions listed in Q2 and (2) a discussion of alternatives for three of the questions and justifications for the final decisions.
- **Q4 (30 pt):** A component diagram communicating the scalability-relevant design decisions, including a description of the component responsibilities.