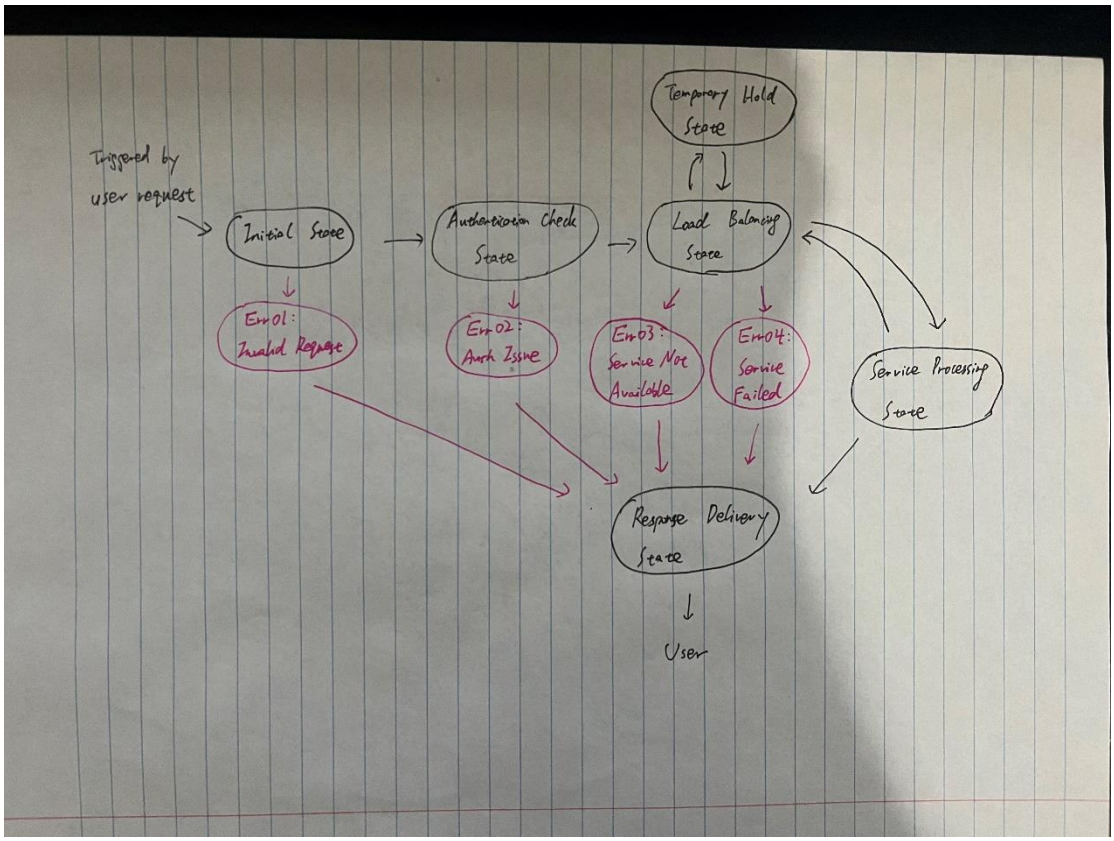
Our gateway is designed to handle incoming user requests (e.g., speech-to-text transcription, code generation from transcribed text) and route them to the appropriate backend services (e.g., speech-to-text, text-to-code services).
The workflow of the gateway includes the following states:

1. **Initial State (Request Received)**
   - **Description**: A user request arrives at the gateway, awaiting processing.
   - **Acceptable Transition**: Handle preliminary checks here (e.g., verify if the requested endpoint exists and is available.)
     - **Passed**: The request moves to *Authentication Check State*.
     - **Otherwise**: The request transitions to an error state with a response indicating the problem.
   - **Error States**: Invalid request.
2. **Authentication Check State**
   - **Description**: he gateway validates user credentials and ensures the user has sufficient permissions for the requested action.
   - **Acceptable Transitions**:
     - **Authorized**: The request moves to the *Load Balancing* state for workload distribution.
     - **Unauthorized**: The request transitions to an error state, triggering a rejection response to the user.
   - **Error States**: Missing or invalid tokens, insufficient permissions.
3. **Load Balancing State**
   - **Description**: The gateway assigns the request to an available service replica, taking into account load, availability, and potentially geographic proximity.
   - **Acceptable Transitions**:
     - **Successfully Assigned**: The request transitions to the *Service Processing* state.
     - **Service Unavailable**: If no replica is available, the gateway transitions to a temporary hold state (e.g., retry after a delay or queueing mechanism).
   - **Error States**: Failure to locate an appropriate replica, exceeding max retries in the hold state.
4. **Service Processing State**
   - **Description**: The selected service replica processes the request according to the specified operation (e.g., transcribe, code generation).
   - **Acceptable Transitions**:
     - **Successful Processing**: The request transitions to the *Response Delivery* state.
     - **Service-Specific Failure**: A replica-specific error may send the request back to the *Load Balancing* state to attempt a different replica or retry in a safe manner.
   - **Error States**: If all replicas fail after multiple attempts, the system moves to an error state, alerting system admins.
5. **Response Delivery State**
   - **Description**: The gateway sends the processed result back to the user. If an error occurs, the gateway sends the error message back to the user.

- **Acceptable Transitions**:
    - **Complete**: The request cycle ends successfully.

The following image shows how the system transfers within different states.




The first three error states are relatively straightforward and can be determined at the discretion of the gateway. However, the fourth error state—where a service fails after several repeated attempts—requires communication between the gateway and other backend services. Therefore, our next key task is to collaborate with teammates working on those services to establish clear criteria for when a service should be considered as failed.