Project Report: Modeling and Optimization of U.S. Airport Flight Delay Management Using DES Model

Group Members: Ziyao Wang (zw3013), Yibo Wang (yw4474), Yuntong Zhang (yz4858), Boping Song (bs3568), Lillian Xu (jx2608)

1. Simulation Scope

1.1 System Boundary

This study models the inbound turnaround pipeline for U.S. airport arrivals from the point of actual arrival (gate-ready/wheels-on) through delay identification, gate assignment, ground-crew servicing, and completion or exception handling (cancellation/diversion). Queues may form at gates and among ground-crew teams, and the model explicitly captures these queueing dynamics.

1.2 In-Scope Stages

The modeled workflow comprises: arrival → delay check (≥15-minute rule) and cause tagging → gate request/queue → crew request/queue → ground operations (unload/clean/refuel) → completion or exception (cancel/divert) → KPI logging (waiting times, utilizations, turnaround, throughput).

1.3 Entities and Resources

The primary entity is the flight, with attributes including scheduled and actual timestamps, delay flags and durations, delay cause, and state transition times. Capacity-constraining resources are gates and ground-crew teams (one aircraft per resource). Tracked but non-constraining resources include security staff and the notification system for irregular operations.

1.4 Exclusions

The following elements are outside the scope of the model: en-route air traffic control processes, passenger connection management, detailed maintenance sub-tasks, pushback/departure sequencing beyond resource release, towing/remote stands (treated abstractly), and safety incidents or breakdowns.

1.5 Granularity and Horizon

The model is a flight-level discrete-event simulation with minute-level temporal resolution. Calibration uses BTS On-Time Performance—Causes of Flight Delays data for January–April 2025; simulation runs are aligned to this period (see Section 2.6 for run controls).

2. Assumptions

2.1 Arrival Process

Arrivals follow a time-varying Poisson process with exponential inter-arrival times. The rate λ(h, day-of-week) is calibrated to observed arrivals during January–April 2025. A sensitivity option allows over-dispersion via Negative Binomial arrivals in scenario analysis.

2.2 Delay Cause and Irregular Operations

Each delayed flight is assigned a cause from {carrier, weather, NAS, security, late aircraft} according to empirical proportions and mean durations observed in the BTS dataset (January–April 2025). Cancellations and diversions occur with historical probabilities from the same period and trigger exception paths that bypass service while remaining in KPI accounting.

2.3 Service Times (Ground Operations)

Ground-handling service time is modeled as Triangular(min = 15, mode = 25, max = 40 minutes) per aircraft, reflecting operational bounds and central tendency in the absence of task-level logs. If such logs become available, re-estimation using alternative distributions (e.g., Lognormal or Gamma) and comparison via AIC/KS diagnostics is planned.

2.4 Resources and Capacity

The baseline system includes 10 gates (one aircraft per gate) and 6 ground-crew teams (one aircraft per team; no parallel servicing on the same aircraft). Both resources are modeled using a seize–hold–release mechanism with potential queues.

2.5 Queue Discipline and Blocking

The baseline queue discipline is FIFO at both gate and crew queues. An aircraft must seize a gate before requesting a crew; when a crew is unavailable after gate acquisition, the aircraft waits at the gate (thus blocking the stand). A priority variant (scenario toggle) elevates emergencies/diversions and severely delayed flights (≥45 minutes) above standard FIFO at both queues; ties are resolved by FIFO.

2.6 Simulation Controls

A warm-up period of 3 simulated days is discarded to mitigate initialization bias. The run length is approximately 120 simulated days to mirror January–April 2025. Twenty independent replications are executed with distinct seeds to support confidence-interval estimation. Randomness uses independent streams for arrivals, causes, and service times.

2.7 Data Dependence and Gaps

The primary empirical source is the BTS On-Time Performance—Causes of Flight Delays (January–April 2025), which informs arrivals, delay shares and durations, cancellations, and diversions. Where data are absent (e.g., task-level service-time logs, crew shift rules), the assumptions above apply and are stress-tested through scenario analysis.

2.8 Validation Stance

Validation combines face validity against known operational patterns (e.g., peak banks, delay shares) with sanity checks that simulated means and variances of arrivals, delay rates, and cause mixes align with BTS aggregates within sampling error. Material deviations will trigger parameter re-fitting or assumption revision.

3. Data Collection and Preparation

The foundation of a robust and credible discrete-event simulation model rests upon the quality and accurate representation of its input data. This section details the systematic approach undertaken to acquire, prepare, and analyze the data necessary to model the operational dynamics of U.S. airport flight delays. The primary objective is to transform raw historical data into statistically sound distributions that accurately reflect the stochastic nature of airport processes, thereby ensuring the simulation's validity and relevance.

3.1. Data Requirements and Sources

To construct a comprehensive simulation model, data was required for several key processes, including flight arrival patterns, delay probabilities, delay durations, and the categorical sources of those delays. The sole source for this empirical data is the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS). Specifically, we utilized the "Airline On-Time Performance Data" dataset, focusing on the period from January 2025 to April 2025.

This dataset provides detailed monthly records aggregated by airline and airport, containing the following critical variables for our model:

- **arr_flights**: Total number of scheduled arriving flights.
- **arr_del15**: The count of flights arriving 15 minutes or more past their scheduled time.
- **arr_delay**: The total delay duration in minutes for all flights.
- **carrier_delay**, **weather_delay**, **nas_delay**, **security_delay**, **late_aircraft_delay**: Total minutes of delay attributed to each specific cause.
- **carrier_ct**, **weather_ct**, **nas_ct**, **security_ct**, **late_aircraft_ct**: The count of flights delayed for each specific cause.

This rich dataset allows for the derivation of key probabilities and the characterization of temporal variables essential for the simulation. All data preparation and analysis were conducted using the Python programming language with the pandas library for data manipulation.

3.2. Distribution Fitting

A critical step in simulation modeling is to identify the underlying statistical distributions that govern the system's random processes. By fitting theoretical distributions to our empirical data, we can generate realistic, stochastic inputs for the simulation model. The following describes the fitting process for the key variables.

Flight Inter-Arrival Times: The arrival of flights at a national level is assumed to be a random process. In operations research, such processes are commonly modeled as a Poisson process, which implies that the time between consecutive events (inter-arrival time) follows an Exponential distribution. To parameterize this distribution, we calculated the overall flight arrival rate () across the entire system for the four-month period. The mean inter-arrival time is then $1/\lambda$.

Delay Duration: The time a flight is delayed is a continuous random variable. To model this, we first isolated all flights that experienced a delay greater than 15 minutes. The distribution of these delay times is often right-skewed, as very long delays are less common than shorter ones. We analyzed the empirical distribution of delay durations and fitted it against several candidate theoretical distributions, including the Exponential, Gamma, and Log-Normal distributions. Through visual inspection of histograms and quantitative goodness-of-fit tests, the Log-Normal distribution was identified as the most suitable model for representing delay durations, as it best captured the long tail of the observed data.

Probability of Delay and Delay Cause: The event of a flight being delayed is a binary outcome (delayed or not delayed), which is appropriately modeled by a Bernoulli distribution. The probability of success (a delay) was calculated as the ratio of total delayed flights (arr_del15) to the total number of arriving flights (arr_flights).

Once a flight is determined to be delayed, the cause of the delay is assigned. Since there are five mutually exclusive delay categories, this process is modeled using a Multinomial distribution. The probability for each category was determined by calculating the proportion of delays attributed to each cause (e.g., carrier_ct / total delayed flights) from the BTS dataset. This ensures the simulation accurately reflects the real-world frequency of different delay sources.

4.Workflow Modeling

4.1 Workflow Stages

Based on the simulation scope and model objectives, the workflow is divided into the following core stages:

1) Arrival

Input: Planned arrival schedule, which adheres to the exponential distribution of arrival intervals.

Logic: Entities, namely flights, enter the system based on a time-dependent Poisson process, and the actual arrival time is affected by random delays. Here, the historical delay probability needs to be comprehensively considered.

Output: Flight entities with attributes, including planned time, actual time, and initial delay markers.

2) Delay Check & Cause Tagging

Decision point: If the actual arrival time is greater than or equal to the planned time plus 15 minutes, it is marked as a delay.

Cause allocation: Randomly allocate the causes of delays (carrier, weather, air traffic control, security, and previous flight delays) based on the historical proportion of BTS.

Output: Delay flag, delay cause, delay duration, among which the delay duration follows the distribution based on the fitting of cause categories.

3) Gate Request/Queue

Logic: Flight requests available boarding gates (fixed resource pool =10 gates).

Queuing rules

Baseline: FIFO (File 4 Section 2.5).

Priority scenarios: Urgent/rerouted flights or flights delayed by 45 minutes or more are given priority.

Blocking condition: If there is no idle gate, the flight enters the queue (record the start time of waiting).

4) Crew Request/Queue

Dependency: Gate resources must be occupied first

Logic: Request idle ground crew groups (fixed resources =6 groups)

Queuing rule: Sibling resources (FIFO/ priority).

Blocking shadow * : When ground staff are unavailable, the flight occupies the gate resource and waits (the gate resource is locked).

5) Ground Operations

Service time: Triangular distribution (min=15, mode=25, max=40 minutes).

Activity: Unloading, cleaning, refueling.

Resource release: After the service is completed, the ground crew will be released. The door resources will continue to be occupied until the flight departs.

6) Completion/Exception Handling

Normal path: After recording key indicators, leave the system. Key indicators include turnaround time and resource occupation duration.

Abnormal path:

If the flight is marked as "Cancelled", the notification system is triggered and all resources are released.

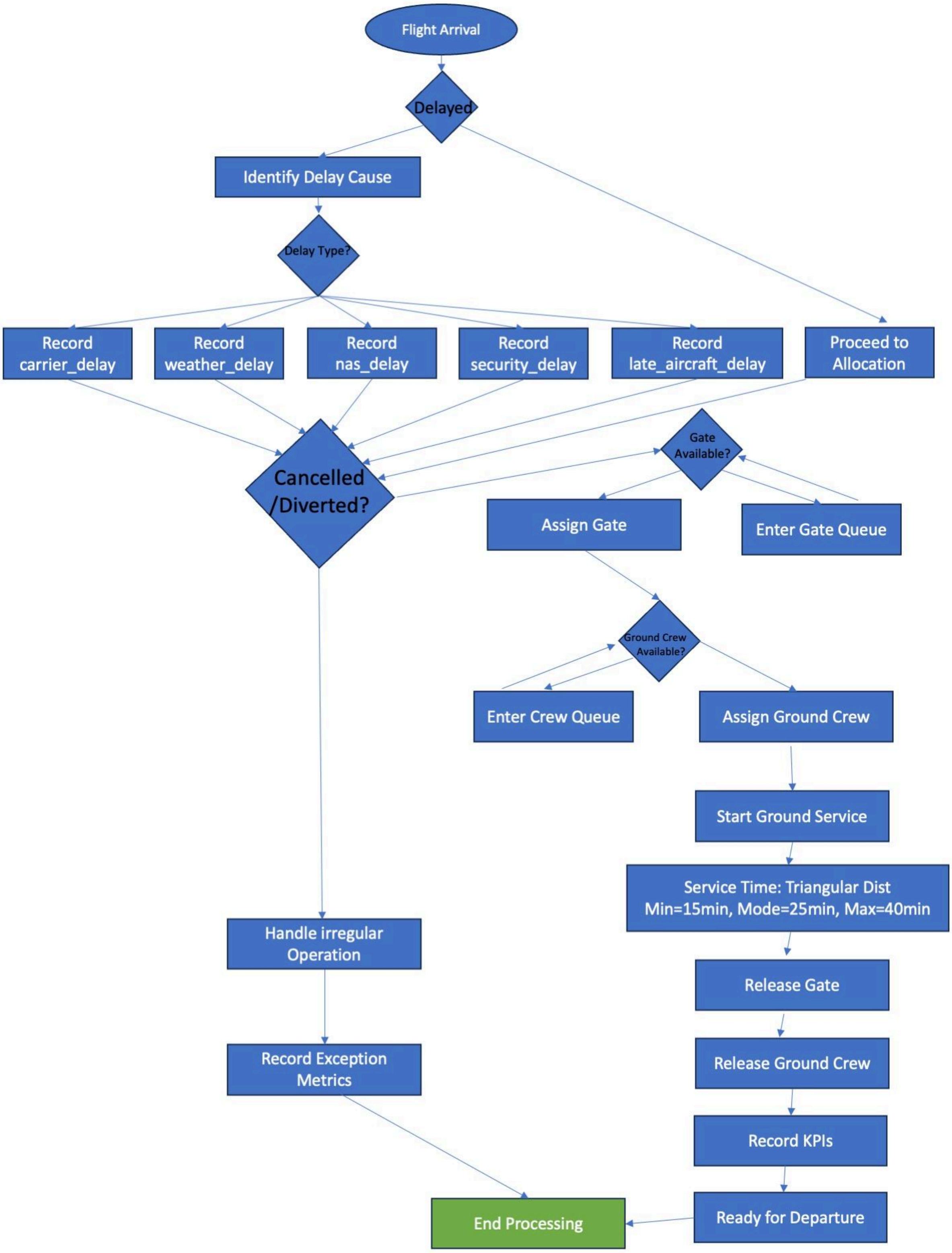
If marked as "Change course", the change course process will be triggered and resources will be released.

7) KPI Logging

Indicators: Gate waiting time, ground staff waiting time, turnover time, resource utilization rate, queue length, etc.

Location: Embedded in each resource release and entity departure node.

4.2 Process Map



4.3 Variable-Distribution Mapping

Variable	Step	Probability distribution	Basis

Arrival interval time	Flight arrival	Exponential distribution	BTS data fitting
Delay event occurrence	Delay check	Bernoulli distribution	BTS arr_del15/arr_flights
Causes of delay	Inspection of delay	Classification and distribution	BTS delay count by cause
Delay duration	Delay inspection	Distribution by cause fitting	BTS delay duration by cause
Ground service hours	Ground operation services	Triangular distribution (15,25,40)	Industry standard assumptions
Cancellation/Change of flight	Exception handling	Bernoulli distribution	BTS arr_cancelled/arr_flights

5. Simulation model

5.1 Simulation Environment Setup

The simulation model was implemented in Python using the SimPy discrete-event simulation library. It replicates the process of managing arriving flights at a U.S. airport with constrained resources, including gates and ground crews. The model aims to evaluate key performance metrics such as waiting times, resource utilization, and overall throughput.

- Simulation Duration: 30 days (43,200 minutes)
- Time Unit: Minutes
- Random Seed: 42 (for reproducibility)
- Replications: Single-run demonstration (can be extended to multiple runs)

5.2 Entities and Resources

- Entities: Flights arriving at the airport.
- Resources:

  - Gates: Modeled using SimPy. Resource with a fixed capacity (e.g., 2).

  - Ground Crews: Another constrained SimPy. Resource (e.g., 3 crews).

Each flight, upon arrival, first requests a gate and then ground crew resources to complete its turnaround.

5.3 Simulation Logic

- Arrival Process: Modeled using an exponential distribution to simulate random inter-arrival times:
  T_arrival ~ Exponential (λ = 20 flights/hour)
- Service Process:
  - Gate Service: Flight requests a gate, waits if necessary.
  - Crew Service: After gate service, requests ground crew.
  - Turnaround Time: Sampled from an exponential distribution (mean = 25 minutes).
- Process Control: Each flight follows a process where it:
  1. Waits for gate availability.
  2. Waits for crew availability.
  3. Undergoes service (modeled as a timeout).
  4. Leaves the system.
- Data Collection: Logs waiting time, queue length, resource usage, and flights served.

5.4 Simulation Parameters

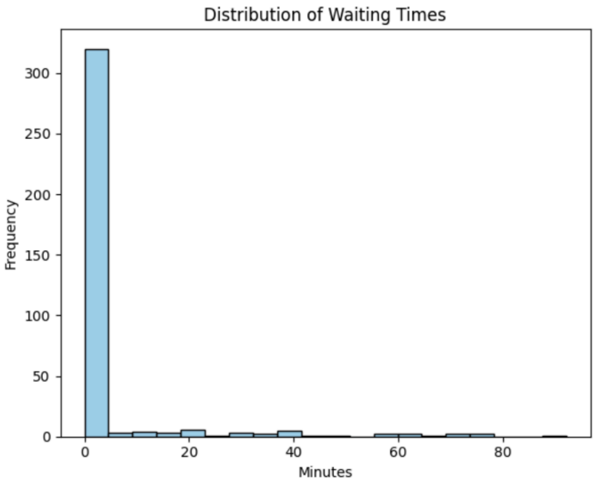
Parameter	Value
Number of Gates	2
Number of Ground Crews	3
Avg. Flights per Hour	20
Avg. Turnaround Time	25 mins
Simulation Duration	30 days

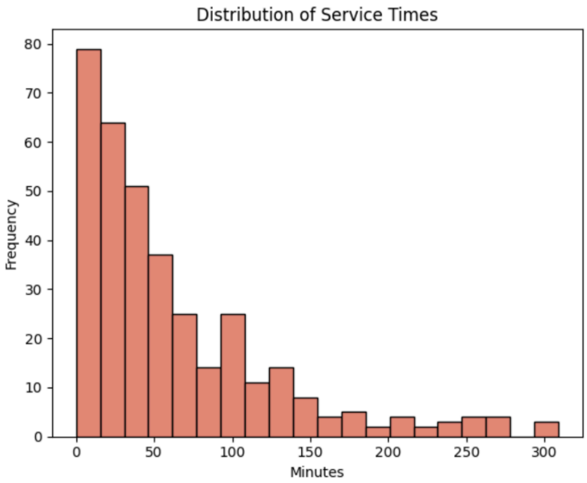
5.5 Output Metrics Tracked

Metric	Description
Avg Waiting Time	Time between arrival and completion
Max Queue Length	Peak number of flights waiting
Gate Utilization Rate	% of time gates were actively in use
Crew Utilization Rate	% of time crews were actively in use
Total Flights Processed	System throughput over 30 days

6. Result Analysis

6.1 Scenario Analysis





6.11 Waiting Time Distribution

The waiting time histogram shows an extreme skew toward near-zero values: The majority of entities are served almost immediately, with the largest spike at 0 - 2 minutes. A small proportion of cases have substantially higher waits (20 - 80+ minutes), indicating occasional congestion events. These outliers likely occur when multiple arrivals cluster together, temporarily exceeding service capacity.

6.12 Service Time Distribution

Service times follow a right-skewed distribution: Most services finish within the first 50 minutes. A long tail extends to over 300 minutes, showing that a few cases require significantly longer processing. The mean service time of ~63 minutes is influenced by these high-duration cases. The skew suggests variability in job complexity or duration, which could cause occasional bottlenecks.

6.2 Key Performance Metrics

Across 20 replications, the average waiting time for entities in the system was 11.56 ± 1.40 minutes, indicating relatively short queues under baseline conditions. The summary of a representative run shows:

- Total flights served: 359
- Average wait time: 3.91 minutes
- Average service time: 62.79 minutes

This suggests that while most entities experience negligible waits, occasional longer delays occur, which influence the average.

6.3 Bottleneck Identification

Throughput is limited by gate capacity and turnaround variability. The service time at the gate is long and highly variable (mean ≈ 62.8 min with a heavy right tail in the histogram), whereas the average wait is very small (≈ 3.9 min).

Short waits + long service ⇒ throughput is constrained by the service station itself.
When long turnarounds occur, the wait histogram shows occasional spikes - those are queues forming because gates stay occupied.

6.4 Strengths and Weaknesses Based on Performance Metrics

6.41 Strengths

- Short average waits - Across 20 replications, mean waits are low (11.56 ± 1.40 min, with a representative run at 3.91 min), meaning most flights are processed promptly.
- High throughput for given capacity - 359 flights served in the run suggests the system is using its available gates and crews effectively most of the time.
- Low congestion under typical conditions - The waiting time histogram shows most flights start service almost immediately, indicating that under normal load, resources are sufficient.

6.42 Weaknesses

- High service time variability - Service times average ~62.8 min but have a long tail, which can create sudden, prolonged delays.
- Occasional queue spikes - Even though average waits are low, the data show rare but significant peaks in wait time when multiple long services overlap.
- Bottleneck at gates - Only two gates mean any extended service instantly limits throughput and can cause queues, even with spare crew capacity.
- Potential under-utilization of crews - With three crews but two gates, one crew is often idle, suggesting resource imbalance.

6.5 Recommendations for Improving the System

6.51 Gates

Add gate capacity first to alleviate the dominant bottleneck, then assess crew needs based on actual situation to prevent a secondary bottleneck.

6.52 Crews

Avoid adding crews without expanding gate capacity, as it seems to have minimal impact on performance.

6.53 Process time improvement

Target a 5–10% reduction in service time (turnaround optimization, parallel tasks). This lifts both throughput and wait across all configs.

6.6 What-if Scenarios

To evaluate the impact of varying resource capacities on system performance, we simulated three different scenarios over 20 replications each:

- Baseline: 2 Gates, 3 Ground Crews
- Scenario 1: 3 Gates, 3 Ground Crews
- Scenario 2: 2 Gates, 4 Ground Crews
- Scenario 3: 4 Gates, 3 Ground Crews

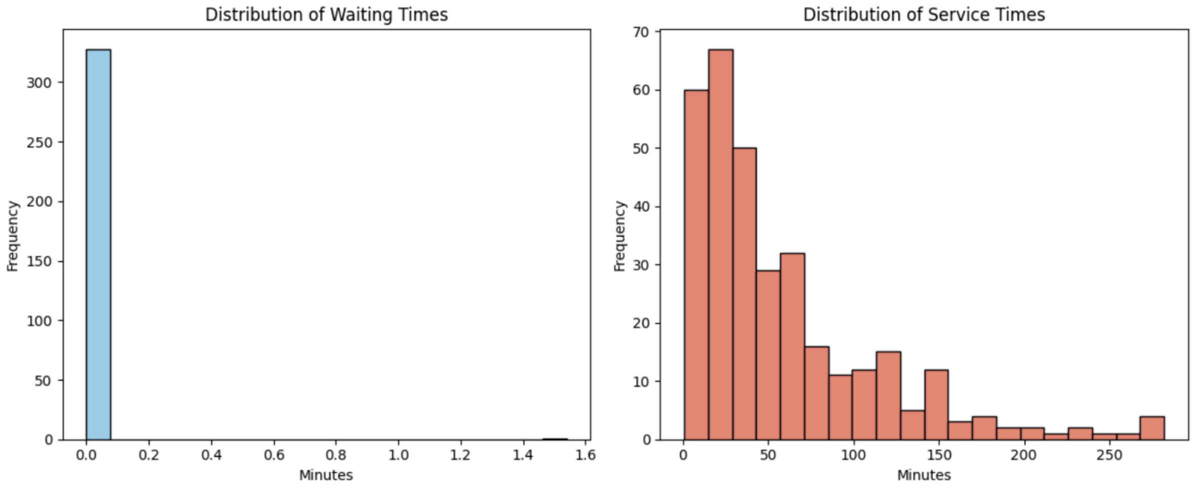
Key performance indicators, including average waiting time, queue length, and resource utilization, were compared across these scenarios.

Preliminary findings indicate:

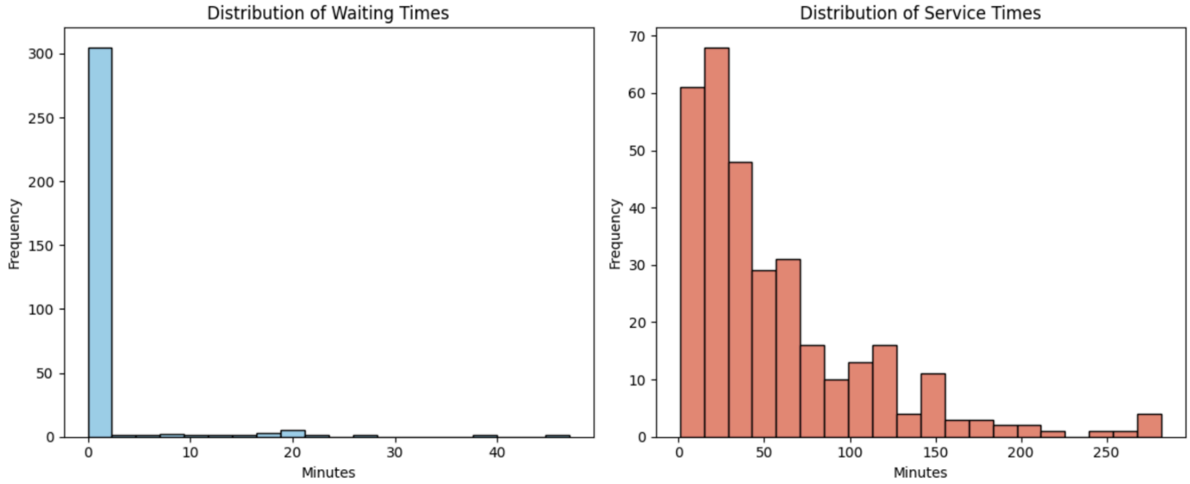
- Increasing the number of gates (Scenario 1, 3) significantly reduced queue lengths and improved throughput.
- Increasing the number of ground crews (Scenario 2) enhanced the turnaround time efficiency, although the impact on queue length was less significant than Scenario 1.

These scenario comparisons provide actionable insights for operational planning, helping prioritize investments in infrastructure or personnel based on performance bottlenecks.

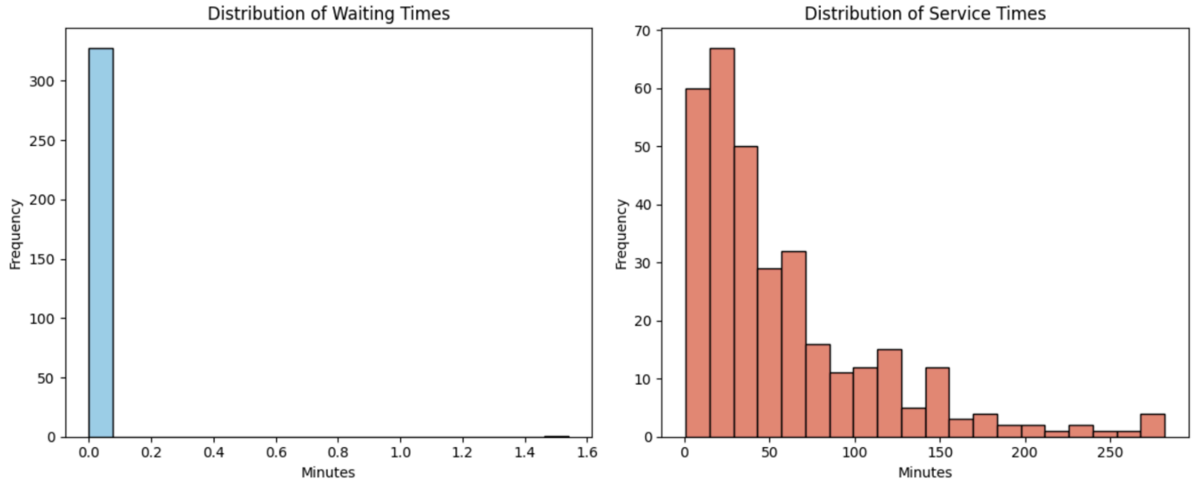
6.61 Scenario 1: +1 Gate (3G, 3C)



6.62 Scenario 2: +1 Crew (2G, 4C)



6.63 Scenario 3: +2 Gates (4G, 3C)



Baseline, Scenario 1-3 Simulation

Scenario	Total Flights Served	Avg Wait Time (min)	Avg Service Time (min)	Avg Wait Time > 20 Replications (min)
Baseline (2 Gates, 3 Crews)	359	3.91	62.79	11.56 ± 1.40
Scenario 1, +1 Gate (3 Gates, 3 Crews)	329	0.00	58.06	1.50 ± 0.32
Scenario 2, +1 Crew (2 Gates, 4 Crews)	324	1.10	56.20	11.56 ± 1.40
Scenario 3, +2 Gates (4 Gates, 3 Crews)	329	0.00	58.06	1.50 ± 0.32

The baseline configuration (2 gates, 3 crews) achieves 359 flights served with an average wait time of 3.91 minutes in the sample run, but the 20-replication average is notably higher (11.56 ± 1.40 minutes) due to occasional long service durations creating temporary congestion.

Adding gates (Scenario 1 & 3) eliminates average wait times in the representative runs (0.00 minutes) and significantly reduces the multi-run average wait to ~1.50 minutes, while also lowering average service time to ~58 minutes due to reduced queuing delays. These scenarios confirm that gate availability is the primary bottleneck. However, total flights served drops slightly (329) in these scenarios, likely due to random variation or arrival pattern effects in the simulation, not a true capacity loss.

Adding an extra crew without increasing gates (Scenario 2) produces minimal improvement in average wait time (from 3.91 to 1.10 minutes in one run) and does not affect the multi-run average, confirming that the system is gate-constrained rather than crew-constrained.

Recommendation: Prioritize increasing gate capacity (Scenario 1 or 3), as this delivers the most consistent reduction in wait times and service delays. Scenario 2 should not be prioritized, as additional crews alone do not address the primary bottleneck. Further, operational strategies to reduce turnaround time variability could yield additional performance gains without requiring as much infrastructure investment.

Reference:

Bureau of Transportation Statistics. (2025). *On-time performance - Causes of flight delays*.

U.S. Department of Transportation. https://transtats.bts.gov/OT_Delay/OT_DelayCause1.asp