

No more queueing at the ladies' room: modeling, analysis and simulation of waiting times at toilet facilities with 'ladies first' scheduling policies

Stan De Troyer

Student number: 02109634

Supervisors: Prof. dr. ir. Sabine Wittevrongel, Prof. dr. ir. Wouter Rogiest

Counsellor: Dr. Arnaud Devos

Master's dissertation submitted in order to obtain the degree of Master of Science in Industrial Engineering and Operations Research
- Manufacturing and Supply Chain Engineering

Academic year 2025-2026

Preface

Looking back on the work that went into this dissertation, I would like to thank the people who supported me along the way.

I am sincerely grateful to my promotors, Prof. Dr. ir. Sabine Wittevrongel and Prof. Dr. ir. Wouter Rogiest, for their guidance and feedback. I would like to thank Prof. Wittevrongel for her support on the more practical and theoretical aspects of the work, particularly in the area of queueing theory and its implementation. My thanks also go to Prof. Rogiest for his broad expertise and vision, which shaped many of the major decisions and assumptions throughout this project and helped define its overall direction.

I also would like to thank my supervisor, Dr. Arnaud Devos, whose contribution has been invaluable. His expertise in queueing theory helped me make sense of difficult concepts and apply them to this work. I am equally grateful for his extensive feedback, the many one-on-one sessions, his critical remarks, and his constant willingness to help improve both the research and the writing throughout this process.

Finally, I would like to thank my girlfriend Louise for tolerating the many evenings that started with “I will only be five more minutes” and ended several simulation runs later. Thank you for pretending to care when I talked about queues, for accepting that my laptop fan became a permanent background noise, and for being there at the end of every long day.

The author gives permission to make this master dissertation available for consultation and to copy parts of this master dissertation for personal use. In all cases of other use, the copyright terms have to be respected, in particular with regard to the obligation to state explicitly the source when quoting results from this master dissertation.

Ghent, May 2026

Stan De Troyer

Remark on the master's dissertation and oral presentation

This master's dissertation is part of an exam. Any comments formulated by the assessment committee during the oral presentation of the master's dissertation are not included in this text.

Use of generative artificial intelligence

Generative artificial intelligence tools were used in a supporting capacity throughout this dissertation, in accordance with the documentation requirements of the Faculty of Engineering and Architecture at Ghent University for the academic year 2025–2026. Three tools were used: **Claude** (Anthropic), **ChatGPT** (OpenAI), and **NotebookLM** (Google). Throughout this dissertation, the latest publicly released models were used as they became available between September 2025 and the submission date.

Claude was used for the following purposes:

- **Writing assistance and style improvement:** The author drafted all content independently. Claude was used to refine phrasing, sentence structure, and register, particularly to achieve the appropriate balance between technical precision and readability when writing in English as a second language.
- **Grammar and spelling correction.**
- **Code development and debugging:** The simulation design and logic were developed by the author. Claude assisted with debugging, efficiency improvements, and implementation suggestions for specific technical challenges, most notably the configuration of Bayesian optimisation via Optuna after an initial Gurobi-based approach proved difficult to adapt to the stochastic objective function.
- **L^AT_EX formatting:** Claude assisted with table layouts, figure cross-references, and structural formatting throughout the thesis.
- **Hypothesis testing and brainstorming:** When simulation results behaved unexpectedly, Claude served as a sounding board to verify or challenge the author’s hypotheses about the underlying causes.
- **Bibliography formatting:** Claude assisted with ensuring bibliographic entries conformed to the required reference style.

ChatGPT was used to generate two schematic illustrations: the two-state MMPP arrival process shown in Figure 1.2, and the Layout 6 worked example shown in Figure A.1. In each

case the author specified the structure, labels, and annotations, and reviewed and lightly edited the output before inclusion. All other figures in this dissertation were produced directly from simulation output by the author's own Python code, or adapted from cited sources.

NotebookLM was used for academic source discovery. Relevant background material was uploaded to the platform, which suggested potentially related academic sources. Each suggested source was then individually retrieved, read, and evaluated by the author before any decision on inclusion in the literature review was made.

The author is aware of the risks associated with generative AI tools, including the possibility of hallucinated references, unintended bias in generated content, and misrepresentation of intent. All AI-generated suggestions were critically reviewed and, where necessary, corrected. The research design, analytical framework, interpretation of results, and all conclusions in this dissertation are entirely the author's own.

No more queueing at the ladies’ room: modeling, analysis and simulation of waiting times at toilet facilities with ‘ladies first’ scheduling policies

by

STAN DE TROYER

Master’s dissertation submitted in order to obtain the academic degree of
Master of Science in Industrial Engineering and Operations Research
– Manufacturing and Supply Chain Engineering

Promotors: Prof. dr. ir. Sabine Wittevrongel, Prof. dr. ir. Wouter Rogiest
Supervisor: Dr. Arnaud Devos

Faculty of Engineering and Architecture
Ghent University

Academic year 2025–2026

Abstract. At large public events, women often face longer restroom queues than men. Part of this imbalance is structural: equal floor area does not imply equal service capacity, because women’s occupancy times are longer and traditional men’s facilities include fast-serving urinals. This dissertation studies whether the gender waiting-time gap can be reduced, and what this costs in overall efficiency and in delays for male users.

A discrete-event simulation compares six restroom layouts under a common budget of 20 stall-equivalents. Demand is modelled by a two-state Markov-modulated Poisson process (MMPP), with calm and busy regimes. The simulation is checked through warm-up and observation-window analysis and verified against M/M/c (Erlang C) benchmarks. On the best-performing layout, four “ladies-first” priority policies are evaluated: Strict Head-of-Line, Token Bucket, Threshold Priority, and Dynamic Aging. The last three are calibrated by Bayesian optimisation.

The best layout is a mixed design with 14 shared stalls and 8 men’s urinals. It reduces the overall mean wait to 31 s and the gender gap to 11 s, reductions of 43% and 84% relative to the best feasible separated design. Women’s mean wait falls by 58%. The mixed layout is the only configuration with sub-minute waits for both groups.

On this layout, the optimised priority policies close the residual gap to within statistical noise, at the cost of a 26% increase in men’s mean wait, consistent with Kleinrock’s conservation law. Dynamic Aging lowers the worst-case male wait below the FCFS baseline and is recommended when arrival times can be recorded reliably. The Token Bucket is the practical fallback. Robustness checks show that priority is useful mainly near balanced demand: male-skewed demand already reaches near-parity under FCFS, while female-skewed demand calls for more shared-stall capacity. With fixed floor area, layout choice drives most of the improvement; priority scheduling closes the residual gap, and only when demand is near-balanced.

Keywords: potty parity, DES, MMPP, priority queueing, Bayesian optimisation.

No more queueing at the ladies’ room: modeling, analysis and simulation of waiting times at toilet facilities with ‘ladies first’ scheduling policies

Stan De Troyer

Promotors: Prof. dr. ir. Sabine Wittevrongel, Prof. dr. ir. Wouter Rogiest

Supervisor: Dr. Arnaud Devos

Master’s dissertation submitted in partial fulfilment of the requirements
for the degree of Master of Science in Industrial Engineering and Operations Research

Department of Telecommunications and Information Processing

Faculty of Engineering and Architecture, Ghent University

Academic year 2025–2026

Abstract—At large public events, women often face longer restroom queues than men. Part of this imbalance is structural: equal floor area does not imply equal service capacity, because women’s occupancy times are longer and traditional men’s facilities include fast-serving urinals. This dissertation studies whether the gender waiting-time gap can be reduced, and what this costs in overall efficiency and in delays for male users. A discrete-event simulation compares six restroom layouts under a common budget of 20 stall-equivalents. Demand is modelled by a two-state Markov-modulated Poisson process (MMPP), with calm and busy regimes. The simulation is checked through warm-up and observation-window analysis and verified against M/M/c (Erlang C) benchmarks. On the best-performing layout, four “ladies-first” priority policies are evaluated: Strict Head-of-Line, Token Bucket, Threshold Priority, and Dynamic Aging. The last three are calibrated by Bayesian optimisation. The best layout is a mixed design with 14 shared stalls and 8 men’s urinals. It reduces the overall mean wait to 31 s and the gender gap to 11 s, reductions of 43% and 84% relative to the best feasible separated design. Women’s mean wait falls by 58%. The mixed layout is the only configuration with sub-minute waits for both groups. On this layout, the optimised priority policies close the residual gap to within statistical noise, at the cost of a 26% increase in men’s mean wait, consistent with Kleinrock’s conservation law. Dynamic Aging lowers the worst-case male wait below the FCFS baseline and is recommended when arrival times can be recorded reliably. The Token Bucket is the practical fallback. Robustness checks show that priority is useful mainly near balanced demand: male-skewed demand already reaches near-parity under FCFS, while female-skewed demand calls for more shared-stall capacity. With fixed floor area, layout choice drives most of the improvement; priority scheduling closes the residual gap, and only when demand is near-balanced.

Index Terms—potty parity, discrete-event simulation, Markov-modulated Poisson process, priority queueing, Bayesian optimisation, gender equity.

I. INTRODUCTION

At concerts, festivals and sporting events, the queue for the women’s restroom is often longer than the queue for the men’s [1]–[3]. The cause is structural. Building codes typically

allocate equal floor area, but equal area does not produce equal service capacity. Women’s average occupancy times are longer than men’s [4], [5], while traditional men’s facilities combine stalls with urinals that serve users much faster. During busy periods, this difference makes the women’s queue grow faster than the men’s.

Event organisers face two main design choices. The first is the layout: how a fixed floor area is divided between women’s stalls, men’s stalls and urinals, or pooled into shared facilities. The second is the queue discipline: how users are selected from a shared queue, and whether women receive priority.

The layout determines how much shared capacity exists and which users can access it. The queue discipline then determines how that shared capacity is used. A priority rule may perform well when enough men are present in the shared queue, but have little effect when men mostly choose a shorter urinal queue. Conversely, a layout may remove the mean gender gap by pooling capacity, while still creating long waits for everyone if the pooled system is close to saturation. The relevant comparison must consider layout and policy together under a fixed floor-area budget.

This dissertation follows that structure. First, six layouts are compared under the same floor-area budget and under bursty event demand. The best-performing layout is then used to evaluate four “ladies-first” priority policies. The three parameterised policies are calibrated by Bayesian optimisation, using the gender waiting-time gap as the main objective.

II. RELATED WORK

The classic potty-parity literature documents the inequity and argues that equal floor allocation does not produce equal access [1]–[3]. Van Hautegeem and Rogiest [6] compare the same six layouts that this thesis uses, Huh et al. [7] evaluate building-code fixture ratios for gender-specific waits, and Bovens and Marcoci [8] compare gender-separated and gender-neutral provision. All three studies assume stationary

Poisson arrivals. Farajollahzadeh and Hu [9] extend the question with a utility-parity model that adds safety perceptions and gender-identity preferences to waiting times. Converting only a subset of male facilities to shared use improves service for some users without making any group worse off, summarised as “a little flexibility dominates full flexibility”. Operational policies remain outside their scope.

Classical queueing theory provides the building blocks for the policies tested here. Cobham [10] laid the foundation for priority queues in M/G/1, with the multi-server M/M/c extension [11] as the analytical reference point used in this study. The accumulating priority queue, proposed by Kleinrock [12] as a delay-dependent discipline and analysed by Walraevens et al. [13] for its tail behaviour and equity-starvation trade-off, is the basis for Dynamic Aging. Threshold Priority adapts a threshold-based policy analysed by Bruneel [14], with the activation logic inverted to suspend (rather than trigger) priority under heavy load. Kleinrock’s conservation law [15] shows that the total wait is the same under any work-conserving priority rule.

Two more recent UGent theses on the same topic move in different directions. De Jonghe [16] distinguishes sitters from uprights and models dynamic nudging from stalls to urinals. Duysens [17] models flexible-signage layouts. Both keep arrivals stationary and exclude priority policies.

The present work combines realistic burst demand, a fixed floor-area budget across competing layouts, and a calibrated family of ladies-first priority policies under one common modelling framework.

III. MODEL AND DEMAND PROCESS

A. Simulation framework

The restroom system is modelled as a discrete-event simulation in Python. Customers arrive, join a queue, receive service at a stall or urinal, and depart. Per-customer waiting times are recorded separately for women and men, together with queue lengths and per-facility utilisations. When priority policies are active, bypass counts are also tracked.

B. Demand process

Demand at large events is highly variable, especially around scheduled breaks such as half-time or the end of an opening act. To capture this, arrivals follow a two-state Markov-modulated Poisson process (MMPP) [18]. The system alternates between a calm state with 12 arrivals per minute and a busy state with 24 arrivals per minute. Mean dwell times are 10 minutes in the calm state and 2 minutes in the busy state (Fig. 1). Unless stated otherwise, arrivals split equally between women and men.

C. Service times

Service times are exponential, with means that depend on both gender and facility type, namely 1.5 min for women’s stalls, 1.25 min for men’s stalls, and 1.0 min for urinals. Real service times are unlikely to be memoryless: if they are less variable than an exponential distribution, the model may

2-State Markov Modulated Poisson Process (MMPP)

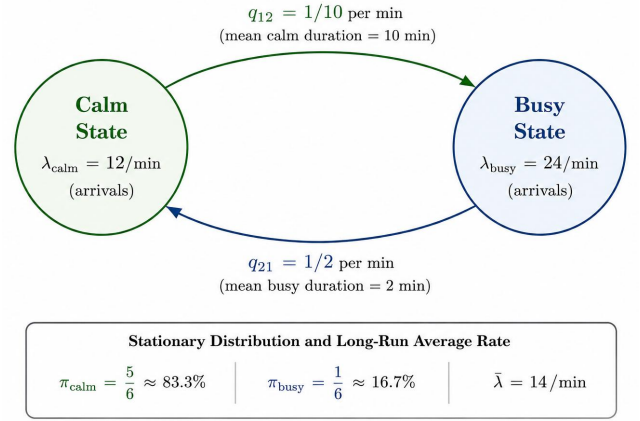


Figure 1. Two-state MMPP arrival rate: short, intense busy periods overlay a longer calm baseline.

overstate waiting times. Male customers choose between an available stall and an available urinal based on the shortest expected waiting time at the moment of arrival.

D. Capacity constraint

All six layouts share a common floor-area budget of 20 stall-equivalents, with each stall counting as 1.0 unit and each urinal as 0.75 units.

IV. SIX LAYOUTS

The six layouts follow Van Hautegeem and Rogiest [6] and span the design space from traditional separated provision to a fully shared pool. The names follow their classification, but under the MMPP demand of this thesis some labels (notably Layout 2’s “Equal Wait”) no longer fit. Table I lists each, and Fig. 2 shows them schematically.

Table I
LAYOUTS COMPARED IN THIS STUDY. EACH ROW RESPECTS THE 20 STALL-EQUIVALENT BUDGET.

#	Name	W stalls	M stalls	Urinals	Shared
1	Baseline	10	2	10	0
2	Equal Wait	13	2	6	0
3	Min Avg Wait	12	2	8	0
4	Mixed	0	0	10	12
5	Gender Neutral	0	0	0	20
6	Optimal Mixed	0	0	8	14

Layouts 1–3 are separated designs that differ only in how the budget is split between women’s stalls and urinals. Layout 1 represents typical equal-area provision. Layout 2 was designed for approximately equal gender waits under stationary Poisson assumptions. Layout 3 minimises the overall mean wait under the same assumptions and is the best feasible separated design.

Layouts 4–6 introduce shared capacity. Layout 4 pools the stalls into a single mixed-gender queue served by 12 shared stalls, while keeping 10 male urinals separate. Layout 5 removes urinals entirely and pools all capacity into 20 shared

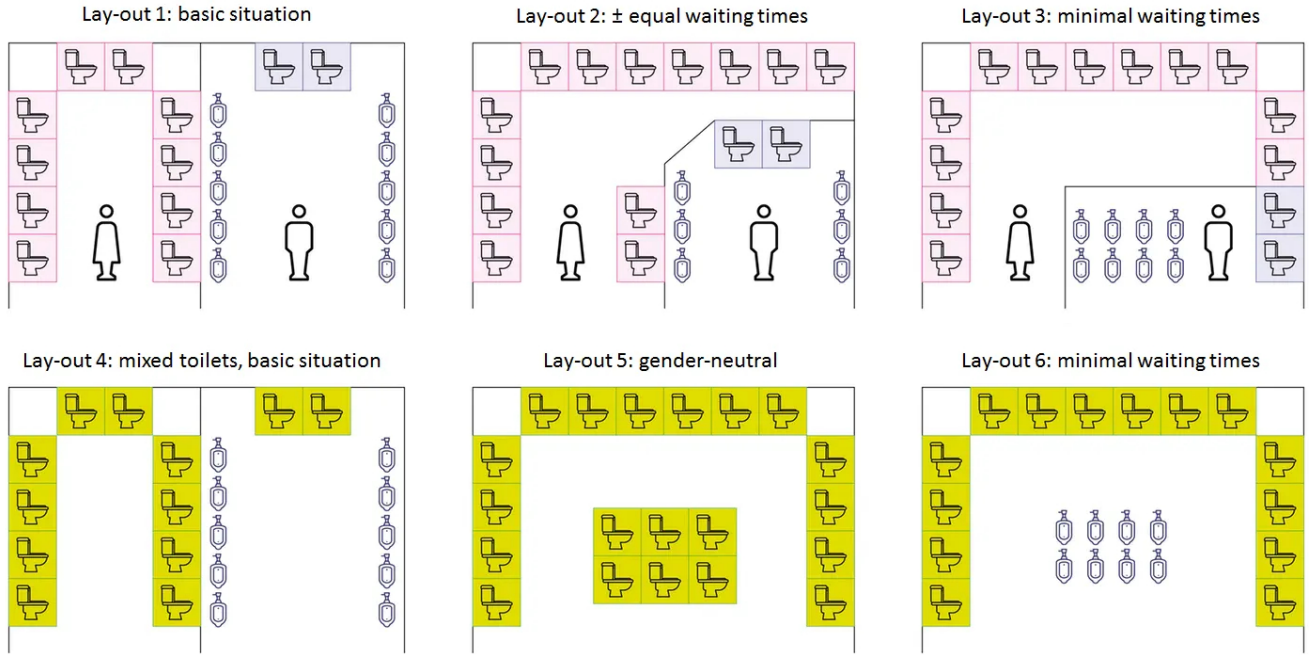


Figure 2. The six layouts, drawn to a common 20 stall-equivalent budget. Layouts 1–3 are separated, Layout 4 pools the stalls, Layout 5 pools everything, and Layout 6 pools only the stalls while keeping a smaller bank of urinals.

stalls. Layout 6 keeps 8 male urinals and pools the remaining capacity into 14 shared stalls, balancing the two design ideas.

V. STATISTICAL PROTOCOL AND ANALYTICAL VERIFICATION

A. Warm-up determination

Each layout is evaluated over 1 000 independent replications, each consisting of a warm-up period followed by a 90-minute observation window. The warm-up length of 1 778 minutes is the global maximum across layouts, selected by a multi-criterion convergence test on the across-replication cumulative mean of the per-customer waiting time. Three stability conditions must hold over a 30-minute look-ahead window (absolute stability, relative stability, and closeness to the final level), and a 90-minute post-convergence buffer must remain beyond the candidate cutoff. The common warm-up removes initialisation bias uniformly across layouts and makes the layout comparison directly meaningful. All reported quantities carry 95% confidence intervals based on the empirical between-replication variance.

B. Analytical verification

The full model is too complex for closed-form analysis, but in a simplified verification mode (MMPP disabled, deterministic routing for men in mixed layouts) the layouts reduce to standard M/M/c queues with closed-form (Erlang C) predictions for all 12 queue-gender pairs. A 30 000-minute observation window per replication narrows the confidence intervals enough to distinguish persistent residuals from finite-window noise. Eight pairs match the analytical prediction within their 95% confidence interval. The four pairs that fall

just outside lie within $\pm 5\%$ relative error and correspond to mixed queues where stalls and urinals are pooled into a single male queue with a heterogeneous service-time mixture (coefficient of variation above one). The closed-form prediction does not capture this mixture, which explains the residual gap. A binomial test on the eight M/M/c-exact pairs shows no detectable bias ($p \approx 0.73$). The agreement is a strong indication that the simulator behaves as expected wherever an analytical reference is available, though it is not a formal proof of correctness.

VI. LAYOUT COMPARISON

Table II reports the mean women's wait, the mean men's wait, and the gender gap for each layout under the 50/50 baseline. Fig. 3 summarises the same data graphically together with per-facility utilisation.

Layouts 1–3 illustrate the limits of separated provision. Layout 1, the equal-area baseline, is infeasible: with only 10 women's stalls and a 50% female arrival fraction, the women's load is $\rho_{\text{women}} = 1.050 > 1$ and the queue grows without bound. Layout 2 keeps the women's wait short by allocating 13 stalls, but the saved capacity has to come from men's urinals, and the men's wait rises to nearly three minutes. Layout 3 minimises the overall mean wait under the separated constraint, but still leaves a 1 minute 8 second gender gap.

Pooling capacity into a shared queue closes the gap, but can be expensive. Layout 5, the fully gender-neutral pool, makes the gap zero by design. Its overall mean wait, however, rises to more than five minutes because every customer competes for the same 20 stalls and the pool operates at a utilisation

Table II
MEAN WAITING TIMES BY LAYOUT UNDER THE 50/50 BASELINE (MMPP DEMAND, 1 000 REPLICATIONS, 95% CONFIDENCE INTERVALS).

#	Layout	Women's wait	Men's wait	Gap ΔW	Status
1	Baseline	unstable	5 s $\pm$ < 1 s	n/a	Women's queue grows without bound
2	Equal Wait	39 s $\pm$ 3 s	2 min 49 s $\pm$ 12 s	2 min 10 s $\pm$ 11 s	Feasible but imbalanced
3	Min Avg Wait	1 min 28 s $\pm$ 6 s	20 s $\pm$ 1 s	1 min 8 s $\pm$ 6 s	Best separated
4	Mixed	1 min 31 s $\pm$ 6 s	12 s $\pm$ 1 s	1 min 19 s $\pm$ 6 s	Shared stalls, urinals separate
5	Gender Neutral	5 min 21 s $\pm$ 24 s	5 min 21 s $\pm$ 24 s	0 s $\pm$ 1 s	Perfect parity, slow overall
6	Optimal Mixed	37 s $\pm$ 2 s	25 s $\pm$ 2 s	11 s $\pm$ 1 s	Best overall, smallest gap

Layout performance comparison (50/50 baseline)

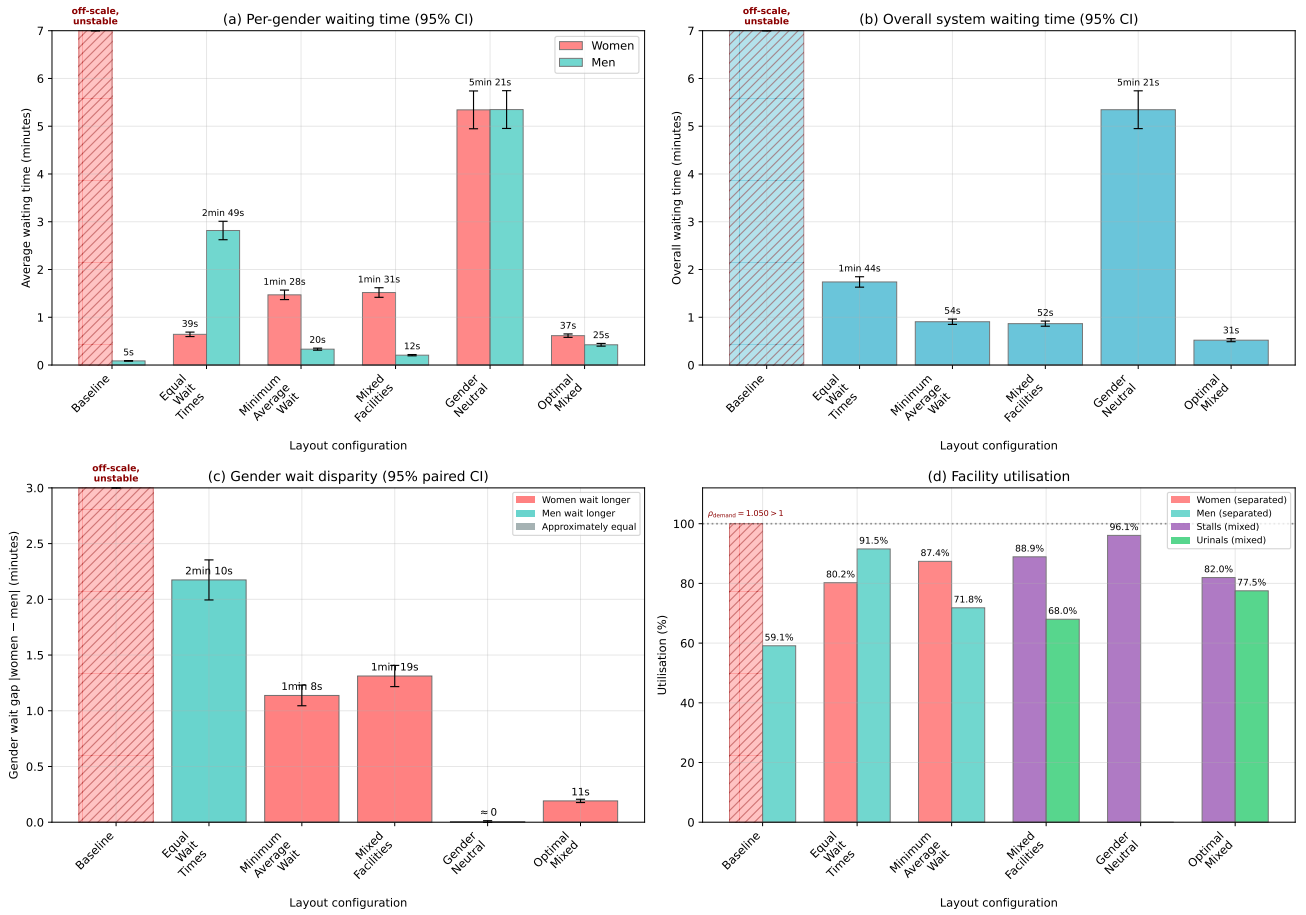


Figure 3. Performance comparison across layouts: per-gender mean waits, overall mean wait, gender gap, and per-facility utilisation under MMPP demand (1000 replications, 95% confidence intervals). Layout 1 is shown off-scale because its women's queue is unstable.

of around 0.96, close to saturation, leaving no buffer for burst demand.

Layout 6 keeps 8 urinals separate and pools only the stalls. Male customers route to whichever side offers the shortest expected wait. The urinals absorb most male demand, leaving the stalls largely available to women. Stalls and urinals operate at comparable utilisations (around 0.82 and 0.78), both well below saturation, and the layout retains a buffer for MMPP bursts. Against Layout 3, the best feasible separated design, Layout 6 cuts the overall mean wait by 43% and the gender gap by 84%. Women's mean wait drops by 58%.

VII. LADIES-FIRST PRIORITY ON THE WINNING LAYOUT

Four ladies-first priority policies are evaluated on Layout 6.

A. Strict Head-of-Line

Any waiting woman is served before any waiting man, with no protection against starvation. Used as a reference ceiling, not as a practical recommendation. On Layout 6 at the 50/50 baseline it reverses the gap: women's mean wait falls to 28 s while men's rises to 37 s, a 48% cost that leaves men 9 s slower than women. On Layout 5 men's mean wait rises from

5 min 0 s under FCFS to 11 min 27 s, creating practical starvation.

B. Token Bucket

The system holds a bucket of tokens that refills at a constant rate up to a maximum capacity. When a woman arrives at a shared queue with men ahead of her, she can spend one token to bypass the men in front. The bucket caps how often reordering happens, smoothly interpolating between FCFS (rate near zero) and Strict Head-of-Line (rate large enough never to run out). The calibrated parameters on Layout 6 are a token refill rate of 6.91 tokens per minute and a bucket capacity of 19.

C. Threshold Priority

Token Bucket reordering applies only while the shared queue is at or below a threshold K . Above the threshold, the policy reverts to FCFS, which protects men from priority delays when the queue is congested. The calibrated parameters on Layout 6 are a refill rate of 9.55 tokens per minute, a bucket capacity of 20, and a threshold $K = 34$.

D. Dynamic Aging

Each waiting man accumulates priority linearly with time, at a rate r_{age} , so that no man can be indefinitely bypassed. Above a crossover wait time, men reach women's priority and are served in turn. The calibrated aging rate on Layout 6 is $r_{\text{age}} = 0.1271$ per minute, which corresponds to a crossover wait of 7 min 52 s.

E. Calibration

The parameters of the three parameterised policies are calibrated using Bayesian optimisation with the Tree-structured Parzen Estimator algorithm [19], allocating 100 trials with 500 replications per trial. The objective is the gender gap, with a small tie-breaker on the overall mean wait:

$$\mathcal{L} = |\Delta W| + 10^{-4} \bar{W}. \quad (1)$$

The noise floor at 500 replications is around 1.8 s, so a robust-neighbourhood criterion selects the optimum from the most reliable region of the search space rather than the single best-scoring trial. A separate study with overall mean wait as the objective found no parameter set that beats FCFS, consistent with Kleinrock's conservation law: priority cannot reduce the aggregate work in a work-conserving queue [15].

F. Results

Table III reports the calibrated outcomes.

The three optimised policies are statistically indistinguishable on the mean gap and on the mean wait, but they differ in how they protect individual men. Fig. 4 shows the empirical cumulative distribution of male waiting times under each policy.

Dynamic Aging guarantees that no man can be indefinitely bypassed. Above the crossover wait time, men reach higher priority as women, so the upper tail is bounded. On Layout 6, the mean per-replication maximum male wait drops to 6

Table III
CALIBRATED POLICIES ON LAYOUT 6 (50/50 BASELINE, MMPP DEMAND, 1 000 REPLICATIONS, 95% CI).

Policy	Women's wait	Men's wait	Gap ΔW
FCFS (no priority)	37 s $\pm$ 2 s	25 s $\pm$ 2 s	11 s $\pm$ 1 s
Token Bucket	32 s $\pm$ 2 s	32 s $\pm$ 2 s	< 1 s
Threshold Priority	33 s $\pm$ 2 s	32 s $\pm$ 2 s	< 1 s
Dynamic Aging	33 s $\pm$ 2 s	32 s $\pm$ 2 s	< 1 s

min 21 s under Dynamic Aging, against 7 min 55 s under Token Bucket and 7 min 57 s under Threshold Priority. The single worst case across 1000 replications is 10 min 52 s under Dynamic Aging, against an FCFS baseline of 12 min 15 s. Worst-case bypass counts tell the same story: 89 under Dynamic Aging, 207 under Token Bucket, 319 under Threshold Priority. All three policies are indistinguishable from parity. Dynamic Aging adds tighter tail protection at no extra cost. It is the recommended policy when arrival times can be recorded reliably, with the Token Bucket as the practical fallback. Threshold Priority is not recommended for deployment, as it adds operational complexity without improving the men's tail over the Token Bucket.

VIII. ROBUSTNESS AND LIMITS

Two robustness checks vary the gender split away from 50/50. Together they show that priority is useful mainly near balanced demand.

When men dominate arrivals (60% men, 40% women), the women's queue is already short under FCFS on Layout 6 and the residual gap is statistically zero. Priority policies are unnecessary.

When women dominate arrivals (60% women, 40% men), the shared-stall queue contains no men in 99.2% of snapshots, leaving too few men for any priority rule to act on. The remaining gap reflects a layout limit, not an ordering one. Layout redesign, not priority, is the only remaining lever in this regime. The same logic suggests that for venues with a strongly female user base, an even larger fraction of shared stalls would be preferable to the 14/8 split used here.

A general conservation law underlies these findings. Priority rules can redistribute waiting time between user groups, but cannot reduce the aggregate waiting time [15].

IX. CONCLUSION AND OUTLOOK

The practical message is simple. First, get the layout right. Then, if a small gap remains, priority can help as a second step.

The first step is to allocate capacity well. Switching from the best feasible separated design to a partial-unisex design with 14 shared stalls and 8 men's urinals, on the same floor area, cuts the overall mean wait by 43% and the gender gap by 84%. Women's mean wait drops by 58%. No additional floor area is required.

The second step is a smaller but useful refinement. On the winning layout, an optimised priority policy closes the residual

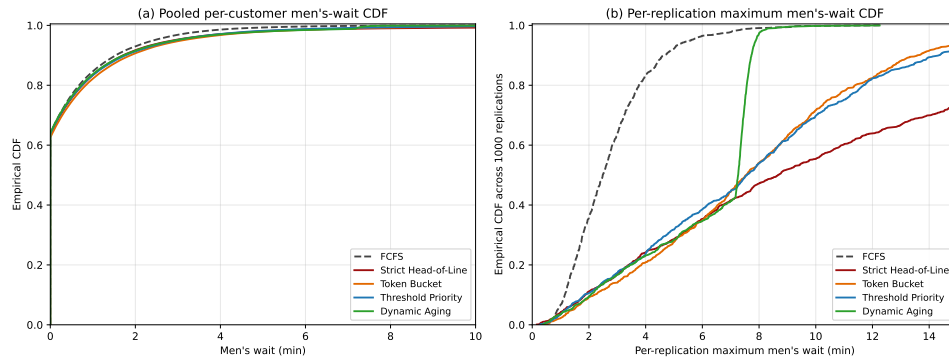


Figure 4. Empirical cumulative distribution of male waiting times on Layout 6 under FCFS and the three optimised priority policies. Dynamic Aging compresses the upper tail.

11 s gap to within statistical noise at a 26% rise in men's mean wait. Dynamic Aging additionally lowers the worst-case male wait below the FCFS baseline. With fixed floor area, layout choice drives most of the improvement; priority scheduling closes the residual gap.

The framework generalises to other public venues with bursty demand, such as train stations, stadiums, shopping centres, and airports.

A. Limits and future work

Three caveats limit the scope of the recommendation. The model treats gender as a binary classifier, which excludes non-binary users and the identity-cost dimensions that a real design must address. The simulation has not been validated against a field experiment, and if real occupancy times are less variable than the exponential distribution, the model may overstate waiting times. The relative ranking of layouts and policies is likely to be more stable than the exact values. The implementation of any ladies-first rule has to handle gender classification with care, since self-declaration respects users but is open to misuse, and visual classification is not an acceptable option.

Future work includes a field deployment at a controlled event, the use of empirical service-time distributions in place of the exponential proxy, and an extension of the user model to include non-binary users alongside the safety and identity dimensions of restroom choice.

REFERENCES

- [1] C. Greed, "Public toilet provision for women in Britain: An investigation of discrimination against urination," *Women's Studies International Forum*, vol. 18, no. 5–6, pp. 573–584, 1995.
- [2] K. H. Anthony and M. Dufresne, "Potty parity in perspective: Gender and family issues in planning and designing public restrooms," *Journal of Planning Literature*, vol. 21, no. 3, pp. 267–294, 2007.
- [3] C. Greed, "Join the queue: Including women's toilet needs in public space," *The Sociological Review*, vol. 67, no. 4, pp. 908–926, 2019.
- [4] M. A. Baillie, S. Fraser, and M. J. Brown, "Do women spend more time in the restroom than men?" *Psychological Reports*, vol. 105, no. 3, pp. 789–790, 2009.
- [5] S. M. V. Gwynne, A. L. E. Hunt, J. R. Thomas, A. J. L. Thompson, and L. Séguin, "The toilet paper: Bathroom dwell time observations at an airport," *Journal of Building Engineering*, vol. 24, p. 100751, 2019.
- [6] K. Van Hautegeem and W. Rogiest, "Nooit meer file aan het damestoilet," Eos Wetenschap, 2017, originally published June 2017; webpage last revised 20 June 2023. Accessed: 2025-10-12. Technical appendix available at <https://www.eoswetenschap.eu/technischebijlagetoiletten>. [Online]. Available: <https://www.eoswetenschap.eu/festivaltoiletten>
- [7] W. T. Huh, J. Lee, H. Park, and K. S. Park, "The potty parity problem: Towards gender equality at restrooms in business facilities," *Socio-Economic Planning Sciences*, vol. 68, p. 100693, 2019.
- [8] L. Bovens and A. Marcoci, "The gender-neutral bathroom: A new frame and some nudges," *Behavioural Public Policy*, vol. 7, no. 1, pp. 1–24, 2020, online-first July 2020; volume published January 2023.
- [9] S. Farajollahzadeh and M. Hu, "Potty parity," *Management Science*, 2025.
- [10] A. Cobham, "Priority assignment in waiting line problems," *Journal of the Operations Research Society of America*, vol. 2, no. 1, pp. 70–76, 1954.
- [11] R. B. Cooper, *Introduction to Queueing Theory*, 2nd ed. North Holland, 1981.
- [12] L. Kleinrock, "A delay dependent queue discipline," *Naval Research Logistics Quarterly*, vol. 11, no. 3-4, pp. 329–341, 1964.
- [13] J. Walraevens, T. Van Giel, S. De Vuyst, and S. Wittevrongel, "Asymptotics of waiting time distributions in the accumulating priority queue," *Queueing Systems*, vol. 101, pp. 221–244, 2022.
- [14] H. Bruneel, "Analysis of a threshold-based priority queue," *Queueing Systems*, vol. 109, no. 8, pp. 1–45, 2025.
- [15] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. Wiley-Interscience, 1975.
- [16] L. De Jonghe, "Genderneutrale toiletten en nudging: de sleutel tot kortere wachtrijen," Master's thesis, Ghent University, 2024.
- [17] A. Duysens, "Modeling and analysis of waiting times at toilet facilities with dynamic signage," Master's thesis, Ghent University, 2025.
- [18] W. Fischer and K. S. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," *Performance Evaluation*, vol. 18, no. 2, pp. 149–171, 1993.
- [19] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.

Contents

Preface	i
Remark on the master’s dissertation and oral presentation	ii
Use of generative artificial intelligence	iii
Abstract	v
List of figures	xvii
List of tables	xix
List of abbreviations and symbols	xx
1 Introduction and methodology	1
1.1 Problem context	1
1.2 Contributions	2
1.3 Simulation framework	3
1.3.1 Model components	3
1.3.2 Scheduling policies	4
1.3.3 Performance metrics	5
1.4 Experimental design	6
1.4.1 Restroom layouts	6
1.4.2 Simulation parameters	8
1.4.3 The MMPP arrival process	10
1.5 Thesis outline	11
1.6 Chapter summary	12
2 Literature review	13
2.1 The potty parity problem	13
2.2 Queueing-theoretic models for restroom systems	14
2.3 Priority scheduling	16
2.4 Discrete-event simulation and warm-up analysis	17

2.5	Simulation-based optimisation	17
2.6	Chapter summary	18
3	Warm-up period and observation window analysis	19
3.1	Motivation	19
3.2	Methodology for warm-up period determination	19
3.2.1	Cumulative mean waiting time	19
3.2.2	Across-replication aggregation	20
3.2.3	Convergence criteria	20
3.3	Results of warm-up period analysis	22
3.3.1	Interpretation and implications	25
3.4	Observation period validation	26
3.5	Adopted experimental protocol	29
3.6	Chapter summary	29
4	Analytical verification using queueing theory	30
4.1	Goal of verification	30
4.2	Theoretical foundation: M/M/ c queues	30
4.2.1	The Erlang C formula	31
4.2.2	Capacity-weighted average service rate (heterogeneous servers)	31
4.3	Methodology: reducing to known models	32
4.3.1	Verification-mode runtime	32
4.4	Worked examples: Layouts 1 and 5	32
4.4.1	Layout 1: women's queue (M/M/10)	33
4.4.2	Layout 1: men's queue (approximate M/M/12)	34
4.4.3	Layout 5: gender-neutral pooled queue (M/M/20)	34
4.5	Analytical predictions for all layouts	35
4.6	Verification results	36
4.6.1	Discussion of results	36
4.7	Model validation and limitations	38
4.7.1	Scope and limitations	38
4.7.2	Stability under MMPP arrival rates	38
4.8	Chapter summary	40
5	Simulation results: layout comparison	41
5.1	Experimental setup	41
5.2	Performance by layout	41
5.2.1	Infeasible baseline: Layout 1	43
5.2.2	Alternative configurations	44
5.3	Utilisation analysis	46
5.4	Sensitivity analysis: robustness to gender distribution	47
5.4.1	Patterns across the gender splits	48
5.5	Chapter summary	50

6	Ladies-first priority policies	51
6.1	Motivation	51
6.2	Strict Head-of-Line: the theoretical limit	52
6.2.1	Definition	52
6.2.2	Results	53
6.2.3	Implications for policy design	54
6.3	Optimisation methodology	54
6.3.1	Equity objective	54
6.3.2	Noise floor and robust optimum	55
6.4	Token Bucket	55
6.4.1	Mechanism	55
6.4.2	Parameters	56
6.4.3	Search ranges	56
6.4.4	Optimal parameters and achieved parity	57
6.4.5	Cross-layout transfer	57
6.4.6	Parameter sensitivity and importance	58
6.5	Threshold Priority	59
6.5.1	Motivation for a threshold	59
6.5.2	Decision rule	60
6.5.3	Optimisation	60
6.5.4	Results	61
6.5.5	Cross-layout transfer	63
6.6	Dynamic Aging	64
6.6.1	Definition and starvation-free property	64
6.6.2	Optimisation	65
6.6.3	Results	65
6.6.4	Summary	67
6.7	Cross-policy tail and starvation diagnostic	68
6.7.1	Men's tail diagnostic	68
6.7.2	Women's tail diagnostic	70
6.7.3	Other observations	71
6.8	Sensitivity analysis: robustness to gender distribution	72
6.8.1	Transfer of 50/50-optimised parameters	72
6.8.2	Re-optimisation under skewed splits	74
6.9	Real-world implementation	76
6.9.1	Common operational constraints	76
6.9.2	Strict Head-of-Line	77
6.9.3	Token Bucket	77
6.9.4	Threshold Priority	78
6.9.5	Dynamic Aging	78
6.9.6	Handling different gender splits	79
6.10	Comparative summary and recommendation	79

6.11 Chapter summary	81
7 Conclusion	83
7.1 Summary of findings	83
7.2 Layout design	84
7.3 Priority policies	84
7.4 Robustness to gender split	86
7.5 Limitations	86
7.6 Future work	87
7.7 Closing remarks	88
8 Societal reflection	89
8.1 Public restrooms as gendered infrastructure	89
8.2 From simulation results to public-space decisions	90
8.3 Ethical limits and implementation risks	90
8.4 Concluding reflection	91
Bibliography	92
A Accuracy of the rational-agent heuristic	95
A.1 Simulator state and the form of W_{queue} and W_{service}	95
A.2 Comparison with the M/M/c closed-form expected wait	95
A.3 Disagreement window	96
A.4 Worked example: Layout 6	97
B Observation-window convergence study for verification residuals	100
B.1 Design of the diagnostic	100
B.2 Results	101
C Robust optimum selection methodology	103
C.1 Three-step procedure	103
C.2 Hyperparameter choices	104
D Warm-up convergence diagnostic for Layout 5 priority policies	105
D.1 Methodology	105
D.2 Results	105
D.3 Implication for the chapter tables	108
E Implementation details and computational aspects	109
E.1 Discrete-event simulation engine	109
E.1.1 Event scheduling	109
E.1.2 State variables	109
E.1.3 Random number generation	110
E.2 Statistical analysis pipeline	111

E.2.1	Data collection	111
E.2.2	Aggregation and inference	111
E.3	Simulation scale	111
E.4	Software dependencies	112

List of Figures

1.1	Overview of the six restroom layouts	7
1.2	Two-state MMPP arrival process	10
3.1	Warm-up validation for Layout 1	22
3.2	Warm-up validation for Layout 2	23
3.3	Warm-up validation for Layout 3	23
3.4	Warm-up validation for Layout 4	23
3.5	Warm-up validation for Layout 5	24
3.6	Warm-up validation for Layout 6	24
3.7	Observation-window validation for Layout 1	27
3.8	Observation-window validation for Layout 2	27
3.9	Observation-window validation for Layout 3	27
3.10	Observation-window validation for Layout 4	28
3.11	Observation-window validation for Layout 5	28
3.12	Observation-window validation for Layout 6	28
4.1	Analytical verification at $T = 30\,000$ min	36
4.2	Queue utilisation by arrival regime	39
5.1	Performance comparison across layouts	42
5.2	Layout sensitivity by gender split	49
6.1	Token Bucket cross-layout transfer	57
6.2	Token Bucket optimisation summary	59
6.3	Threshold Priority optimisation summary	61
6.4	Threshold Priority cross-layout transfer	63
6.5	Dynamic Aging optimisation summary	66
6.6	Dynamic Aging cross-layout transfer	66
6.7	Men’s wait distributions by policy	69
6.8	Women’s wait distributions by policy	71
6.9	Priority sensitivity on Layout 6	73

A.1	Layout 6 rational-agent worked example	98
B.1	Verification residual by observation window	101
D.1	Layout 6 warm-up under Strict Head-of-Line	106
D.2	Layout 5 warm-up under priority policies	107

List of Tables

1.1	Detailed specifications of restroom layouts.	8
3.1	Recommended warm-up periods by layout.	25
4.1	Analytical model and routing assumptions for each layout type.	33
4.2	M/M/c predictions at calm-state demand	35
4.3	Verification summary against M/M/c benchmarks	37
4.4	Utilisation of the women’s queue (or pooled queue) at different MMPP arrival rates.	39
5.1	Average waiting times by layout	43
5.2	Layout sensitivity to gender arrival distribution	48
6.1	Strict Head-of-Line results by layout	53
6.2	Token Bucket results on Layout 6	57
6.3	Token Bucket cross-layout results	58
6.4	Three-policy comparison on Layout 6	61
6.5	Threshold Priority cross-layout results	64
6.6	Dynamic Aging cross-layout results	67
6.7	Men’s tail-wait diagnostics	68
6.8	Priority-policy sensitivity by gender split	73
6.9	Re-optimised policy parameters by gender split	75
6.10	Four-policy comparison on Layout 6	80
A.1	Disagreement windows by layout	97
B.1	Relative residuals by observation window	101

List of abbreviations and symbols

Abbreviations

CDF	Cumulative distribution function
CI	Confidence interval
CRN	Common Random Numbers
CTMC	Continuous-time Markov chain
DES	Discrete-event simulation
FCFS	First-come, first-served
HoL	Head-of-line (priority discipline)
M/M/c	Markovian arrivals, Markovian service, c servers (Kendall notation)
MMPP	Markov-modulated Poisson process
NHPP	Non-homogeneous Poisson process
QBD	Quasi-birth-death
SDG	Sustainable Development Goal
TB	Token Bucket
TH	Threshold Priority
TPE	Tree-structured Parzen Estimator

Symbols

λ	Customer arrival rate (customers per minute)
λ_w, λ_m	Per-gender arrival rates (women, men)
$\lambda_{\text{calm}}, \lambda_{\text{busy}}, \bar{\lambda}$	Calm-state, busy-state, and long-run average MMPP arrival rates
μ	Service rate per facility (customers per minute)

μ_w, μ_m, μ_u	Service rates for women's stalls, men's stalls, and urinals
$\bar{\mu}$	Capacity-weighted service rate
ρ	Traffic intensity, defined as $\lambda/(c\mu)$
c	Number of parallel servers (general)
c_w, c_m	Counts of women's and men's stalls (separated layouts)
$\pi_{\text{calm}}, \pi_{\text{busy}}$	MMPP steady-state probabilities
q_{12}, q_{21}	MMPP transition rates between calm and busy states
Q	CTMC generator matrix for the MMPP background process
Λ	Diagonal matrix of MMPP state-dependent arrival rates
$Z(t)$	MMPP background state at time t
n	Number of independent simulation replications
$\bar{X}$	Sample mean across replications
s	Sample standard deviation across replications
$t_{n-1, 1-\alpha/2}$	$(1 - \alpha/2)$ -quantile of Student's t -distribution with $n - 1$ degrees of freedom
$t_{\text{warm-up}}$	Warm-up duration discarded before measurement
t_{obs}	Post-warm-up observation period
T	Simulation horizon or observation-window endpoint
W_w, W_m	Mean waiting time, women and men
$\bar{W}$	Gender-weighted overall mean waiting time across customers
ΔW	Gender gap, $ W_w - W_m $
W_q	Mean queue waiting time for a queueing subsystem
C_s^2	Squared coefficient of variation of service times
P_W	Erlang C probability that an arriving customer must wait
$\mathcal{Q}(t)$	Queue contents or queue length at time t
$\mathcal{L}$	Optimisation objective for policy parameters
θ	Vector of policy parameters being optimised
P_{90}, P_{95}, P_{99}	Empirical 90th, 95th, and 99th percentiles
$r_{\text{tok}}^{\text{TB}}, C^{\text{TB}}$	Token Bucket token rate (per minute) and capacity
$r_{\text{tok}}^{\text{TH}}, C^{\text{TH}}$	Threshold Priority token rate (per minute) and capacity
K	Threshold Priority queue-length gate (integer customers)
r_{age}	Dynamic-aging rate (per minute)
τ	Crossover time of the dynamic-aging policy, $\tau = 1/r_{\text{age}}$

CHAPTER 1

Introduction and methodology

1.1 Problem context

Imagine attending your favourite festival and arriving early to secure the best possible spot for the headlining act. As the concert unfolds, you start to feel the growing need to use the restroom, but leaving would mean giving up your hard-earned place. The moment the performance ends, you rush to the toilets, only to find yourself facing a queue that stretches far beyond reason, while the line at the adjacent restroom moves with barely any delay.

This scenario is familiar to women at concerts, sporting events, airports, and other public venues worldwide. Disproportionate restroom waiting times have become an accepted inconvenience. The frustration points to a real design question: how should restroom facilities be configured to serve all users efficiently and equitably?

Building codes typically allocate equal floor space to women’s and men’s facilities [1], but equal floor area does not produce equal service capacity [2]. Women face longer waiting times than men in conventional configurations, a phenomenon termed *potty parity* in the literature [2, 3]. The observed disparity in waiting times can be attributed to two main factors. On the user side, women generally have longer restroom occupancy times due to physiological factors, clothing-related needs, and, in some cases, the responsibility of accompanying children [3]. On the design side, men’s facilities typically combine stalls with urinals, which can provide greater service capacity than only using stalls occupying the same floor area [3].

The persistence of the problem despite being so visible suggests that intuitive equal-space allocations and the fixture ratios prescribed by current building codes are not enough (see Chapter 2 for evidence from prior work). This thesis models, analyses, and simulates operational policies with a focus on a family of “ladies-first” priority policies. These are designed to narrow the fairness gap while preserving aggregate efficiency in shared-capacity restroom settings.

This thesis addresses the following central research question:

How do different “ladies-first” priority policies perform in shared-capacity restroom facilities? Do they successfully narrow the gender waiting time gap, and at what cost to overall system efficiency and individual male users?

To answer this central question, we also examine the following supporting questions:

- **Layout design:** Under a fixed total floor area, which restroom layout minimises overall waiting time while simultaneously achieving the greatest gender equity? We examine this question across multiple layouts as a foundation for policy evaluation, building on prior work by Van Hautegeem & Rogiest [4] but implementing the analysis under slightly different modelling assumptions.
- **Robustness:** How sensitive are the performance and policy conclusions to deviations from the 50/50 gender split (60/40 and 40/60) in arrival composition?

1.2 Contributions

This thesis makes four contributions.

- **Comparative layout evaluation under matched capacity.** Six restroom configurations are simulated under a fixed total stall-equivalent budget. The framework follows the comparative analysis of Van Hautegeem & Rogiest [4] but uses a two-state Markov-modulated Poisson process (MMPP) for arrivals and assigns distinct mean service times to women’s stalls, men’s stalls, and urinals.
- **Analytical verification against M/M/c benchmarks.**¹ For two reference layouts that admit a closed-form solution, the steady-state simulation outputs are checked against Erlang C predictions. The match confirms that the simulator behaves as expected in the limit cases.
- **A family of ladies-first priority policies.** Four priority policies are designed: Strict Head-of-Line (the parameter-free theoretical extreme that serves as the upper-bound reference) and three parameterised alternatives, the Token Bucket, Threshold Priority, and Dynamic Aging. Their parameters are optimised using Bayesian optimisation. The thesis then maps how each policy trades aggregate efficiency for gender equity.
- **Robustness analysis.** The layout and policy conclusions are re-evaluated under 60/40 and 40/60 gender splits. The split separates results that hold across demand compositions from those that do not.

¹M/M/c is Kendall notation [5] for a queue with Poisson arrivals (M), exponential service times (M), and c parallel servers. The Erlang C formula gives the expected waiting time. Chapter 4 treats both in full.

1.3 Simulation framework

A discrete-event simulation (DES) model is used to assess how facility layout, queue discipline, and gender equity interact under time-varying demand. The model follows each customer from arrival through queueing, service, and departure. Waiting times are recorded directly from post-warm-up samples, producing an empirical waiting-time distribution per layout. Chapter 3 confirms that the chosen warm-up period is long enough for the system to reach steady state before measurement starts.

1.3.1 Model components

The simulation model comprises the following components:

- **Arrival Process:** Within each MMPP regime, arrivals form a Poisson process with combined rate λ . In the baseline scenario, gender is drawn independently for each arrival as Bernoulli(0.5), so the per-gender rates are $\lambda_w = \lambda_m = \lambda/2$ and the combined arrival rate is $\lambda = \lambda_w + \lambda_m$. The Poisson-within-regime assumption reflects the independent nature of individual arrival decisions and the large population of potential users. The MMPP regime-switching dynamics are introduced in Section 1.4.3. The Bernoulli(0.5) assumption is relaxed in the robustness analysis (Section 5.4), which repeats the study under Bernoulli(0.6) and Bernoulli(0.4) splits.
- **Service Process:** Service times (i.e. toilet occupancy durations) are modelled as exponentially distributed random variables and are facility- and gender-specific: women's stalls with mean μ_w^{-1} , men's stalls with mean μ_m^{-1} , and men's urinals with mean μ_u^{-1} (so μ_w , μ_m , μ_u are the corresponding facility- and gender-specific service rates). The exponential assumption represents the memoryless property of service durations.
- **Service Requirements:** The model assumes that all men arrive exclusively for urination, meaning they can choose between a stall or a urinal based on availability and expected wait time. The case where some men require a stall with certainty is left out of scope to keep the baseline model clean. This is left for future work.
- **Queue Discipline:** Baseline service selection follows First-Come-First-Served (FCFS). Chapter 6 introduces priority policies that modify this discipline for the shared stall pool.
- **Server Configuration:** Each layout uses either (a) separate stalls for women (c_w capacity) and men (c_m capacity), or (b) shared stalls accessible to all users, plus optional urinals reserved for men.
- **Rational Agent Decision Logic:** Each arriving man chooses between a stall and a urinal using a workload-based heuristic under a fluid approximation:

$$\widehat{E[W]} = \frac{W_{\text{queue}} + W_{\text{service}}}{c} \quad (1.1)$$

where

$\widehat{E[W]}$ is the estimated expected wait the arriving man assigns to the candidate facility;
 c is the number of servers (stalls or urinals) at that facility;
 $\mathcal{Q}$ is the set of customers currently waiting at that facility;
 $W_{\text{queue}} = \sum_{i \in \mathcal{Q}} \mu_i^{-1}$ is the sum of the gender-specific mean service times of the customers currently in the queue;
 W_{service} is the total residual service time across busy servers.

The residual is computed as

$$W_{\text{service}} = k \cdot \frac{E[S^2]}{2E[S]} \quad (1.2)$$

where k is the number of busy servers, S denotes the service time at the facility (a random variable), $E[S]$ its mean and $E[S^2]$ its second moment. For mixed-stall facilities, where women and men have distinct service-time distributions, the simulation evaluates this residual per gender and combines the two by arrival probability, which coincides with the mixture-distribution residual under the exponential service assumption used throughout this thesis. This heuristic models the arriving man as satisficing (satisfy + suffice, after Simon [6]) rather than optimising: he selects an adequate option from the visible queue state rather than computing the closed-form optimum. Appendix A quantifies its accuracy relative to the M/M/c closed-form benchmark and works through an example in detail.

Once an arriving man joins a queue, he does not revisit his choice, even if the other facility later becomes available. His position in the chosen queue is fixed until he enters service.

For analytical verification (Chapter 4), the simulator runs in a *verification mode* that disables the MMPP and, where needed to keep the gender queues independent, replaces the rational-agent rule of Equation (1.1) with deterministic urinal routing for men. Each gender queue then reduces to an M/M/c analysis, either exact or with a capacity-weighted mean service rate.

1.3.2 Scheduling policies

This section focuses on the policies implemented in the simulation for mixed and gender-neutral configurations, where men and women draw from a common set of stalls (see Section 1.4.1 for the precise layout definitions). All policies are non-preemptive: a customer already in service is never interrupted. In this thesis, the following scheduling policies are studied:

- **FCFS (baseline):** Both genders join one shared queue and are served in order of arrival.
- **Strict Head-of-Line:** When a stall becomes free, if a man is at the head of the queue and any woman is waiting behind him, the first such woman is promoted to the head and served immediately. The remaining queue order is preserved. This represents the theoretical upper bound of ladies-first priority.

- **Token Bucket:** Women receive priority in the shared stall queue only while tokens are available. Tokens accumulate at a configured rate $r_{\text{tok}}^{\text{TB}}$ (tokens per minute) in a finite-capacity bucket of size C^{TB} (tokens), limiting persistent priority and moderating the cost imposed on men. In the limit $r_{\text{tok}}^{\text{TB}} \rightarrow \infty$ the discipline coincides with the Strict Head-of-Line policy above. Finite parameters and the empirical trade-offs are studied in Chapter 6.
- **Threshold Priority:** Let $|Q(t)|$ denote the number of customers in the mixed-stall queue at time t . While $|Q(t)| < K$ (an integer customer-count threshold), Token Bucket priority applies. When $|Q(t)| \geq K$, the system reverts to FCFS to reduce the burden on men during heavy demand. The threshold is observable as a customer count, simplifying the real-world deployment.
- **Dynamic Aging:** Each waiting man’s effective priority grows linearly with his elapsed waiting time at rate r_{age} , while women hold a constant baseline priority. A man overtakes any waiting woman after his wait exceeds the crossover time $\tau = 1/r_{\text{age}}$. The policy is deterministically starvation-free: for any finite $r_{\text{age}} > 0$, no man can be bypassed indefinitely.

1.3.3 Performance metrics

The simulation tracks four primary performance metrics, together with two diagnostic metrics used to characterise the priority policies introduced above:

- **Average waiting time:** The mean time users spend in queue before service, measured separately for women (W_w) and men (W_m).
- **Overall waiting time:** The gender-weighted mean across all users.
- **System utilisation** (ρ , defined in Chapter 3): The average long-run busy fraction of the servers in a given service pool.²
- **Equity metric:** The gender waiting-time gap, defined below.
- **Residual queue at simulation end** (diagnostic): The number of customers, reported per gender, still waiting when the 90-minute observation window closes. Used as a starvation indicator for the priority policies.
- **Bypass count** (diagnostic): The number of times a woman moves ahead of a man under a ladies-first policy. Reports mechanism activity rather than service quality.

²Strictly, “utilisation” is a single-server notion. In multi-server or multi-resource systems, it is defined differently, or not at all. We include ρ as an indicator of system load. For an M/M/ c queue with identical servers, symmetry makes $\rho = \lambda/(c\mu)$ the long-run busy fraction at each server [7]. For layouts that combine fixture types or when applying routing restrictions, ρ is reported separately for each relevant service pool (e.g., stalls, urinals, or the women’s and men’s subsystems in separated layouts).

Defining service equity

This study evaluates both the *efficiency* of restroom configurations (minimising overall waiting times) and their *equity*, i.e. the fairness of service delivery between genders. Traditional facility design has prioritised efficiency or relied on simple space-allocation rules, which typically produce disparities where one user group waits much longer than another.

Service equity is measured by the *gender waiting-time gap*:

$$\Delta W = |W_w - W_m| \quad (1.3)$$

where W_w and W_m are the average waiting times for women and men. The gap is zero when both groups wait the same amount on average.

Balancing efficiency and equity Does improving equity necessarily compromise overall system efficiency? This study examines this trade-off by evaluating layouts and operational policies across both dimensions:

- **Efficiency Objective:** Minimise the overall average waiting time $\bar{W} = \frac{\lambda_w W_w + \lambda_m W_m}{\lambda_w + \lambda_m}$, where λ_w and λ_m are the arrival rates for women and men. In most cases, it is assumed that $\lambda_w = \lambda_m = \lambda/2$. For this case, the expression reduces to the unweighted average $\bar{W} = (W_w + W_m)/2$. Gender proportions are thus not treated as a design variable in the baseline experiments. Robustness to shifts in gender proportion is examined separately in Section 6.8.
- **Equity Objective:** Minimise the gender waiting time gap $\Delta W = |W_w - W_m|$.

The two objectives sometimes coincide: a better-allocated layout can lower the overall wait and shrink the gap at the same time. They can also conflict, and then the design or the queue discipline has to favour one over the other. Chapters 5 and 6 sort out which case applies.

1.4 Experimental design

1.4.1 Restroom layouts

This study evaluates six restroom layouts, following the designs proposed by Van Hautegeem & Rogiest [4]. The layouts vary the allocation and type of facilities (stalls and urinals) between women's and men's sections, covering traditional gender-separated, mixed-use, and fully gender-neutral designs. All layouts respect a common floor-area budget of 20.0 units, with stalls counted as 1.0 unit and urinals as 0.75 units. Because urinals occupy a fractional unit, the achievable totals are 19.5 (Layouts 1, 2, 4) or 20.0 (Layouts 3, 5, 6) under integer counts of facilities. The fixed-area constraint reflects how event and venue planners typically work: the available footprint is decided first, and the question is how to subdivide it.

Figure 1.1 provides a visual overview of all six configurations, illustrating the spatial arrangement and facility allocation for each design.

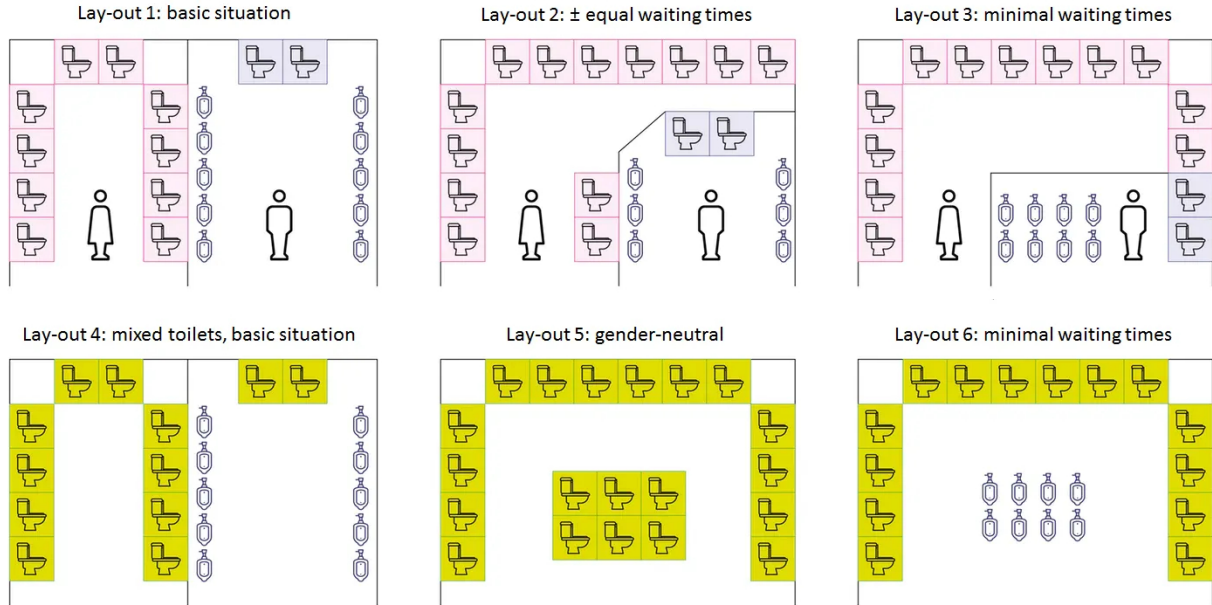


Figure 1.1: Visual overview of the six restroom layouts evaluated in this study. Layouts progress from traditional separated facilities (Layout 1) through optimised separated configurations (Layouts 2 and 3), mixed-use facilities (Layouts 4 and 6), to fully gender-neutral design (Layout 5). Pink icons represent women's facilities, blue icons represent men's facilities, and yellow icons indicate mixed-use stalls. (Adapted from Van Hautegeem & Rogiest [4]).

Table 1.1 presents the detailed specifications for each layout, including the allocation of stalls and urinals, total facility counts, and surface area utilisation.

Each layout reflects a different allocation choice. The names follow Van Hautegeem & Rogiest [4], and the outcomes observed under the MMPP demand of this thesis may differ from those in their study (see Chapter 5).

- **Layout 1 (Baseline):** The traditional equal-space split. With 10 women's stalls and 12 men's facilities (2 stalls + 10 urinals), it is the typical space-based allocation that produces disparate waiting times.
- **Layout 2 (Equal Wait Times):** Optimised for roughly equal waits across genders while keeping the separation. It gives women 13 stalls and reduces men to 2 stalls and 6 urinals, in line with the rule of thumb of 1.5 to 2 women's stalls per men's facility.
- **Layout 3 (Minimum Average Wait):** Optimised for minimum overall mean wait under separation. It allocates 12 women's stalls and 10 men's facilities (2 stalls + 8 urinals).
- **Layout 4 (Mixed Facilities):** 12 shared stalls accessible to either gender, plus 10 men-only urinals. This approximates the conversion of existing separated facilities to mixed use without structural changes.

Layout	Type	Women's Stalls	Men's Stalls	Men's Urinals	Mixed Stalls	Surface Area
Layout 1 <i>Baseline</i>	Separated	10	2	10	n/a	19.5
Layout 2 <i>Equal Wait</i>	Separated	13	2	6	n/a	19.5
Layout 3 <i>Min. Avg. Wait</i>	Separated	12	2	8	n/a	20.0
Layout 4 <i>Mixed Facilities</i>	Mixed	n/a	n/a	10	12	19.5
Layout 5 <i>Gender Neutral</i>	Neutral	n/a	n/a	0	20	20.0
Layout 6 <i>Optimal Mixed</i>	Mixed	n/a	n/a	8	14	20.0

Note: Surface area calculations assume 1.0 unit per stall and 0.75 units per urinal. Mixed/neutral stalls are accessible to all users.

Table 1.1: Detailed specifications of restroom layouts.

- **Layout 5 (Gender Neutral):** 20 shared stalls and no urinals. A single pooled queue guarantees equal waits at the cost of losing the throughput advantage of urinals for men.
- **Layout 6 (Optimal Mixed):** 14 shared stalls and 8 men's urinals. This “festival-style” layout balances shared capacity against the throughput of urinals.

Together, these six layouts span the design space from fully separated to fully shared, and they form the setting for the scheduling policies of Chapter 6.

1.4.2 Simulation parameters

The simulation uses the following parameters, chosen to reflect realistic event conditions:

- **Arrival Process:** Arrivals are generated by a 2-state Markov Modulated Poisson Process (MMPP) [8] that stochastically alternates between two intensity regimes (see Section 1.4.3 for the formal definition):
 - *Calm state:* Combined arrival rate $\lambda_{\text{calm}} = 0.2$ arrivals/second (12 arrivals/minute)
 - *Busy state:* Combined arrival rate $\lambda_{\text{busy}} = 0.4$ arrivals/second (24 arrivals/minute)

Sojourn times in each regime are exponentially distributed, with transition rates:

- Calm $\rightarrow$ Busy: $q_{12} = 1/10$ per minute (mean calm duration = 10 min)

- Busy $\rightarrow$ Calm: $q_{21} = 1/2$ per minute (mean busy duration = 2 min)
- **Service Times:** Service times are modelled as exponentially distributed random variables with gender- and facility-specific mean durations:
 - Women (stalls): $\mu_w^{-1} = 1.5$ minutes (90 seconds)
 - Men (stalls): $\mu_m^{-1} = 1.25$ minutes (75 seconds)
 - Men (urinals): $\mu_u^{-1} = 1.0$ minute (60 seconds)

These values are based on the work of Van Hautegeem & Rogiest [4], who report women's stall occupancy at 1.5 min and men's urinal occupancy at 1.0 min. The men's stall occupancy time of 1.25 min was chosen in consultation with Prof. W. Rogiest, who judged this intermediate value to be more realistic than using the urinal time for stall visits.

- **Temporal Structure:** The simulation uses a two-phase structure [9, 10]:
 - *Warm-up Period:* An initial transient period during which the MMPP runs but performance statistics are not collected, allowing the system to reach steady state before measurement begins. The warm-up duration $t_{\text{warm-up}}$ is common across all six layouts and set to the value required by the slowest-converging layout. The per-layout convergence analysis used to derive it is presented in Chapter 3.
 - *Observation Period:* 90 minutes of post-warm-up operation during which performance metrics are recorded. Performance metrics are recorded for every customer who arrives during the observation window, $t_{\text{warm-up}} \leq a_i \leq t_{\text{warm-up}} + 90$ min, even if the corresponding service completion occurs after the window closes. This convention follows Law [9] and ensures that a service initiated within the observation period is fully accounted for. The MMPP continues its stochastic regime switching throughout both the warm-up and observation windows. Only the recording of waiting times is gated by the warm-up cutoff.
- **Number of replications:** Each layout was simulated with $n = 1000$ independent replications, each with its own random seed [9]. From this point onwards (and consistently across all subsequent chapters), the symbol n denotes the number of independent simulation replications, in the standard statistical sense of sample size. This reuses the letter that appeared earlier as a local index for system occupancy in the worked example of Equation (1.1), but the two meanings never coexist within a single derivation. The count of $n = 1000$ follows Van Hautegeem and Rogiest [4]. It is used here because it gives confidence intervals narrow enough to separate the layouts under comparison.
- **Confidence level:** All confidence intervals are constructed at the 95% level ($\alpha = 0.05$) [9]. For a sample mean $\bar{X}$ with sample standard deviation s over the $n = 1000$ replications, the interval is

$$\text{CI}_{0.95} = \bar{X} \pm t_{n-1, 1-\alpha/2} \cdot \frac{s}{\sqrt{n}}, \quad (1.4)$$

where $t_{n-1, 1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the Student t -distribution with $n - 1$ degrees of freedom (so $t_{999, 0.975} \approx 1.962$).

1.4.3 The MMPP arrival process

The simulation employs a 2-state Markov Modulated Poisson Process (MMPP) to model time-varying arrival intensities. The MMPP captures the stochastic nature of demand fluctuations at real-world events, where busy periods occur at random times and last for random durations (see Figure 1.2).

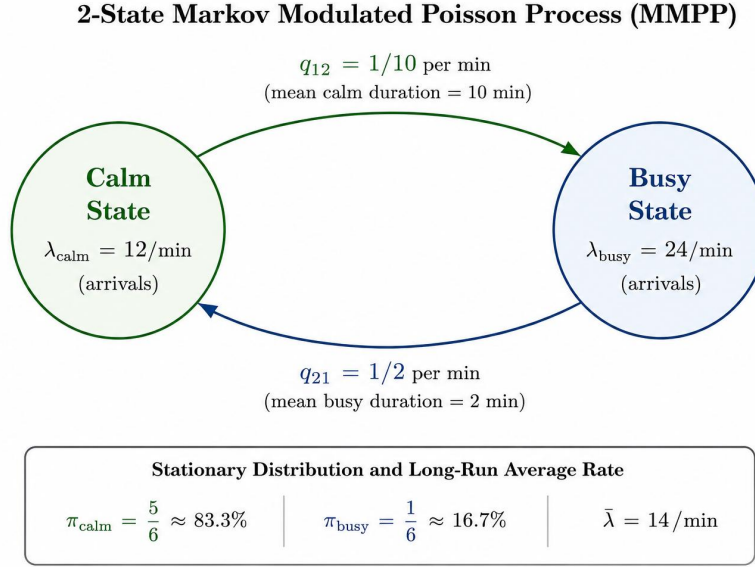


Figure 1.2: Schematic of the two-state Markov Modulated Poisson Process. Arrivals occur at rate $\lambda_{\text{calm}} = 12/\text{min}$ in the calm state and $\lambda_{\text{busy}} = 24/\text{min}$ in the busy state. Transitions occur at rates q_{12} (calm to busy) and q_{21} (busy to calm), with mean sojourn times of 10 min and 2 min respectively. The summary panel reports the stationary distribution $(\pi_{\text{calm}}, \pi_{\text{busy}})$ and the long-run average rate $\bar{\lambda}$.

Formal definition

The MMPP is defined by an underlying continuous-time Markov chain (CTMC) $\{Z(t)\}_{t \geq 0}$ on state space $\mathcal{S} = \{\text{calm}, \text{busy}\}$ with infinitesimal generator matrix:

$$Q = \begin{pmatrix} -q_{12} & q_{12} \\ q_{21} & -q_{21} \end{pmatrix} = \begin{pmatrix} -1/10 & 1/10 \\ 1/2 & -1/2 \end{pmatrix} \quad (\text{rates per minute}) \quad (1.5)$$

Conditional on the current state $Z(t) = s$, arrivals follow a Poisson process with rate λ_s . Collecting the two state-dependent arrival rates into a diagonal rate matrix Λ gives:

$$\Lambda = \text{diag}(\lambda_{\text{calm}}, \lambda_{\text{busy}}) = \text{diag}(12, 24) \quad (\text{arrivals per minute}) \quad (1.6)$$

The stationary distribution $\boldsymbol{\pi} = (\pi_{\text{calm}}, \pi_{\text{busy}})$ satisfies $\boldsymbol{\pi}Q = \mathbf{0}$, giving:

$$\pi_{\text{calm}} = \frac{q_{21}}{q_{12} + q_{21}} = \frac{5}{6}, \quad \pi_{\text{busy}} = \frac{q_{12}}{q_{12} + q_{21}} = \frac{1}{6} \quad (1.7)$$

The initial MMPP state for each replication is drawn from this stationary distribution, which avoids transient bias from a deterministic start in either regime.

The long-run average arrival rate is $\bar{\lambda} = \pi_{\text{calm}}\lambda_{\text{calm}} + \pi_{\text{busy}}\lambda_{\text{busy}} = \frac{5}{6} \cdot 12 + \frac{1}{6} \cdot 24 = 14$ arrivals/min.

Justification and design rationale

The parameter choices reflect a simple intuition: at a concert or festival, the crowd does not arrive at a perfectly steady rate. Most of the time things are calm, but every now and then a rush hits: the opening act ends, people pour out at half-time, or the headliner is about to take the stage. The MMPP captures this: a system that is quiet most of the time but occasionally spikes.

- **How often is it busy?** The system is in the busy state only about 1 in 6 minutes on average ($\pi_{\text{busy}} \approx 16.7\%$). True peak moments are short-lived. The calm state is the norm.
- **How intense is the rush?** During a busy period, arrivals double from 12 to 24 per minute.
- **How long does each period last?** Calm periods last around 10 minutes on average. Busy periods last around 2 minutes. The system typically sees several rush-and-recover cycles during a single 90-minute observation window, giving each layout and scheduling policy a fair chance to show its behaviour under varying load.

The 10 and 2 minute means are illustrative rather than empirically calibrated. They give a calm baseline broken by short, intense bursts, the qualitative pattern at events with scheduled break points such as set changes or intermissions. Smaller fluctuations also appear throughout the event: groups arriving together from public transport, visitors moving between stages, or temporary queues at food stands and bars. An empirical fit of the regime-switching rates to live-event traffic data is left for future work.

Together, the two states create an average cycle of 12 minutes, meaning that roughly seven full cycles occur within the 90-minute observation window. As a result, each simulation replication captures multiple shifts between calm and busy periods, rather than remaining in a single regime for the entire run.

1.5 Thesis outline

The remainder of this thesis is organised as follows. Chapter 2 reviews prior work on the potty parity problem, queueing-theoretic restroom models, priority scheduling, and simulation-based optimisation. Chapter 3 determines the warm-up period required for the simulation to reach steady state, using a replication-based convergence test. Chapter 4 then verifies the simulation outputs against closed-form M/M/c predictions for two reference layouts. Chapter 5 presents

the main simulation results, comparing the six layouts on waiting time and gender equity and assessing robustness to deviations from the 50/50 gender split. Chapter 6 develops four ladies-first priority policies and characterises the trade-off between efficiency and equity. Chapter 7 summarises the findings, discusses limitations, and proposes future research directions. Chapter 8 closes the thesis by relating the contribution to the relevant UN Sustainable Development Goals and reflecting on ethical considerations, including the limits of a binary model.

1.6 Chapter summary

With the simulation model and the statistical protocol now in place, the modelling and experimental choices that shape all subsequent results are summarised below.

- Women consistently wait longer than men at public restrooms: the so-called “potty parity” problem. This thesis uses simulation to study whether smarter facility design and priority policies can close that gap.
- Six layouts and a family of “ladies-first” priority policies are evaluated on two criteria: how long people wait on average, and how fairly that wait is distributed between genders.
- Arrivals are not constant: the simulation alternates between calm periods (12 arrivals/min on average) and short rush spikes (24 arrivals/min), the way demand really moves at a live event.
- Each arriving man chooses the facility type (stall or urinal) that minimises his estimated expected wait under the fluid approximation of Equation (1.1). Once he commits, he remains in the chosen queue without re-evaluating.
- Results are based on 1 000 independent simulation runs per configuration, with 95 % confidence intervals throughout. A warm-up period of 1 778 min, derived in Chapter 3 from a cumulative-mean convergence analysis, is discarded before measurement.

CHAPTER 2

Literature review

2.1 The potty parity problem

Differences in restroom waiting times between women and men at public venues have been documented for decades. Greed [11] was among the first to study the issue systematically and showed that planning authorities in Britain had historically treated public toilet provision as a secondary concern, often overlooked in master-planning documents. Although unintended, the resulting design norm leaves women with slower and less reliable access to facilities. The phrase “potty parity”, though informal, has become the standard term in the academic literature [3].

Anthony and Dufresne [3] identify three causes. A single visit takes longer for women on average, partly because of physiology and partly because they more frequently accompany children or companions with reduced mobility. Building codes also treat an enclosed stall and a urinal as one fixture each, even though urinals have higher throughput. Event demand peaks around intermissions, which steady-state models underestimate [3]. Equal floor space does not guarantee equal service.

Two field studies put numbers on this asymmetry. Baillie et al. [12] observed 120 college students at a public restroom and reported mean dwell times of 178.9s for women and 118.4s for men, a statistically significant difference ($p = 0.001$, $d = 0.34$). Gwynne et al. [13] report a comparable pattern from 502 observations at a North American airport, with female users averaging 166s and male users 136s under uncongested conditions. These absolute durations include secondary activities such as hand-washing, mirror use, and clothing adjustment. They cannot be transferred directly into a queueing model, in which service time is the period during which a fixture is occupied. The female-to-male ratio, however, is of the same order as the ratio implied by the service rates used in the present thesis.

For large event venues such as theatres, concert halls, and stadiums, the International Plumbing Code requires more toilets per occupant in women’s facilities than in men’s [1]. The ratios depend on the type of building, and applying them to older buildings is often difficult. Greed’s later work [2] reframes the issue as one of inclusive urban design, broadening the discussion

beyond gender to include disability access and age-related needs. In a later critique of gender-neutral facilities, Greed [14] argues that simply relabelling restrooms without increasing capacity can worsen competition for stalls and reduce the already limited provision available to women. A review by Lewkowitz and Gilliland [15] reaches a similar conclusion: across 19 studies, equal floor allocation consistently resulted in unequal access in practice.

In the Belgian context, federal workplace legislation requires gender-separated facilities [16]. Public events and temporary venue structures have different regulations, which allow mixed or gender-neutral restroom provision. This makes a comparison of different layouts directly relevant for the large-event setting that motivates this thesis: it quantifies the performance cost of permanent gender separation and the alternatives available where those regulations do not apply.

2.2 Queueing-theoretic models for restroom systems

The standard analytical baseline for this setting is the M/M/ c queue [5, 17]. In the M/M/ c formulation, a Poisson stream of arrivals is served by c identical, parallel servers under exponential service times. Applied to a restroom, the servers represent the available stalls or urinals. The Erlang-C formula then gives the probability that an arriving customer must wait. Via Little’s law, this can be translated directly into the corresponding mean waiting time. Each gender’s facilities can be modelled as a separate M/M/ c system, and comparing the resulting waiting times makes the disparity explicit. The studies reviewed below build on this foundation, varying mainly along three dimensions: whether arrivals are treated as stationary, whether user classes are differentiated, and whether layout flexibility is modelled explicitly.

Huh et al. [18] produced the first quantitative comparison of gender-specific waiting times against the fixture ratios in building codes. They model the men’s and women’s facilities as two independent M/M/ c queues and compare the genders by the probability of waiting longer than one minute. The ratios do not equalise that probability. Their model, however, is stationary, so it cannot reproduce the demand surges typical of large events.

Farajollahzadeh and Hu [19] approach the problem from a different angle with their “utility parity” framework. Rather than looking solely at waiting times, they factor in safety perceptions and gender-identity preferences. Their central result is counterintuitive: converting *all* facilities to gender-neutral reduces overall utility. This happens because the safety disutility (experienced by users who prefer gender-separated facilities) offsets the service-time gains. Converting only a subset of male facilities to shared use avoids this penalty and results in faster service for some users without leaving any user group worse off. The authors summarise this as “a little flexibility dominates full flexibility”. This thesis adopts a narrower scope: it considers only waiting time under stochastic demand, and does not model user preferences or safety perceptions.

Van Hautegeem and Rogiest [4] applied this framework to festival restrooms. Their comparison of several layouts, using both Erlang-C calculations and discrete-event simulation, confirms that equal-space allocation leads to unequal waiting times. That study, however, used a fixed Poisson

arrival rate and did not evaluate priority policies.

All of the $M/M/c$ models above share one simplification: the waiting area is unbounded, and customers never abandon the queue. Real restroom layouts at large events are not. Once the queue is visibly long, some users balk: they refuse to join in the first place. Rahman et al. [20] model this with an $M/M/c/K$ extension, where the balking probability rises sharply as the finite buffer approaches saturation. This thesis keeps the infinite-buffer assumption, which may understate congestion under heavy load. Chapter 7 revisits this.

Bovens and Marcoci [21] simulate both gender-separated and gender-neutral layouts under stationary Poisson arrivals. They find that conversion to gender-neutral facilities reduces women’s average waiting time under most stationary operating conditions, while men’s waiting time remains approximately unchanged or slightly higher, depending on urinal availability and the female-to-male occupancy ratio. Under the time-varying MMPP demand used in this thesis, that conclusion may no longer hold. When arrivals exhibit busy-state surges and urinal throughput is removed, fully gender-neutral layouts can saturate during peaks. Chapter 5 returns to this question.

De Jonghe [22] extends this line of research. Her thesis builds three queueing models that distinguish *sitters* (users who require an enclosed stall) from *uprights* (users who can use either a stall or a urinal). The thesis also models dynamic nudging: a parameter captures the probability that a man who would have chosen a stall is instead steered to a urinal.

Duysens [23] analyses flexible-signage restroom layouts using Markov chain models. The thesis studies two designs that both assume stationary Poisson arrivals. One is a curtain-partitioned layout modelled on the Ziggo Dome concert hall, where stalls can be reassigned between the men’s and women’s sections. The women’s section reduces to an $M/M/c$ queue, while the men’s section is solved as a continuous-time Markov chain over the (stalls, urinals) state. The other is a mixed layout combining women-only stalls, gender-neutral stalls, and urinals, analysed as a three-dimensional continuous-time Markov chain.

The stationary Poisson assumption shared by all of these models matters in the event setting. Visitors do not arrive at a constant rate throughout a large event: crowds thin during performances and surge around intermissions. A common relaxation is the non-homogeneous Poisson process (NHPP), where the arrival rate varies deterministically over time [24]. If the rate function is set to match a scheduled peak, an NHPP can reproduce it. But the busy periods at a large event are not on a fixed schedule: they start and end at random. The Markov-Modulated Poisson Process (MMPP) [8] handles this by letting the arrival rate jump between states, with the timing of each jump drawn from a continuous-time Markov chain. As a result, both when a busy period starts and how long it lasts are random. Fischer and Meier-Hellstern [8] establish its analytical properties and discuss estimation procedures. Numerically exact solutions for the MMPP/M/1 queue, and by extension multi-server variants, are available via quasi-birth-death (QBD) methods [10]. However, they are less suited to the present setting: they are computationally expensive, and cannot handle non-Markovian priority policies such as the Token Bucket developed in Chapter 6. Once a non-Markovian priority policy is layered onto MMPP arrivals,

simulation is the practical evaluation method.

One result from the stationary case still holds, whatever the arrival process. Kleinrock's conservation law [25] says that, for a fixed layout and a fixed arrival and service process, the time-averaged total work waiting in the system is the same under any work-conserving discipline (a rule that never lets a server idle while a compatible customer is waiting). Under those conditions, priority policies can shift waiting time between men and women but cannot reduce the total. This is why the scheduling policies in the next section target the gender gap rather than the overall mean.

2.3 Priority scheduling

Cobham [26] laid the theoretical foundation for priority queues, deriving the mean waiting time for each priority class under non-preemptive scheduling in the $M/G/1$ queue. The implication is that reducing delay for one class necessarily increases delay for another. The extension to multi-server systems is standard [27], and the $M/M/c$ case provides the simplest analytical reference point for the present restroom setting.

Classical priority theory shows how a priority policy redistributes waiting times between classes, but not how to choose a policy that achieves a specific equity outcome. A strict ladies-first policy reduces women's waiting time, but men's waiting time can grow sharply in heavy traffic. Krawczyk et al. [28] formalise this through their queue-intrusion model: when later arrivals can jump ahead of earlier ones, the expected waiting time of those who wait their turn grows rapidly, and its variance scales superlinearly in the intrusion rate. This is why Strict Head-of-Line ladies-first priority produces severe male delays (see Chapter 6), and why the priority policies that follow regulate priority instead of applying it strictly. The design question is how to grant priority in a bounded way, not whether to grant it at all.

A first approach controls the rate at which priority is granted. Turner [29] introduced the leaky bucket to control bursty traffic by limiting the output rate. The related token bucket generalises this rate-control idea by allowing short bursts: tokens accumulate at a steady rate in a finite-capacity bucket, and each event consumes one token. When the bucket is empty, no further events pass until tokens accumulate again. This caps the long-run event rate while still letting short bursts through.

A second approach makes priority conditional on queue state rather than on a rate budget. Bruneel [30] analyses a threshold-based priority policy in which the server gives priority to one customer class only when its queue length exceeds a fixed threshold. Below the threshold, the alternative class receives priority. This activates the priority policy only when one class has accumulated a substantial backlog, leaving normal queue ordering undisturbed when both queues are short.

A third approach raises priority continuously with waiting time. The accumulating priority queue, first proposed by Kleinrock [31] as a delay-dependent discipline, assigns each waiting customer a priority that grows linearly with its waiting time at a class-dependent rate. At

each service opportunity, the customer with the highest accumulated priority is served. Walraevens et al. [32] analyse the tails of the waiting time distributions under this scheme, and show that the trade-off between equity and starvation depends only on the ratio of the two classes' accumulation rates.

Chapter 6 adapts each of these three parameterised policies to the present setting. Strict Head-of-Line is retained separately as the limiting reference case. In the Token Bucket version, a waiting woman is promoted ahead of a man at the head of the shared queue whenever a token is available, and the system falls back to first-come, first-served otherwise. The Threshold version suspends this promotion when the mixed queue exceeds a certain length, reverting to first-come, first-served under heavy load. This is the opposite of Bruneel's design, where heavy load activates priority for the backed-up class. Here, heavy load suspends priority to protect men from starvation. Under the Dynamic Aging policy, women keep a fixed baseline priority and each waiting man's priority rises with the time he has spent in the queue. As mentioned in the previous section, the optimisation target for each parameterised policy is the gender waiting time gap ΔW . Because the model combines MMPP arrivals, these priority policies, and heterogeneous service rates across stalls and urinals, closed-form analysis is out of reach. They are evaluated using discrete-event simulation.

2.4 Discrete-event simulation and warm-up analysis

Discrete-event simulation can represent non-exponential arrivals, heterogeneous servers, and state-dependent routing without the Markovian simplifications used in closed-form models [9, 33]. This matches the model used in the next chapters: MMPP arrivals, gender-dependent service rates across stalls and urinals, a queue-length heuristic for men's stall-versus-urinal choice, and the priority policies introduced in the previous section.

A common methodological issue is initialisation bias. If a simulation starts from an empty system, it does not yet reflect the typical operating regime of a busy venue. The first part of each run is a transient warm-up phase, and including these early observations can bias the performance estimates. Determining where this transient period ends, and removing all output before that point, is a well-studied problem. Law [9] provides a general framework based on cumulative means across replications, while Welch's moving-average procedure for identifying the point of convergence [34] is a standard alternative. Chapter 3 follows Law's cumulative-means framework across 1000 replications per layout and sets a conservative warm-up period valid for all layouts.

2.5 Simulation-based optimisation

Each of the three parameterised priority policies has parameters: the Token Bucket has a token rate and a bucket capacity, Threshold Priority has a queue-length activation level, and Dynamic Aging has an accumulation rate that controls how fast men's priority catches up with women's.

Finding the parameter values that minimise ΔW requires optimising a function that cannot be evaluated analytically and is expensive to compute because each evaluation requires a full simulation run. Grid search wastes evaluations on uninformative points, random search is less sample-efficient, and gradient-based methods cannot be applied because the objective is noisy and lacks a closed-form derivative. This is exactly the setting for Bayesian optimisation [35, 36]. Instead of trying every possible parameter setting, the method learns from the simulation runs that have already been completed. It builds a surrogate model that predicts how good a new setting is likely to be, while also keeping track of how uncertain that prediction is. The next simulation run is then chosen based on both pieces of information: the method tests settings that already look promising, but also explores uncertain regions where a better solution may still be found. The specific Bayesian-optimisation algorithm used in this thesis is the Tree-structured Parzen Estimator (TPE), introduced by Bergstra et al. [37]. TPE is computationally efficient for the small search spaces of this study, with between one and three parameters per policy. It is implemented via the Optuna framework [38].

2.6 Chapter summary

The literature reviewed in this chapter leads to three main conclusions for the present thesis. First, the potty parity problem is structural and can be quantified using queueing theory. Second, in event settings the assumption of stationary Poisson arrivals is inadequate, so an MMPP-driven simulation is the appropriate evaluation framework. Third, priority scheduling can redistribute waiting time between gender groups, but only under the hard constraint that total workload is conserved. To work within this constraint, Chapter 6 evaluates Strict Head-of-Line as a reference case and optimises the parameters of three priority policies (the Token Bucket, Threshold Priority, and Dynamic Aging) through simulation-based optimisation.

Within the literature reviewed in this chapter, no study has combined all of these elements in one framework: separated, mixed, and gender-neutral layouts, compared under the same fixed floor-area budget. Van Hautegeem and Rogiest [4] come closest, but their model uses constant Poisson arrivals and does not include priority scheduling. Two more recent UGent theses on the same topic move in different directions but stop short of the present setting. De Jonghe [22] refines the male population by distinguishing sitters from uprights, and Duysens [23] models flexible-signage layouts as Markov chains. Both, however, keep demand stationary and do not include priority scheduling. This thesis takes the layout comparison from Van Hautegeem and Rogiest and adds two ingredients absent from all three prior studies: time-varying demand and priority policies, so that the results are closer to the conditions found at large events.

CHAPTER 3

Warm-up period and observation window analysis

3.1 Motivation

As introduced in Chapter 2, each replication starts from an empty system and only reaches steady state after an initial transient period. A warm-up cutoff is needed to decide which early observations should be excluded from the performance estimates. If the warm-up period is too short, transient observations still affect the collected data. If it is too long, useful steady-state data is discarded without improving accuracy. This chapter determines where to place the cutoff by identifying the shortest warm-up period after which all layouts have converged.

3.2 Methodology for warm-up period determination

To determine the warm-up period for each layout, the simulation tracks the cumulative mean waiting times across 1 000 replications [9, 39]. The warm-up phase is taken to be over once these means have stabilised, by the criteria stated in Section 3.2.3. From that point on, the remaining data reflects the system's steady-state behaviour.

3.2.1 Cumulative mean waiting time

The cumulative mean waiting time at time t is defined as:

$$\bar{W}(t) = \frac{1}{N(t)} \sum_{i=1}^{N(t)} W_i \quad (3.1)$$

where W_i is the waiting time of the i -th customer served up to time t , and $N(t)$ is the total number of customers served by time t .

The cumulative mean $\bar{W}(t)$ is a running average over all customers served up to time t ,

updated each time a customer enters service. Early on, the queue is still filling and $\bar{W}(t)$ has not yet settled. Once the system reaches steady state, $\bar{W}(t)$ converges to its long-run mean.

This convergence is slow. Every transient observation stays in the average, so $\bar{W}(t)$ only approaches its long-run value as later arrivals dilute the early ones. Because each replication starts empty, the early waits are short and $\bar{W}(t)$ rises toward its long-run mean from below. The convergence criteria below detect this stabilisation, not the actual start of steady state. The resulting cutoff is conservative: it lies after the system has settled, which protects the reported means from any residual transient bias.

A standard alternative is Welch's moving-window plot [34], which avoids the lag of cumulative means but requires choosing a window size and judging the cutoff visually. This thesis instead works directly with the cumulative mean and uses the criteria described below to decide when it has stabilised.

3.2.2 Across-replication aggregation

The warm-up analysis used the $n = 1\,000$ independent replications adopted in Chapter 1. Each replication j produces its own curve $\bar{W}_j(t)$. These are combined across replications into a mean curve and a percentile envelope:

- **Mean curve:** The across-replication average of cumulative means at each time point:

$$\bar{W}_{\text{mean}}(t) = \frac{1}{n} \sum_{j=1}^n \bar{W}_j(t) \quad (3.2)$$

where n is the number of replications.

- **Percentile envelope:** The 10th and 90th percentiles of the cumulative mean distribution across replications, which give the spread:

$$\bar{W}_{p10}(t), \quad \bar{W}_{p90}(t) \quad (3.3)$$

These percentiles are not confidence intervals for the mean. They mark the range within which the central 80% of replication curves fall at each time point. The width reflects between-replication variability from the bursty MMPP arrival process, not transient bias. It remains substantial even after $\bar{W}_{\text{mean}}(t)$ has settled.

3.2.3 Convergence criteria

The warm-up period is identified as the earliest time t^* at which the mean curve $\bar{W}_{\text{mean}}(t)$ has stabilised. Stabilisation is declared at time t when the following three local conditions hold at t and at every grid point after it, up to the end of the analysis horizon t_{end} :

1. **Absolute stability:** The mean value must change only slightly over a Δt -minute look-ahead window. In absolute terms, this change must remain below ϵ_{abs} :

$$|\bar{W}_{\text{mean}}(t + \Delta t) - \bar{W}_{\text{mean}}(t)| \leq \epsilon_{\text{abs}} \quad (3.4)$$

2. **Relative stability:** The same change is also measured relative to the value at the end of the look-ahead window. This relative change must remain below ϵ_{rel} :

$$\frac{|\bar{W}_{\text{mean}}(t + \Delta t) - \bar{W}_{\text{mean}}(t)|}{\bar{W}_{\text{mean}}(t + \Delta t)} \leq \epsilon_{\text{rel}} \quad (3.5)$$

3. **Closeness to final level:** The value at time t must also be close to its final level $\bar{W}_{\text{mean}}(t_{\text{end}})$:

$$|\bar{W}_{\text{mean}}(t) - \bar{W}_{\text{mean}}(t_{\text{end}})| \leq \max(\epsilon_{\text{abs}}, \epsilon_{\text{tail}} \cdot \bar{W}_{\text{mean}}(t_{\text{end}})) \quad (3.6)$$

The threshold on the right-hand side combines a fixed absolute tolerance ϵ_{abs} with a relative tolerance ϵ_{tail} based on the final level. The larger of the two is used. This criterion prevents the algorithm from accepting a locally flat part of the curve that has not yet reached its eventual value.

A post-convergence buffer is added as a final safeguard. At least b minutes of data must remain after the candidate cutoff, so $t_{\text{end}} - t \geq b$. This prevents the algorithm from selecting a point that only appears stable because it lies too close to the end of the analysis horizon.

The specific parameter values used in this study are:

- Look-ahead window: $\Delta t = 30$ minutes
- Absolute tolerance: $\epsilon_{\text{abs}} = 0.01$ minutes
- Relative tolerance: $\epsilon_{\text{rel}} = 1\%$
- Tail tolerance: $\epsilon_{\text{tail}} = 2\%$
- Post-convergence buffer: $b = 90$ minutes ($= 3\Delta t$). After the candidate cutoff, at least three look-ahead windows must remain. This gives the convergence check enough data to confirm that the cumulative mean stays stable beyond the candidate point.

These values are specifically chosen for this study and are not universal constants. The 30-minute look-ahead window smooths out short-run fluctuations, but is still short compared with the full analysis horizon. The relative and tail tolerances are chosen to match the small amount of noise that remains in $\bar{W}_{\text{mean}}(t)$ with $n = 1\,000$ replications. They are also smaller than the waiting-time differences that matter when comparing layouts. The tail tolerance is slightly looser than the relative one. It has to cover the entire remaining horizon, while the relative tolerance only checks one look-ahead window at a time.

The absolute tolerance ϵ_{abs} is used in two places. First, in the absolute-stability criterion, it limits how much $\bar{W}_{\text{mean}}$ may change over the look-ahead window. This mainly matters for slow subsystems, where $\bar{W}_{\text{mean}} > 1$ min. Second, in the closeness criterion, it provides a minimum threshold. This prevents the threshold from becoming too small when the final mean waiting time is very low. This matters only when $\bar{W}_{\text{mean}}(t_{\text{end}}) < 0.5$ min, which occurs for the fastest-service subsystems.

A subsystem is called *stable* when its server utilisation $\rho = \lambda/(c\mu)$ stays below one, where λ is the arrival rate, c the number of servers, and μ the service rate per server. The quantity ρ is the long-run fraction of time each server is busy. A stable subsystem ($\rho < 1$) has a finite long-run mean queue length. An overloaded subsystem ($\rho \geq 1$) grows without bound. Chapter 4 gives the full derivation.

The analysis horizon t_{end} was chosen adaptively per layout. Each layout was first run with $t_{\text{end}} = 500$ minutes. If any stable subsystem had not converged, t_{end} was extended by 500 minutes and the analysis repeated until all stable subsystems had converged.

The post-convergence buffer makes sure that enough data remains after the candidate cutoff for the convergence check to be meaningful. The procedure as a whole follows the standard warm-up deletion approach for steady-state simulation [9, 39].

3.3 Results of warm-up period analysis

The warm-up analysis was conducted separately for women's and men's subsystems across all layouts. Figures 3.1 through 3.6 show, for each layout, the mean curve $\bar{W}_{\text{mean}}(t)$ from simulation start to the end of the analysis horizon, together with the 10th–90th percentile envelope and the recommended cutoff. Each figure contains two panels: the women's subsystem (left) and the men's subsystem (right). In every panel, the solid line is the across-replication mean, the shaded region is the 10th–90th percentile range of cumulative means across 1 000 replications, and the vertical green line marks the identified warm-up cutoff.

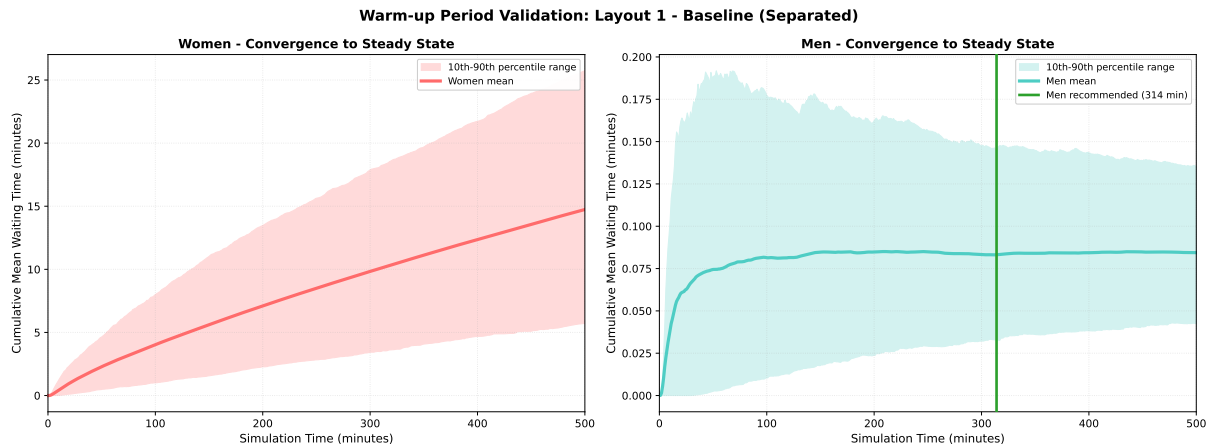


Figure 3.1: Warm-up period validation for Layout 1 (Baseline configuration). The women's subsystem is overloaded ($\rho_{\text{women}} > 1$) and never converges. No cutoff line is shown in the left panel.

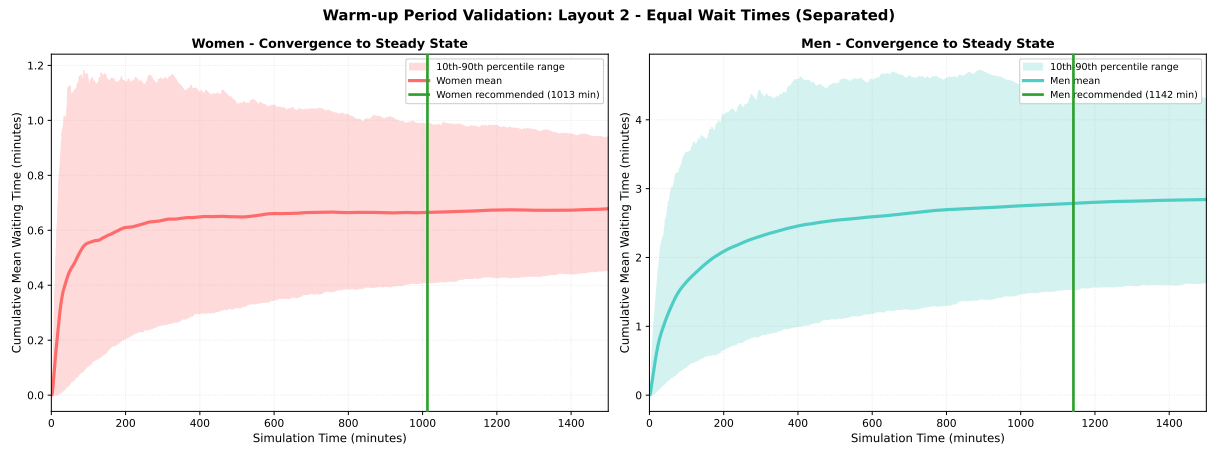


Figure 3.2: Warm-up period validation for Layout 2 (Equal Wait Times).

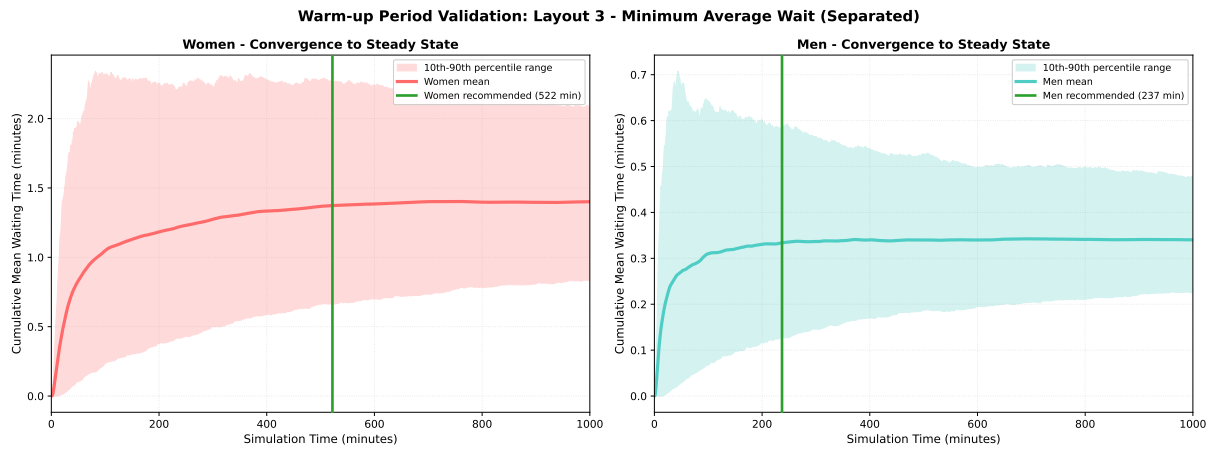


Figure 3.3: Warm-up period validation for Layout 3 (Minimum Average Wait).

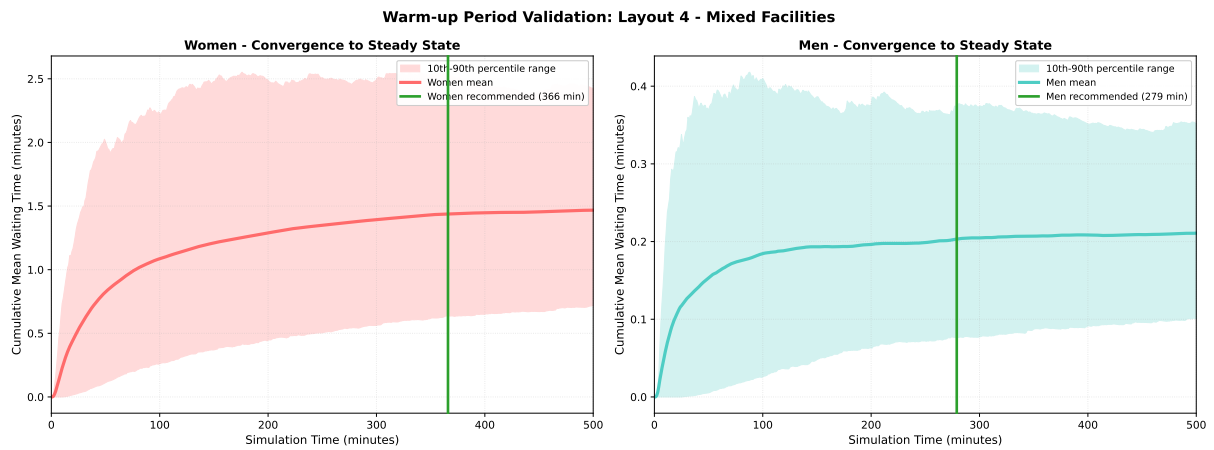


Figure 3.4: Warm-up period validation for Layout 4 (Mixed Facilities).

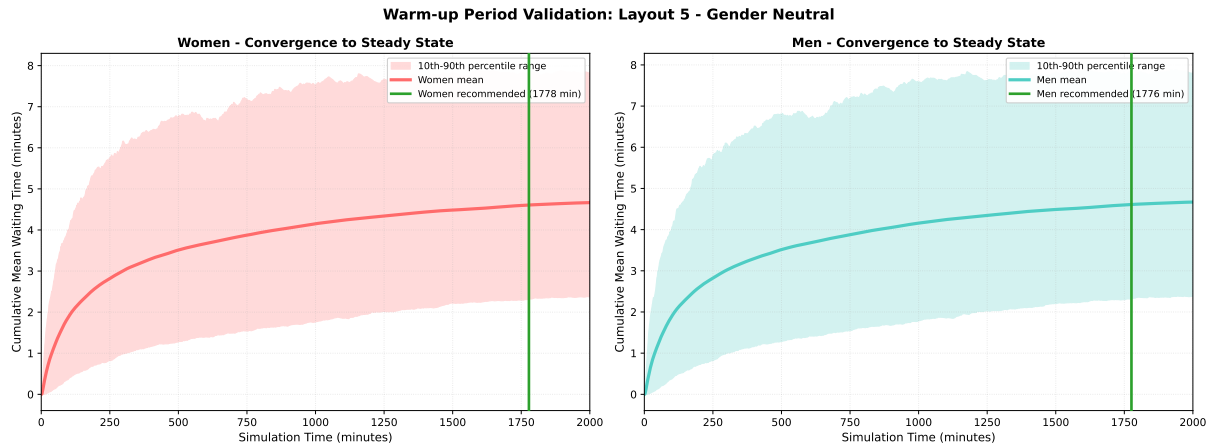


Figure 3.5: Warm-up period validation for Layout 5 (Gender Neutral). This layout sets the binding global warm-up of 1 778 minutes, the latest cutoff across all layouts.

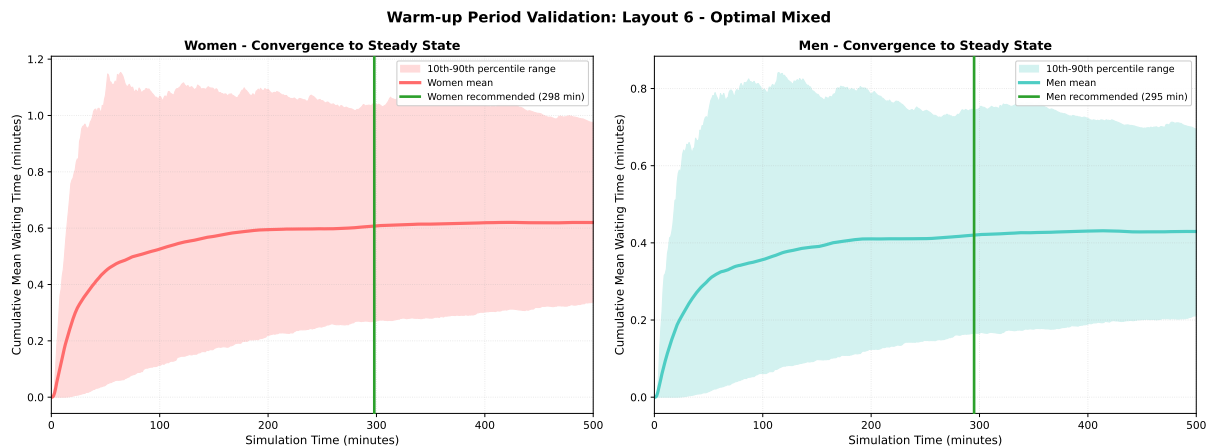


Figure 3.6: Warm-up period validation for Layout 6 (Optimal Mixed). The 298-minute cutoff is the shortest across all layouts.

Table 3.1 summarises the recommended warm-up periods determined for each layout.

Layout	Women's Convergence (min)	Men's Convergence (min)	Recommended Warm-up (min)
Layout 1 [†]	not stable	314	314
Layout 2	1 013	1 142	1 142
Layout 3	522	237	522
Layout 4	366	279	366
Layout 5	1 778	1 776	1 778
Layout 6	298	295	298
Global warm-up period adopted:			1 778 min

Note: The global warm-up period is $t_{\text{warm-up}} = \max_{\ell} t_{\ell}^*$, where ℓ ranges over layouts and the maximum is taken over all stable subsystems ($\rho < 1$).

[†]Layout 1: the women's subsystem is overloaded ($\rho_{\text{women}} = 1.050 > 1$) so its queue does not converge. The recommended cutoff reflects only the men's convergence time (314 min).

Table 3.1: Recommended warm-up periods by layout.

3.3.1 Interpretation and implications

Convergence times range from 298 minutes (Layout 6) to 1 778 minutes (Layout 5). Higher or more imbalanced utilisation drives slower convergence. Layout 5 is the slowest because its shared stall pool sits very close to saturation, while Layout 6 is the fastest because its two subsystems run well below saturation.

In separated layouts, women's and men's subsystems converge at different times. The binding constraint within a layout is consistently the gender with higher utilisation. This matches the standard M/M/c result that queues near saturation have slow-converging transients [9].

Layout 1's women's subsystem is unstable ($\rho_{\text{women}} = 1.050 > 1$) and is excluded from steady-state analysis. It does not constrain the warm-up period. Among the feasible layouts, Layout 5 is the binding case, with a convergence time of 1 778 minutes. Adopting this global maximum means every feasible configuration (Layouts 2–6) enters its post-warm-up observation window in steady state. Layout 1 is still included in the comparison to illustrate the consequence of equal-space allocation under realistic demand. The cost is a longer simulation runtime per replication. In exchange, a single experimental protocol applies to every layout and removes initialisation bias.

3.4 Observation period validation

A correct warm-up period is necessary but not sufficient for valid performance estimation. The 90-minute observation window that follows must itself be in steady state. Three signals confirm this: the cumulative mean stays approximately constant, batch means (waiting times averaged over non-overlapping 30-minute windows) fluctuate around a stable value, and queue-length profiles are stationary.

Figures 3.7 through 3.12 provide this confirmation for each layout. The post-warm-up observation window is highlighted in yellow in each panel. Each figure shows three panels:

- **Left:** The left panel shows the across-replication cumulative mean waiting time from simulation start (time 0), with the 10th–90th percentile band. A vertical line marks the warm-up cutoff for this specific layout. A flat mean curve through the yellow post-warm-up observation window confirms that no initialisation bias remains.
- **Middle:** The full simulation timeline is split into non-overlapping 30-minute intervals. For each interval and each replication, the simulator averages the waiting times of customers who arrived during that interval. This average is the interval’s batch mean. Each customer counts toward the interval of their arrival, even if their service starts or completes later in the run, or after the simulation’s nominal end. Without this rule, customers still waiting at the simulation’s end would be excluded, making the final intervals biased. Each replication contributes one batch mean per interval, plotted at the interval’s centre. Across the 1000 replications, the thick coloured line at each interval is the mean of the 1000 batch values, and the shaded band is their 10–90th percentile (separately for women and men). The curve rises during the warm-up phase as the queue fills, then plateaus once the system reaches steady state. A flat plateau through the yellow post-warm-up observation window confirms that no systematic drift remains.
- **Right:** The right panel shows queue length evolution across the full simulation (across-replication mean with 10th–90th percentile band). The two curves count the total number of women and men currently waiting, summed across every queueing location they have access to. For mixed layouts (4 and 6) the men’s count includes men in the urinal queue together with men in the shared stall queue. The queue lengths rise as the system fills during the warm-up phase, then settle to a steady-state distribution. A stationary profile through the yellow post-warm-up observation window confirms that the queue is not still building up.

For Layouts 2–6, all three checks pass: the cumulative means are flat, the batch means do not drift, and the queue lengths are stable. During the 90-minute observation window, the system is in steady state. Layout 1 is the exception. Its women’s subsystem is unstable ($\rho_{\text{women}} > 1$), so none of the three checks pass: the cumulative mean rises, the batch means drift upward, and the queue length keeps growing.

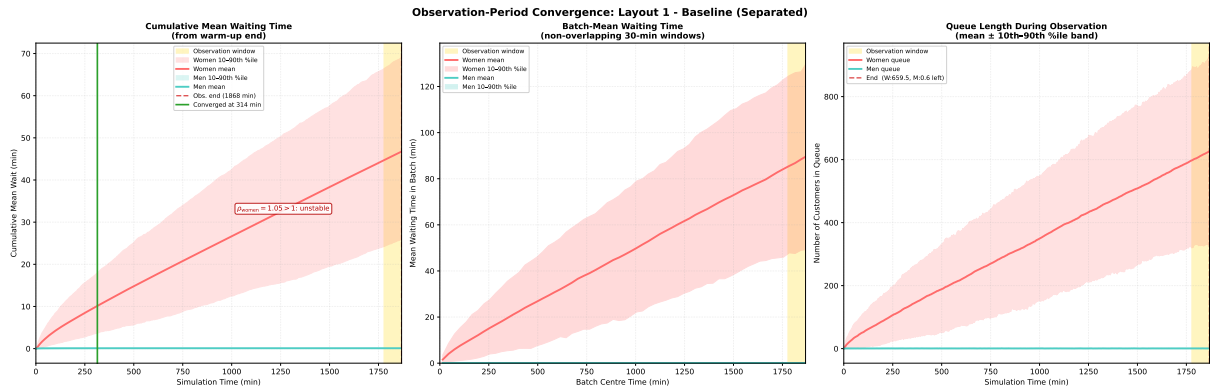


Figure 3.7: Observation-period validation for Layout 1 (Baseline configuration). The women's subsystem is unstable ($\rho_{\text{women}} > 1$), so none of the three checks pass: the cumulative mean rises, the batch means drift upward, and the queue length keeps growing. No steady state is reached.

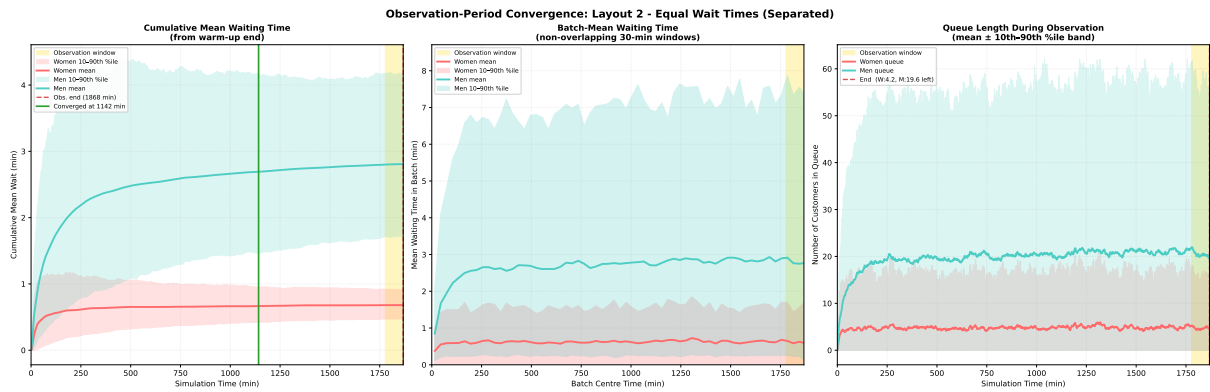


Figure 3.8: Observation-period validation for Layout 2 (Equal Wait Times). All three checks pass.

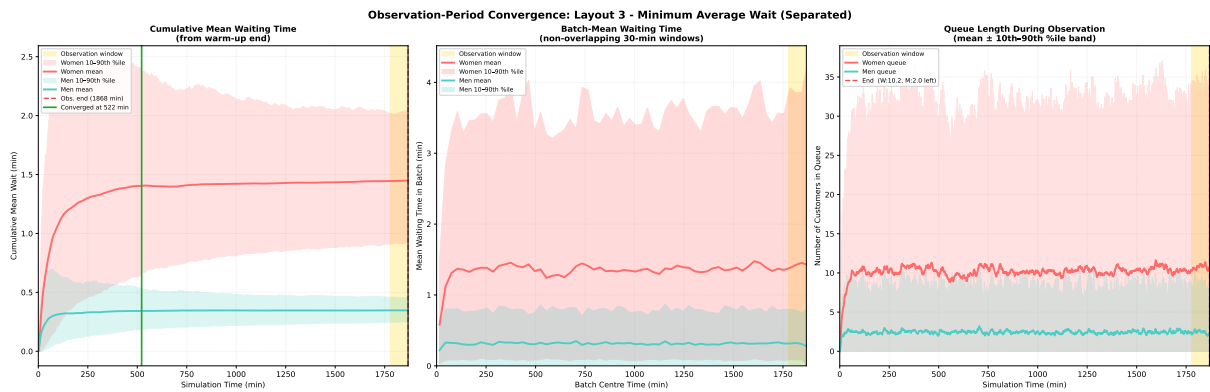


Figure 3.9: Observation-period validation for Layout 3 (Minimum Average Wait). All three checks pass.

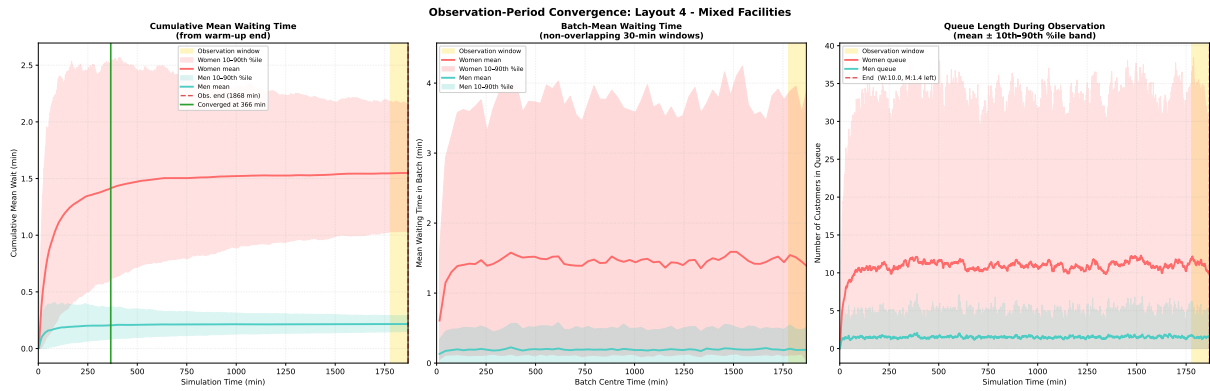


Figure 3.10: Observation-period validation for Layout 4 (Mixed Facilities). All three checks pass.

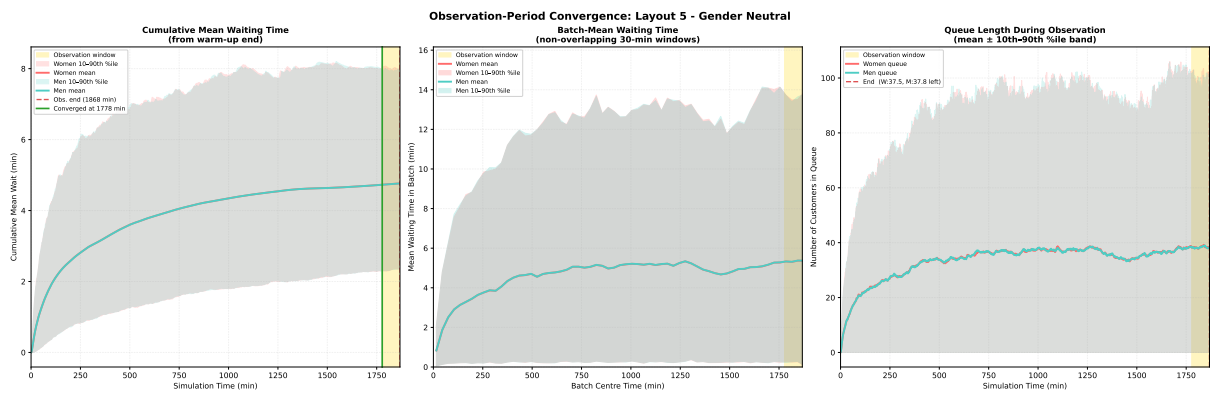


Figure 3.11: Observation-period validation for Layout 5 (Gender Neutral). All three checks pass.

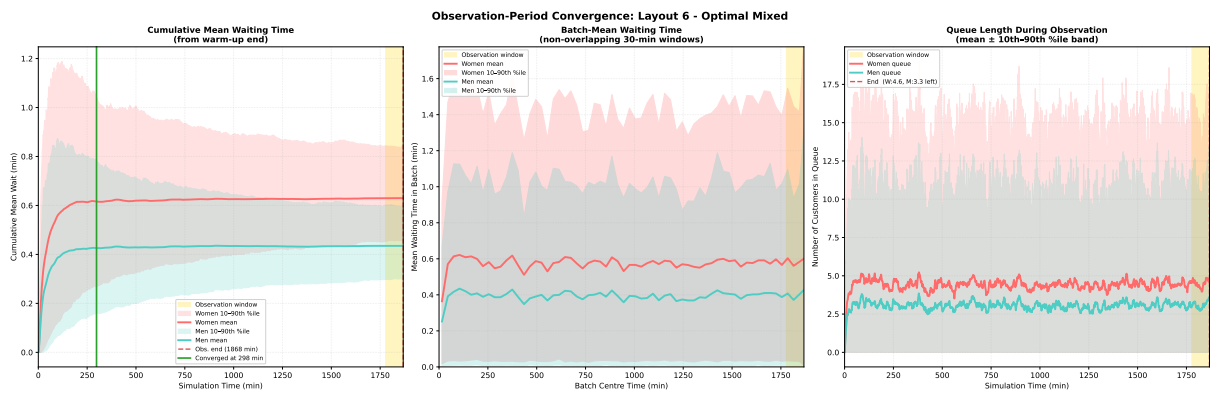


Figure 3.12: Observation-period validation for Layout 6 (Optimal Mixed). All three checks pass. The narrow percentile bands reflect this layout's low utilisation.

3.5 Adopted experimental protocol

All feasible layouts share a single warm-up duration, set to the largest layout-specific convergence time: $t_{\text{warm-up}} = \max_{\ell} t_{\ell}^*$. The following protocol is applied to each of the $n = 1\,000$ replications:

1. Discard all observations collected during the warm-up period $[0, t_{\text{warm-up}}]$.
2. Collect statistics for every customer whose arrival time falls in the post-warm-up observation window $[t_{\text{warm-up}}, T]$:

$$T = t_{\text{warm-up}} + t_{\text{observation}} = 1\,778 \text{ minutes} + 90 \text{ minutes} = 1\,868 \text{ minutes}$$

where T is the end of the observation window and $t_{\text{observation}}$ is its length.

Experiments outside the baseline 50/50 demand mix (the 40/60 and 60/40 gender-split robustness checks) use a fixed warm-up of 3 000 minutes instead of per-scenario convergence detection. The same bound applies to all priority-policy evaluations under those mixes.

The 3 000-minute bound sits well above the 1 778-minute baseline maximum, leaving room for the higher subsystem utilisations that asymmetric demand creates. Detecting convergence separately for every variant would be computationally expensive without changing the comparisons. The fixed bound avoids that cost.

Every reported metric is a steady-state measurement, free of initialisation bias. The MMPP keeps switching throughout the run, as defined in Section 1.4.3. The cutoff only determines which observations are recorded. It does not change the simulation itself.

3.6 Chapter summary

This chapter establishes the experimental protocol used in the rest of the thesis: a 1 778-minute warm-up followed by a 90-minute observation window, with stationarity verified per layout.

- The convergence criteria of Section 3.2.3 combine absolute, relative, and tail tolerances with a post-convergence buffer. They detect when the cumulative mean curve has stabilised per layout and per gender.
- Convergence times range from 298 min (Layout 6) to 1 778 min (Layout 5). Higher utilisation drives slower convergence, and women's and men's subsystems within a layout converge at different rates.
- The global warm-up of 1 778 min, set by Layout 5 (Gender Neutral), applies uniformly to all six layouts.
- Observation-period validation passes for Layouts 2–6: cumulative means are flat, batch means do not drift, and queue lengths are stationary. Layout 1's women's subsystem remains unstable, as expected.

CHAPTER 4

Analytical verification using queueing theory

4.1 Goal of verification

Chapter 5 will compare the six layouts under the full MMPP arrival process (Section 1.4.3). Before that, this chapter checks the simulation engine against analytical benchmarks.

The MMPP arrival process is too complex for a closed-form solution. However, in the special case where the system is *always* in the calm state, the layouts collapse to standard M/M/ c queues, subject to two simplifications: a capacity-weighted mean service rate for heterogeneous server pools (Section 4.2.2), and deterministic routing of men to urinals in mixed layouts to decouple the gender queues (Section 4.3). If the simulation is correct, it should reproduce the M/M/ c predictions in that special case. The verification strategy is:

1. Run the simulator in verification mode (Section 4.3).
2. Compute the analytical M/M/ c benchmarks (Erlang C formula) (by hand).
3. Compare the simulation output to the benchmarks.

Layout 1 (women's queue: M/M/10 with homogeneous stalls) and Layout 5 (M/M/20 with a pooled queue) can be worked out by hand, which gives a direct first check on the simulation.

4.2 Theoretical foundation: M/M/ c queues

We model the restroom system as an M/M/ c queue. This is standard Kendall notation [5] in queueing theory [17, 25] where:

- **M** (Markovian Arrivals): Customers arrive according to a Poisson process with rate λ .
- **M** (Markovian Service): Service times are independent and identically distributed (i.i.d.) exponential random variables with rate μ .

- c (Servers): There are c parallel servers (toilet stalls or urinals).

The main performance measures are the server utilisation ρ , the probability that an arriving customer must wait P_W , and the expected waiting time in the queue W_q .

4.2.1 The Erlang C formula

Define the load $a = \lambda/\mu$ and the per-server utilisation $\rho = a/c = \lambda/(c\mu)$. The system is stable if and only if $\rho < 1$. Under stability, the Erlang C formula [27] gives the probability that all c servers are occupied and an arriving customer must join the queue:

$$P_W = C(c, a) = \frac{\frac{a^c}{c!} \cdot \frac{1}{1-\rho}}{\sum_{k=0}^{c-1} \frac{a^k}{k!} + \frac{a^c}{c!} \cdot \frac{1}{1-\rho}} \quad (4.1)$$

Using this probability, the expected waiting time in the queue is:

$$W_q = \frac{P_W}{c\mu(1-\rho)} \quad (4.2)$$

4.2.2 Capacity-weighted average service rate (heterogeneous servers)

Standard M/M/c theory assumes all servers are identical. However, the men's restrooms in separated layouts (Layouts 1–3) contain two distinct server types with different mean service times: urinals at $\mu_u = 1.0/\text{min}$ and stalls at $\mu_m = 0.8/\text{min}$ (Chapter 1). To apply M/M/c theory to these heterogeneous facilities, we compute a capacity-weighted average service rate:

$$\bar{\mu} = \frac{c_{\text{urinals}} \cdot \mu_u + c_{\text{stalls}} \cdot \mu_m}{c_{\text{total}}} \quad (4.3)$$

where $c_{\text{total}} = c_{\text{urinals}} + c_{\text{stalls}}$. The utilisation for the men's queue is then:

$$\rho_{\text{men}} = \frac{\lambda_{\text{men}}}{c_{\text{total}} \cdot \bar{\mu}} \quad (4.4)$$

This formula is exact for homogeneous servers but biased upward for heterogeneous ones. With heterogeneous servers the per-customer service-time distribution has higher variance than exponential at the same mean (a two-component hyperexponential¹ with squared coefficient of variation $C_s^2 > 1$). Higher service-time variance produces longer mean waits in multi-server queues, so the M/M/c approximation under-predicts W_q . Section 4.6 quantifies the resulting residuals.

¹A two-component hyperexponential distribution is a mixture of two exponential distributions: with probability p a service time is drawn from $\text{Exp}(\mu_1)$ and with probability $1 - p$ from $\text{Exp}(\mu_2)$ [17].

4.3 Methodology: reducing to known models

M/M/ c theory requires a *constant* Poisson arrival rate and cannot handle the time-varying MMPP (Section 1.4.3) directly. Numerically exact solutions for the MMPP/M/ c queue are available via quasi-birth-death (QBD) methods [10], though their computational cost makes simulation the practical choice here. To create a meaningful analytical benchmark, we run the simulation in a simplified *verification mode* that reduces each layout to a known M/M/ c model. Verification mode consists of three modifications:

1. **Constant arrival rate (calm-state-only mode).** The MMPP is disabled. Each gender arrives at the calm-state rate, $\lambda_w = \lambda_m = 6/\text{min}$, and the system reduces to a standard steady-state M/M/ c queue.
2. **Simplified routing for mixed layouts.** In mixed layouts (Layouts 4 and 6), the rational-agent rule (Equation (1.1)) lets men choose between shared stalls (also used by women) and men-only urinals. Because the choice depends on current queue lengths, it correlates the women's and men's queues through the shared stall pool, ruling out independent M/M/ c analysis. In verification mode the choice is removed: women use stalls only, men use urinals only, giving two independent M/M/ c queues solvable in closed form.
3. **Unchanged separated and neutral layouts.** In separated layouts (Layouts 1–3), stall pools are already gender-specific, so the women's and men's queues are independent by design. Men still route between stalls and urinals via shortest-queue, but since this occurs entirely within the men's subsystem, it is handled by the capacity-weighted average $\bar{\mu}$ (Section 4.2.2) rather than by a routing simplification. Layout 5 uses a single pooled queue, directly described by one M/M/ c model.

Table 4.1 summarises the analytical model applied to each layout type.

4.3.1 Verification-mode runtime

The verification runs use a 30 000-min observation window and 1 000 replications. This window is longer than the original 1 000-min setting, which was rejected after it produced an unexplained +1.8% residual for Layout 6 men.

Appendix B reports the convergence study across observation windows. At $T = 30\,000$ the Layout 6 men residual is essentially zero, so the +1.8% at $T = 1\,000$ was finite-window noise. The narrower CIs also reveal real $\sim 0.5\%$ residuals on Layout 5, which the wider 1 000-min CI had hidden.

4.4 Worked examples: Layouts 1 and 5

We work through two layouts. Layout 1 has separate queues for women (exact M/M/10) and men (approximate M/M/12 via the capacity-weighted mean). Layout 5 has a single pooled

Layout type	Layouts	Analytical model	Routing assumption
Separated	1, 2, 3	Two independent M/M/c	Women → stalls; Men → stalls + urinals
Mixed	4, 6	Two independent M/M/c	Women → stalls; Men → urinals only
Neutral	5	Single M/M/c	All users → pooled stalls

Table 4.1: Analytical model and routing assumptions for each layout type.

queue, approximated as M/M/20 via $\bar{\mu} = 1/E[S]$ (justified below).

4.4.1 Layout 1: women's queue (M/M/10)

Layout 1 assigns $c = 10$ stalls to women. Each stall has mean service time $\mu_w^{-1} = 1.5$ min, so $\mu_w = \frac{2}{3} \text{ min}^{-1}$. With the calm-state arrival rate:

$$\lambda_w = 6 \text{ /min}, \quad a = \frac{\lambda_w}{\mu_w} = \frac{6}{2/3} = 9, \quad \rho = \frac{a}{c} = \frac{9}{10} = 0.9$$

Step 1: Compute the partial sums. The Erlang C formula (Eq. 4.1) requires the terms $\frac{a^k}{k!}$ for $k = 0, \dots, c - 1$:

k	$9^k/k!$	cumulative sum
0	1.000	1.000
1	9.000	10.000
2	40.500	50.500
3	121.500	172.000
4	273.375	445.375
5	492.075	937.450
6	738.113	1675.563
7	949.002	2624.564
8	1067.627	3692.191
9	1067.627	4759.818

Step 2: Compute P_W (Erlang C).

$$\frac{a^c}{c!} = \frac{9^{10}}{10!} = 960.864, \quad \frac{a^c}{c!} \cdot \frac{1}{1 - \rho} = 960.864 \times 10 = 9608.643$$

$$P_W = \frac{9608.643}{4759.818 + 9608.643} = \frac{9608.643}{14368.461} = 0.6687$$

Step 3: Compute W_q .

$$W_q = \frac{P_W}{c\mu(1 - \rho)} = \frac{0.6687}{10 \times \frac{2}{3} \times 0.1} = \frac{0.6687}{0.6667} = \boxed{1.003 \text{ min} \approx 60.2 \text{ sec}}$$

This is among the highest waiting times across all layouts, driven by the near-saturation utilisation of $\rho = 0.9$.

4.4.2 Layout 1: men's queue (approximate M/M/12)

The men's side of Layout 1 has $c_{\text{urinals}} = 10$ urinals ($\mu_u = 1.0/\text{min}$) and $c_{\text{stalls}} = 2$ stalls ($\mu_m = 0.8/\text{min}$), for a total of $c = 12$ servers. The two server types have different service rates, so the men's subsystem is not strictly an M/M/ c queue. The capacity-weighted approximation (Eq. 4.3) collapses it to a homogeneous M/M/12 with mean service rate

$$\bar{\mu} = \frac{10 \times 1.0 + 2 \times 0.8}{12} = \frac{11.6}{12} = 0.967/\text{min}.$$

This is the same approximation that produces the residuals analysed in Section 4.6. For Layout 1 men the load is low enough that the resulting bias is invisible against the confidence interval:

$$a = \frac{6}{0.967} = 6.207, \quad \rho = \frac{6.207}{12} = 0.517.$$

Applying the Erlang C procedure with 12 partial-sum terms gives $P_W = 0.028$ and:

$$W_q = \frac{0.028}{12 \times 0.967 \times 0.483} = 0.005 \text{ min} \approx 0.3 \text{ sec}.$$

Men's mean wait is below one second and the queue is rarely non-empty. The capacity imbalance is clear: women wait over a minute while men wait under a second.

4.4.3 Layout 5: gender-neutral pooled queue (M/M/20)

Layout 5 pools all 20 stalls into a single queue for both genders. Since men and women arrive in equal proportions but have different (exponential) service times, the service-time distribution is technically a two-component hyperexponential (Section 4.2.2). Layouts 1–3 had heterogeneous servers (urinals and stalls) drawing from a common stream of male customers, so Equation 4.3 averaged the server rates. Layout 5 is the mirror case: one homogeneous stall pool serving two customer classes (women and men). The natural mean service rate is then the reciprocal of the per-customer expected service time:

$$\bar{\mu} = \frac{1}{E[S]} = \frac{1}{0.5 \times 1.5 + 0.5 \times 1.25} = \frac{1}{1.375} \approx 0.7273/\text{min}$$

The total arrival rate is $\lambda = 12/\text{min}$ (both genders), so:

$$a = \frac{12}{0.7273} = 16.5, \quad \rho = \frac{16.5}{20} = 0.825$$

Applying the same Erlang C procedure as Layout 1 (extended to 20 partial-sum terms) gives:

$$P_W = 0.316, \quad W_q = \frac{0.316}{20 \times 0.7273 \times 0.175} = \boxed{0.124 \text{ min} \approx 7.5 \text{ sec}}$$

Although the utilisation is relatively high at $\rho = 0.825$, the large server count keeps waiting times low. This is a feature of M/M/ c at moderate-to-high ρ : as long as c is large enough that $\rho < 1$ with room to spare, the wait stays short even when each server is busy most of the time.

4.5 Analytical predictions for all layouts

Applying the same M/M/c methodology to all six layouts gives the results in Table 4.2. These are the exact theoretical benchmarks against which the simulation is verified.

Layout	Women's queue				Men's queue			
	c	μ	ρ	W_q (sec)	c	$\bar{\mu}$	ρ	W_q (sec)
1 – Baseline ^a	10	0.667	0.900	60.19	12	0.967	0.517	0.30
2 – Equal Wait ^a	13	0.667	0.692	3.54	8	0.950	0.789	16.32
3 – Min. Avg. Wait ^a	12	0.667	0.750	7.98	10	0.960	0.625	2.09
4 – Mixed ^b	12	0.667	0.750	7.98	10	1.000	0.600	1.52
5 – Gender Neutral ^c	$c = 20, \bar{\mu} = 0.727, \rho = 0.825, W_q = 7.45$ sec				(single pooled queue)			
6 – Optimal Mixed ^b	14	0.667	0.643	1.61	8	1.000	0.750	10.71

Service rates ($\mu, \bar{\mu}$) in min^{-1} . ^a Men: $\bar{\mu}$ via Eq. 4.3 (heterogeneous servers). ^b Verification mode: women $\rightarrow$ stalls, men $\rightarrow$ urinals ($\mu_u = 1.0/\text{min}$). ^c $\bar{\mu} = 1/(0.5\mu_w^{-1} + 0.5\mu_m^{-1})$.

Table 4.2: M/M/c analytical predictions at the calm-state arrival rate ($\lambda = 12/\text{min}$, $\lambda_g = 6/\text{min}$ per gender).

Observations.

- **Layout 1 (Baseline)** exhibits the most extreme gender disparity: women wait ≈ 60 sec while men wait < 1 sec at $\rho_w = 0.90$. Even a small increase in traffic would push the women's queue past the stability limit.
- **Layout 2 (Equal Wait)** redistributes space from men to women (13 stalls vs. 8 men's facilities). The women's wait drops to 3.5 sec while men's rises to 16.3 sec. The reallocation overshoots equality at the analytical level: instead of equal waits, women now wait considerably less than men.
- **Layout 3 (Min. Avg. Wait)** balances the two queues to minimise $\bar{W} = (7.98 + 2.09)/2 \approx 5.0$ sec, outperforming both Layouts 1 and 2 on overall delay.
- **Layout 5 (Gender Neutral)** reaches $W_q \approx 7.5$ sec at $\rho = 0.825$. The pooling effect of a single M/M/20 queue cuts individual waits compared with splitting the same servers across two gender-separated queues.
- **Layouts 4 and 6 (Mixed)** route men exclusively to faster urinals ($\mu = 1.0/\text{min}$) in verification mode. Layout 4 (12 mixed stalls, 10 urinals) gives 8.0 sec for women and 1.5 sec for men. Layout 6 (14 mixed stalls, 8 urinals) gives 1.6 sec for women and 10.7 sec for men, the lowest women's wait of any layout under the calm-state arrival rate.

4.6 Verification results

Each layout was simulated 1 000 times in verification mode and the mean waiting times were compared against the analytical M/M/c predictions. Throughout this section, W_q^{ana} denotes the analytical Erlang-C prediction and $\hat{W}_q$ the corresponding simulation estimate. Figure 4.1 presents the results. Table 4.3 lists the simulation means alongside the analytical benchmarks and the relative residuals.

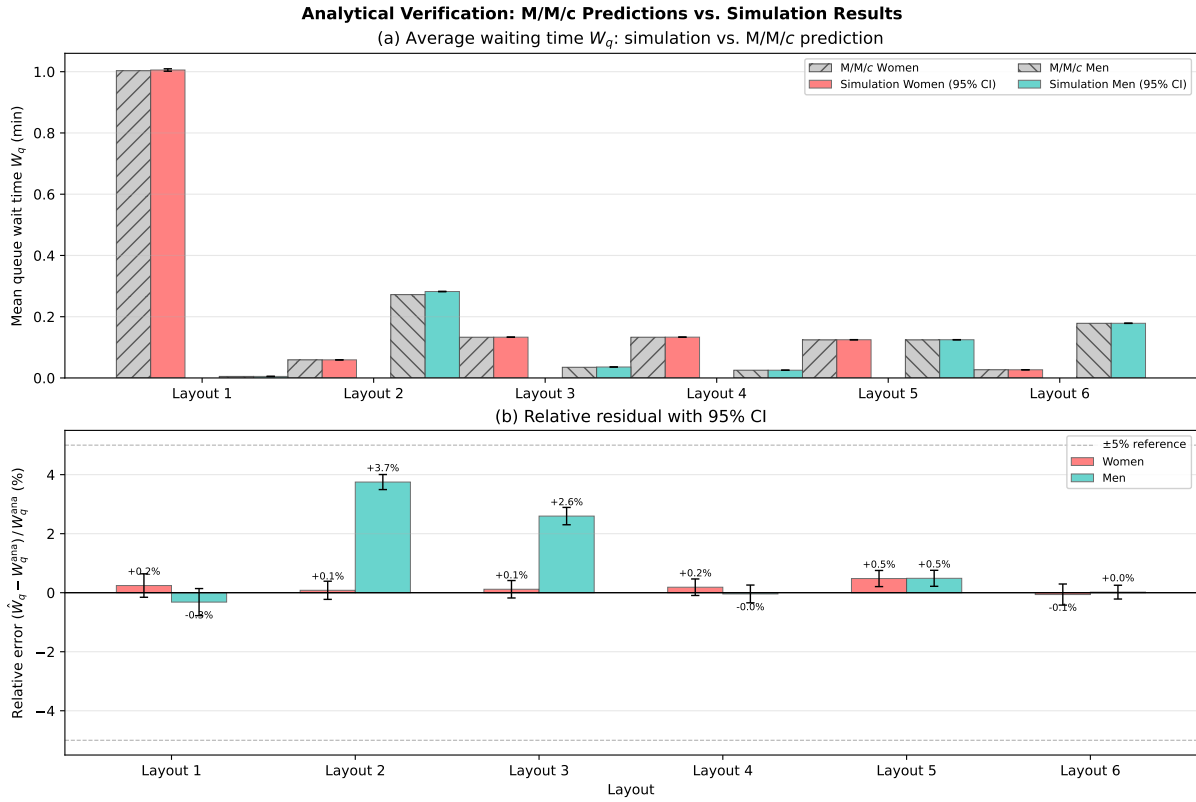


Figure 4.1: Verification at $T = 30\,000$ min: simulated W_q (coloured bars, 95% CI error bars) against analytical M/M/c predictions (grey hatched bars). The top panel shows mean waiting times. The bottom panel shows the relative residual $(\hat{W}_q - W_q^{\text{ana}})/W_q^{\text{ana}}$ in percent, with dashed lines marking the $\pm 5\%$ reference band.

4.6.1 Discussion of results

The verification criterion is that the analytical W_q^{ana} falls within the 95% confidence interval of the simulated mean. The $\pm 5\%$ band shown alongside is a reference scale for the size of any deviation. Eight of twelve queue–gender combinations meet the criterion. The four exceptions (Layouts 2 and 3 men, and both halves of Layout 5’s pooled queue) sit just outside the CI but inside the $\pm 5\%$ band, with a maximum relative error of $+3.75\%$ (Layout 2 men).

Layout	Gender	W_q^{ana} (min)	$\hat{W}_q$ (min)	95% CI	Rel. error	ρ
1 – Baseline	Women	1.0031	1.0055	[1.0015, 1.0095]	+0.24%	0.90
	Men	0.0051	0.0051	[0.0050, 0.0051]	−0.32%	0.52
2 – Equal Wait	Women	0.0591	0.0591	[0.0589, 0.0593]	+0.08%	0.69
	Men [†]	0.2720	0.2822	[0.2815, 0.2829]	+3.75%	0.79
3 – Min. Avg. Wait	Women	0.1330	0.1332	[0.1328, 0.1336]	+0.12%	0.75
	Men [†]	0.0349	0.0358	[0.0357, 0.0359]	+2.60%	0.63
4 – Mixed	Women	0.1330	0.1333	[0.1329, 0.1336]	+0.19%	0.75
	Men	0.0253	0.0253	[0.0252, 0.0254]	−0.04%	0.60
5 – Gender Neutral	Women [†]	0.1242	0.1248	[0.1245, 0.1251]	+0.48%	0.82
	Men [†]	0.1242	0.1248	[0.1245, 0.1251]	+0.49%	0.82
6 – Optimal Mixed	Women	0.0268	0.0267	[0.0266, 0.0268]	−0.06%	0.64
	Men	0.1785	0.1785	[0.1781, 0.1789]	+0.02%	0.75

[†] Analytical value outside 95% CI. The mechanism is the heterogeneous service-time mixture (Section 4.2.2).

Per-layout details follow.

Table 4.3: Verification summary: analytical M/M/c predictions versus simulated means (1 000 runs, 30 000-min observation window, 95% CI). Daggers mark cases where the analytical value falls outside the CI.

Hyperexponential service-time mixture. Wherever the analytical model approximates a queue with non-constant per-customer service rates as a single M/M/c with capacity-weighted mean $\bar{\mu}$ (Section 4.2.2), the underlying service-time distribution is a two-component hyperexponential with $C_s^2 > 1$. Higher service-time variance produces a positive bias on the mean wait. Two situations in this study trigger the mechanism:

- Men’s queues in separated layouts (Layouts 1–3) draw from a heterogeneous server pool of stalls ($\mu_m = 0.8/\text{min}$) and urinals ($\mu_u = 1.0/\text{min}$) under shortest-queue routing. Layout 2 men ($\rho = 0.79$) gives +3.75% and Layout 3 men ($\rho = 0.63$) gives +2.60%, both outside the CI. Layout 1 men shares the same heterogeneous pool but at low utilisation ($\rho = 0.52$), where $W_q \approx 0.005$ min is too small for the bias to be distinguished from CI noise.
- The pooled queue of Layout 5 mixes class-dependent service times (women $\mu_w = 2/3/\text{min}$, men $\mu_m = 0.8/\text{min}$). The variance gap is narrower than for the urinal/stall pool, so the residual is correspondingly smaller: +0.48% for the women’s wait and +0.49% for the men’s wait at $\rho = 0.82$, both outside their (narrow) 95% CIs.

In each case the slower servers occupy capacity for longer than the weighted-average rate implies. This raises the probability that all servers are simultaneously busy, so the queues run slightly longer than the approximation predicts. The residual grows with both the variance gap and the utilisation.

Layout 6 men. Layout 6 men is a pure M/M/8 in verification mode (urinals only, $\mu_u = 1.0/\text{min}$), so the hyperexponential mechanism does not apply. Its residual of +0.02% is indistin-

guishable from zero. At the original short window of $T = 1\,000$ min the residual was +1.8%. The decay to zero by the longer observation window of $T = 30\,000$ min identifies it as finite-window estimator variance, not an error in the model (Appendix B).

Of the eight cases where the M/M/c model is exact, five residuals are positive and three negative. A two-sided exact binomial test of $H_0: p = 0.5$ gives $p \approx 0.73$, well above the conventional $\alpha = 0.05$ threshold. The simulation engine shows no detectable bias. The four cases marked † are excluded here because they are known to be biased positive by the hyperexponential mechanism above.

4.7 Model validation and limitations

4.7.1 Scope and limitations

- **Validation scope.** The verification covers steady-state operation at constant arrival rates (calm period only, no MMPP surges). The simulation was configured to match the M/M/c assumptions: Poisson arrivals, exponential service times, and stable traffic intensity ($\rho < 1$).
- **Approximation limits.** The capacity-weighted mean $\bar{\mu}$ (Eq. 4.3) biases the analytical prediction upward whenever the server pool is heterogeneous (the hyperexponential mechanism of Section 4.2.2). The bias grows with ρ .
- **Mixed-layout routing.** For Layouts 4 and 6, the verification routes men deterministically to urinals to create independent M/M/c queues. The full simulation lets men choose dynamically between stalls and urinals. The resulting correlations between queues rule out closed-form M/M/c analysis.
- **Time-varying arrivals.** The validation does not cover the MMPP. Stability under MMPP is treated in Section 4.7.2.

4.7.2 Stability under MMPP arrival rates

The verification above tests the simulator at the calm-state rate. Under the MMPP (Section 1.4.3), the time-average arrival rate $\bar{\lambda} = 14/\text{min}$ raises the load on every layout. Table 4.4 and Figure 4.2 report per-layout utilisations at the three rates of interest (calm, time-average, busy). Layout 1 fails the stability condition $\rho < 1$ already at the time-average rate ($\rho_{\text{avg}} = 1.050$). All six layouts fail at the busy-state rate. Chapter 5 reports the full-MMPP wait times.

Layout	ρ_{calm}	ρ_{avg}^*	ρ_{busy}
1 – Baseline	0.900	1.050 [†]	1.800 [†]
2 – Equal Wait	0.692	0.808	1.385 [†]
3 – Min. Avg. Wait	0.750	0.875	1.500 [†]
4 – Mixed	0.750	0.875	1.500 [†]
5 – Gender Neutral	0.825	0.963	1.650 [†]
6 – Optimal Mixed	0.643	0.750	1.286 [†]

* MMPP-weighted average: $\lambda_{\text{avg}} = \pi_{\text{calm}}\lambda_{\text{calm}} + \pi_{\text{busy}}\lambda_{\text{busy}} \approx 0.233/\text{sec}$.

[†] Unstable ($\rho \geq 1$): queue grows without bound.

Table 4.4: Utilisation of the women’s queue (or pooled queue) at different MMPP arrival rates.

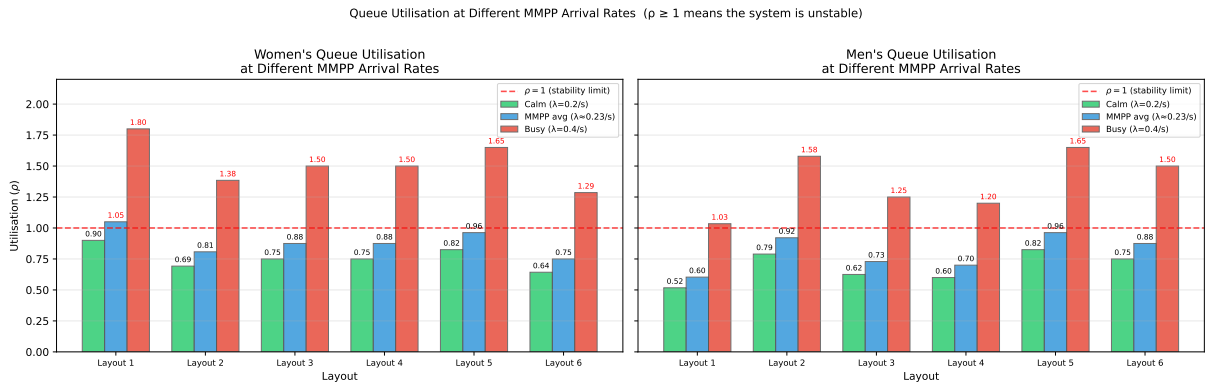


Figure 4.2: Queue utilisation ρ for each layout at the calm-state, MMPP-weighted, and busy-state arrival rates, separated by gender subsystem. The dashed red line marks the M/M/c stability limit ($\rho = 1$). Bars exceeding it identify configurations whose queue grows without bound at the corresponding rate.

4.8 Chapter summary

The chapter has compared the simulator against closed-form M/M/ c predictions at the calm-state rate, under simplifications that reduce each layout to a known M/M/ c queue, and identified where the predictions stop matching. The agreement is a strong indication that the simulator behaves as expected wherever an analytical reference is available, though it is not a formal proof of correctness.

- Under constant calm-state arrivals and the verification-mode routing, every layout reduces to independent M/M/ c queues. Eight of twelve queue–gender combinations pass verification (analytical value inside 95% CI). The four exceptions are men’s queues in Layouts 2 and 3 and both halves of Layout 5’s pooled queue. The maximum relative error is +3.75% (Layout 2 men).
- All four out-of-CI residuals share one mechanism: heterogeneous service times give $C_s^2 > 1$, so the mean wait sits above the matched-mean M/M/ c value. The offset grows with the variance gap and the utilisation. Layouts 2 and 3 men have the wider stall–urinal rate ratio of 1.25, with residuals of +3.75% at $\rho = 0.79$ and +2.60% at $\rho = 0.63$. Layout 5’s pooled queue has the narrower rate ratio of 1.20 and smaller residuals: +0.48% for women and +0.49% for men, both at $\rho = 0.82$. Restricted to the eight cases where the M/M/ c model is exact, five residuals are positive and three negative ($p \approx 0.73$ on a two-sided binomial test). The simulation engine shows no detectable bias.
- The M/M/ c predictions at the calm-state rate are a *lower bound* on full-MMPP waiting times. The busy-state rate $\lambda_{\text{busy}} = 24/\text{min}$ exceeds the service capacity of every layout ($\rho > 1$), so queues build up during bursts. Layout 1 is the most extreme case, already exceeding stability at the MMPP time-average rate ($\rho_{\text{avg}} = 1.050$). These transient overloads produce the higher waits reported in Chapter 5.
- The pooling effect is evident in the predictions. Layout 5’s M/M/20 pooled queue cuts the calm-state women’s wait by a factor of about 8 relative to Layout 1’s M/M/10 subsystem ($W_q \approx 7.5$ s vs. ≈ 60 s). The cost is that all men must use (the less efficient) stalls.

Simulation results: layout comparison

5.1 Experimental setup

Now that we have the correct warm-up duration and a verified simulation engine, we can present the layout-comparison results. This chapter uses the simulation framework from Chapter 1 (Sections 1.4.1 and 1.4.3). The warm-up duration $t_{\text{warm-up}} = 1\,778$ min was derived in Chapter 3. Each layout is simulated with 1 000 independent replications and each replication uses this warm-up followed by a 90-min observation window. Arrivals follow the MMPP throughout the whole run. The baseline gender split is 50/50. Service times are exponential. The mean is 1.5 min at women’s stalls, 1.25 min at men’s stalls, and 1.0 min at urinals. Men choose between men’s stalls and urinals using the shortest-queue rule from Section 1.3.2. All confidence intervals use the 95 % level.

This chapter first reports the FCFS baseline for all six layouts. A sensitivity analysis at the end assesses how robust those results are under different gender splits. Priority policies are studied later, in Chapter 6.

5.2 Performance by layout

Table 5.1 reports the average waiting times and 95 % confidence intervals for each layout, averaged across the calm and busy MMPP states during the observation window. Figure 5.1 shows the same data graphically, alongside the gender-wait gap and per-facility utilisation.

The utilisation values shown in Figure 5.1(d) and used in the per-layout discussion below are simulation-measured time-averages. They match the $M/M/c$ analytical predictions of Chapter 4 to within 1.2 percentage points (see Section 5.3 for the full comparison).

Three patterns stand out. Layout 1 is infeasible at the modelled demand: its women’s queue runs above capacity. Among the feasible configurations, Layout 6 (Optimal Mixed) achieves the lowest overall wait (31 s) and the smallest non-zero gender gap (11 s), while Layout 5 (Gender

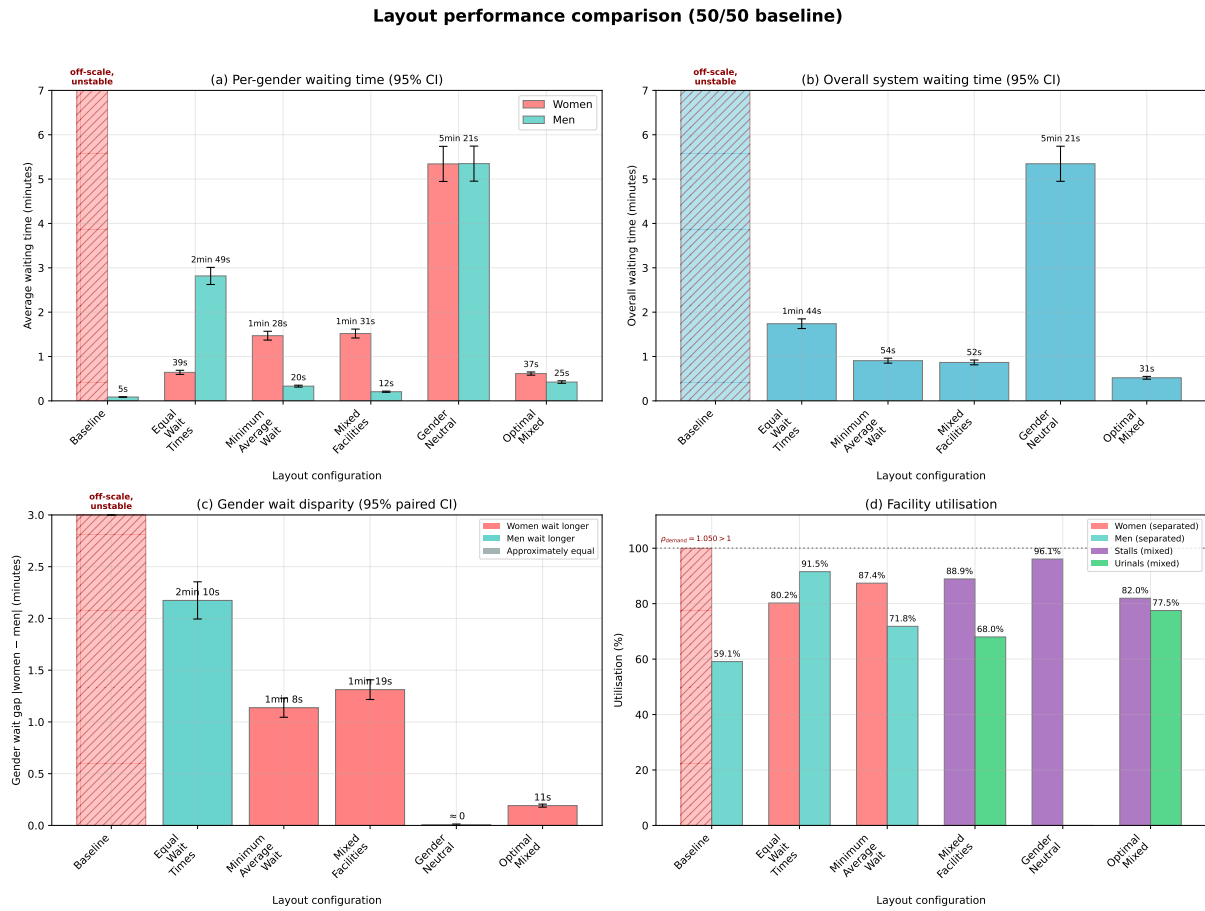


Figure 5.1: Performance comparison across layouts. Layout 1 is shown off-scale (hatched) in all panels because its women’s subsystem is unstable ($\rho_{\text{women}} = 1.050$, see Section 5.2.1). (a) Per-gender average waiting times with 95 % confidence intervals. (b) Overall system waiting time with 95 % confidence intervals. (c) Gender wait disparity by layout, computed as the absolute difference between women’s and men’s average waits. Error bars show 95 % confidence intervals on the per-replication difference $W_w - W_m$. They are tighter than independent intervals because both genders share the same random seed in each replication (Common Random Numbers; see Section E.1.3). Bar colour shows direction: pink when women wait longer, blue when men wait longer, grey when the two are approximately equal. (d) Per-facility utilisation.

Layout	Women's Mean $\pm$ 95 % CI	Men's Mean $\pm$ 95 % CI	Overall Mean $\pm$ 95 % CI
Layout 1 [†]	<i>Infeasible</i>	5s $\pm$ <1s	n/a
Layout 2	39s $\pm$ 3s	2min 49s $\pm$ 12s	1min 44s $\pm$ 7s
Layout 3	1min 28s $\pm$ 6s	20s $\pm$ 1s	54s $\pm$ 3s
Layout 4	1min 31s $\pm$ 6s	12s $\pm$ 1s	52s $\pm$ 3s
Layout 5	5min 21s $\pm$ 24s	5min 21s $\pm$ 24s	5min 21s $\pm$ 24s
Layout 6	37s $\pm$ 2s	25s $\pm$ 2s	31s $\pm$ 2s

[†]Layout 1 is infeasible under the modelled demand ($\rho_{\text{women}} = 1.050 > 1$), so the women's queue has no steady-state mean. See Section 5.2.1 for the diagnostic value and discussion.

Table 5.1: Average waiting times by layout with 95 % confidence interval half-widths (post-warm-up period, MMPP). Means and CIs are computed across 1 000 independent replications.

Neutral) achieves perfect parity (by design) but the highest overall wait (5 min 21 s). The subsections below examine each layout.

As a reminder, the six layouts are (see Section 1.4.1 for the full specifications):

- Layout 1 (separated): 10 women's stalls, 2 men's stalls, 10 urinals
- Layout 2 (separated): 13 women's stalls, 2 men's stalls, 6 urinals
- Layout 3 (separated): 12 women's stalls, 2 men's stalls, 8 urinals
- Layout 4 (mixed): 12 shared stalls, 10 men's urinals
- Layout 5 (gender neutral): 20 shared stalls, no urinals
- Layout 6 (mixed): 14 shared stalls, 8 men's urinals

5.2.1 Infeasible baseline: Layout 1

Layout 1 (the traditional equal-space allocation) is presented first because it is the standard design for restroom facilities and the one the gender-equity debate has long argued against. Women typically spend more time per visit than men, and a 50/50 space split also gives them fewer facilities (women's stalls take more space than men's urinals). Both factors put more load on the women's subsystem than on the men's.

Layout 1's women's subsystem is unstable: analytical $\rho_{\text{women}} = \lambda_{\text{women}} \cdot E[S_{\text{women}}]/c_{\text{women}} = 7 \cdot 1.5/10 = 1.050 > 1$ (the men's subsystem is fine, analytical $\rho_{\text{men}} = 0.603$). The simulation confirms this. The women's stalls run at 100 % utilisation throughout the observation window (always occupied), and the simulated men's utilisation $\rho_{\text{men}} = 0.591$ matches the analytical prediction within ~ 1.2 pp (see Section 4.2.2 for the small gap due to rational-agent routing across heterogeneous servers). Men wait 5 s on average, while the women's wait has no steady-state

value and grows without bound as the simulation horizon increases. The 90-min observation records 91 min 59 s on average, and a 3 000-min warm-up raises that to about 149 min for the same configuration. This shows how inefficient the most popular design is.

5.2.2 Alternative configurations

The five feasible alternatives are examined below.

Layout 2: Equal Wait Times

Layout 2 reallocates capacity to balance women's and men's waiting times more. The name comes from Van Hautegeem & Rogiest [4], where the allocation achieves approximately equal waits under their modelling assumptions (which differ here on arrival pattern and men's stall service time).

- Overall average waiting time: 1 min 44 s $\pm$ 7 s
- Women's average waiting time: 39 s $\pm$ 3 s
- Men's average waiting time: 2 min 49 s $\pm$ 12 s
- Gender gap: 2 min 10 s $\pm$ 11 s
- Utilisation: $\rho_{\text{women}} = 0.802$, $\rho_{\text{men}} = 0.915$

Layout 2 keeps the women's queue stable ($\rho_{\text{women}} = 0.802$) but produces severe congestion for men. The men's average wait of 2 min 49 s pushes the overall wait to 1 min 44 s. The layout only shifts the bottleneck from women to men. Under the model of this thesis, the "Equal Wait Times" label no longer fits.

Layout 3: Minimum Average Wait

Layout 3 has one fewer women's stall than Layout 2 and is optimised for minimum overall wait among separated layouts.

- Overall average waiting time: 54 s $\pm$ 3 s
- Women's average waiting time: 1 min 28 s $\pm$ 6 s
- Men's average waiting time: 20 s $\pm$ 1 s
- Gender gap: 1 min 8 s $\pm$ 6 s
- Utilisation: $\rho_{\text{women}} = 0.874$, $\rho_{\text{men}} = 0.718$

Unlike Layout 2, the name fits here: Layout 3 has the lowest overall wait among the separated configurations. Cutting women's stalls from 13 to 12 reduces men's wait by far more than it raises women's wait. Both Layout 2 and Layout 3 are clear improvements over the infeasible Layout 1, but neither delivers a low overall wait and a small gender gap together. The mixed layouts that follow pursue both.

Layout 4: Mixed Facilities

Layout 4 abandons the separated structure: women and men share a single stall queue, with urinals reserved for men.

- Overall average waiting time: 52 s $\pm$ 3 s
- Women's average waiting time: 1 min 31 s $\pm$ 6 s
- Men's average waiting time: 12 s $\pm$ 1 s
- Gender gap: 1 min 19 s $\pm$ 6 s
- Utilisation: $\rho_{\text{stalls}} = 0.889$, $\rho_{\text{urinals}} = 0.680$

Layout 4 performs close to Layout 3 overall (52 s vs. 54 s). Sharing the stalls in a mixed queue improves men's wait (12 s vs. 20 s), while women's wait barely changes (1 min 31 s vs. 1 min 28 s for Layout 3). The 10 urinals absorb enough male demand to shorten men's waits but leave too little stall capacity to shorten women's.

Layout 5: Gender Neutral

Layout 5 is the fully gender-neutral design: all facilities are shared stalls.

- Overall average waiting time: 5 min 21 s $\pm$ 24 s
- Women's average waiting time: 5 min 21 s $\pm$ 24 s
- Men's average waiting time: 5 min 21 s $\pm$ 24 s
- Gender gap: 0 s $\pm$ 1 s
- Utilisation: $\rho_{\text{pooled}} = 0.961$

Layout 5 has the worst overall performance of any feasible configuration. Without the faster urinals, the effective service rate drops sharply. Near-equal waiting times come at a steep cost: everyone waits about 5 min 21 s.

Pool utilisation $\rho_{\text{pooled}} = 0.961$ is very close to 1, with no headroom to absorb the bursts. During a burst, the arrival rate exceeds the pool's service capacity and the queue grows. Between

bursts the queue shrinks, but often doesn't fully clear before the next burst. Most of the average wait builds up during the bursts. A steady arrival pattern at the same average rate would give much shorter waits.

This is the setting that Bovens & Marcoci [21] study. Under stationary Poisson demand, they report that gender-neutral provision reduces women's waiting time without substantially harming men. Under the MMPP demand of this thesis, the burst dynamics increase the wait, so the parity gain comes at a steep cost.

Layout 6: Optimal Mixed

Layout 6 is the best layout overall: in contrast to Layout 4, it manages to make the utilisation of the stalls and the urinals close to equal.

- Overall average waiting time: $31 \text{ s} \pm 2 \text{ s}$
- Women's average waiting time: $37 \text{ s} \pm 2 \text{ s}$
- Men's average waiting time: $25 \text{ s} \pm 2 \text{ s}$
- Gender gap: $11 \text{ s} \pm 1 \text{ s}$ (computed from unrounded replication means)
- Utilisation: $\rho_{\text{stalls}} = 0.820$, $\rho_{\text{urinals}} = 0.775$

Layout 6 achieves the best performance of any feasible layout. The urinals take most men ($\rho_{\text{urinals}} = 0.775$), and the stall queue stays at moderate load ($\rho_{\text{stalls}} = 0.820$) with room for MMPP bursts. The 31-second overall wait is 43% lower than the best separated layout (Layout 3) and 90% lower than the gender-neutral pool (Layout 5). Layout 6 is the only configuration with sub-minute waits for both genders. This layout also produces the smallest non-zero gender gap (11 s).

5.3 Utilisation analysis

The utilisation values shown in Figure 5.1(d) and in the per-layout bullet lists are measured directly from the simulation. For each facility pool, utilisation is the average share of facilities that are occupied during the observation window, averaged over the 1 000 replications. For the separated and pooled layouts (1, 2, 3, and 5), they match the M/M/c analytical predictions of Chapter 4 to within 1.2 percentage points. The largest residuals (1.2 pp on Layout 1 men, 1.1 pp on Layout 3 men) sit in the heterogeneous men's subsystems, where shortest-queue routing sends more men to the faster urinals than the capacity-weighted M/M/c formula assumes (see Section 4.2.2 for the underlying mechanism). The remaining comparisons fit within 0.6 pp.

Inside the men's pool, the routing effect is visible directly. The capacity-weighted M/M/c formula treats stalls and urinals as one homogeneous pool with a single ρ -value. The simulation shows them at different levels: within Layout 2's men's subsystem, the urinals run at $\rho = 0.93$

while the men's stalls sit at $\rho = 0.87$, and in Layout 3 the split is even wider ($\rho = 0.77$ on urinals against 0.52 on men's stalls). Layouts 4 and 6 route male demand dynamically between stalls and urinals, so the M/M/c benchmark does not predict their per-pool values.

Utilisation and waiting time are coupled. Higher utilisation makes better use of capacity but shrinks the buffer against demand peaks, so queues grow disproportionately when MMPP bursts arrive. The six layouts fall into four operating regimes.

- **Beyond capacity** ($\rho \geq 1$): Layout 1's women's subsystem is unstable. The queue grows without bound and the expected waiting time is infinite.
- **Near-capacity** (ρ close to 1): Layout 5's pooled utilisation stays below 1 on the time-average, but the buffer against MMPP bursts is too small. Queues build during each burst and only partially clear in the calmer intervals, producing long average waits. Feasibility is not enough on its own. A layout must keep ρ well below 1 to absorb demand peaks.
- **Imbalanced subsystems**: Layouts 2, 3, and 4 have one facility pool with much higher utilisation than the other (a gap of 10 to 20 pp). The system stays feasible, but the more heavily loaded pool drives the overall wait.
- **Balanced mixed-capacity**: Layout 6 uses two pools (shared stalls and urinals). Routing part of male demand to urinals leaves enough stall buffer for both genders. The urinal count is the design degree of freedom that keeps both pools below saturation with room to spare.

Across the four regimes, the constraint is the same: every pool must stay below $\rho = 1$ for feasibility, and well below 1 for low average waits, since MMPP bursts need spare capacity. Layout 6 is the only configuration where both pools satisfy the second condition.

5.4 Sensitivity analysis: robustness to gender distribution

To assess sensitivity to the gender mix, the full MMPP simulation was repeated for all six layouts under three arrival splits: 50/50 (baseline), 60/40 (female-skewed), and 40/60 (male-skewed). Each scenario used 1 000 replications, and 95 % confidence intervals were computed.

A warm-up of 3 000 min was used in every case, so all reported results reflect steady state. Because the gender-skewed scenarios push some queues to higher utilisation than the 50/50 baseline, the conservative 3 000-min buffer keeps the simulation safely above the worst-case convergence time across all skews.

All utilisation values ρ quoted in this section are analytical demand-side values, $\rho = \lambda \cdot E[S]/c$, not simulation-measured time averages. Simulated utilisation cannot exceed 1.0, so the analytical ρ is the only honest indicator of overload for the layout and scenario combinations where demand exceeds capacity.

For the gender gap $\Delta W = |W_w - W_m|$, the $\pm$ half-width refers to the paired CI on the signed per-replication difference $W_w - W_m$, computed before taking absolute value (the standard paired-CI convention under Common Random Numbers). The half-width can exceed the absolute point estimate when ΔW is close to zero, as in Layout 5’s near-parity rows.

Table 5.2 reports the mean waiting times and their 95% confidence intervals for each layout across the three scenarios. Figure 5.2 shows the same data visually, together with the absolute gender gap $\Delta W = |W_w - W_m|$.

Layout	60/40 (female-skewed)			50/50 (baseline)			40/60 (male-skewed)		
	W_w	W_m	ΔW	W_w	W_m	ΔW	W_w	W_m	ΔW
Layout 1	<i>Infeasible</i>	1s $\pm$ <1s	<i>Infeasible</i>	<i>Infeasible</i>	5s $\pm$ 1s	<i>Infeasible</i>	1min 0s $\pm$ 4s	16s $\pm$ 1s	43s $\pm$ 4s
Layout 2	7min 28s $\pm$ 30s	26s $\pm$ 1s	7min 2s $\pm$ 29s	40s $\pm$ 3s	2min 50s $\pm$ 11s	2min 11s $\pm$ 10s	8s $\pm$ 1s	<i>Infeasible</i>	<i>Infeasible</i>
Layout 3	<i>Infeasible</i>	5s $\pm$ 1s	<i>Infeasible</i>	1min 22s $\pm$ 6s	19s $\pm$ 1s	1min 3s $\pm$ 6s	15s $\pm$ 1s	1min 22s $\pm$ 5s	1min 7s $\pm$ 5s
Layout 4	<i>Infeasible</i>	4s $\pm$ <1s	<i>Infeasible</i>	1min 30s $\pm$ 6s	11s $\pm$ 1s	1min 18s $\pm$ 6s	27s $\pm$ 2s	17s $\pm$ 1s	10s $\pm$ 1s
Layout 5	9min 21s $\pm$ 40s	9min 22s $\pm$ 40s	1s $\pm$ 1s	5min 0s $\pm$ 21s	5min 0s $\pm$ 21s	0s $\pm$ 1s	3min 14s $\pm$ 15s	3min 14s $\pm$ 15s	0s $\pm$ 1s
Layout 6	1min 58s $\pm$ 8s	14s $\pm$ 1s	1min 43s $\pm$ 8s	36s $\pm$ 2s	25s $\pm$ 2s	11s $\pm$ 1s	24s $\pm$ 1s	23s $\pm$ 1s	0.6s $\pm$ 0.3s

Table 5.2: Sensitivity of mean waiting times to gender arrival distribution across all six layouts (1000 replications, 3000-min warm-up to cover every gender split under one cutoff). This warm-up is why Layout 5 at 50/50 reads 5 min 0 s here, against 5 min 21 s in Table 5.1 (see Appendix D). Feasible cells report mean $\pm$ 95 % CI half-width. ΔW CIs use paired replication differences (Common Random Numbers; see Section E.1.3). *Infeasible*: $\rho_{\text{avg}} > 1$, the queue grows without bound and no true steady-state exists. The observed simulation value is not meaningful.

5.4.1 Patterns across the gender splits

Several layout and scenario combinations are infeasible. Layout 1 fails for women under both 60/40 ($\rho = 1.26$) and 50/50 ($\rho = 1.05$). Layout 2 fails for men under 40/60 ($\rho = 1.11$), the mirror image of Layout 1: too few facilities for men when male demand dominates. Layouts 3 and 4 fail for women under 60/40 ($\rho = 1.05$). Even 12 stalls cannot handle 60 % of a 14/min average stream at 1.5 min per visit. Each case is marked *Infeasible* in Table 5.2 and in Figure 5.2.

Layout 5 stays feasible under every split, but at uniformly high cost. Its pooled utilisation sits between $\rho_{\text{avg}} = 0.945$ and 0.980, producing waits between roughly 3 and 9 min across the three scenarios. The fully shared design eliminates gender imbalance and pays for it with increased waiting times.

Layout 6 is the only configuration that stays both feasible and efficient under every split. Even in its worst case (60/40 female-skew), women’s wait rises only to 1 min 58 s $\pm$ 8 s, while Layouts 3 and 4 collapse to infeasibility. Under the 40/60 male-skew, Layout 6 reaches near-perfect equity ($\Delta W \approx 0.6$ s): lower female demand frees stall capacity, and the urinals absorb the extra male traffic. Where Layouts 3 and 4 remain feasible, Layout 6 still beats them on both overall wait and gender gap. Where they fail, Layout 6 is the only efficient alternative to Layout 5. Its combination of shared stalls and urinals provides the flexibility that the separated layouts lack.

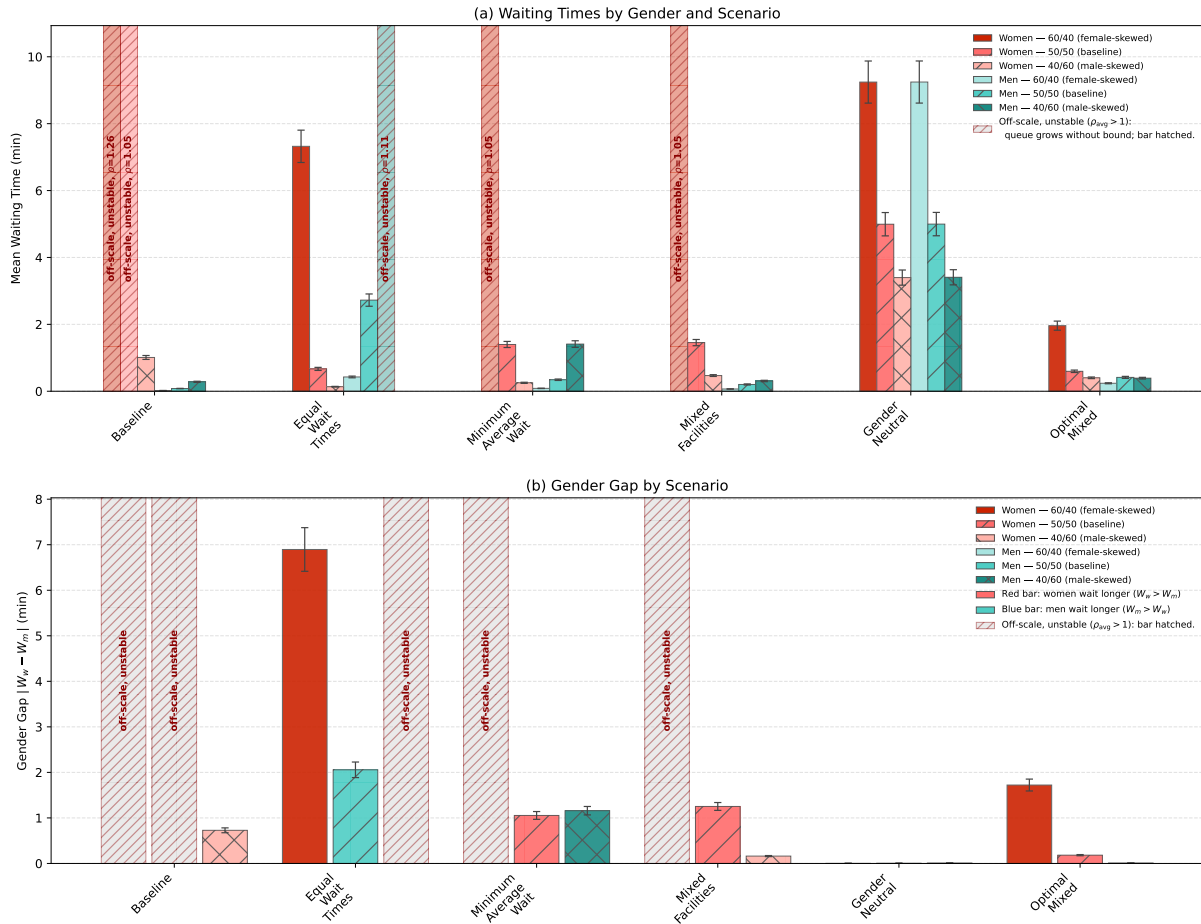


Figure 5.2: Sensitivity of mean waiting times and gender gap to gender arrival distribution across all six layouts (1 000 replications, 95 % confidence intervals). Within each gender, darker shades correspond to higher same-gender demand (dark red = 60/40, red = 50/50, light red = 40/60 for women; light blue = 60/40, blue = 50/50, dark blue = 40/60 for men). Panel (a): mean waiting times by gender. Panel (b): absolute gender gap $\Delta W = |W_w - W_m|$. Bar colour indicates who waits longer (red = women disadvantaged, blue = men disadvantaged). No bar is drawn for infeasible cells ($\rho_{avg} > 1$) because the observed waiting time does not represent a true steady-state.

5.5 Chapter summary

This chapter compared the six layouts under MMPP demand at the 50/50 baseline and tested their robustness under female-skewed (60/40) and male-skewed (40/60) arrivals.

- Layout 1 (traditional equal-space allocation) is infeasible under the modelled demand. With equal floor space and women's service times of 1.5 min (versus 1.0 to 1.25 min for men), the women's subsystem sits at $\rho_{\text{women}} = 1.050 > 1$ and its queue grows without bound.
- Among the feasible layouts at 50/50, Layout 6 (Optimal Mixed: 14 shared stalls + 8 men's urinals) has the lowest overall wait (31 s) and the smallest non-zero gender gap (11 s). It is the only configuration where both pools (shared stalls and urinals) sit well below saturation, leaving room for MMPP bursts.
- Layout 5 (gender-neutral pool of 20 stalls) has zero gender gap by design ($\Delta W \approx 0$), but its overall wait is the highest of any feasible layout (5 min 21 s). The pooled utilisation $\rho_{\text{avg}} = 0.961$ leaves not enough buffer for MMPP bursts. The cost comes mostly from removing the faster urinals.
- Under the three tested gender splits, Layouts 1 through 4 each fail in at least one scenario: Layout 1 at 50/50 and 60/40, Layout 2 at 40/60, and Layouts 3 and 4 at 60/40. Layout 5 stays feasible everywhere but always near saturation (ρ_{avg} between 0.945 and 0.980). Only Layout 6 stays both feasible and efficient at every split, and at 40/60 its gender gap drops to $\Delta W \approx 0.6$ s with no operational intervention.
- Layout 6's residual 11 s gender gap at the 50/50 baseline is small but non-zero. Chapter 6 examines whether priority scheduling on the shared stall pool can close it, subject to Kleinrock's conservation law: any work-conserving priority policy can only redistribute waiting time between genders, not reduce the total.

CHAPTER 6

Ladies-first priority policies

6.1 Motivation

The previous chapters showed that Layout 6 (Optimal Mixed: 14 shared stalls and 8 men’s urinals) has the lowest overall waiting time among all feasible layouts. However, even in this optimised layout, a gender disparity remains: under FCFS women have an average wait about 11 s longer than men (Table 5.1). This chapter addresses the research question of the thesis:

How do different “ladies-first” priority policies perform in shared-capacity restroom facilities? Do they successfully narrow the gender waiting time gap, and at what cost to overall system efficiency and individual male users?

We evaluate four such policies: Strict Head-of-Line (the most extreme case), the Token Bucket, the Threshold Priority, and Dynamic Aging, applied to the shared stall queues of Layouts 4, 5, and 6. Layout 5 is included as a reference case: because all users share a single queue under FCFS, that layout already has perfect gender parity, so a priority policy is both unnecessary and counterproductive. Optimisation of the Token Bucket, Threshold Priority, and Dynamic Aging is run only on Layout 6, since the aim is to combine the best physical layout with the best policy.

The main objective is to reduce the gender gap $\Delta W = |W_w - W_m|$, where W_w and W_m are the women’s and men’s mean waits. The overall average wait $\bar{W}$ is also reported, as a check that the equity gain does not come at the cost of a slower system. In practice $\bar{W}$ should not move much either way: Kleinrock’s conservation law shows that in a stable system, a work-conserving priority discipline cannot meaningfully reduce the aggregate mean wait, and waiting time taken from one class is approximately passed to another [25].

We are looking for a policy that reduces the gap but also avoids extremes. Kleinrock fixes the total wait time shifted to men, but each policy decides how that time is shared. A policy where one man walks straight through while another waits an hour would just trade one unfairness

for another. Section 6.7 examines how each policy distributes that time across individual men, looking at how long the worst-affected men wait and how often each is bypassed, alongside the mean waiting time comparisons. Women’s tail statistics are reported in the same section for completeness, although bypass counts do not apply to women, who are the bypassers under every priority policy considered here.

Section 6.2 starts with Strict Head-of-Line, the unconditional rule that serves any waiting woman before any waiting man. It has no parameters to optimise and is the theoretical limit. The remaining three policies relax that rule in different ways. The Token Bucket (Section 6.4) limits how often a man can be skipped: each bypass spends one token from a budget that accumulates over time. Threshold Priority (Section 6.5) keeps that token mechanism and adds a load-aware gate, switching bypassing off once the queue grows past a chosen length. Dynamic Aging (Section 6.6) drops the token state altogether and assigns each waiting man a priority that grows with his wait, so no man can be bypassed indefinitely. Because the last three policies require parameter choices, Section 6.3 presents the shared parameter optimisation procedure before the policies are discussed.

Unless stated otherwise, the results tables use 1 000 MMPP replications. Optimisation trials use 500 replications per candidate parameter setting, followed by 1 000-replication confirmation runs at the selected optima. The defaults are a 50/50 gender split and a 1 778-min warm-up (Chapter 3). On Layout 5, this warm-up understates men’s mean wait by 2–6 % depending on policy, and Dynamic Aging’s gap by 41 % (Appendix D). All Layout 5 rows in the priority tables use a 3 000-min warm-up to address this. Layouts 4 and 6 use the canonical 1 778-min value across every priority table. The cross-policy tail diagnostic (Section 6.7) and the gender-split sensitivity analysis (Section 6.8) use 3 000 min throughout, so the 60/40 and 40/60 robustness splits share a single warm-up. The sensitivity analysis additionally sweeps 60/40 and 40/60 alongside the 50/50 baseline.

6.2 Strict Head-of-Line: the theoretical limit

6.2.1 Definition

Strict Head-of-Line priority is the unconditional ladies-first rule. When a mixed stall becomes free, the shared queue is checked. If a man is at the front and at least one woman is waiting behind him, the first such woman moves to the front and is served. The other customers keep their order. If no woman is waiting, the man at the front is served under FCFS. Urinals are unaffected, since only men use them.

In queueing terms, the policy is a non-pre-emptive head-of-line priority discipline with two priority classes.

6.2.2 Results

Table 6.1 reports the results for Layouts 4, 5, and 6 under Strict Head-of-Line priority, alongside the corresponding FCFS baselines.

Layout	Policy	W_w	W_m	ΔW	Bypasses (obs window)	Note
Layout 4	FCFS	1min 31s	12s	1min 19s	n/a	women slower
Layout 4	Strict	1min 30s	15s	1min 15s	75	women still slower
Layout 5	FCFS	5min 0s	5min 0s	0s	n/a	parity
Layout 5	Strict	9s	11min 27s	11min 18s	600	men slower (starvation)
Layout 6	FCFS	37s	25s	11s	n/a	women slower
Layout 6	Strict	28s	37s	9s	170	men slower (over-correction)

Table 6.1: FCFS baseline versus Strict Head-of-Line priority for Layouts 4, 5, and 6 (1000 MMPP replications; L4 and L6 use the 1778-min warm-up, L5 the 3000-min warm-up). Bypasses are the mean women-first count over the 90-min observation window. Layout 5 values use the 3000-minute warm-up correction introduced in Section 6.1; this is why its FCFS baseline differs slightly from the 5 min 21 s value in Table 5.1.

Layout 4 (12 mixed stalls, 10 urinals). The gender gap falls from 1 min 19 s to 1 min 15 s, a 5% reduction. Layout 4’s gap is structural, not an ordering problem. Reordering can only help when men and women wait in the same queue, so a woman can be moved ahead of a waiting man. On Layout 4 the urinals absorb most male demand, leaving the shared queue almost all women. With no men to bypass, the stalls cannot keep up with women’s demand. The 75 bypasses per 90-min observation window show the policy is firing, but there are not enough men in the mixed queue for bypassing to matter. Because Strict Head-of-Line is the strongest women-priority discipline possible, this 5% is also the ceiling on what any priority policy can achieve on Layout 4. The cross-layout discussions for the three optimised policies (Sections 6.4.5, 6.5.5, and 6.6.3) are capped by this ceiling and refer back to the observation rather than re-deriving it.

Layout 5 (20 shared stalls, no urinals). This outcome is the most extreme. Women’s average wait falls from 5 min 0 s to 9 s (−97%), while men’s average wait more than doubles from 5 min 0 s to 11 min 27 s (+129%). The mean bypass count of 600 per 90-min observation window is by far the highest of any layout: with no urinals, every departure event triggers a bypass when a woman is in the queue, and at the 50/50 arrival split a woman is almost always present. This is not instability in the queueing-theoretic sense (the system as a whole is stable, $\rho_{\text{pooled}} = 0.96 < 1$), but men effectively starve on Layout 5: the ratio of men’s wait to women’s wait under Strict Head-of-Line priority is approximately 76:1, while under FCFS it is about 1:1. Krawczyk et al. [28] predict this: expected wait grows rapidly as bypass probability approaches one.

Layout 6 (14 mixed stalls, 8 urinals). The policy gives a small improvement over FCFS, but with a slight over-correction in favour of women: the gap is 9 s, with men now waiting longer. Men’s average wait rises from 25 s to 37 s (+48%). The 8 urinals divert enough male demand that Strict Head-of-Line priority does not trigger at every departure (170 bypasses per 90-min observation window, well below Layout 5’s 600), so men avoid the starvation seen on Layout 5. Still, unconditional bypassing transfers more delay to men than equity requires. The optimised policies in Sections 6.4–6.6 try to get rid of this over-correction.

6.2.3 Implications for policy design

Two conclusions follow from this analysis. The outcome is strongly layout-dependent: Strict Head-of-Line priority gives a slight improvement in Layout 6, is irrelevant in Layout 4, and causes severe starvation in Layout 5. The deciding factor is whether the layout has enough urinal capacity. Even on Layout 6, the 9 s over-correction and the +48% cost to men show that unconditional bypassing is too aggressive for equity.

These limitations motivate the three policies that follow (Sections 6.4, 6.5, and 6.6). They share a common optimisation framework, presented in Section 6.3.

6.3 Optimisation methodology

Three policies need parameter values: the Token Bucket, the Threshold Priority, and Dynamic Aging. The same optimisation procedure is applied to all three, with policy-specific objectives and search ranges given in Sections 6.4, 6.5, and 6.6. The goal is to bring the gender gap ΔW as close as possible to zero on Layout 6, with a small tie-breaking penalty for lower overall waits.

The optimiser is Optuna’s Tree-structured Parzen Estimator (TPE) [38], a Bayesian-optimisation algorithm. Each trial runs the simulation at one parameter setting, and TPE uses past results to decide where to sample next. The search focuses on promising regions and finds good solutions with far fewer trials than a grid search would.

Before turning to equity, we verified Kleinrock’s conservation law [25] empirically. An Optuna study targeting overall wait $\bar{W}_{\text{total}}$ (100 trials, 500 replications each) found no parameter set that beat FCFS, confirming that priority cannot reduce aggregate wait time.

6.3.1 Equity objective

The equity optimisation minimises an objective function $\mathcal{L}$, defined as the gender gap with a small tie-breaking penalty on overall wait:

$$\mathcal{L}(\boldsymbol{\theta}) = |W_w - W_m| + 10^{-4} \bar{W}, \quad (6.1)$$

where $\boldsymbol{\theta}$ denotes the parameter vector of whichever policy is being optimised, and W_w ,

W_m , $\bar{W}$ are expressed in minutes. The small coefficient 10^{-4} on $\bar{W}$ acts as a tie-breaker: among configurations with the same gender gap, those with a marginally lower overall wait are preferred. The coefficient is small enough that it cannot influence the ranking of solutions with meaningfully different gap values: the penalty contribution is on the order of milliseconds, orders of magnitude below typical gap differences.

Each trial (one candidate parameter combination) evaluates 500 replications on Layout 6 with common random seeds (CRN) for variance reduction, and each study uses 100 trials in total. Policy-specific search ranges and their rationale are presented in each policy’s section below.

6.3.2 Noise floor and robust optimum

Each Optuna trial estimates ΔW from 500 replications and validation runs use 1 000 replications. Both are noisy: re-running with different random seeds would give slightly different point estimates. The 95% confidence-interval (CI) half-width measures this noise, and scales as $1/\sqrt{N}$ [9]. The 1 000-replication validation runs reported later in this chapter give a half-width of about 1.3 s for ΔW . At 500 replications, the half-width is larger by $\sqrt{1000/500} = \sqrt{2}$, giving a noise floor of $\epsilon_{500} \approx 1.8$ s.

Trials whose ΔW falls below this floor are statistically equivalent: their reported objective values differ only by Monte-Carlo noise. We call this set the *noise-floor basin* [40]. Calling any single trial the optimum would be misleading, since a different random seed would identify a different basin member as best, and both would behave equivalently.

Each policy’s reported parameters come from a *robust optimum*. For each candidate parameter combination, we compute the objective at the candidate and at a few nearby combinations. The candidate’s score is the worst (highest) of these values. We pick the candidate with the lowest score, so the chosen parameters perform well across a small parameter region rather than only at one exact point. The chosen optimum then sits in a flat region of the objective surface, so a slightly miscalibrated r_{tok} , K , or r_{age} does not meaningfully worsen the outcome. The full procedure is in Appendix C.

The basin is shown explicitly in the Threshold Priority convergence diagnostics (Section 6.5.3), where the in-basin trial count and parameter span are reported alongside the single-best trial.

6.4 Token Bucket

6.4.1 Mechanism

The Token Bucket adapts an idea from network traffic shaping [29]. Tokens accumulate at a fixed rate in a bucket of fixed size, and each token grants permission for one event. In our adaptation, that event is the promotion of a woman ahead of waiting men. The bucket caps how often these bypasses can happen, not how often anyone is served. The policy operates on

the shared stall queue only.

When a mixed stall becomes free, the policy checks the head of the shared stall queue:

1. If the head customer is a woman, she enters service. No token is spent.
2. If the head customer is a man, at least one woman waits behind him, and the bucket has at least one token, the first such woman bypasses the men ahead of her and enters service. One token is spent.
3. Otherwise the head customer (a man) enters service under FCFS. No token is spent.

The policy is non-preemptive (no customer in service is interrupted) and non-destructive (no one is dropped from the queue). Tokens are spent only on actual bypasses, not on every departure. As both the token rate and the bucket capacity grow, tokens are always available and the policy approaches the Strict Head-of-Line priority of Section 6.2.

6.4.2 Parameters

Two parameters govern the system:

1. **Token generation rate** $r_{\text{tok}}^{\text{TB}}$: The rate at which priority tokens build up, in tokens per minute.
2. **Bucket capacity** C^{TB} : The maximum number of tokens that can accumulate, limiting burst bypasses.

6.4.3 Search ranges

The optimisation procedure of Section 6.3 is applied to the parameter vector $(r_{\text{tok}}^{\text{TB}}, C^{\text{TB}})$ with the objective of Equation (6.1).

The token rate $r_{\text{tok}}^{\text{TB}}$ is searched on a log scale over $[0.01, 12.0]$ tokens per minute. At the lower bound of 0.01, at most one token accumulates during a typical 90-minute observation window, so the policy behaves like FCFS. At the upper bound of 12, a token is generated for every arriving customer on average (matching the calm-state arrival rate), so the bucket is almost always full and the policy approaches Strict Head-of-Line priority for women regardless of C^{TB} . We use a log scale because we did not know in advance where the optimum would sit. The policy behaves very differently at each scale of $r_{\text{tok}}^{\text{TB}}$ (FCFS-like at very low rates, Strict-like at very high), and linear sampling would have given too little attention to the low end, risking a missed low-rate optimum. The eventual optimum landed in the high-rate region, but this was discovered by the search, not assumed beforehand.

The bucket capacity C^{TB} is searched as an integer over $\{1, \dots, 30\}$. The lower bound of 1 is the most restrictive setting (at most one stored token at a time). The upper bound of 30 was set high enough to prevent any optimum sitting at the search range boundary, and the actual optimum landed well inside the range.

6.4.4 Optimal parameters and achieved parity

Optuna’s single-best trial gave $r_{\text{tok}} = 6.806/\text{min}$, $C = 19$. The robust optimum (Section 6.3.2) is very close: $r_{\text{tok}}^{*,\text{TB}} = 6.91/\text{min}$, $C^{*,\text{TB}} = 19$. We report results at the robust optimum (Table 6.2).

Metric	FCFS	Token Bucket	Change
Women’s wait W_w	37s	32s	−12.9%
Men’s wait W_m	25s	32s	+25.5%
Overall wait $\bar{W}$	31s	32s	+2.9%
Gap ΔW	11s	0.2s	−98%

Table 6.2: Standard FCFS versus equity-optimised Token Bucket on Layout 6 (MMPP arrivals, 1 000 replications at the robust optimum $r_{\text{tok}}^{*,\text{TB}} = 6.91$, $C^{*,\text{TB}} = 19$).

The Token Bucket closes the gap from 11 s to 0.2 s, with a 1 s rise in overall mean wait (+2.9%) that sits within the simulation’s noise floor. The residual gap is small but in the same direction as FCFS (women still slightly slower), also well inside the noise floor.

Women’s wait falls by 5 s (−12.9%) and men’s rises by 7 s (+25.5%). This matches Kleinrock’s conservation law: priority can shift wait time between classes but cannot reduce the total, so women’s gain is paid for by men. The full four-policy comparison is summarised in Section 6.10.

6.4.5 Cross-layout transfer

To test how widely the optimised parameters apply, we used them without re-optimising them on Layouts 4 and 5. Figure 6.1 visualises the cross-layout waits and gender gaps under the Token Bucket at the robust optimum; Table 6.3 reports the underlying numbers.

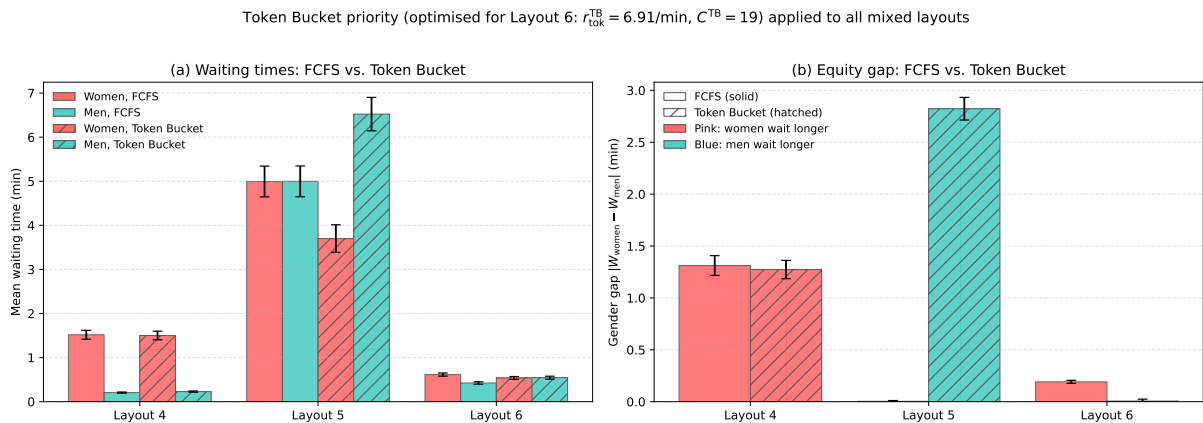


Figure 6.1: Token Bucket cross-layout transfer (robust optimum $r_{\text{tok}}^{*,\text{TB}} = 6.91/\text{min}$, $C^{*,\text{TB}} = 19$; 1 000 MMPP replications per layout; Layout 5 uses the 3 000-min warm-up). **(a)** Mean wait by gender under FCFS (solid) and Token Bucket (hatched); 95% CI half-widths shown. **(b)** Gender gap $|\Delta W|$, coloured by which gender is slower; paired-difference CIs.

Layout	Policy	W_w	W_m	ΔW	Bypasses (obs window)	Note
Layout 4	FCFS	1min 31s	12s	1min 19s	n/a	women slower
Layout 4	Token Bucket	1min 25s	13s	1min 13s	39	women still slower
Layout 5	FCFS	5min 0s	5min 0s	0s	n/a	parity
Layout 5	Token Bucket	3min 52s	6min 13s	2min 22s	552	men slower (gap created)
Layout 6	FCFS	37s	25s	11s	n/a	women slower
Layout 6	Token Bucket	32s	32s	0.2s	132	parity (optimised)

Table 6.3: FCFS baseline versus Token Bucket priority at the robust optimum ($r_{\text{tok}}^{*,\text{TB}} = 6.91$ tokens/min, $C^{*,\text{TB}} = 19$) for Layouts 4, 5, and 6 (1000 MMPP replications; Layout 5 uses the 3000-min warm-up, Layouts 4 and 6 the 1778-min baseline). Bypasses are the mean women-first count over the 90-min observation window; FCFS rows use the same replication seeds as Tables 6.5 and 6.6.

On Layout 4 the policy barely moves the gap. It does not matter if the parameters are optimised for this layout or not. No parameters could get beyond the ceiling already reached by Strict Head-of-Line (Section 6.2.2): the urinal bank absorbs most male stall demand, so the mixed queue holds too few men for any reordering to matter. Priority policies are the wrong tool here, no matter which one.

Layout 5 (20 shared stalls, no urinals) reaches perfect parity under FCFS without any intervention. Men and women share one queue, and both wait about 5 min 0 s (Table 6.1). There is no bias to correct. Adding female-priority bypasses does not fix a fairness problem. It creates one.

The Token Bucket works when the gender gap is mainly an ordering problem in a moderately loaded shared-stall queue, as in Layout 6. It does little when the gap comes from a capacity imbalance (Layout 4), and it opens a new gap when the queue is already balanced (Layout 5).

6.4.6 Parameter sensitivity and importance

The optimisation history and hyperparameter importance analysis are shown in Figure 6.2.

The right panel of Figure 6.2 attributes nearly all of the objective’s variance to the token rate $r_{\text{tok}}^{\text{TB}}$. The bucket capacity C^{TB} accounts for almost none. A token is usually available when a bypass opportunity appears, so the capacity rarely binds. The policy then reduces to a *rate-limited bypass* at the token rate. Each bypass spends one token, and tokens regenerate at a steady $r_{\text{tok}}^{*,\text{TB}} = 6.91$ per minute, so the bucket size matters little.

A 1000-replication measurement at the robust optimum confirms this. Across roughly 2799 bypass-eligible events per replication (a man at the queue head with a woman waiting), tokens were available in $94.69\% \pm 0.08\%$ of attempts (95% CI [94.61%, 94.78%]). The token rate alone supplies nearly every bypass, so the bucket does not need a large reserve.

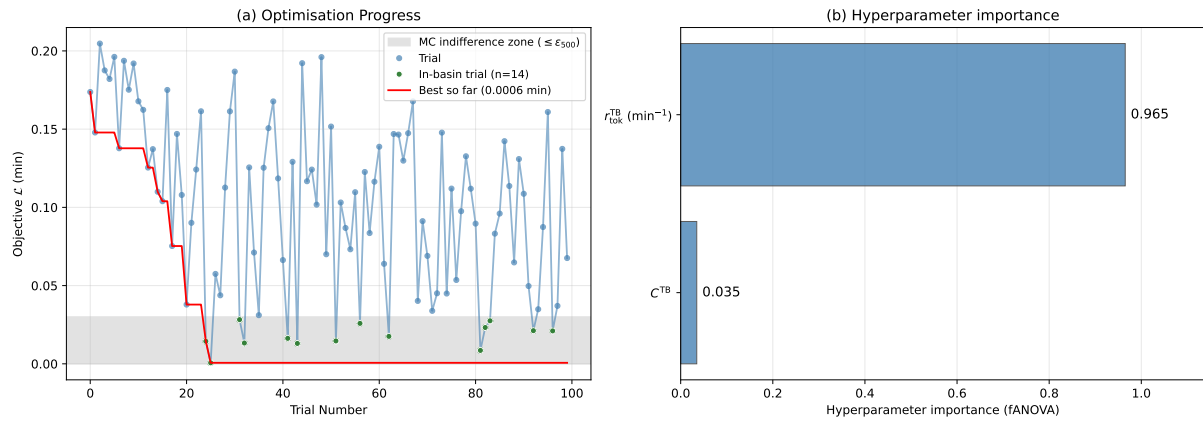


Figure 6.2: Bayesian optimisation convergence for the Token Bucket equity-objective study (left) and hyperparameter importance analysis (right). The token generation rate $r_{\text{tok}}^{\text{TB}}$ accounts for nearly all variance in objective value. Bucket capacity C^{TB} has negligible importance.

The noise-floor basin in Section 6.3.2 shows the same pattern. Trials with $\Delta W \leq \epsilon_{500}$ have similar token rates but very different capacities. Within this basin, only the token rate meaningfully affects performance. The robust optimum (Section 6.3.2) lies inside the basin, so it has the same property. Capacity mainly provides a buffer during busy bursts. Since a token is almost always available at the optimum, the policy fires on most bypass opportunities, similar to Strict Head-of-Line (Section 6.2). The small fraction of held-back bypasses brings the residual gap into the noise band rather than the over-correction Strict produces.

6.5 Threshold Priority

6.5.1 Motivation for a threshold

The pure Token Bucket of Section 6.4 ignores queue length. Tokens are spent whether the queue is very short or very long. During MMPP peak-demand bursts, men near the head of a long queue can be bypassed many times in a row. As introduced in Section 6.1, this kind of extreme behaviour is something a good priority policy should avoid.

Threshold Priority adds a load-aware gate. Priority is active only while the mixed-stall queue holds fewer than K waiting customers. Once the queue grows past K , the system reverts to strict FCFS, so men are not bypassed when bypassing would hurt most.

Queue length is used because it is directly observable to a steward or a sensor, needs no service-rate estimate, and matches the signal operators already use to spot congestion. It also updates only at arrivals and departures, so the gate needs no extra state.

6.5.2 Decision rule

Let $|\mathcal{Q}(t)|$ denote the number of customers waiting in the mixed-stall queue at time t , counting both genders. When a mixed stall becomes free and a bypass would otherwise be considered, the gate is evaluated against a single integer threshold K :

$$\begin{aligned} |\mathcal{Q}(t)| < K & \quad \textbf{(light load):} \text{ Token Bucket bypass logic applies.} \\ |\mathcal{Q}(t)| \geq K & \quad \textbf{(heavy load):} \text{ strict FCFS enforced, no tokens consumed.} \end{aligned}$$

6.5.3 Optimisation

The optimisation procedure of Section 6.3 is applied with one addition. A third parameter, the integer queue-length threshold $K \in \{1, \dots, 40\}$, is searched as an Optuna integer variable alongside the token rate and bucket capacity. To keep the two policies apart, the Threshold gate's token rate and capacity are written $r_{\text{tok}}^{\text{TH}}$ (tokens per minute) and C^{TH} (tokens). Their roles inside the bypass logic match $r_{\text{tok}}^{\text{TB}}$ and C^{TB} in Section 6.4, but the optimal values differ because the load-aware gate changes the token dynamics. The objective is the shared $\mathcal{L}$ of Equation (6.1), evaluated over 100 Optuna TPE trials with 500 replications each on Layout 6.

A search-space sensitivity check was needed because the integer K has no obvious upper bound. An initial study over $K \in [0, 20]$, $C^{\text{TH}} \in [1, 20]$ converged to $K^* \approx 19$, $C^{*,\text{TH}} = 20$, $\Delta W = 6$ s, with both parameters pinned to the upper edges of their ranges. Widening the search to $K \in [1, 40]$, $C^{\text{TH}} \in [1, 30]$ moved K^* to 39 (still on the upper edge), brought $C^{*,\text{TH}}$ to 21 (now inside its range), and cut ΔW to 0.2 s on the search seeds.

K keeps saturating against the upper bound of its search range, which matches the $K \rightarrow \infty$ limit: there the gate never fires and Threshold Priority collapses to the Token Bucket. The optimiser is pulling the policy toward that limit. The safety gate fires only in the high-congestion bursts at the tail of the MMPP distribution, which are rare in steady state.

With 500 replications, the estimated noise floor is still about $\epsilon_{500} \approx 1.8$ s. This is much larger than the best observed trial value of 0.2 s, so that trial should not be interpreted as a clearly better solution. In fact, 27 of the 100 trials are within ϵ_{500} of perfect parity (panel (a) of Figure 6.3).

The robust optimum procedure of Section 6.3.2 handles this. The basin spans $K \in [32, 40]$, $r_{\text{tok}}^{\text{TH}} \in [7.0, 12.0]$, $C^{\text{TH}} \in [16, 30]$, centred on $K = 36$, $r_{\text{tok}}^{\text{TH}} = 9.55$, $C^{\text{TH}} = 21$. A $5^3 = 125$ -point sweep of that basin picks the interior point with the best worst-case score across itself and its neighbours:

$$K^* = 34, \quad r_{\text{tok}}^{*,\text{TH}} = 9.55, \quad C^{*,\text{TH}} = 20.$$

At this point the gap is within the noise floor of parity, so these are the values used below for the analysis.

The hyperparameter importance analysis in Figure 6.3(b) shows that bucket capacity again has little effect on the objective value. Instead, the result is mainly driven by the token rate $r_{\text{tok}}^{\text{TH}}$ and the threshold K . Together, these two parameters explain almost all of the objective

variance, with each contributing roughly half. The capacity parameter C^{TH} explains only about 2%.

By comparison, for the Token Bucket policy, the token rate almost fully determines the objective value (Figure 6.2). With the threshold extension, K becomes equally important, but mainly because it controls how close the policy is to the original Token Bucket. Low values of K activate the threshold gate more often, while high values of K make the policy behave more like a standard Token Bucket. The importance of K mostly reflects the difference between these low- K and high- K regimes. Bucket capacity remains a secondary parameter under both policies.

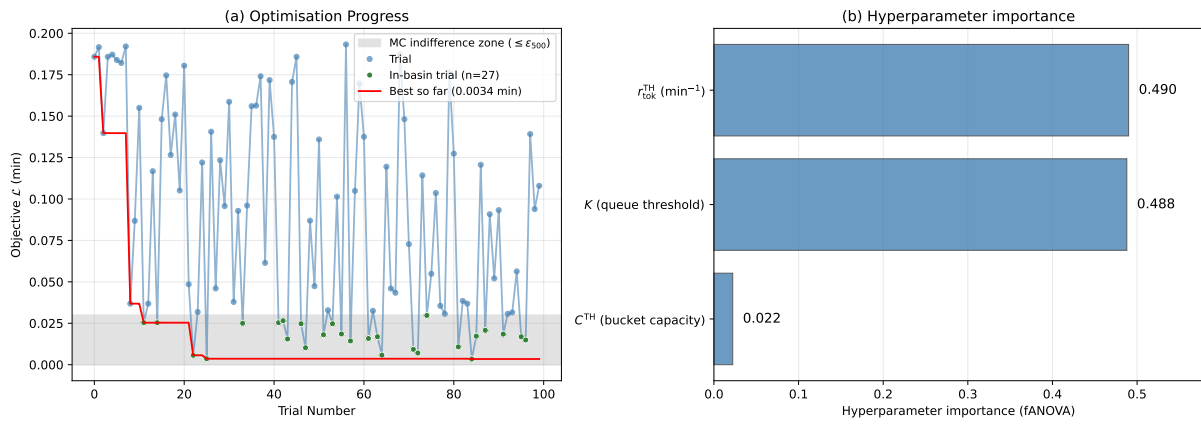


Figure 6.3: Bayesian optimisation convergence for the Threshold Priority widened search ($K \in [1, 40]$, $C^{\text{TH}} \in [1, 30]$; 100 Optuna TPE trials, 500 replications each) on a linear objective axis (left) and hyperparameter importance analysis (right). Compare with Figure 6.2 for the Token Bucket.

6.5.4 Results

The results below use the robust optimum found above: $K^* = 34$ customers, $r_{\text{tok}}^{*,\text{TH}} = 9.55$ tokens/min, and $C^{*,\text{TH}} = 20$.

A confirmation run with these settings gives a remaining gender gap of $\Delta W = 0.6$ s. At the precision used in this chapter, this is statistically indistinguishable from zero.

Policy	W_w	W_m	Gap ΔW	Men vs. FCFS
FCFS	37s	25s	11s	n/a
Token Bucket	32s	32s	0.2s	+25.5%
Threshold Priority	33s	32s	0.6s	+26.9%

Table 6.4: Three-policy comparison on Layout 6 at the robust optima (1000 MMPP replications). Token Bucket uses parameters from Table 6.2; Threshold Priority uses $K^* = 34$, $r_{\text{tok}}^{*,\text{TH}} = 9.55/\text{min}$, $C^{*,\text{TH}} = 20$.

Interpreting the optimal threshold. The optimal threshold, $K^* = 34$ customers, is a queue length the mixed-stall queue reaches during MMPP busy-state bursts. In the calm state, the arrival rate rarely pushes the queue past this point.

When the queue becomes very long, the gate limits additional bypasses that would otherwise delay men further. At $K = 0$, the gate is always closed and the policy reduces to FCFS. At large K , it never closes and the policy reduces to the pure Token Bucket. The main difference is the token rate. Threshold Priority uses $r_{\text{tok}}^{*,\text{TH}} = 9.55$ tokens/min, about 38 % higher than the Token Bucket's 6.91 tokens/min. That higher rate matters more than how often the gate closes. The Token Bucket has no separate starvation safeguard, so its rate must stay low enough to protect men's tail waits during the worst MMPP bursts. Threshold Priority puts that protection in the queue-length gate: when the queue reaches $K = 34$, the gate handles the overload and lets the policy use a more aggressive token rate.

The capacity result is also stable. The robust optimum uses $C^{*,\text{TH}} = 20$, well inside the search range and almost the same as the Token Bucket's $C^{*,\text{TB}} = 19$. The basin spans $C^{\text{TH}} \in [16, 30]$, with the robust optimum in the lower-middle of that range.

Token accumulation and bypass bursts. A concern with the threshold gate is what happens just after a busy period. While the gate is closed, no bypass tokens are spent, so tokens build up in the bucket. When the queue later falls below K and the gate opens again, these saved tokens can be used in a burst. This can create a long sequence of consecutive bypasses against the customer at the front of the queue.

Across 1 000 replications on Layout 6 (3 000-min warm-up), the gate reopens 30.7 times on average. Most of those reopenings produce short bursts (median run length 3), but the tail is heavy: P90 = 41, P99 = 111, the average of the per-replication maxima is 158, and the longest observed run is 319 consecutive bypasses. A man at the queue head when the gate reopens can be passed by dozens of women before he is served.

Queue-length data (86 756 samples) show the queue is empty 61 % of the time (mean 4.9 customers), with P90 = 17, P95 = 26, and P99 = 47. Since $K^* = 34$ lies between P95 and P99, the gate is closed during roughly the highest 2–3 % of observations. Section 6.7 compares this with the other three policies.

Future-work directions for tail mitigation. Two refinements would reduce the burst that appears when the gate reopens. Both are outside this thesis's scope.

The simpler option is to *drain* the bucket on every gate-close transition: tokens are discarded and bypasses must build up from scratch when the gate reopens. This kills the burst but weakens equity recovery, since every gate closure throws away tokens that were meant to compensate women later.

The second option protects the customers who are already waiting when the gate reopens. These customers are marked FCFS-protected and cannot be bypassed. Customers who arrive later still follow the normal token-bucket rule. Accumulated tokens are then used for new arrivals

instead of for bypassing the group that already waited through the closed period. This preserves the long-run equity while protecting the front of the queue from the burst.

There is one complication. Under MMPP load, the queue can move back and forth around K , causing repeated short closures that lock more and more customers and push the policy toward FCFS. A possible fix is hysteresis, with two thresholds $K_{\text{close}} > K_{\text{open}}$ to prevent flicker. The cost is a fourth parameter to optimise. How these variants should balance mean waiting time, tail waiting time, and gender equity, and how they should be optimised, is left for future work.

6.5.5 Cross-layout transfer

The Threshold policy is applied without re-optimising its parameters on Layouts 4 and 5, following the same cross-layout test used for the Token Bucket (Section 6.4.5), to check whether the parameters generalise to other mixed-stall layouts. Figure 6.4 shows the cross-layout waiting times and gender gaps. Table 6.5 reports the underlying numbers, including Layout 6, for which the parameters were optimised.

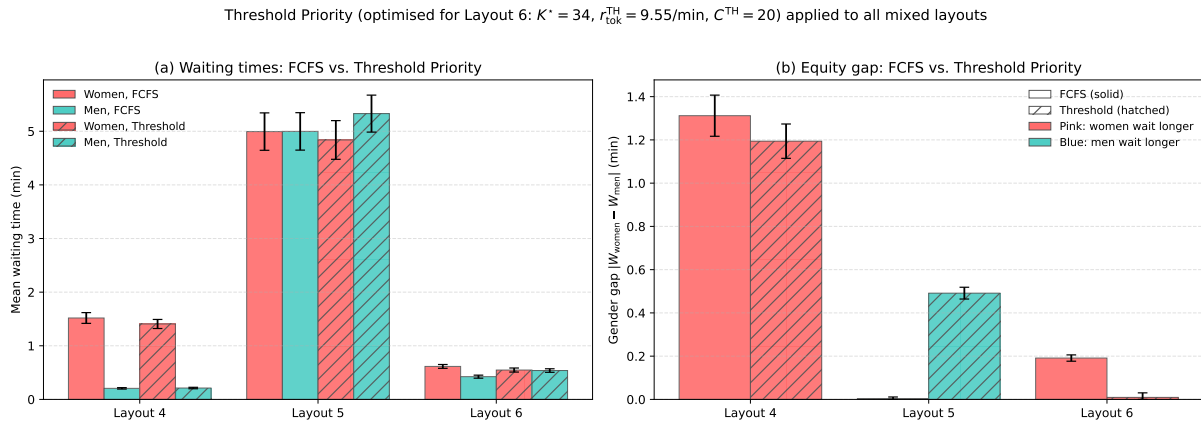


Figure 6.4: Threshold Priority cross-layout transfer (robust optimum $K^* = 34$, $r_{\text{tok}}^{*,\text{TH}} = 9.55/\text{min}$, $C^{*,\text{TH}} = 20$; 1 000 MMPP replications per layout; Layout 5 uses the 3 000-min warm-up). **(a)** Mean wait by gender under FCFS (solid) and Threshold (hatched); 95% CI half-widths shown. **(b)** Gender gap $|\Delta W|$, coloured by which gender is slower; paired-difference CIs.

The result is similar to the Token Bucket transfer (Section 6.4.5). On Layout 4, the gap is reduced by only 7 s. On Layout 5, the policy creates a 29-second gap favouring women, again showing we should not use priority policies for this layout. Layout 6 reaches parity. Section 6.10 compares Threshold Priority on Layout 6 with the other priority policies.

Layout	Policy	W_w	W_m	ΔW	Bypasses (obs window)	Note
Layout 4	FCFS	1min 31s	12s	1min 19s	n/a	women slower
Layout 4	Threshold	1min 24s	13s	1min 12s	33	women still slower
Layout 5	FCFS	5min 0s	5min 0s	0s	n/a	parity
Layout 5	Threshold	4min 44s	5min 13s	29s	177	men slower (gap created)
Layout 6	FCFS	37s	25s	11s	n/a	women slower
Layout 6	Threshold	33s	32s	0.6s	129	parity (optimised)

Table 6.5: Threshold Priority cross-layout transfer at the robust optimum ($K^* = 34$, $r_{\text{tok}}^{\star, \text{TH}} = 9.55/\text{min}$, $C^{\star, \text{TH}} = 20$). 1 000 MMPP replications per layout; Layout 5 uses the 3 000-min warm-up, Layouts 4 and 6 the 1 778-min baseline (Section 6.1). Bypasses are the mean women-first count over the 90-min observation window; FCFS rows use the same replication seeds as Tables 6.2 and 6.6.

6.6 Dynamic Aging

6.6.1 Definition and starvation-free property

Dynamic Aging priority replaces the binary women-first rule of Strict Head-of-Line priority with a continuous, time-dependent function. Each waiting customer’s *effective priority* at time t is:

$$p(t) = \begin{cases} 1 & \text{gender} = \text{woman}, \\ r_{\text{age}} \cdot (t - t_{\text{arr}}) & \text{gender} = \text{man}, \end{cases} \quad (6.2)$$

where t_{arr} is the customer’s arrival time and $r_{\text{age}} > 0$ is the only policy parameter (units: priority per minute). At each mixed-stall departure, the customer with the highest effective priority enters service.

A man’s priority grows linearly with his waiting time and overtakes the constant women’s baseline of 1 after a *crossover time*:

$$\tau = \frac{1}{r_{\text{age}}}. \quad (6.3)$$

No man waits indefinitely. Once a man has waited longer than τ minutes, he has higher priority than any woman, regardless of how many women are present. This makes the Aging policy starvation-free, in contrast to Strict Head-of-Line priority (Section 6.2), where a continuous stream of women can keep men waiting for a very long time (see Layout 5 in that section).

Aging also gives a stronger individual guarantee than the Token Bucket (Section 6.4) and Threshold policy (Section 6.5). Both are starvation-free as long as the token generation rate stays below service throughput, so the bucket can run dry between bursts. These two policies still admit runs of consecutive bypasses. The length of each run scales with C^{TB} plus any tokens built up during the burst, so the worst-case wait for an individual customer depends on both the bucket capacity and the token rate.

Under Aging, the bound is tighter: no man is bypassed once his wait exceeds $\tau = 1/r_{\text{age}}$. As $r_{\text{age}} \rightarrow 0$ the crossover time $\tau \rightarrow \infty$ and the policy approaches strict women-first. As $r_{\text{age}} \rightarrow \infty$ the crossover time $\tau \rightarrow 0$ and the policy approaches strict men-first priority. Unlike the Token Bucket, Aging keeps no token count: each priority decision uses only the clock and the arrival times of waiting customers.

6.6.2 Optimisation

The optimisation procedure of Section 6.3 is applied to the single parameter r_{age} with the shared objective of Equation (6.1). The search space for r_{age} is $[0.01, 10.0]$ on a log scale. The lower bound corresponds to $\tau = 100$ min, which exceeds the 90-minute observation window and so approximates Strict Head-of-Line priority in practice. The upper bound corresponds to $\tau = 6$ s, far shorter than any single service cycle, so virtually every queued man crosses τ before being served and the policy approximates strict men-first priority.

The study's single-best trial is at $r_{\text{age}} = 0.1288 \text{ min}^{-1}$. This corresponds to a crossover time of $\tau = 7 \text{ min } 46 \text{ s}$. The robust optimum from the basin-centred neighbourhood search (Section 6.3.2) is slightly lower at $r_{\text{age}}^* = 0.1271 \text{ min}^{-1}$, $\tau^* = 7 \text{ min } 52 \text{ s}$. The 1 000-replication confirmation at the robust optimum gives $\Delta W = 0.6 \text{ s}$ on Layout 6. The cross-layout transfer table below applies the robust parameters to all three layouts to keep the comparison consistent across rows.

Figure 6.5 shows the optimisation progress and the objective surface across the search range. With a single policy parameter, the objective collapses to a one-dimensional curve. This makes the search behaviour easier to interpret than the two- and three-parameter studies of Sections 6.4 and 6.5. TPE concentrates its trials near $r_{\text{age}} \approx 0.13 \text{ min}^{-1}$, and the surface dips below the noise floor only in a narrow band around the single-best and robust-optimum values. A hyper-parameter importance analysis is unnecessary because the variance attribution is trivially 100% to r_{age} .

6.6.3 Results

Figure 6.6 visualises the cross-layout waits and gender gaps under Dynamic Aging at the robust optimum; Table 6.6 reports the underlying numbers at the robust optimum $r_{\text{age}}^* = 0.1271 \text{ min}^{-1}$.

Layout 4 (12 mixed stalls, 10 urinals). Aging reduces the gap from 1 min 19 s to 1 min 14 s, again the same structural problem as before: too few men in the shared queue for reordering to matter. The 24 bypass events per 90-min observation window confirm this.

Layout 5 (20 shared stalls, no urinals). This layout illustrates the starvation-free property most clearly. Under Strict Head-of-Line priority the men's wait reaches 11 min 27 s, more than double the FCFS baseline of 5 min 0 s (Table 6.1). Under Aging priority applied at the Layout-6-optimised rate r_{age} (which was *not* re-optimised for Layout 5), both genders sit close to the

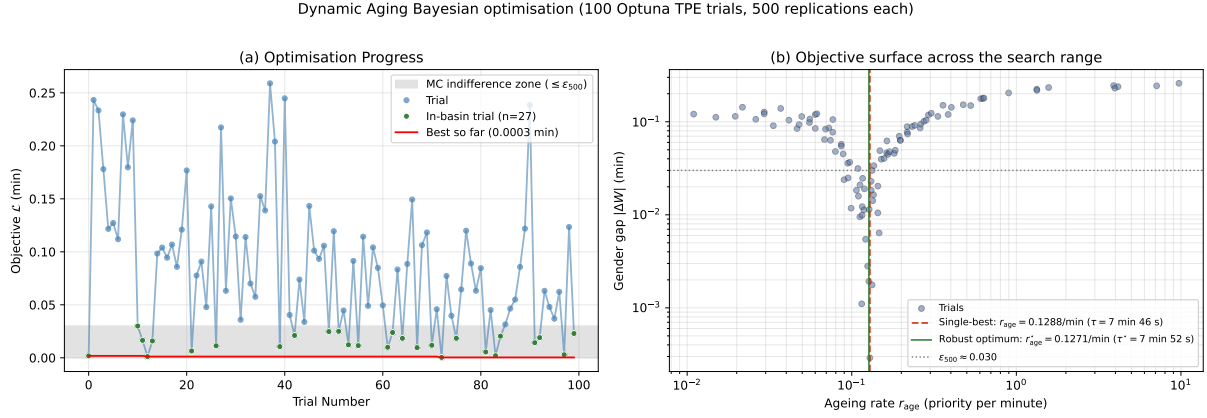


Figure 6.5: Dynamic Aging Bayesian optimisation summary (100 Optuna TPE trials, 500 replications each). **(a)** Convergence trace: trial-by-trial objective (blue) and running best (red), with noise floor ϵ_{500} for reference. **(b)** Objective surface $|\Delta W|$ vs r_{age} on log-log axes; single-best and robust optimum marked.

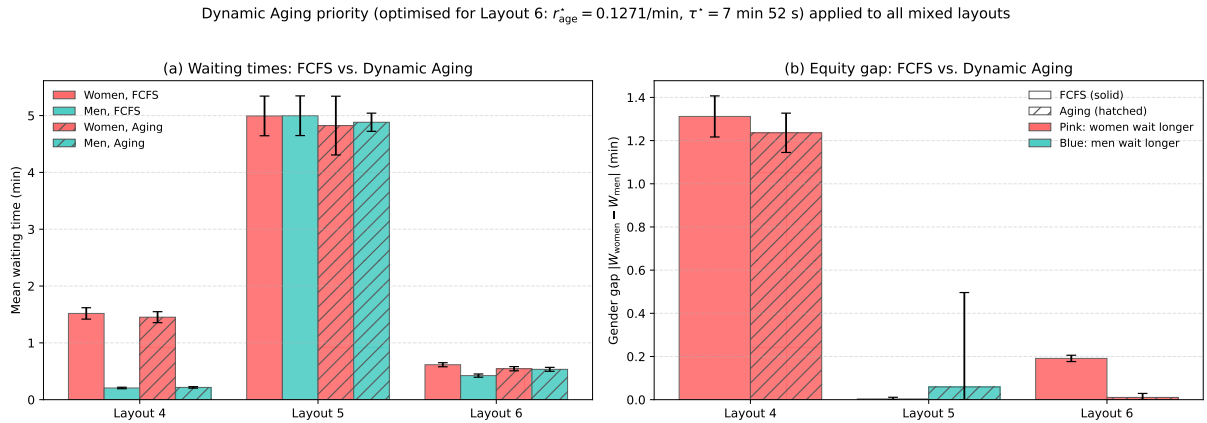


Figure 6.6: Dynamic Aging cross-layout transfer (robust optimum $r_{\text{age}}^* = 0.1271/\text{min}$, $\tau^* = 7 \text{ min } 52 \text{ s}$; 1000 MMPP replications per layout; Layout 5 uses the 3000-min warm-up). **(a)** Mean wait by gender under FCFS (solid) and Dynamic Aging (hatched); 95% CI half-widths shown. **(b)** Gender gap $|\Delta W|$, coloured by which gender is slower; paired-difference CIs.

Layout	Policy	W_w	W_m	ΔW	Bypasses (obs window)	Note
Layout 4	FCFS	1min 31s	12s	1min 19s	n/a	women slower
Layout 4	Aging	1min 27s	13s	1min 14s	24	women still slower
Layout 5	FCFS	5min 0s	5min 0s	0s	n/a	parity
Layout 5	Aging	4min 49s	4min 53s	3.5s	566	near-parity
Layout 6	FCFS	37s	25s	11s	n/a	women slower
Layout 6	Aging	33s	32s	0.6s	134	parity

Table 6.6: FCFS baseline versus Dynamic Aging priority at the robust optimum ($r_{\text{age}}^* = 0.1271 \text{ min}^{-1}$, $\tau^* = 7 \text{ min } 52 \text{ s}$) for Layouts 4, 5, and 6 (1 000 MMPP replications; Layout 5 uses the 3 000-min warm-up, Layouts 4 and 6 the 1 778-min baseline). Bypasses are the mean women-first count over the 90-min observation window; FCFS rows use the same replication seeds as Tables 6.2 and 6.5.

FCFS baseline: women wait 4 min 49 s and men 4 min 53 s, leaving a residual gap of $3.5 \text{ s} \pm 26 \text{ s}$ that is statistically indistinguishable from zero. The crossover time $\tau^* = 7 \text{ min } 52 \text{ s}$ prevents the unbounded bypass accumulation seen under Strict Head-of-Line priority on this layout. The wide confidence interval reflects Layout 5’s near-saturated regime: 47 % of mixed-stall departures trigger a bypass, against 5 % on Layout 4 and 38 % on Layout 6. Because Layout 5’s mean wait is already around 5 minutes, many men cross $\tau^* = 7 \text{ min } 52 \text{ s}$ before service, where Aging prioritises them over newly arriving women. This keeps the residual gap small.

Aging has a wider confidence interval ($\pm 26 \text{ s}$) than the Token Bucket ($\pm 7 \text{ s}$) or Threshold ($\pm 2 \text{ s}$). This is because small changes in MMPP timing can move a man’s wait just below or just above τ^* . That changes whether Aging protects him in that replication. The Token Bucket and Threshold policies are less sensitive to this kind of shift, because they depend on more aggregate state variables, such as token count and queue length. The starvation-free guarantee still holds: a waiting man eventually outranks newly arriving women.

Layout 6 (14 mixed stalls, 8 urinals). The result is again parity: $\Delta W = 0.3 \text{ s}$ at the single-best parameters, and $\Delta W = 0.6 \text{ s}$ at the robust optimum, both indistinguishable from zero. Men’s average wait rises from 25 s to 32 s at both optima (+25.9% at the robust optimum).

6.6.4 Summary

Dynamic Aging differs from the earlier policies in one important way.

Layouts 4 and 6 follow the same pattern as before. Queue reordering helps when the gap is mainly an ordering problem, as in Layout 6, where Aging reaches parity with $\Delta W = 0.6 \text{ s}$. It does not solve structural problems, like the one in Layout 4.

Layout 5 is the exception. The Token Bucket and Threshold Priority policies both opened

a new gap on this layout, while Strict Head-of-Line priority caused starvation. Aging does not show this problem. It preserves parity, with a remaining gap of $\Delta W = 3.5$ s, which is statistically indistinguishable from zero.

This comes from the starvation-free guarantee in Equation 6.2. Bypass accumulation is limited by the crossover time $\tau^* = 7$ min 52 s. This prevents the loss of parity observed for the other three policies on Layout 5.

Aging has three main advantages over the Token Bucket and Threshold Priority policies: it has only one policy parameter, it gives a direct crossover-time guarantee of $\tau^* = 7$ min 52 s, and it produces the lightest men's tail among the priority policies. Sections 6.7 and 6.10 provide the empirical evidence and the full four-policy comparison.

6.7 Cross-policy tail and starvation diagnostic

Sections 6.2–6.6 reported each policy's results in isolation. This section compares the four priority policies and the FCFS baseline on their tail-wait performance, looking at the extremes.

6.7.1 Men's tail diagnostic

The diagnostic uses common random numbers: 1 000 replications of each policy on Layout 6, with the same 3 000-min warm-up convention. For each run, every male customer's waiting time is recorded. Table 6.7 summarises the male wait distributions, and Figure 6.7 shows the corresponding empirical CDFs.

Mean wait and the pooled individual-wait percentiles are stable across replications. The mean and 90th-percentile per-replication maximum capture a typical bad session. The global (across 1 000 replications) maximum is a single observation and noisy.

Policy	Mean	P99	Mean per-rep max	P90 per-rep max	Global max	Bypasses/man	
						(P99)	(max)
FCFS	25s	4min 27s	2min 45s	4min 39s	12min 15s	0	0
Strict	37s	7min 45s	11min 17s	23min 27s	1h 7min 10s	128	1 001
Token Bucket	32s	7min 5s	7min 55s	13min 35s	27min 53s	76	207
Threshold Priority	32s	6min 34s	7min 57s	14min 20s	34min 38s	86	319
Dynamic Aging	32s	7min 52s	6min 21s	8min 22s	10min 52s	60	89

Table 6.7: Men's tail-wait diagnostics on Layout 6 (1 000 MMPP replications, 3 000-min warm-up). Mean and P99 are the mean and 99th-percentile of pooled individual men's waits. Mean and P90 per-rep max take the mean and 90th-percentile of the worst-wait man in each replication. Global max is the single largest wait observed across all replications. Bypasses/man (P99) and (max) are the 99th-percentile and maximum per-customer bypass count across all male arrivals.

Table 6.7 and Figure 6.7 show us several things.

FCFS looks like the best policy in the table and figure: it wins almost every column. The

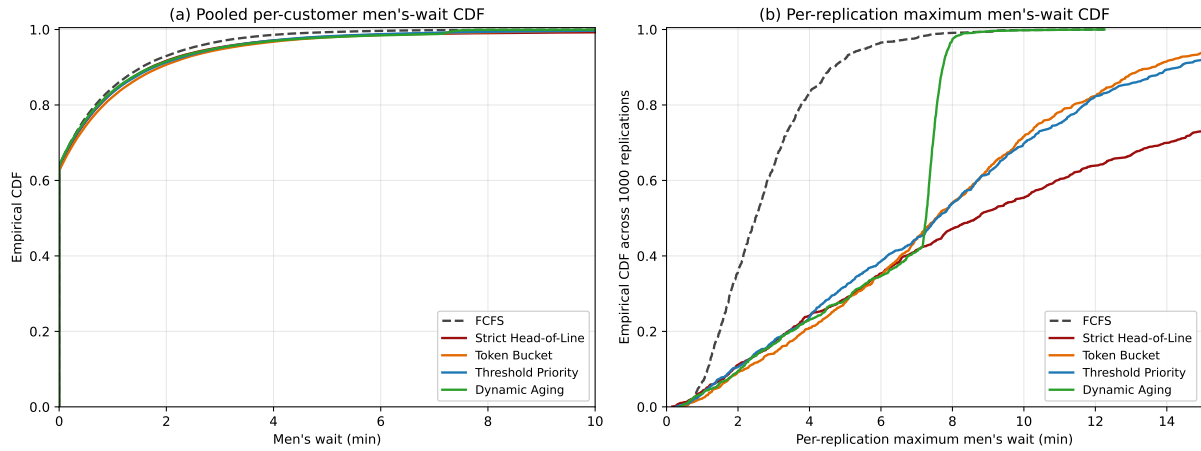


Figure 6.7: Empirical CDFs of men's waits on Layout 6 (1000 MMPP replications per policy, 3000-min warm-up). (a) Pooled individual-wait CDF $P(W \leq x)$. (b) CDF of the per-replication maximum $P(\max W \leq x)$. See Table 6.7 for summary statistics.

catch is the 11-second gender gap that motivates this chapter. At the other extreme, Strict Head-of-Line is the worst on almost every column, which fits its Section 6.2 role as a theoretical reference, not a policy you would implement in practice. The three optimised policies close the gender gap.

Aging has the cleanest tail of the four priority policies. Its mean per-replication maximum (6 min 21 s) is over a minute and a half below Token Bucket (7 min 55 s) and Threshold Priority (7 min 57 s). The 90th-percentile per-replication maximum (8 min 22 s) is also well below the 13 min 35 s and 14 min 20 s of the other two.

This is the starvation-free property of Section 6.6 in action. A waiting man's priority grows with his wait, and after the crossover time $\tau^* = 7$ min 52 s he outranks any newly arrived woman, which caps how often he can be bypassed. Figure 6.7(b) shows this very beautifully: Aging's per-replication-max CDF rises sharply around τ^* and reaches 1 a few minutes later. The other priority policies extend their tail well beyond.

The mean wait is the same across the three optimised policies (32 s), which follows from gender-parity optimisation plus the conservation argument of Section 6.1. What changes is the shape of the tail, which the table above brings out.

The global maximum also shows this. Aging's worst observation across 1000 replications is 10 min 52 s, below the FCFS baseline of 12 min 15 s. In this experiment, Aging never makes the worst observed wait worse than FCFS would. Token Bucket's worst case is 27 min 53 s, and Threshold's is 34 min 38 s, between two and three times the FCFS worst case. Threshold sits above Token Bucket because its safety gate, designed to prevent extreme bypass during bursts, instead creates a backlog of waiting women that releases at once when the gate reopens (Section 6.5.4). The introduction of the threshold gate fails to prevent extreme bypasses. Strict Head-of-Line at 1h 7min 10s confirms what Layout 5 already showed in Section 6.2: without a rate cap or a queue-length guard, the worst case is essentially unbounded. One caveat: the

global maximum is a single observation in 1 000 and is sensitive to noise, but the policy ordering matches the stable tail metrics.

The same pattern appears in how often individual men get bypassed. Each man's bypass count is the number of times a woman jumped ahead during his wait. The 99th-percentile of those counts (the threshold for the worst 1% of men) is 60 under Aging, 76 under Token Bucket, 86 under Threshold Priority, and 128 under Strict. Aging's worst-1% threshold is roughly $1.3\times$ tighter than Token Bucket's and $1.4\times$ tighter than Threshold's.

The single worst-bypassed man shows the same ordering more strongly: 89 bypasses under Aging, compared with 207 under Token Bucket, 319 under Threshold Priority, and 1 001 under Strict Head-of-Line. This comparison comes with the same caveat: it is based on one extreme observation and is sensitive to simulation noise.

Across most of the distribution, the four priority policies look much more similar. The mean per-man bypass count ranges from 5.4 to 7.0, and P90 ranges from 13 to 17. The difference only becomes clear in the upper tail, which is exactly the region we wanted to inspect.

6.7.2 Women's tail diagnostic

The same tail statistics for women show a different picture. The four priority policies barely change the women's upper-tail waiting times. Across all five policies, the women's 99th-percentile individual wait is about 5 min, and the global maximum stays between 10 and 15 min.

The priority policies reduce the women's mean wait, but only modestly. Under Token Bucket, Threshold Priority, and Dynamic Aging, the mean is about 32 s, compared with 38 s under FCFS. This matches the men's mean and is the gender-parity outcome. The reduction comes from the bulk of the distribution, not the upper tail: the women's longest waits remain similar across policies.

Figure 6.8 confirms this visually. In panel (a), the per-customer CDFs are almost indistinguishable across policies. In panel (b), the per-replication maximum CDFs are also much closer together than the men's curves in Figure 6.7.

The small differences that remain have the expected direction. FCFS performs slightly worse for women because it leaves the original gender gap in place: women wait longer on average. Strict Head-of-Line appears slightly better for women, but this is not a desirable improvement. It occurs because the policy overcompensates, shifting delay onto men and creating a gender gap in the opposite direction.

Bypass-count diagnostics are not reported for women, because women are the bypassers in all priority policies considered here. They are never the bypassed group. The promise in Section 6.1 that no individual should wait much longer than the rest is mainly about men under priority. The women's tail results are included as a check that the priority rules do not create a new tail problem on the women's side.

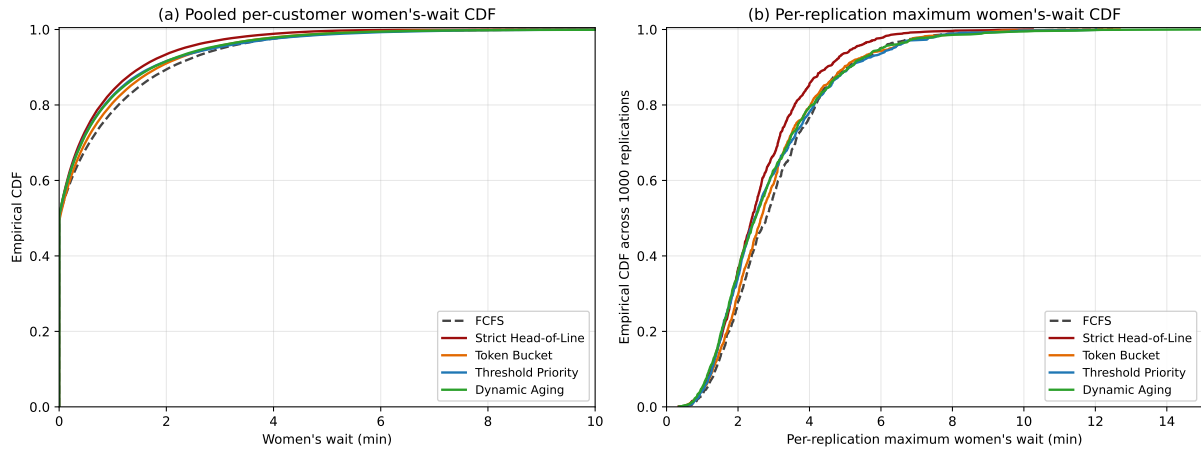


Figure 6.8: Empirical CDFs of women's waits on Layout 6 (1000 MMPP replications per policy, 3000-min warm-up). (a) Pooled individual-wait CDF $P(W \leq x)$. (b) CDF of the per-replication maximum. The three optimised policies are nearly indistinguishable.

6.7.3 Other observations

Calm versus busy demand. A second pattern is about when people arrive, not which policy runs. The simulation alternates between calm and busy demand periods. Customers who arrive during a busy period wait much longer than during a calm period: across all five policies, men's busy-state mean wait is 3.2 to $3.9\times$ their calm-state mean, and women's ratio is 2.7 to $2.9\times$.

Priority does not change this gap. It shifts a small amount of waiting from women to men during busy periods (0.1 to 0.3 minutes, depending on the policy), but the calm-versus-busy ratio stays the same. This is Kleinrock's conservation law (Section 6.1) at work again: priority can shift wait between men and women, but it cannot change the total wait. The law applies within each demand state separately, so the calm-versus-busy gap is set by the arrival pattern, not by the policy.

Per-customer worst case. The four priority policies differ in what they promise about an individual man's worst case. The main distinction is whether the policy controls bypassing only in aggregate, or whether it protects each customer directly.

Strict Head-of-Line offers no protection. A continuous stream of arriving women can keep men waiting indefinitely. In the Layout 6 diagnostic, the worst-bypassed man was bypassed 1001 times. On Layout 5, this same policy produced the practical starvation documented in Section 6.2.

The Token Bucket improves on this by capping the long-run bypass rate at $r_{\text{tok}}^{\star, \text{TB}}$. This is still an aggregate guarantee. It does not bound how many times one specific man can be bypassed, because tokens refill while he waits. The diagnostic worst case under the Token Bucket was 207 bypasses.

Threshold Priority adds a queue-length gate to the Token Bucket. When the queue exceeds

K , the gate suspends bypasses. This helps during high-load periods, but it does not create a per-customer bound. Tokens build up while the gate is closed, and once the queue drops below K , they can be released in a burst. In the diagnostic, the worst-bypassed man under Threshold Priority was bypassed 319 times, worse than under the Token Bucket despite the added safeguard.

Dynamic Aging gives the strongest guarantee. After a man has waited the crossover time $\tau^* = 7 \text{ min } 52 \text{ s}$, his priority exceeds any woman's, so no further bypasses are possible. This is a direct per-customer bound, not an aggregate rate limit. Consistent with this, Aging has the smallest diagnostic worst case among the four priority policies: 89 bypasses.

The next question is whether these conclusions still hold when the gender arrival split changes.

6.8 Sensitivity analysis: robustness to gender distribution

Section 5.4 showed that Layout 6 is the only layout that remains both feasible and close to equitable across the three gender arrival distributions tested: 60/40 female-skewed, 50/50 baseline, and 40/60 male-skewed. The priority policies in this chapter, however, were optimised only for the 50/50 baseline. This leaves two questions. First, do the 50/50-optimised policies still reduce the gender waiting-time gap when the arrival split changes? Second, if they do not, is it worth re-optimising the policy parameters for the new split?

This section answers both questions on Layout 6. The comparison includes FCFS and the four priority policies studied in this chapter: Strict Head-of-Line, Token Bucket, Threshold Priority, and Dynamic Aging. All experiments use the 3 000-minute warm-up from Section 5.4. This is the conservative upper bound used for the gender-split robustness analysis, so the same warm-up convention is kept here for consistency.

6.8.1 Transfer of 50/50-optimised parameters

First, we keep the priority-policy parameters fixed at their 50/50 robust optima and apply them to the 60/40 and 40/60 gender splits. This checks whether the baseline optimisation transfers to different gender splits.

Table 6.8 reports the mean waiting times and gender gap for FCFS and the four priority policies on Layout 6. For the priority policies, the parameter values are the 50/50 robust optima reported in Sections 6.4.4, 6.5.4, and 6.6.3. Each cell reports $\bar{W} \pm$ the 95% confidence interval half-width, based on 1 000 replications with a 3 000-minute warm-up. The same results are shown visually in Figure 6.9. Because this table uses the 3 000-minute warm-up convention from the gender-split sensitivity analysis, the 50/50 reference values can differ slightly from the main 1 778-minute baseline tables.

1. **Layout 6 remains feasible across all splits.** The FCFS row confirms that Layout 6 remains feasible at every split tested, as shown earlier in Section 5.4.1. This gives a valid

Policy	60/40 (female-skewed)			50/50 (baseline)			40/60 (male-skewed)		
	W_w	W_m	ΔW	W_w	W_m	ΔW	W_w	W_m	ΔW
FCFS (baseline)	1min 58s $\pm$ 8s	14s $\pm$ 1s	1min 43s $\pm$ 8s	36s $\pm$ 2s	25s $\pm$ 2s	11s $\pm$ 1s	24s $\pm$ 1s	23s $\pm$ 1s	0.6s $\pm$ 0.3s
Strict Head-of-Line	1min 54s $\pm$ 8s	18s $\pm$ 1s	1min 36s $\pm$ 7s	27s $\pm$ 1s	35s $\pm$ 3s	8s $\pm$ 2s	11s $\pm$ 1s	37s $\pm$ 3s	26s $\pm$ 2s
Token Bucket	1min 54s $\pm$ 8s	15s $\pm$ 1s	1min 39s $\pm$ 8s	33s $\pm$ 2s	34s $\pm$ 2s	0.8s $\pm$ 1s	13s $\pm$ 1s	36s $\pm$ 3s	23s $\pm$ 2s
Threshold Priority	1min 52s $\pm$ 7s	16s $\pm$ 1s	1min 36s $\pm$ 7s	32s $\pm$ 2s	31s $\pm$ 2s	0.3s $\pm$ 1s	15s $\pm$ 1s	33s $\pm$ 2s	18s $\pm$ 1s
Dynamic Aging	1min 51s $\pm$ 7s	15s $\pm$ 1s	1min 36s $\pm$ 7s	31s $\pm$ 2s	31s $\pm$ 2s	0.06s $\pm$ 1s	14s $\pm$ 1s	34s $\pm$ 2s	20s $\pm$ 1s

Table 6.8: Sensitivity of mean waiting times to the gender arrival distribution on Layout 6 (1 000 replications, 3 000-minute warm-up). Priority-policy parameters are fixed at their 50/50 robust optima. Values are mean $\pm$ 95 % CI half-width; ΔW is the absolute gender gap.

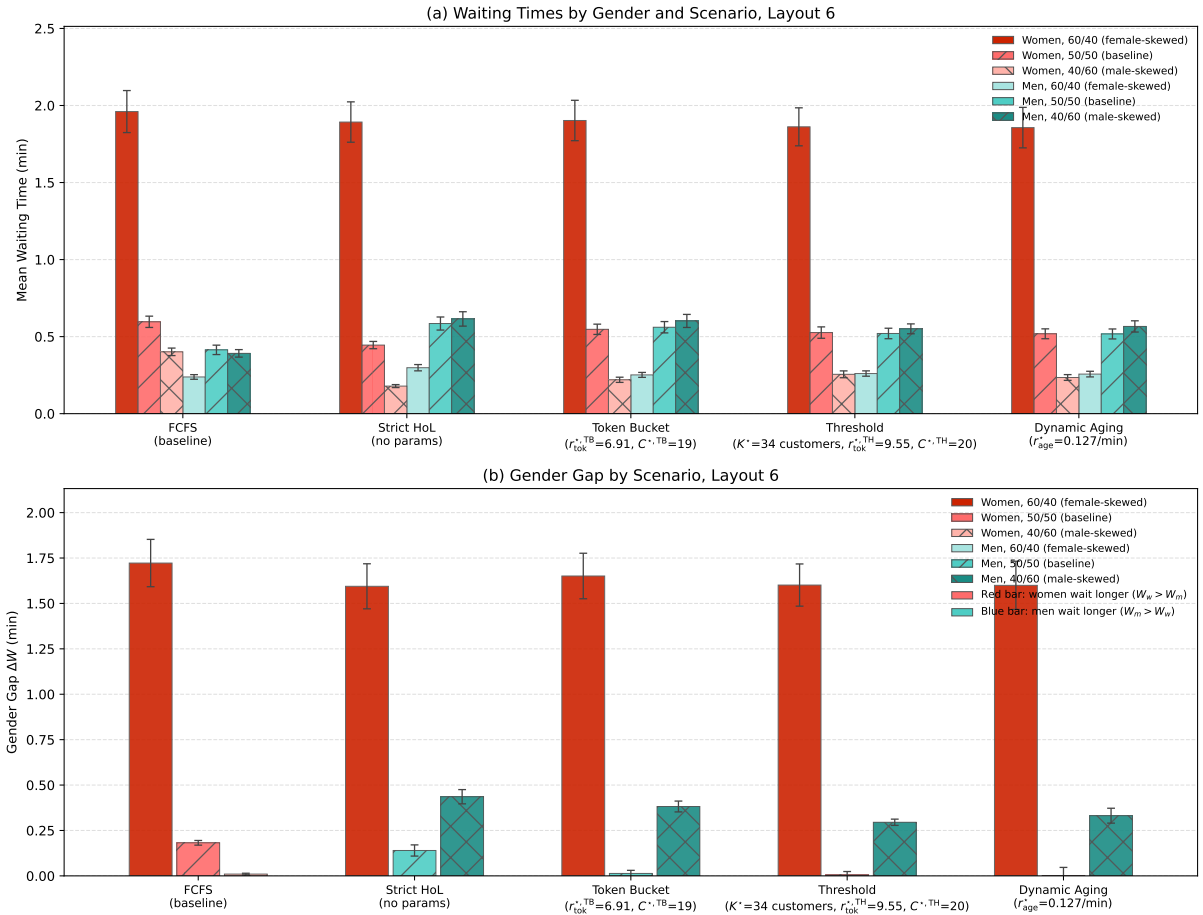


Figure 6.9: Sensitivity of mean waiting times in panel (a) and gender gap ΔW in panel (b) to the gender arrival distribution on Layout 6. The priority-policy parameters are fixed at their 50/50 robust optima. Red/pink indicate women, and blue indicates men. In panel (b), the bar colour indicates which gender waits longer.

baseline for comparing the priority policies across gender splits.

2. **The 50/50 parity results do not transfer to the 60/40 scenario.** Under the female-skewed split, FCFS produces a gap of approximately 1 min 43 s. The Token Bucket, at its 50/50 robust optimum, leaves this nearly unchanged at 1 min 39 s, a difference from FCFS within the 8 s confidence interval half-width. Threshold Priority and Dynamic Aging ($r_{\text{age}}^* = 0.1271/\text{min}$) both reach 1 min 36 s, also within one confidence interval half-width of FCFS and of each other. All three optimised policies share the same limitation: their parameters were optimised for a balanced arrival stream. At 60/40, women's arrivals exceed what those parameters were built to offset.
3. **The 40/60 scenario produces reversed gaps.** When men are the majority, women's arrival rate is lower and the shared stalls face less women-side pressure. Under FCFS, the gender gap almost disappears: $\Delta W = 0.6$ s.

The 50/50-optimised policies still apply priority, so they shift too much delay onto men. This creates reversed gaps of 23 s under the Token Bucket, 18 s under Threshold Priority, and 20 s under Dynamic Aging. These gaps are larger than the 50/50 FCFS gap of 11 s, but much smaller than the 60/40 female-skewed FCFS gap of 1 min 43 s. The issue is over-correction, not a capacity failure.

Threshold Priority over-corrects slightly less than the Token Bucket (18 s versus 23 s). A plausible explanation is that the threshold gate makes the policy partly revert to FCFS when the shared queue exceeds $K = 34$. This matters because FCFS is already close to parity at 40/60. The gate reduces unnecessary women-priority reordering, so the reversed gap is smaller than under the pure Token Bucket.

4. **Strict Head-of-Line behaves as expected.** Strict Head-of-Line shifts delay from women to men at every split. At 60/40, this is still not enough to close the gap. At 50/50 and 40/60, it over-corrects and leaves men waiting longer.

Overall, the transfer test shows that 50/50 optimisation is not robust to changes in the gender arrival split. Under 60/40, the optimised policies are too weak to offset the additional women's demand. Under 40/60, they are too strong and create reversed gaps.

6.8.2 Re-optimisation under skewed splits

The transfer results point to two different problems. At 40/60, FCFS already produces near-parity, so the optimiser has little to do except weaken the priority intervention and move the policy back towards FCFS. At 60/40, the transferred policies remain close to FCFS, but this may simply mean that the 50/50-optimised parameters are not aggressive enough for the female-skewed split. Re-optimisation tests whether stronger parameters can close the gap, or whether the layout sets a structural capacity ceiling.

To test this, Bayesian optimisation (Optuna TPE sampler) was repeated for the Token Bucket, Threshold Priority, and Dynamic Aging on Layout 6 under the 60/40 and 40/60 split.

Each study uses 100 Optuna trials with 500 replications per trial. The best parameter set is then confirmed with 1 000 replications.

Policy	Split	r_{tok}^*	C^*	K^*	r_{age}^*	W_w	W_m	ΔW	Men cost (%)
Token Bucket	40/60	0.15	17	n/a	n/a	25s	25s	$0.1\text{s} \pm 0.3\text{s}$	≈ 0
Token Bucket	60/40	8.98	21	n/a	n/a	1min 56s	16s	$1\text{min } 40\text{s} \pm 8\text{s}$	≈ 0
Threshold	40/60	0.12	6	17	n/a	24s	24s	$0.0\text{s} \pm 0.3\text{s}$	≈ 0
Threshold	60/40	4.04	17	8	n/a	1min 51s	14s	$1\text{min } 37\text{s} \pm 7\text{s}$	≈ 0
Dynamic Aging	40/60	n/a	n/a	n/a	0.69	24s	24s	$0.2\text{s} \pm 0.5\text{s}$	≈ 0
Dynamic Aging	60/40	n/a	n/a	n/a	6.03	1min 50s	15s	$1\text{min } 35\text{s} \pm 6\text{s}$	≈ 0
Token Bucket	50/50	6.91	19	n/a	n/a	33s	34s	$0.8\text{s} \pm 1\text{s}$	+35.4
Threshold	50/50	9.55	20	34	n/a	32s	31s	$0.3\text{s} \pm 1\text{s}$	+25.7
Dynamic Aging	50/50	n/a	n/a	n/a	0.13	31s	31s	$0.06\text{s} \pm 1\text{s}$	+24.9

Table 6.9: Re-optimised policy parameters and achieved gender gap on Layout 6 by gender split (1 000-replication confirmation, 3 000-minute warm-up). The 50/50 reference rows apply the robust-optimum parameters from Tables 6.2, 6.4, and 6.6, re-run here under the same 3 000-minute warm-up convention as the skewed-split rows. Men cost is the percentage change in W_m relative to FCFS under the same split. ≈ 0 means that the difference lies within the 95 % CI half-width.

1. Re-optimisation confirms that no active priority policy is needed at 40/60.

Under the male-skewed split, FCFS already gives near-parity: $\Delta W = 0.6$ s, with both genders waiting about 24 s. Re-optimisation mainly weakens the priority intervention. The Token Bucket and Threshold Priority reduce their token rates to 0.15/min and 0.12/min, respectively. Dynamic Aging increases r_{age} from 0.13/min to 0.69/min, shortening the crossover time to about 1.4 min. All three policies then hold both genders near 24–25 s, with residual gaps of at most 0.2 s and negligible men cost. The result confirms the existing FCFS equilibrium.

2. Re-optimisation does not close the 60/40 gap. Under the female-skewed split, the re-optimised gaps remain large: 1 min 40 s for the Token Bucket, 1 min 37 s for Threshold Priority, and 1 min 35 s for Dynamic Aging. The FCFS gap is 1 min 43 s. Dynamic Aging gives the largest apparent reduction, about 8 s, but this remains within the confidence interval of the FCFS estimate. No re-optimised policy gives a clear improvement over FCFS, and none comes close to restoring parity at 60/40.

We again have a structural bottleneck: during MMPP busy periods, $\lambda_{\text{busy}} = 24/\text{min}$. With a 60/40 split, women arrive at $0.6 \times 24 = 14.4/\text{min}$, while the 14 shared stalls can serve only $14 \times 1/1.5 = 9.3$ customers per minute. Women’s demand alone can exceed shared-stall capacity during these periods, so changing the queue order cannot remove the waiting-time gap.

Once the shared-stall queue becomes long, priority also has little to act on. Men are drawn towards the shorter urinal route, leaving almost no men in the shared-stall queue to move behind women. In Layout 6 under the 60/40 split, 99.2% of shared-stall queue snapshots contain no men. This is the same issue we saw in Layout 4 at the 50/50 split (zero men in 98.9% of shared-stall queue snapshots). In both cases, the priority rule has almost no male waiting time in the shared queue to transfer to women.

The unusual Dynamic Aging parameter reflects this lack of reordering opportunity. At 60/40, r_{age}^* rises to 6.03/min, reducing the crossover time to about 10 s. This should not be seen as a useful ladies-first policy parameter. The shared-stall queue is already almost entirely female, so the aging rate has little effect on women's waiting time. The optimiser is optimising noise, resulting in the unexpected parameter values.

At 40/60, no active priority policy is needed. FCFS already produces near-equal waiting times, re-optimisation does find parameters that reach near-parity, but they are in practice very similar to FCFS.

At 60/40, re-optimisation is insufficient. There are too few men left in the shared-stall queue to make a meaningful difference. The appropriate solution is a layout change, such as adding more shared stalls, rather than further adjustment of the priority rule.

6.9 Real-world implementation

The previous sections compared the priority policies on equity, men's tail waits, and robustness to gender-split changes. A real event operator faces a different question: can the rule be enforced clearly at the queue itself?

This section translates the four policies into practical requirements. It considers staffing, signage, timing information, and how easily users can see why a bypass is allowed. The main trade-off is simple: Dynamic Aging gives the best individual fairness guarantee, but it needs arrival-time information. The Token Bucket is easier to run, but it gives a weaker per-customer guarantee.

6.9.1 Common operational constraints

Three constraints apply to every candidate policy.

- **Visibility of the rule.** The rule must be clear to the next person in line. A bypass that looks like queue-jumping will be resisted, even if it is statistically justified.
- **Limited dwell time.** Event attendees spend only a short time in the queue. A system that requires an app download, account creation, or detailed explanation is unlikely to be followed.

- **Staff cost.** Steward time is usually the largest operating cost. A rule that needs continuous supervision is more expensive than one that a steward can monitor while doing other tasks.

The policies differ mainly in the information they need at the queue head.

6.9.2 Strict Head-of-Line

Strict Head-of-Line priority is the easiest rule to explain: whenever a mixed stall becomes free and a woman is anywhere in the queue, she is served next. A single steward and a clear sign, for example “Women have priority for shared stalls”, are enough to enforce it. No timing, counting, or measurement is needed.

This simplicity is also the problem. Section 6.2.2 showed that Strict Head-of-Line can cause starvation on Layout 5 and over-correction on Layout 6. The steward has no rule for switching priority off when it becomes too strong.

6.9.3 Token Bucket

The Token Bucket limits how often women can bypass men. It does not need complex technology. The steward only needs to know whether a bypass token is available.

1. **Visible token counter.** A small display at the queue head shows how many bypass tokens are available. Tokens are added automatically every $1/r_{\text{tok}}^{\star, \text{TB}} \approx 9$ s, up to the cap $C^{\star, \text{TB}} = 19$. The display can be digital, or a simple board with magnets, cards, or sliders that the steward updates when a timer signals a new token.

When a man reaches the head of the queue and a woman waits behind him, the steward checks the counter. If a token is available, the steward removes one token and lets the woman go next. If no token is available, the man is served. The visible token count makes the rule easier to accept, because users can see why a bypass is allowed.

2. **Automated counter or light signal.** A small device can track tokens internally and show when bypassing is allowed. Tokens are added automatically over time. When the steward lets a woman bypass, they press a button to spend one token. The signal then updates immediately. This removes manual token tracking and makes the rule visible to the queue, but it adds equipment cost and a possible failure point.

The Token Bucket is relatively easy to deploy because it is governed mainly by the token rate $r_{\text{tok}}^{\text{TB}}$ (Section 6.4.6). Bucket capacity matters much less, as long as it is large enough to cover short bursts. In practice, the most important setting is the average time between bypass opportunities.

6.9.4 Threshold Priority

Threshold Priority adds a visible threshold point to the Token Bucket. The threshold is marked on the floor at the approximate position where the queue reaches K^* customers. The rule is:

- **Queue below the line:** Token Bucket bypass logic applies.
- **Queue beyond the line:** no bypasses occur. The queue runs under FCFS until it shortens again.

Users can see whether the rule is active. The steward needs no service-time estimates, stall-occupancy data, or waiting-time predictions.

However, the floor marking is an approximation. Queues bend and spread unevenly, so the line will not always correspond to exactly K^* customers. It works as a visual trigger, but is not a precise count. It can be moved on site if the queue behaves differently than expected, but it will always be a noisy approximation.

The deployment cost is about the same as the Token Bucket, with one extra floor marking.

6.9.5 Dynamic Aging

Dynamic Aging has the clearest individual protection rule. After the crossover time $\tau^* = 7 \text{ min } 52 \text{ s}$, a waiting man can no longer be bypassed by newly arriving women. This makes the policy easy to communicate, but only if arrival times are recorded reliably.

The practical challenge is recording those arrival times. Three implementations are possible.

1. **Paper tickets with printed priority time.** A ticket printer at the queue entrance gives each man a numbered ticket showing the time from which he receives priority: $t_{\text{arr}} + \tau^*$. For example, the ticket could say “priority after 14:11”. Only men need this printed time, because women’s priority is constant under the Dynamic Aging rule.

The steward then applies a simple rule: women are served first unless a man has passed the priority time printed on his ticket. After that point, the longest-waiting eligible man goes next. This avoids asking the steward to compute Equation 6.2 directly.

2. **Queue-number display.** The same idea can be automated with a queue-number system. A small kiosk gives each arrival a number, and a screen at the queue head shows the next number to be served, similar to the order-number screens used in fast-food restaurants.

The system records each customer’s arrival time and applies Dynamic Aging in the background. For men, priority starts automatically after $t_{\text{arr}} + \tau^*$. The screen then tells the steward and the queue which number goes next. This reduces manual checking and lets the steward supervise the queue without constantly inspecting tickets.

3. **Wristband scanner.** Many large events already use chipped wristbands for entry or payments. In that setting, Dynamic Aging can be added with little extra effort. Each entrant taps their wristband at the queue entrance, and the system records their arrival time. If gender is registered in the system or selected at the queue entrance, the same tap can also identify which priority rule applies.

A screen at the queue head then shows the current service order. In a more automated version, the wristband could also vibrate or give another signal when the user's number is called, although this requires compatible hardware. This implementation avoids paper tickets and manual time checks. It is most attractive for venues that already have wristband infrastructure in place. For events without such infrastructure, it adds equipment cost and creates a possible failure point at the entry tap.

Dynamic Aging is preferred when reliable arrival time recording is already available. It has one policy parameter, a clear public guarantee, and the best men's tail among the deployable policies. Without arrival time checks, however, the policy is hard to enforce. The steward would no longer have a visible basis for deciding when a man has crossed the aging threshold, and the rule may drift towards Strict Head-of-Line priority in practice.

6.9.6 Handling different gender splits

The implementations above are most useful when the arrival split is close to balanced. Section 6.8 showed that parameters optimised at 50/50 do not transfer well to skewed crowds.

For a male-skewed crowd, close to the 40/60 case, no active priority rule is needed. FCFS already produces near-equal waiting times, and re-optimisation mainly reduces the priority effect for women. In practice, the priority policy can be disabled: the token counter is switched off, the threshold marking is ignored and there is no ticket printer.

For a female-skewed crowd, close to the 60/40 case, priority reordering is the wrong intervention. The shared-stall queue contains almost no men to reorder behind women. Event organisers who expect a female-skewed crowd should use a different layout with more shared-stall capacity, rather than trying to optimise the priority rule more aggressively.

The present thesis does not optimise a separate layout for female-skewed demand. Such a design would require the same type of analysis used earlier in this thesis: a layout comparison under the new demand mix, followed by policy optimisation only if the remaining gap is mainly an ordering problem. This is left for future work.

6.10 Comparative summary and recommendation

The preceding sections evaluated the four priority policies on four dimensions: equity at the 50/50 baseline, men's tail behaviour, robustness to gender-split changes, and practical deployment. The comparison in this section is restricted to Layout 6, the recommended physical

layout from Chapter 5. Table 6.10 summarises the baseline Layout 6 results. The discussion that follows combines these results into a deployment recommendation.

Policy	W_w	W_m	ΔW	Men vs. FCFS	Mean per-rep max (men)	Max bypasses per man
FCFS	37s	25s	11s	n/a	2min 45s	0
Strict Head-of-Line	28s	37s	9s	+48%	11min 17s	1 001
Token Bucket	32s	32s	0.2s	+25.5%	7min 55s	207
Threshold Priority	33s	32s	0.6s	+26.9%	7min 57s	319
Dynamic Aging	33s	32s	0.6s	+25.9%	6min 21s	89

Table 6.10: Four-policy comparison on Layout 6 at the robust optimum of each policy (1 000 MMPP replications).

Equity at 50/50. At the 50/50 baseline, the three optimised policies perform the same on gender equity. The residual gap is 0.2 s for the Token Bucket, 0.6 s for Threshold Priority, and 0.6 s for Dynamic Aging. These values are at the simulation noise floor, so all three policies are indistinguishable from parity.

Strict Head-of-Line is different. It reduces the absolute gap only slightly and overcompensates in the other direction: men wait 9 s longer than women, and men’s mean wait rises by 48 % relative to FCFS.

When priority is useful. The recommendation in this section applies only to Layout 6. Earlier cross-layout results showed that Layout 4 and Layout 5 are not suitable cases for priority-policy deployment: Layout 4 has too few men in the shared queue for reordering to matter, while Layout 5 already reaches parity under FCFS. The remaining question is which priority rule, if any, should be used on Layout 6.

On Layout 6, priority is useful only when the arrival split is close to balanced. At 50/50, the gender gap mainly comes from the service order in a moderately loaded shared queue. In that setting, all three optimised policies close the gap to the simulation noise floor.

The pattern changes once the split becomes skewed. At 40/60, FCFS already gives near-parity, so priority is unnecessary. At 60/40, priority is too weak: the shared queue contains almost no men to move behind women. The right response is then more shared-stall capacity, as priority policies have no chance to work when the shared queue is almost entirely female.

The recommendation below is for Layout 6 under a near-balanced arrival split.

Choosing between the deployable policies. The three optimised policies are tied on the mean gender gap (they all reach parity). The real choice is between tail protection and practical implementation.

Dynamic Aging is the best option if arrival times can be recorded reliably. Its rule is easy to state: once a man has waited $\tau^* = 7$ min 52 s, newly arriving women can no longer bypass him. This direct guarantee gives the lightest men's tail of the deployable policies. The mean per-replication maximum is 6 min 21 s, the worst per-customer bypass count is 89, and the global maximum is 10 min 52 s, below the FCFS baseline of 12 min 15 s. This policy might be a little bit more difficult to implement: the venue needs printed priority times, a queue-number display, wristband scanning, or another way to record arrival times.

The Token Bucket is the practical fallback when that infrastructure is not available. It is governed mainly by one parameter, $r_{\text{tok}}^{\text{*,TB}} = 6.91/\text{min}$, and can be run with a visible token counter or a steward-side app. It also reaches parity, but it only limits bypassing over time. It does not protect each man directly. Its men's tail is heavier than Dynamic Aging's, with a mean per-replication maximum of 7 min 55 s and a worst per-customer bypass count of 207.

Threshold Priority is a more visible version of the Token Bucket. A floor marking shows when the queue is long enough for the rule to switch back to FCFS, so users can see when priority is active and when it is suspended. This is useful operationally, but it does not improve the men's tail in the Layout 6 diagnostic: the mean per-replication maximum is 7 min 57 s, almost the same as the Token Bucket. Threshold Priority is not recommended for deployment, as it adds complexity without a clear benefit over the Token Bucket.

Strict Head-of-Line should not be deployed. It is easy to explain, but it has no protection against over-correction or starvation. Its role is as a theoretical reference point.

Final recommendation. For Layout 6 under a near-balanced gender split, Dynamic Aging is the preferred policy when arrival-time tracking is already available or can be easily implemented. It reaches parity, gives the strongest individual guarantee, and has the lightest men's tail.

Without arrival-time tracking, the Token Bucket is the most practical fallback. It is simple to explain, needs mainly one optimised parameter, and can be enforced by a single steward with a visible counter or app. Threshold Priority is not recommended since it adds complexity without a clear benefit over the Token Bucket, but can also reach parity. Strict Head-of-Line should not be deployed.

For clearly male-skewed events, FCFS is sufficient. For clearly female-skewed events, priority is not enough, and the layout should be adjusted by adding or reallocating shared-stall capacity.

6.11 Chapter summary

This chapter tested whether ladies-first priority policies can remove the remaining gender gap in Layout 6 without creating unacceptable delays for men. The starting point was FCFS on Layout 6, where women wait about 11 s longer than men.

The first result is that priority cannot reduce aggregate waiting time. This matches Kleinrock's conservation law and was confirmed empirically by an Optuna study that found no param-

eter setting with a meaningful efficiency gain over FCFS. Priority can only redistribute waiting time between groups.

Strict Head-of-Line shows why an unconditional ladies-first rule is too strong. It over-corrects on Layout 6, raising men's mean wait by 48 %, and it causes practical starvation on Layout 5. The three optimised policies avoid this failure. The Token Bucket, Threshold Priority, and Dynamic Aging all reduce the Layout 6 gender gap from 11 s to below 1 s, which is at the simulation noise floor. Men's mean wait increases between +25.5 % and +26.9 % relative to FCFS.

The main difference between the optimised policies is not the gender gap, but the men's tail. Dynamic Aging gives the strongest individual guarantee: after the crossover time $\tau^* = 7 \text{ min } 52 \text{ s}$, a waiting man can no longer be bypassed by newly arriving women. This gives the lightest men's tail, with a mean per-replication maximum of 6 min 21 s, compared with about 7 min 56 s for the Token Bucket and Threshold Priority. Dynamic Aging is the preferred policy when arrival-time tracking is available. Without that infrastructure, the Token Bucket is the practical fallback. Threshold Priority adds complexity, but it does not improve the men's tail relative to the Token Bucket, so it is not recommended for deployment.

The sensitivity analysis shows that this recommendation only applies in the right setting. Priority policies do not solve every gender split or every layout. At 40/60, FCFS already gives near-parity, so priority is unnecessary. At 60/40, re-optimisation does not close the gap, because the problem is shared-stall capacity rather than queue order. Across layouts, Layout 4 has too few men in the shared queue for reordering to matter, while Layout 5 already reaches parity under FCFS. Priority is useful only when the gap is mainly an ordering problem in a shared queue, as in Layout 6 under a near-balanced arrival split.

CHAPTER 7

Conclusion

7.1 Summary of findings

This chapter returns to the research questions introduced in Chapter 1. The central research question asked whether “ladies-first” priority policies can narrow the gender waiting-time gap in shared-capacity restroom facilities, and what this costs in overall system efficiency and in delays for individual male users. Two supporting questions also had to be answered. The first concerned the layout itself: under a fixed floor-area budget, which restroom configuration performs best? The second concerned robustness: do the layout and policy conclusions still hold when the gender split changes from the 50/50 baseline to 60/40 or 40/60?

The thesis answers these questions using the simulation framework introduced in Chapter 1 and validated in Chapters 3 and 4. Six fixed-area restroom layouts are tested under two-state Markov-modulated Poisson process (MMPP) demand. Service times depend on both gender and facility type, and men choose between stalls and urinals using a rational-agent routing rule. The priority-policy parameters are then tuned with Bayesian optimisation in Chapter 6. This framework links four elements that have to be considered together in the event setting studied here: layout design, bursty demand, routing behaviour, and queue discipline. Previous restroom-queueing studies have looked at these elements separately. This thesis evaluates them together, under a fixed floor-area budget.

Before interpreting the main results, the simulation framework was checked for warm-up bias and compared with analytical M/M/c benchmarks. The warm-up analysis selected a common warm-up period of 1 778 min for the main MMPP layout and policy studies, followed by a 90-min observation window. The MMPP demand and the men’s stall-versus-urinal routing rule both fall outside the M/M/c setting, so closed-form formulas cannot replace the simulation in the main study. For the verification step, the simulator was run in a simplified mode (MMPP switched off, mixed-layout routing decoupled), reducing each layout to a known M/M/c queue. Simulation estimates in this mode were then compared with M/M/c predictions using 1 000 replications and a 30 000-min observation window. Eight of the twelve queue-gender combinations had simulation

confidence intervals that included the analytical benchmark. The four mismatches occur in mixed-stall queues with gender-specific service rates. These queues are close to the $M/M/c$ setting, but not identical to it: they combine the men's and women's exponential service-time distributions into a hyperexponential mixture. The benchmark comparison in Chapter 4 is not a formal proof of simulator correctness, but a strong indication that the simulator behaves as expected wherever an analytical reference is available.

7.2 Layout design

The first step is to allocate capacity well. Capacity here means how the fixed floor area is divided between stalls and urinals, and between women's and men's facilities. Queue discipline only matters after that: if the layout is not suited to the demand, priority policies cannot work as intended. At the baseline 50/50 gender split, Layout 6, with 14 shared stalls and 8 men's urinals, is the best-performing configuration among the six layouts tested. It has the lowest overall mean wait among feasible layouts, 31 s, and the smallest non-zero gender gap, 11 s. It is also the only layout in which both women and men wait less than one minute on average.

The comparison also shows that the floor-area budget alone does not determine performance. The allocation of that area matters. Layout 1 follows the traditional equal-space allocation, but the women's subsystem is unstable under the modelled demand: $\rho_{\text{women}} = 1.050 > 1$. Its women's queue has no steady-state mean. Layout 5 takes the opposite approach by pooling all 20 stalls. This removes the mean gender gap by design, but it also removes the faster urinals and leaves the pooled system close to saturation. Both groups then wait about 5 min 21 s. Equal waiting times are not a useful outcome if the common wait is too long.

Layouts 2, 3, and 4 sit between these endpoints. None combine a low overall wait with a small gender gap: Layout 2 overcompensates and pushes men's wait above 2 min 30 s, while Layouts 3 and 4 leave women waiting around 1 min 30 s.

Layout 6 avoids both extremes. The shared stall pool keeps the system flexible by serving both genders, while the urinals keep much of the male demand out of the stall queue. In the baseline 50/50 FCFS run, both pools stay below saturation: $\rho_{\text{stalls}} = 0.820$ for the shared stalls and $\rho_{\text{urinals}} = 0.775$ for the urinals. This leaves headroom for MMPP bursts. As a result, Layout 6 reduces the gender gap while also reducing overall waiting time, all within the same floor-area budget.

7.3 Priority policies

Once the layout is well chosen, the priority policies target the residual gap that remains in the shared queue. They should not be used as a substitute for capacity design. They are useful only when the remaining gap is mainly caused by the order in which customers are served. In this thesis, that situation occurs for Layout 6 under a (near-)balanced gender split.

At the 50/50 baseline, the three optimised policies all close the remaining Layout 6 gap. The Token Bucket, Threshold Priority, and Dynamic Aging reduce the gap from 11 s under FCFS to below 1 s, which is at the simulation noise floor. The equity gain is not free: men’s mean wait increases by about +25.5 % to +26.9 % relative to FCFS.

This trade-off is consistent with Kleinrock’s conservation law [25]. A work-conserving priority rule does not add service capacity. It can only change the order in which waiting time is distributed between user groups. In the simulations, the reduction in women’s waiting time is matched by extra delay for men, while the aggregate waiting time remains almost unchanged. The optimisation run that targeted overall waiting time reduction found no meaningful improvement over FCFS, which confirms the same mechanism empirically.

Strict Head-of-Line illustrates the risk of an unconditional ladies-first rule. On Layout 6, it reduces the absolute gap only slightly and reverses its direction: men wait 9 s longer than women, and men’s mean wait rises by 48 % relative to FCFS. On Layout 5, the effect is much stronger. Men’s mean wait increases from 5 min 0 s to 11 min 27 s, creating practical starvation. Strict Head-of-Line is useful as a reference case, but it should not be deployed at large events.

Among the deployable policies, Dynamic Aging is the preferred option when arrival-time recording is easy to implement. It reaches parity like the other optimised policies, but it also protects individual male users from excessive delays. At the robust optimum, $r_{\text{age}}^* = 0.1271 \text{ min}^{-1}$, which corresponds to $\tau^* = 7 \text{ min } 52 \text{ s}$. Once a man has waited longer than this crossover time, newly arriving women can no longer bypass him. This gives Dynamic Aging the lightest men’s tail among the deployable policies. The mean per-replication maximum is 6 min 21 s, the worst per-customer bypass count is 89, and the global maximum is 10 min 52 s, below the FCFS baseline of 12 min 15 s. The operational drawback is that the venue needs a reliable way to record arrival times, such as printed priority times, a queue-number display, or wristband scanning.

When arrival-time recording is not available, the Token Bucket is the most practical fallback. It also reaches parity and only needs a token counter or a simple app controlled by the steward. Its behaviour is driven mainly by the token rate $r_{\text{tok}}^{*,\text{TB}} = 6.91/\text{min}$. The robust optimum also uses a bucket capacity of $C^{*,\text{TB}} = 19$. This makes the policy simpler to operate than Dynamic Aging because it does not require individual arrival times. The trade-off is that it regulates bypassing over time rather than protecting each man directly. A single man can still be bypassed many times during a burst, so its men’s tail remains heavier than under Dynamic Aging.

Threshold Priority also reaches parity, and its visible queue-length gate is easy to explain in practice. However, it does not improve the men’s tail over the Token Bucket on Layout 6. Its mean per-replication maximum is 7 min 57 s, almost the same as the Token Bucket, and the gate can create bypass bursts when it reopens after a busy period. Threshold Priority is not the main recommendation. Dynamic Aging is the primary recommendation, with the Token Bucket as a fallback when arrival times cannot be recorded.

The parameters that were optimised on Layout 6 do not transfer cleanly to other layouts. On Layout 4, all three policies barely move the gap: the urinals already serve most of the male demand, so the shared queue has too few men to reorder. On Layout 5, the Token Bucket and

Threshold Priority open a new gap favouring women, while Dynamic Aging holds parity thanks to its starvation-free crossover. Whether priority is the right tool depends on the layout it sits in.

7.4 Robustness to gender split

The robustness analysis shows that the policy recommendation depends on the gender split. Under the male-skewed 40/60 case, Layout 6 already gives near-parity under FCFS. An active priority rule is not needed. Re-optimising the priority parameters confirms this: the selected settings move the policies back toward FCFS and remove most of the women-priority effect.

The female-skewed 60/40 case is different. Bayesian re-optimisation of the priority parameters does not close the gap, because the gap is no longer mainly caused by service order in the shared queue. The shared queue simply contains too few men to move behind women. A stronger priority setting cannot fix that. For events with persistently female-skewed demand, the venue should change the capacity allocation instead of pushing the priority rule further. This would require a separate layout optimisation, which is outside the scope of this thesis and left for future work.

These skewed splits show that one priority rule does not fit every crowd. At 40/60, the system is already close to parity under FCFS, so priority is not needed. At 60/40, the gap comes from too little shared-stall capacity, not from service order. Priority should only be used when the arrival split is close enough to balanced for reordering to have an effect.

7.5 Limitations

The numerical results are useful as model-based estimates, but they should not be treated as design standards for every venue.

The first limitation is the service-time assumption. The model uses exponential service times, which supports the analytical verification and keeps the simulation aligned with M/M/c queueing theory. Real restroom service times, however, are unlikely to be memoryless. If they are less variable than an exponential distribution, the model may overstate waiting times. The relative ranking of layouts and policies is likely to be more stable than the exact values. This limitation can be revisited at low cost: unlike closed-form M/M/c analysis, the simulation is not tied to the exponential assumption, so any other service-time distribution can be swapped in without redesigning the framework.

The demand model is also simplified. The two-state MMPP captures an important feature of large-event demand: calmer periods interrupted by short bursts. In practice, event demand may follow a different pattern. Peaks may be fewer but larger, or demand may build up and fade more gradually than the two-state model assumes.

The model also treats arrivals as independent of the current queue state. In practice, cus-

tomers may avoid joining a long queue or leave after waiting for some time. Such behaviour would change the effective arrival pattern observed by the system.

The user model treats users within each gender category as homogeneous. It does not represent different service requirements within the same group. In particular, all men are assumed to be able to choose between a stall and a urinal. Cases where some men require a stall with certainty fall outside the model.

The model also uses two gender categories. This is a modelling simplification, not a statement about how restroom facilities should accommodate users in practice. Inclusive design also needs to consider non-binary and transgender users, as well as privacy, safety, accessibility, and local regulations. These aspects fall outside the quantitative model used in this thesis.

Spatial and queue-choice effects are also simplified. Once customers choose a queue, they stay in that queue until service. The model does not include switching queues after the initial choice, walking distance, facility visibility, crowd density, physical accessibility, or movement between other restroom locations in the venue.

Finally, the results have not been tested in a real deployment. The arrival and service parameters come from prior work and modelling assumptions, not from measurements at a specific venue. A field study would be needed before using the reported waiting times as design standards.

7.6 Future work

Some useful extensions fall outside the scope of this thesis. They would show how far the conclusions hold under more realistic assumptions and in real venues.

The first extension is to move beyond exponential service times. Real restroom use varies across users, but it is still bounded in practice. More general service-time distributions would capture this more realistically and would show how much the waiting-time estimates depend on the service-time assumption. The simulation supports this change at low cost: only the service-time draw needs to be replaced, while the rest of the framework remains unchanged.

A second extension is to model user differences in more detail. The model could allow some men to require a stall, instead of assuming that every man can choose between a stall and a urinal. It could also include group arrivals, accessibility needs, and different willingness to use mixed or gender-neutral facilities.

A future version of the model could also use arrival data from real events. Arrival patterns and service times could be measured at specific event sites. These data could then be used to fit a more detailed time-varying arrival process, such as a field-calibrated non-homogeneous Poisson process (NHPP). This would show whether the two-state MMPP is detailed enough for practical planning.

Future work could also include behavioural responses. In real facilities, customers may avoid a long queue, leave after waiting for some time, or move to another queue. Modelling balking,

reneging, and jockeying would bring the simulation closer to real queueing behaviour, especially during busy periods.

Chapter 6 also showed possible improvements to Threshold Priority. When the queue-length gate reopens, stored tokens can be released in a burst, causing many consecutive bypasses against the customer at the queue head. The gate can also switch on and off repeatedly when the queue stays close to the threshold. Three variants were suggested but not tested: emptying the bucket whenever the gate closes, protecting customers who were already waiting when the gate reopens, and using separate close and reopen thresholds. Future work could compare these variants on mean wait, men's tail, and gender equity, while optimising their parameters.

The female-skewed case should be studied as a separate layout problem. This thesis shows that the priority policies tested on Layout 6 do not close the 60/40 gap. Future work could optimise a layout for that demand mix first, and add priority policies if a residual ordering gap remains.

The most direct practical extension is a pilot implementation. Layout 6 could be tested in a real or temporary event facility. If a near-balanced residual gap remains, Dynamic Aging could be added as a second-stage policy. The pilot should record arrival times, user compliance, perceived fairness, and waiting times. This would show whether the simulation results carry over to a real queue.

7.7 Closing remarks

The gender waiting-time gap is not just a matter of behaviour or longer service times. In the event setting modelled in this thesis, much of it comes from how capacity is allocated. Layout 6 shows this clearly: with the same floor-area budget, it removes most of the gap and lowers the overall waiting time. If a small gap remains under a near-balanced arrival split, Dynamic Aging can close it when arrival times can be recorded reliably.

The practical message is simple. First, get the layout right, so the system has enough room to handle demand bursts. If a near-balanced shared queue still leaves a small gap, priority can then help as a second step. In the model studied here, this combination gives the best balance between shorter waits, gender equity, and protection of individual users.

The thesis opened with a visitor leaving a favourite festival performance and finding a restroom queue that stretches far beyond reason. The results show that this queue is not inevitable. With better capacity allocation and a controlled priority rule, the wait at the ladies' room can be shorter, fairer, and more predictable.

CHAPTER 8

Societal reflection

8.1 Public restrooms as gendered infrastructure

Public-restroom design rarely receives the same attention as other parts of public-space planning. This thesis shows why that matters. The allocation of floor area between stalls and urinals affects who waits, how long they wait, and whether a public facility offers comparable service to different users. Waiting time is not only an operational performance measure. It is also a measure of service equity.

The results of this thesis show that equal floor area does not necessarily mean equal service. At first sight, giving women's and men's facilities the same amount of space may seem fair. Stalls and urinals have different throughputs, and service times differ between user groups. Once these factors are included, the same floor area can produce very different effective capacities. A rule that appears neutral can still lead to unequal access.

Earlier work makes the same point. Anthony and Dufresne [3] describe public restrooms as one of the few remaining sex-segregated spaces in the American public landscape, and show how restroom provision has been shaped by gender, class, race, physical ability, and sexual orientation. Restroom design is not only a technical question about fixtures and queues, but also part of a longer history of unequal access to public space.

For that reason, a purely efficiency-oriented design is not enough. A layout could minimise total average waiting time and still leave one group systematically worse off. The gender gap in waiting time should not be treated as a secondary statistic, but as a design objective in its own right. The engineering question is not only how to make the system faster. It is also how the waiting time is distributed between users. This also links the thesis to three Sustainable Development Goals. SDG 5 is about gender equality. In this thesis, the issue is not formal exclusion, but unequal service quality: women can have access to a facility and still wait much longer. SDG 10 is about reducing inequalities. The results show that a rule can treat groups equally on paper while still producing unequal access in practice. SDG 11 is about inclusive and sustainable cities and communities. Public infrastructure should not only exist, it should also

work fairly for the people who depend on it. This thesis studies one small but very visible part of that larger question.

8.2 From simulation results to public-space decisions

The results show that a partially shared layout can substantially reduce the gender waiting-time gap within the same floor-area budget. In the modelled setting, the problem is not simply how much space is available. It also depends on how that space is organised. Layout choice shapes who gets access to capacity and who is left waiting.

The practical value of the work goes beyond the simulation model. Festival organisers, architects, infrastructure providers, municipalities, and building-code regulators can use the results to compare layouts and to question whether fixed gender ratios and equal-area rules deliver equal service quality. In Belgium, federal workplace legislation requires gender-separated facilities [16]. Temporary and festival structures are less restricted in this respect, so the mixed-layout results are most directly relevant for large-event settings.

The modelling framework is not limited to festival restrooms. The same approach could be used for other public venues with time-varying demand, such as train stations, stadiums, shopping centres, and airports. The broader lesson is that equitable public-space design does not always start with more space or more fixtures. It starts with understanding how the available capacity is used, which groups it serves, and where the bottlenecks remain.

8.3 Ethical limits and implementation risks

The main ethical limitation of the model is its binary treatment of gender. This reflects the categories used in current building codes and in much of the available data, but it excludes users for whom neither category fits. It also leaves out identity-related costs, such as discomfort, safety concerns, and the stress of choosing between gendered facilities. Waiting time alone cannot capture these experiences. Availability, safety, privacy, dignity, and accessibility also matter [15]. Users with disabilities, older adults, and adults accompanying small children may have service needs that differ from the average user [2].

The literature also cautions against treating shared provision as automatically inclusive. Farajollahzadeh and Hu [19] show that user preferences, safety perceptions, and identity-related disutility can change how gender-neutral facilities are evaluated. Greed [14] makes a similar point from a planning perspective: changing the sign on a restroom door is not enough. If the same limited stalls remain available, the same users may still face long waits. Shared provision only helps when the capacity itself is better allocated.

The findings of this thesis should be read with that caution in mind. The mixed layout performs well in the model because it changes how capacity is allocated. It is not evidence that every facility should become fully gender-neutral.

A further responsibility is to move from simulation to evidence. The arrival and service parameters used in this thesis are based on prior modelling work by Van Hautegeem and Rogiest [4], not on measurements from one specific event site. A pilot deployment of Layout 6 at a cooperating event would test whether the simulated gains also appear in practice, while revealing social effects that the model cannot fully predict, such as privacy concerns, signage clarity, whether users understand how the queue is organised, and user acceptance.

Even the partially shared layout does not remove the gender gap completely. A small residual disparity remains. In this thesis, that residual gap is addressed through the priority policies studied in Chapter 6. The simulation can show what priority would change, but it cannot decide whether the remaining gap is acceptable without it.

The priority policies themselves raise sharper implementation concerns. A Token Bucket or Dynamic Aging rule may reduce the measured waiting-time gap, but a real queue still needs a fair way to decide who the rule applies to. Self-declaration is the least intrusive option, but it may be open to misuse. Visual classification would be intrusive, unreliable, and ethically unacceptable. This shows a broader risk: a rule that is mathematically fair in a simulation can become socially problematic when it has to be applied in a real queue.

Data protection is part of the same issue. If eligibility or arrival times are recorded digitally, the deployment creates personal data where none may have existed before. Such data need a clear purpose, a retention period, and a way for users to decline recording without losing access to service.

This points to a broader engineering responsibility: not only to optimise a system once an objective has been chosen, but also to ask whether that objective captures what matters socially.

This is why average waiting time and gender gap should not be the only objective. It is easy to measure, well defined, and useful in queueing theory, which makes it a good starting point for this thesis. Hygiene, water use, accessibility, comfort, privacy, and perceived safety could all lead to different design choices. Choosing the metric is itself a value judgement, made early in the modelling process and carried through every result.

8.4 Concluding reflection

Long queues at women's restrooms are not only caused by behaviour or unavoidable demand peaks. They are also shaped by design rules, fixture choices, and assumptions about the default user. The engineering contribution of this thesis is to make that inequality measurable, so that it can be questioned rather than accepted as normal.

The findings are not a complete social solution. Privacy, safety, accessibility, gender identity, and user acceptance must still be considered alongside waiting-time performance. The role of the engineer is to provide quantitative evidence for better decisions, while also checking whether the objective being optimised reflects the people affected by the design.

Bibliography

- [1] *International Plumbing Code*. International Code Council. Washington, DC, 2021.
- [2] Clara Greed. *Inclusive Urban Design: Public Toilets*. Architectural Press, 2003.
- [3] Kathryn H. Anthony and Meghan Dufresne. “Potty Parity in Perspective: Gender and Family Issues in Planning and Designing Public Restrooms”. In: *Journal of Planning Literature* 21.3 (2007), pp. 267–294. DOI: 10.1177/0885412206295846.
- [4] Kurt Van Hautegeem and Wouter Rogiest. *Nooit meer file aan het damestoilet*. Eos Wetenschap. Originally published June 2017; webpage last revised 20 June 2023. Accessed: 2025-10-12. Technical appendix available at <https://www.eoswetenschap.eu/technischebijlage/toiletten>. 2017. URL: <https://www.eoswetenschap.eu/festivaltoiletten>.
- [5] David G. Kendall. “Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of the Imbedded Markov Chain”. In: *The Annals of Mathematical Statistics* 24.3 (1953), pp. 338–354. DOI: 10.1214/aoms/1177728975.
- [6] Herbert A. Simon. “A Behavioral Model of Rational Choice”. In: *The Quarterly Journal of Economics* 69.1 (1955), pp. 99–118. DOI: 10.2307/1884852.
- [7] Mor Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.
- [8] Wolfgang Fischer and Kathleen S. Meier-Hellstern. “The Markov-Modulated Poisson Process (MMPP) Cookbook”. In: *Performance Evaluation* 18.2 (1993), pp. 149–171. DOI: 10.1016/0166-5316(93)90035-S.
- [9] Averill M. Law. *Simulation Modeling and Analysis*. 5th. McGraw-Hill Education, 2015.
- [10] Joris Walraevens and Sabine Wittevrongel. *Queueing Theory*. Course text, Faculty of Engineering and Architecture. Ghent University, 2025.
- [11] Clara Greed. “Public Toilet Provision for Women in Britain: An Investigation of Discrimination Against Urination”. In: *Women’s Studies International Forum* 18.5–6 (1995), pp. 573–584. DOI: 10.1016/0277-5395(95)80094-6.
- [12] Michelle A. Baillie, Shawndel Fraser, and Michael J. Brown. “Do Women Spend More Time in the Restroom than Men?” In: *Psychological Reports* 105.3 (2009), pp. 789–790. DOI: 10.2466/pr0.105.3.789-790.

- [13] Steven M. V. Gwynne et al. “The Toilet Paper: Bathroom Dwell Time Observations at an Airport”. In: *Journal of Building Engineering* 24 (2019), p. 100751. DOI: 10.1016/j.jobe.2019.100751.
- [14] Clara Greed. “Join the Queue: Including Women’s Toilet Needs in Public Space”. In: *The Sociological Review* 67.4 (2019), pp. 908–926. DOI: 10.1177/0038026119854274.
- [15] Shawna Lewkowicz and Jason Gilliland. “A Feminist Critical Analysis of Public Toilets and Gender: A Systematic Review”. In: *Urban Affairs Review* 61.1 (2025), pp. 282–309. DOI: 10.1177/10780874241233529.
- [16] Federale Overheidsdienst Werkgelegenheid, Arbeid en Sociaal Overleg. *Codex over het welzijn op het werk*. Vastgesteld bij koninklijk besluit van 28 april 2017 (B.S. 2.6.2017). Geconsolideerde versie van 1 januari 2026. 2017.
- [17] Donald Gross et al. *Fundamentals of Queueing Theory*. 4th. John Wiley & Sons, 2008.
- [18] Woonghee Tim Huh et al. “The Potty Parity Problem: Towards Gender Equality at Restrooms in Business Facilities”. In: *Socio-Economic Planning Sciences* 68 (2019), p. 100693. DOI: 10.1016/j.seps.2018.11.003.
- [19] Setareh Farajollahzadeh and Ming Hu. “Potty Parity”. In: *Management Science* (2025). DOI: 10.1287/mnsc.2021.04075.
- [20] Khalidur Rahman et al. “An M/M/c/K State-Dependent Model for Pedestrian Flow Control and Design of Facilities”. In: *PLOS ONE* 10.7 (2015), e0133229. DOI: 10.1371/journal.pone.0133229.
- [21] Luc Bovens and Alexandru Marcoci. “The Gender-Neutral Bathroom: A New Frame and Some Nudges”. In: *Behavioural Public Policy* 7.1 (2020). Online-first July 2020; volume published January 2023, pp. 1–24. DOI: 10.1017/bpp.2020.23.
- [22] Lisa De Jonghe. “Genderneutrale toiletten en nudging: de sleutel tot kortere wachtrijen”. Master’s thesis. Ghent University, 2024.
- [23] Arthur Duysens. “Modeling and analysis of waiting times at toilet facilities with dynamic signage”. Master’s thesis. Ghent University, 2025.
- [24] Stephen G. Eick, William A. Massey, and Ward Whitt. “The Physics of the $M_t/G/\infty$ Queue”. In: *Operations Research* 41.4 (1993), pp. 731–742. DOI: 10.1287/opre.41.4.731.
- [25] Leonard Kleinrock. *Queueing Systems, Volume 1: Theory*. Wiley-Interscience, 1975.
- [26] Alan Cobham. “Priority Assignment in Waiting Line Problems”. In: *Journal of the Operations Research Society of America* 2.1 (1954), pp. 70–76. DOI: 10.1287/opre.2.1.70.
- [27] Robert B. Cooper. *Introduction to Queueing Theory*. 2nd. North Holland, 1981.
- [28] Małgorzata J. Krawczyk et al. “If Others Jump to the Queue Front, How Long I Will Wait?” In: *Acta Physica Polonica A* 127.3-A (2015), A–95–A–98. DOI: 10.12693/APhysPolA.127.A–95.

- [29] Jonathan S. Turner. “New Directions in Communications (or Which Way to the Information Age?)” In: *IEEE Communications Magazine* 24.10 (1986), pp. 8–15. DOI: 10.1109/MCOM.1986.1092946.
- [30] Herwig Bruneel. “Analysis of a Threshold-Based Priority Queue”. In: *Queueing Systems* 109.8 (2025), pp. 1–45. DOI: 10.1007/s11134-025-09936-3.
- [31] Leonard Kleinrock. “A Delay Dependent Queue Discipline”. In: *Naval Research Logistics Quarterly* 11.3-4 (1964), pp. 329–341. DOI: 10.1002/nav.3800110306.
- [32] Joris Walraevens et al. “Asymptotics of Waiting Time Distributions in the Accumulating Priority Queue”. In: *Queueing Systems* 101 (2022), pp. 221–244. DOI: 10.1007/s11134-022-09839-7.
- [33] Jerry Banks et al. *Discrete-Event System Simulation*. 4th. Pearson Education, 2005.
- [34] Averill M. Law. “Statistical Analysis of Simulation Output Data”. In: *Operations Research* 31.6 (1983), pp. 983–1029. DOI: 10.1287/opre.31.6.983.
- [35] Peter I. Frazier. “A Tutorial on Bayesian Optimization”. In: *arXiv preprint arXiv:1807.02811* (2018).
- [36] Eric Brochu, Vlad M. Cora, and Nando de Freitas. “A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning”. In: *arXiv preprint arXiv:1012.2599* (2010).
- [37] James Bergstra et al. “Algorithms for Hyper-parameter Optimization”. In: *Advances in Neural Information Processing Systems*. Vol. 24. 2011, pp. 2546–2554.
- [38] Takuya Akiba et al. “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2623–2631. DOI: 10.1145/3292500.3330701.
- [39] Stewart Robinson. *Simulation: The Practice of Model Development and Use*. John Wiley & Sons, 2004.
- [40] Victor Picheny, Tobias Wagner, and David Ginsbourger. “A Benchmark of Kriging-Based Infill Criteria for Noisy Optimization”. In: *Structural and Multidisciplinary Optimization* 48.3 (2013), pp. 607–626. DOI: 10.1007/s00158-013-0919-4.
- [41] Charles R. Harris, K. Jarrod Millman, et al. “Array programming with NumPy”. In: *Nature* 585.7825 (2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.
- [42] Pauli Virtanen, Ralf Gommers, et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17.3 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

APPENDIX A

Accuracy of the rational-agent heuristic

The rational-agent decision rule of Equation (1.1),

$$\widehat{E[W]} = \frac{W_{\text{queue}} + W_{\text{service}}}{c},$$

is a workload-based heuristic rather than the M/M/ c closed-form expected wait. This appendix justifies that choice, derives the heuristic’s bias against the closed-form rule, tabulates the resulting disagreement window for every layout in which men face a stall-versus-urinal decision, and works through a representative example for Layout 6. The six layouts referenced throughout this appendix are defined in Section 1.4.1 and Table 1.1.

A.1 Simulator state and the form of W_{queue} and W_{service}

The simulator retains the gender of customers waiting in queue but not of those currently in service (because an arriving man does not know the gender of those in service). This is why W_{queue} uses gender-specific service rates μ_i , while W_{service} is computed per gender and probability-weighted to give a single aggregate residual. For exponentially distributed service times, the memoryless property implies $E[S^2]/(2E[S]) = E[S]$, so the residual reduces to the mean and $W_{\text{service}} = k \cdot E[S]$.

A.2 Comparison with the M/M/ c closed-form expected wait

Equation (1.1) approximates the remaining workload per server, rather than the M/M/ c closed-form expected wait used here as the benchmark. This benchmark is exact only when all service at the facility follows a single exponential distribution. It applies directly to urinals and to single-gender stalls in separated layouts. For mixed stalls (Layouts 4 and 6) the service distribution is hyperexponential: a 50/50 mixture of the women’s 1.5-min and men’s 1.25-min exponentials. The 50/50 weighting reflects the arrival mix. The steady-state stall occupancy is in reality

skewed toward women, since men can also be routed to urinals, but the arrival mix is used here for simplicity. The bias quantified below measures the gap between the heuristic and the M/M/ c closed form under these assumptions. Sensitivity to the weighting is mild: the disagreement window (see A.3) widens from approximately 0.40 min to at most 0.47 min as occupancy skews toward women. For an M/M/ c queue with $n \geq c$ customers in the system on arrival, the closed-form expected wait is $(n - c + 1) E[S]/c$, while the heuristic gives $\widehat{E[W]} = n E[S]/c$. The two formulas differ by the constant

$$\text{bias} = \frac{(c - 1) E[S]}{c}, \quad (\text{A.1})$$

which depends only on the facility's parameters and not on the queue length. The heuristic overestimates the M/M/ c closed-form expected wait by less than one mean service time. The relative gap also vanishes in heavy traffic, since the queue term grows with n while the bias remains bounded by $E[S]$. In behavioural terms, the heuristic is a satisficing (satisfy + suffice) decision rule in the sense of Simon [6]: the arriving man accepts an adequate workload estimate instead of computing the M/M/ c closed-form optimum, which would require him to know the system is Markovian and to evaluate the formula on arrival.

A.3 Disagreement window

The arriving man's decision compares the heuristic estimate on the stall side against the heuristic estimate on the urinal side and joins the shorter one. If the heuristic added the same constant to both sides, the comparison would be unaffected and the heuristic would always agree with the closed-form rule. The two rules can disagree only within a *disagreement window*, defined as the absolute difference between the two per-facility biases:

$$\text{window} = \left| \frac{(c_{\text{stall}} - 1) E[S]_{\text{stall}}}{c_{\text{stall}}} - \frac{(c_{\text{urinal}} - 1) E[S]_{\text{urinal}}}{c_{\text{urinal}}} \right|. \quad (\text{A.2})$$

The two rules pick the same facility unless the two closed-form expected waits at the stalls and the urinals fall within this window. Table A.1 reports the disagreement window for each layout in which men face a choice. For separated layouts (1, 2, 3) the stall is men-only with $E[S]_{\text{stall}} = 1.25$ min. For mixed layouts (4, 6) the shared stall is population-weighted with $\bar{S}_{\text{mix}} = 0.5 \cdot 1.5 + 0.5 \cdot 1.25 = 1.375$ min. Urinals have $E[S]_{\text{urinal}} = 1.0$ min throughout. Layout 5 is fully gender-neutral (all 20 stalls pooled into a single bank, no urinals), so men have only one facility type and the heuristic does not enter their decision.

Every layout has a disagreement window below 0.41 min. Whenever the closed-form waits at the two facilities differ by more than the disagreement window, the heuristic and the closed-form rule pick the same facility. In the close-call states where they disagree, the arriving man's wait differs by at most the window's width, which remains under half a minute throughout. Arguably this small residual mismatch also reflects real-world behaviour, where an arriving man's quick judgement between two roughly equal queues need not coincide with the optimal closed-form choice.

Layout	c_{stall}	$E[S]_{\text{stall}}$	$\text{bias}_{\text{stall}}$	c_{urinal}	$\text{bias}_{\text{urinal}}$	window	larger side
1	2	1.25	0.625	10	0.900	0.275	urinal
2	2	1.25	0.625	6	0.833	0.208	urinal
3	2	1.25	0.625	8	0.875	0.250	urinal
4	12	1.375	1.260	10	0.900	0.360	stall
6	14	1.375	1.277	8	0.875	0.402	stall

Table A.1: Per-facility bias and disagreement window for each layout in which men face a stall-versus-urinal decision. Layout 5 is omitted because men have only one facility type. All values in minutes, rounded to three decimal places.

The larger-bias side flips with layout type. In separated layouts (1, 2, 3) the urinal bank has more servers than the men's stall bank, so the urinal bias is the larger of the two and the heuristic over-routes men to the stalls in close-call states. In mixed layouts (4, 6) the shared stall bank is the larger and serves both genders, so the stall bias dominates and the heuristic over-routes men to the urinals. The magnitude of any such re-routing is bounded by the window itself and remains under half a minute throughout.

A.4 Worked example: Layout 6

To illustrate, consider Layout 6 with all $c = 14$ mixed stalls busy and a queue containing six women and three men, while the eight urinals are also fully busy with twelve men waiting. Figure A.1 depicts this state and the decision the arriving man faces.

Using $\bar{S}_{\text{mix}} = 1.375$ min for the mixed stalls, the heuristic gives

$$\begin{aligned}\widehat{E[W]}_{\text{stall}} &= \frac{14 \cdot 1.375 + 6 \cdot 1.5 + 3 \cdot 1.25}{14} \approx 2.29 \text{ min}, \\ \widehat{E[W]}_{\text{urinal}} &= \frac{8 \cdot 1.0 + 12 \cdot 1.0}{8} = 2.5 \text{ min}.\end{aligned}$$

The arriving man joins the stall queue, since $\widehat{E[W]}_{\text{stall}} < \widehat{E[W]}_{\text{urinal}}$.

For comparison, the M/M/ c -equivalent prediction on the stall side with $n = 23$ and $E[S] = 1.375$ min gives

$$(23 - 14 + 1) \cdot \frac{1.375}{14} \approx 0.98 \text{ min}.$$

The 1.31-min gap between this prediction and the heuristic's 2.29 min matches the predicted per-facility bias $(c - 1)E[S]/c \approx 1.28$ min from Equation (A.1). On the urinal side, the same prediction with $n = 20$ and $E[S] = 1.0$ min gives

$$(20 - 8 + 1) \cdot \frac{1.0}{8} = \frac{13}{8} = 1.625 \text{ min}.$$

The closed-form ordering, $E[W]_{\text{stall}} \approx 0.98$ min $<$ $E[W]_{\text{urinal}} = 1.625$ min, agrees with the heuristic, so the man's choice of stall is unchanged. The closed-form waits differ by $1.625 - 0.98 =$

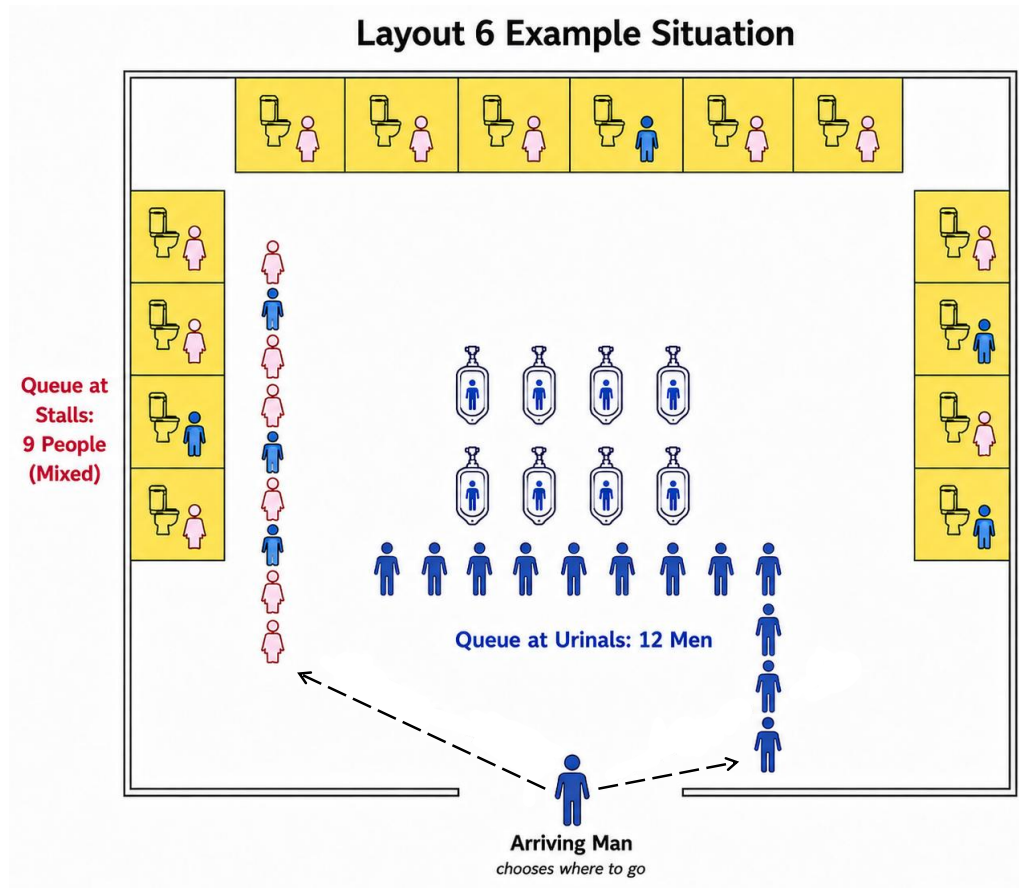


Figure A.1: Worked example for Layout 6 used to illustrate the rational-agent decision rule of Equation (1.1). All 14 mixed stalls and all 8 urinals are busy. The icon inside each facility marks the current occupant (pink for women, blue for men). The shared stall queue holds six women and three men, and the urinal queue holds twelve men. The dashed arrows indicate the two facility types the arriving man compares; under the heuristic he joins the stall queue, since its estimated wait is shorter.

0.645 min, well above Layout 6's 0.40-min disagreement window (Table A.1): the two rules agree, as already visible from the ordering.

APPENDIX B

Observation-window convergence study for verification residuals

Chapter 4 reports the analytical verification at an observation window of $T = 30\,000$ min. This appendix explains why that window length was chosen.

An earlier setting used a shorter observation window of $T = 1\,000$ min. Under that setting, Layout 6 men showed a residual of +1.8% that could not be explained by the closed-form model. The window scan in this appendix showed this was finite-window estimator variance, and it disappears when the observation window is increased to $T = 30\,000$ min.

The same scan clarifies the residuals on Layouts 2, 3, and 5. The residuals for Layout 2 and 3 men, and for Layout 5's pooled queue, do not disappear as T increases. They are caused by the hyperexponential service-time mixture. Layout 1 men shares this mechanism, but at $\rho = 0.52$ the predicted bias stays below CI noise at every T .

Layout 5 is reclassified by the longer window. At $T = 1\,000$ min its residual of approximately +1% lies inside the wide 95% CI and is seen as an analytical match. At $T = 30\,000$ min the CI is narrow enough to reveal a persistent residual of about +0.5%. This is the same hyperexponential mechanism as in Layouts 2 and 3 men, at a smaller magnitude.

B.1 Design of the diagnostic

The verification was rerun at four observation windows $T \in \{1\,000, 3\,000, 10\,000, 30\,000\}$ min, with 1 000 replications per layout at each T . The warm-up duration ($t_{\text{warm-up}} = 1\,778$ min) and the per-replication seeds were held fixed across T , so longer-window runs extend the shorter-window samples.

A residual driven by finite-window estimator variance shrinks as the observation window grows. The standard error of the sample mean scales as $1/\sqrt{T}$, so its relative size against the asymptotic value decays at the same rate. A residual driven by a closed-form mechanism, such

as the hyperexponential service-time mixture, instead converges to a non-zero asymptotic value as the estimator sharpens. The two signatures look qualitatively different.

B.2 Results

Figure B.1 plots the relative residual $(\hat{W}_q - W_q^{\text{ana}})/W_q^{\text{ana}}$ against T for each subsystem. Table B.1 reports the numerical values for the five subsystems with non-trivial residuals.

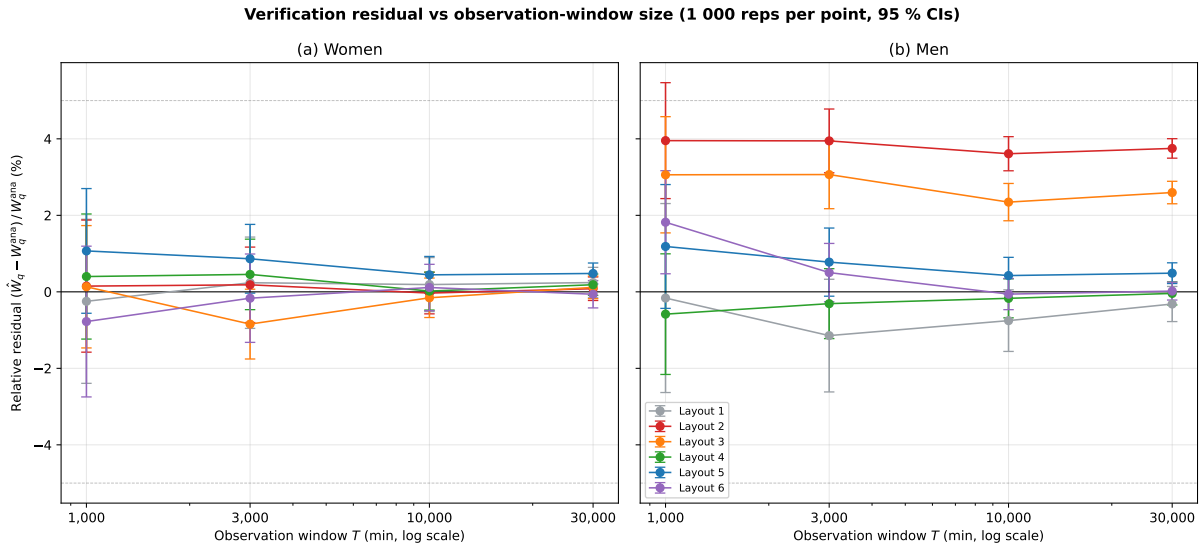


Figure B.1: Verification residual against observation-window size T . Each point is the mean across 1 000 independent replications at the indicated T . Vertical bars are the 95% confidence intervals expressed as a fraction of the analytical value. Dashed grey lines mark the $\pm 5\%$ reference band.

Subsystem	$T = 1\,000$	$T = 3\,000$	$T = 10\,000$	$T = 30\,000$
Layout 2 men	+3.95	+3.95	+3.61	+3.75
Layout 3 men	+3.06	+3.07	+2.35	+2.60
Layout 5 women	+1.07	+0.86	+0.44	+0.48
Layout 5 men	+1.19	+0.78	+0.43	+0.49
Layout 6 men	+1.82	+0.50	−0.06	+0.02

Table B.1: Relative residual (%) at each observation window for the five subsystems with non-trivial residuals at $T = 1\,000$.

Layout 6 men: finite-window confirmed. The residual collapses from +1.82% at $T = 1\,000$ to +0.02% at $T = 30\,000$. The 95% confidence interval narrows by a factor of approximately $\sqrt{30} \approx 5.5$ between the smallest and largest T , as expected from the $1/\sqrt{T}$ decay of the sample-mean standard error. At the same time, the observed residual of +0.02% is consistent with an

asymptotic value of zero. The +1.82% seen at $T = 1\,000$ can be explained as finite-window estimator variance, so it is not an error in the model. At the observation window of $T = 30\,000$, Layout 6 men sits comfortably inside the 95% CI (Table 4.3). An additional check supports this. The same verification at $T = 1\,000$ min was repeated under five other seeds. The analytical value fell inside the 95% CI in all five cases, with relative errors between -0.17% and $+0.65\%$. The +1.82% at the original seed may be a chance fluctuation rather than a real bias. A broader robustness study across many seeds falls outside the scope of this thesis.

Layouts 2 and 3 men: hyperexponential asymptote confirmed. Across the 30-fold range of T , Layout 2 men’s residual stays within $[+3.61\%, +3.95\%]$ and Layout 3 men’s within $[+2.35\%, +3.07\%]$. The 95% confidence intervals narrow by a factor of approximately 5.5 over the same range, but the residual point estimates do not move with them. This is the signature of a real asymptotic offset rather than estimator noise. The hyperexponential service-time mixture has $C_s^2 > 1$, which puts a positive bias on the mean wait relative to the matched-mean M/M/c value (Section 4.2.2). The diagnostic confirms that prediction at high precision.

Layout 5: smaller hyperexponential offset. At $T = 1\,000$, Layout 5’s residuals ($+1.07\%$ and $+1.19\%$, Table B.1) sit inside the 95% CI and are seen as analytical matches. At $T = 30\,000$, with confidence intervals narrower by a factor of approximately 5.5, the residuals settle to $+0.48\%$ and $+0.49\%$ and now fall outside the CI. The apparent match at $T = 1\,000$ was caused by the wide CI. The asymptotic residual itself is small ($\approx +0.5\%$) but persistent, consistent with the same hyperexponential mechanism that drives Layouts 2 and 3 men. The variance gap here is narrower than in the urinal–stall pool: women’s $\mu_w = 2/3/\text{min}$ versus men’s $\mu_m = 0.8/\text{min}$ gives a ratio of 1.20, against 1.25 for urinals versus stalls. The smaller hyperexponential coefficient $C_s^2 - 1$ produces a correspondingly smaller offset.

Summary. The convergence study explains the Layout 6 men residual as finite-window estimator variance. It confirms the hyperexponential mechanism for Layouts 2 and 3 men, whose residuals do not shrink as the CIs narrow. The same mechanism extends to Layout 5, where the smaller variance gap gives a residual of about $+0.5\%$. That residual was hidden inside the wide CI at $T = 1\,000$ min and only becomes visible at $T = 30\,000$ min. All four out-of-band residuals at $T = 30\,000$ min share the same hyperexponential cause. No unexplained residual remains.

APPENDIX C

Robust optimum selection methodology

Chapter 6, Section 6.3.2 introduces the robust-optimum procedure used to select the reported priority-policy parameters. This appendix gives the full three-step procedure that derives the robust optimum from the noise-floor basin, together with the parameter values used at each step.

C.1 Three-step procedure

The robust optimum is derived from the basin in three steps:

1. **Basin centre.** The basin is a set of trials, each with its own parameter values. We take the mean parameter values across all basin trials (one mean per parameter), and round to the nearest integer for integer-valued parameters. This gives a centre point in parameter space.
2. **Neighbourhood grid.** Around the centre we build a fixed grid of candidate parameter combinations. For each parameter independently, we pick 5 evenly-spaced values around the centre. The spacing between adjacent values is 15% of the centre value for continuous parameters, or 2 integer units for integer parameters. The grid is the Cartesian product of these per-parameter values. For example, the Token Bucket centre at $r_{\text{tok}} = 6.91$, $C = 19$ gives $r_{\text{tok}} \in \{4.84, 5.87, 6.91, 7.95, 8.98\}$ and $C \in \{15, 17, 19, 21, 23\}$, producing $5 \times 5 = 25$ grid combinations. Threshold (3 parameters) gives $5^3 = 125$ combinations; Dynamic Aging (1 parameter, r_{age}) gives 5. Each grid combination is then evaluated at 500 replications, so we have one ΔW value per grid point.
3. **Worst-case score.** For each grid point P , we compare P 's ΔW against the ΔW values of P 's grid neighbours (the grid points one step away from P along any single parameter axis). P 's worst-case score is the highest ΔW among P and its neighbours. At the grid boundary fewer neighbours exist, so the worst-case score is computed over whichever neighbours are

present; in practice the chosen optima for all three policies turned out to be interior grid points, so this corner case does not affect the reported results. We pick the grid point with the lowest worst-case score and reconfirm it at 1 000 replications.

The chosen optimum has both a good ΔW at its own point and a good ΔW at every immediate neighbour. A grid point with even one bad neighbour gets a high worst-case score and is ruled out. This rewards regions where the objective surface is flat: the reported parameters are insensitive to small parameter changes, so a slightly miscalibrated r_{tok} or K does not meaningfully worsen the outcome.

In practice, many interior grid points end up with similar worst-case scores. The procedure picks whichever comes out lowest, though several others would do equally well. The 1 000-replication reconfirmation then checks the chosen point still performs as expected.

C.2 Hyperparameter choices

We chose $\pm 15\%$ spacing for continuous parameters and ± 2 for integer parameters, giving a total grid span of $\pm 30\%$ (or ± 4 for integers) around the centre. This roughly matches the typical width of the noise-floor basin, so the grid covers the region where parameters perform equivalently. A narrower span would only test tiny changes, so every candidate would pass. A wider span would test points clearly outside the basin, so every candidate would fail. The 1 000-replication reconfirmation makes sure the chosen point still performs well, so the exact spacing is not critical.

APPENDIX D

Warm-up convergence diagnostic for Layout 5 priority policies

Section 6.1 of Chapter 6 sets every Layout 5 row in the priority-policy tables to a 3 000-min warm-up instead of the baseline 1 778-min value from Chapter 3. This appendix shows the diagnostic behind that choice.

D.1 Methodology

To check the warm-up, each priority policy (Strict Head-of-Line, Token Bucket, Threshold Priority, Dynamic Aging) was run on Layout 5 for 1 000 MMPP replications. Every run extended to 5 000 min with no warm-up filter applied. For each customer served, the arrival time and waiting time were logged. The cumulative mean wait was then sampled at 250 points between 50 min and 5 000 min, separately for women and men. Across replications, the mean and the 10th-to-90th percentile band show how quickly that mean wait settles into steady state. Layout 6 under Strict Head-of-Line was added as a contrast case. On Layouts 4 and 6, the milder policies cap the bypass rate and cannot relax more slowly than Strict, so the Strict result covers them.

D.2 Results

Layout 6 converges quickly under every policy. Men’s mean wait reaches steady state by about $t = 500$ min, so the baseline 1 778-min warm-up leaves plenty of margin (Figure D.1). Layout 4 behaves the same way: its urinal bank diverts most male stall demand and keeps the mixed-stall load light.

Layout 5 is different. It has no urinals and a single shared queue. At the 50/50 gender split, the 20 stalls are busy roughly 96 % of the time ($\rho_{\text{pooled}} = 0.96$). Even basic FCFS converges

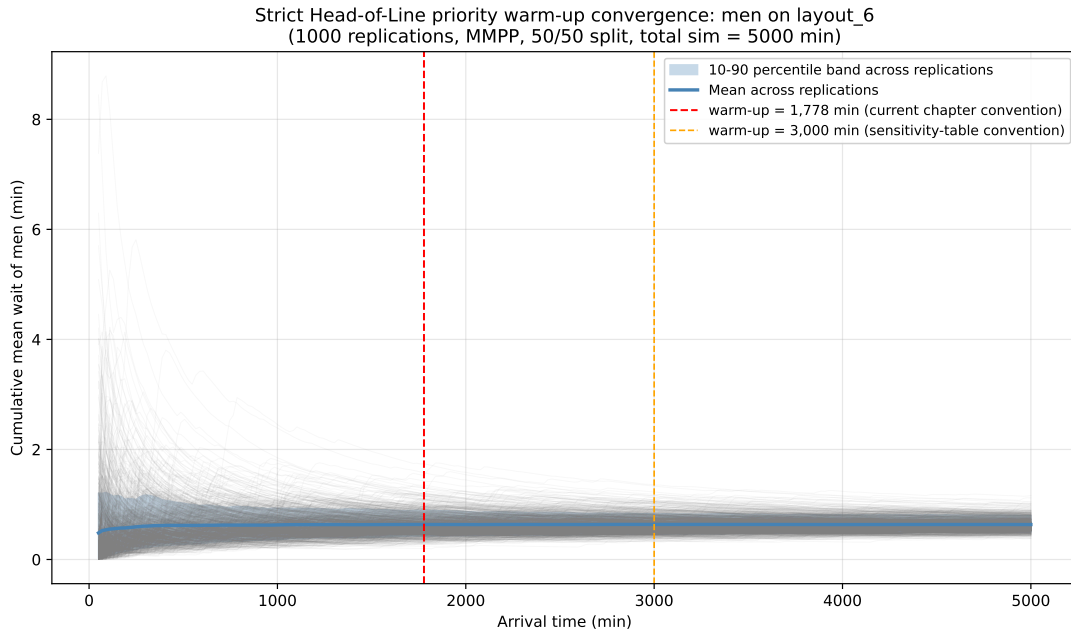


Figure D.1: Cumulative mean wait for men on Layout 6 under Strict Head-of-Line. The mean flattens well before the baseline 1 778-min warm-up (dashed line), confirming that even the most aggressive policy reaches steady state with margin to spare on this layout.

slowly here, and any policy that reorders the shared queue makes it even slower. The diagnostic measured the upward bias in men’s mean wait at 1 778 min relative to the steady-state value:

- **Strict Head-of-Line:** +5.8 % bias on men’s mean (10 min 13 s at 1 778 min, 10 min 51 s in steady state). The trajectory keeps rising past 1 778 min and flattens around $t = 2,500$ to 3,000 min.
- **Token Bucket:** +4.4 % bias on men’s mean (5 min 58 s at 1 778 min, 6 min 15 s in steady state).
- **Threshold Priority:** +5.0 % bias on men’s mean (4 min 57 s at 1 778 min, 5 min 12 s in steady state).
- **Dynamic Aging:** +2.2 % bias on men’s mean (4 min 28 s at 1 778 min, 4 min 34 s in steady state), plus a +41 % bias on the gender gap (24 s at 1 778 min, 40 s in steady state). Aging’s gap is the most sensitive: women’s wait keeps drifting late in the run, and the small absolute gap inflates the relative bias.

Figure D.2 shows the cumulative-mean trajectories behind these numbers. Under all four policies, the men’s mean is still rising at the baseline 1 778-min mark and only flattens between roughly 2 500 and 3 000 min.

Women’s mean wait also shifts upward by 5 to 7 % at 1 778 min under all four policies. For Strict, Token Bucket, and Threshold, men account for most of the gap bias. Aging is the

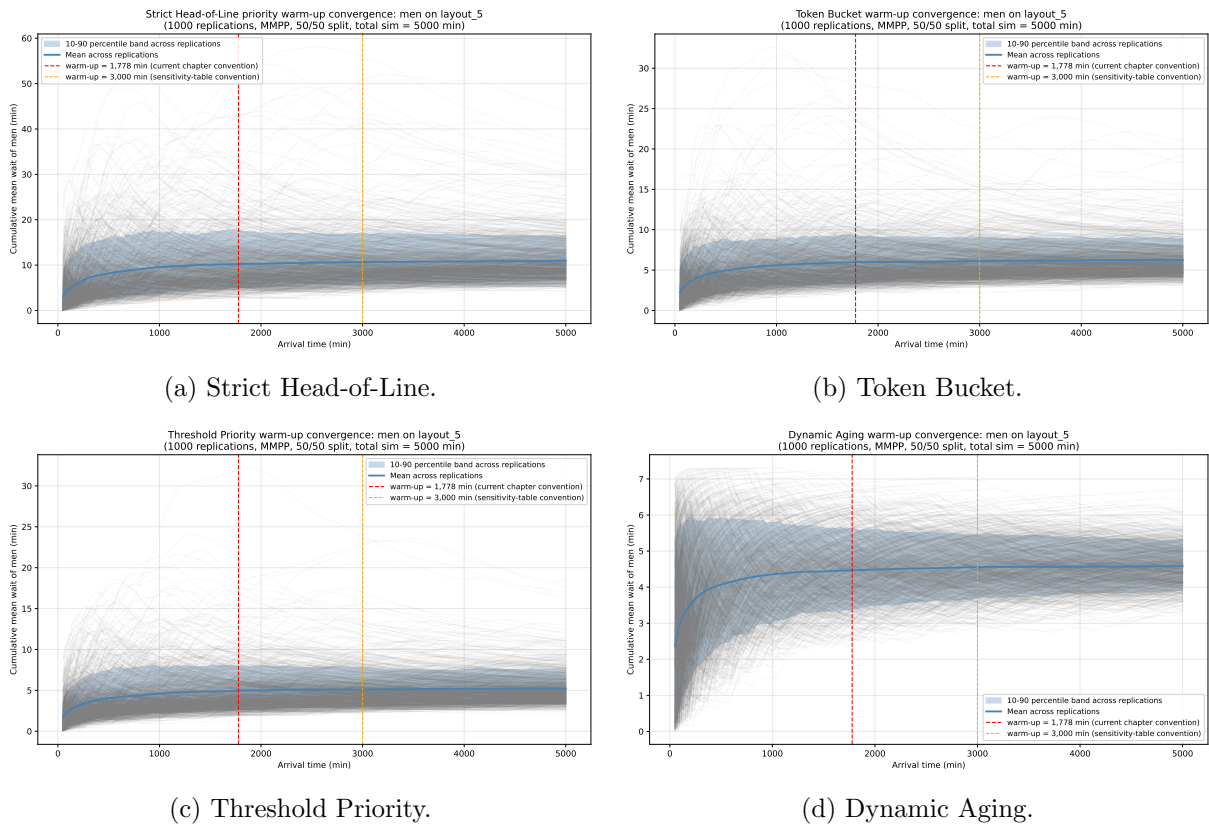


Figure D.2: Cumulative mean wait for men on Layout 5 under the four priority policies. The dashed line marks the baseline 1778-min warm-up and the dotted line the 3000-min warm-up adopted in Chapter 6. The mean is still climbing at 1778 min in every panel.

exception: its men’s bias is small but its women’s bias matches the other policies, so the gap shifts asymmetrically.

D.3 Implication for the chapter tables

Across the chapter, every Layout 5 row uses the 3 000-min warm-up and every Layout 4 and Layout 6 row uses the canonical 1 778-min value. Chapter 5’s Table 5.2 is an exception: it uses 3 000 min across all layouts so the 60/40 and 40/60 robustness splits share a single warm-up.

Implementation details and computational aspects

E.1 Discrete-event simulation engine

E.1.1 Event scheduling

The simulation schedules events on a min-heap, ordered by scheduled time. Five event types are used:

- women's arrival,
- men's arrival,
- women's service completion,
- men's service completion,
- MMPP state change (calm $\leftrightarrow$ busy), with exponentially distributed sojourn times in the current state.

The event-processing loop runs as follows:

```
while event queue not empty and current time < end time do
  Pop next event from queue
  Update current time to event time
  Process event (update system state, schedule future events)
  Record statistics if in post-warm-up period
end while
```

E.1.2 State variables

Each simulation instance maintains:

- queue lists of waiting customers and their arrival times,
- the number of busy servers for each gender,
- per-customer waiting times (lists indexed by gender), from which the mean, percentiles, and maximum are computed at the end of the run,
- queue-length integral (for time-averaged metrics),
- server busy time (for utilisation).

Cumulative-mean trajectories of the waiting time are additionally recorded when running the warm-up convergence study.

E.1.3 Random number generation

To ensure proper statistical properties and reproducibility:

- **Single Stream per Replication:** Each simulation instance uses a single pseudo-random number stream, seeded by the replication seed (see below). All stochastic draws (inter-arrival times, gender assignment, service times, MMPP state transitions, men's stall-vs-urinal routing) consume from this single stream. Drawing from a single stream simplifies seed management and is sufficient for the reproducibility goal of this thesis. This design partially weakens the variance-reduction effect of Common Random Numbers across paired scenarios, because variations in the number of draws per event propagate to all subsequent draws. Stronger CRN would require separate streams per source, which is an option for the simulation engine but is not implemented here.
- **Seeding Strategy:** Each replication on each layout is assigned a deterministic seed, chosen so that no two replications across the study share a seed. This scheme guarantees:
 - Independence across replications within and across layouts.
 - Reproducibility: the same replication on the same layout produces the same trajectory.
 - Variance reduction across paired comparisons (Common Random Numbers): two scenarios that share a seed see the same arrival times and the same service draws. For each replication j we form the difference d_j between the two scenarios. The 95% confidence interval on the mean of d_j across the n replications is $t_{0.975, n-1} \cdot s_d / \sqrt{n}$, where s_d is the standard deviation of the d_j values. Because the two scenarios move together from one replication to the next, s_d is smaller than the unpaired estimate, so the confidence interval is tighter. This pairing produces the gap intervals in Figure 5.1(c) and Table 5.2 (women minus men on the same layout), and the policy-versus-FCFS intervals in Chapter 6.
- **Distribution Generation:**

- Interarrival times conditional on MMPP state: $T \mid Z = \text{calm} \sim \text{Exp}(\lambda_{\text{calm}})$, $T \mid Z = \text{busy} \sim \text{Exp}(\lambda_{\text{busy}})$.
- MMPP sojourn times: $T_{\text{calm} \rightarrow \text{busy}} \sim \text{Exp}(q_{12})$, $T_{\text{busy} \rightarrow \text{calm}} \sim \text{Exp}(q_{21})$.
- Service times: $S \mid (\text{gender}, \text{facility}) \sim \text{Exp}(\mu_{\text{gender}, \text{facility}})$, with means 1.5 min (women, stalls), 1.25 min (men, stalls), 1.0 min (men, urinals).

E.2 Statistical analysis pipeline

E.2.1 Data collection

For each simulation replication:

1. Run simulation with specified parameters
2. Discard observations from warm-up period $[0, t_{\text{warm-up}}]$
3. Collect performance metrics from steady-state period $[t_{\text{warm-up}}, T]$, where T is the end of the simulation horizon
4. Store replication-level summary statistics

E.2.2 Aggregation and inference

Across replications:

1. Compute sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ for each performance metric, where X_i is the value of the metric on the i -th replication
2. Compute sample standard deviation $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$
3. Construct confidence intervals using t-distribution:

$$\text{CI}_{1-\alpha} = \bar{X} \pm t_{n-1, 1-\alpha/2} \cdot \frac{s}{\sqrt{n}} \quad (\text{E.1})$$

where $t_{n-1, 1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of Student's t -distribution with $n - 1$ degrees of freedom (e.g. $t_{999, 0.975} \approx 1.962$ for $\alpha = 0.05$, $n = 1000$, used in this thesis).

4. Report point estimates, standard errors, and confidence bounds

E.3 Simulation scale

The main comparison runs $6 \text{ layouts} \times 1\,000$ replications, for $6\,000$ simulation runs per scenario. Each replication covers a warm-up of $t_{\text{warm-up}} = 1\,778$ min followed by an observation window of $t_{\text{obs}} = 90$ min, giving a simulated horizon of roughly $1\,868$ min per run. Each Bayesian optimisation study uses 100 – 150 trials with 500 replications per trial, repeated for every (policy, gender-split) pair.

E.4 Software dependencies

Key Python packages used:

- NumPy [41]: numerical computation and random-number generation.
- SciPy [42]: statistical functions (Student's t -distribution, confidence intervals).
- Optuna [38]: Bayesian optimisation framework.
- Pandas: data manipulation and aggregation.
- Matplotlib: visualisation and plotting.