





# REAL-TIME CAPACITY PLANNING FOR KYC VIA PETRI-DES

Daliborka Đukić<sup>1\*</sup>,   
Vesna Midić<sup>2</sup> 

<sup>1</sup>Faculty of Organizational Sciences,  
University of Belgrade,  
Belgrade, Republic of Serbia

<sup>2</sup>OTP bank,  
Belgrade, Republic of Serbia

## Abstract:

Banking back-office workflows (loan approval, KYC - know your customer) face volatile demand and regulatory frictions, making service levels sensitive to congestion. We propose an agile capacity-planning workflow that links formal process verification (workflow Petri nets), queueing analytics (M/M/c with Quality-and-Efficiency-Driven (QED) square-root safety), and discrete-event simulation, with time-series forecasts translated into interval design rates via the upper 0.95 quantile (L95) and a utilization cap. Using public euro-area credit volume data as an arrivals proxy, the approach first ensures workflow soundness, then applies rapid analytic sizing, and finally confirms end-to-end effects in a three-stage KYC → Scoring → Contract model. In a scenario that meets L95 under the utilization cap, end-to-end lead times are substantially reduced and the upper-tail delays stabilized while stage-level utilization remains bounded. The framework offers transparent, reproducible rules for agile staffing in banking and can be instantiated either by raising service rates or by adding servers, depending on operational constraints.

## Keywords:

KYC workflow, capacity planning, Petri nets, discrete-event simulation, business agility.

## 1. INTRODUCTION

Loan-approval and related administrative workflows (e.g., KYC, application submission and evaluation) are prone to congestion under variable demand, seasonal peaks, and regulatory constraints. Business agility therefore requires timely, reliable capacity adjustment under cost discipline. Because the flow spans interdependent stages (identification/verification, scoring, contracting), capacity planning is end-to-end; the key question is how to map arrival forecasts into implementable staffing and work-organization choices while controlling overload risk.

Two complementary methodological lines dominate practice and literature. Queueing-analytic models (e.g., M/M/c with Erlang C) provide closed-form expressions for delay probability, expected waiting time, and utilization, enabling rapid sizing in many-server systems, particularly in contact centers and administrative settings (Halfin & Whitt, 1981; Gans, Koole, & Mandelbaum, 2003; Green, Kolesar, & Soares, 2003; van Leeuwen, Mathijssen, & Zwart, 2019). Standard Erlang C excludes abandonment, whereas Erlang A and QED (many-server, square-root safety) adjust staffing when impatience is material and utilization is high (Garnett, Mandelbaum, & Reiman, 2002; Borst, Mandelbaum, & Reiman, 2004; Janssen, van Leeuwen, & Zwart, 2011). Discrete-event simulation (DES) complements analytics by representing multi-stage flows, calendars, and routing rules; it is well suited to validation and scenario analysis. Accordingly, this study combines analytics (for rapid sizing) and simulation (for verification and evaluation).

Correspondence:  
Daliborka Đukić

e-mail:  
dd20245064@student.fon.bg.ac.rs





To preclude structural modeling defects before scenario analysis, workflow nets (WF-nets)—a class of Petri nets with a unique source and sink—are used to check correctness and soundness, providing a reliable foundation for both analytic calculations and simulations (van der Aalst, 1998).

On the demand side, time-series models (e.g., SARIMA) deliver short-term arrival forecasts that workforce management translates into time-varying rates (nonhomogeneous Poisson or piecewise-constant by period), followed by Erlang C/A- and QED-based sizing. The forecast-to-staffing link is well documented in empirical studies and surveys (Taylor, 2012; Ibrahim, Ye, L'Ecuyer, & Shen, 2016). Methodologically, recent contributions stabilize performance under time-varying and “bursty” (non-Poisson) arrivals and connect prediction intervals/quantiles to implementable rules (Liu & Whitt, 2017; Ding & Koole, 2022), while related work addresses time-stable performance with abandonment and nonstationarity (Feldman, Mandelbaum, Massey, & Whitt, 2008; Defraeye & Van Nieuwenhuyse, 2016). Theoretically, QED/square-root safety links forecast uncertainty to the required buffer in server counts (Borst *et al.*, 2004; Janssen *et al.*, 2011; van Leeuwaarden *et al.*, 2019). This forecast-to-planning connection underpins agile capacity management.

This paper contributes an integrated, reproducible framework in R that: (i) verifies workflow correctness via WF-nets; (ii) links M/M/c (Erlang C/A) and QED rules to rapid interval-wise sizing; (iii) validates end-to-end performance with a three-stage DES; and (iv) maps time-series forecasts into time-varying arrival rates used directly for staffing decisions with explicit SLA targets. The focus is on operational rules—for example, “staff to the L95 (upper 0.95-quantile of arrivals) with a utilization target  $\rho \leq 0.90$ ”—that support timely adaptation with risk control.

The remainder of the paper is organized as follows. Section 2 reviews related work (queues, DES, WF-nets, and the forecast-to-planning link). Section 3 presents data and methodology (WF-net model, M/M/c and QED rules, DES, SARIMA and the TS  $\rightarrow$  WFM mapping). Section 4 reports results and discussion. Section 5 concludes with implications and directions for future work.

## 2. LITERATURE REVIEW

### 2.1. FOUNDATIONS OF CAPACITY SIZING (M/M/C, ERLANG C/A, QED)

Queueing models underpin capacity planning in many-server service systems. The classical M/M/c with Erlang C offers transparent expressions for delay probability, waiting time, and utilization; Erlang A extends this by modeling abandonment, which becomes critical when waiting-time sensitivity is non-negligible. In heavy traffic, the QED (Quality-and-Efficiency-Driven) regime explains why a square-root safety buffer can sustain high utilization while keeping delays controlled, and subsequent refinements show how to tune staffing to target SLAs in practice. In call-center and back-office contexts, tutorial and heuristic work illustrates how these analytic results translate into operational sizing and scheduling (Borst, Mandelbaum, & Reiman, 2004; Garnett, Mandelbaum, & Reiman, 2002; Gans, Koole, & Mandelbaum, 2003; Green, Kolesar, & Soares, 2003; Halfin & Whitt, 1981; Janssen, van Leeuwaarden, & Zwart, 2011; van Leeuwaarden, Mathijsen, & Zwart, 2019).

### 2.2. NONSTATIONARY ARRIVALS AND TIME-STABLE PERFORMANCE

Because arrival rates vary over time, organizations rely on dynamic capacity profiles to keep service measures approximately stable throughout the day. A line of work develops heuristics and approximations (e.g., PSA/MOL/SIPP/ISA) and formal methods for staffing under time-varying rates, while more recent contributions address non-Poisson “bursty” arrivals and propose algorithms that stabilize performance and smooth capacity adjustments in many-server environments. These results are especially relevant for peaks and seasonality, where static Erlang-C tends to understate variability (Defraeye & Van Nieuwenhuyse, 2016; Feldman, *et al.*, 2008; Green *et al.*, 2003; Liu & Whitt, 2017).

### 2.3. FORECASTS $\rightarrow$ STAFFING (QUANTILES AND INTERVALS)

Short-term time-series forecasts (e.g., SARIMA, ETS) are routinely converted into intra-day arrival profiles and then into staffing rules. Beyond means, prediction intervals and quantiles (e.g., the 95<sup>th</sup> percentile) “buy down” demand risk and align naturally with QED square-root logic; recent work integrates forecasting and staffing into implementable rules (choice of quantile and target utilization) and documents SLA gains relative to mean-based planning. (Borst *et al.*, 2004; Ding & Koole, 2022; Ibrahim, Ye, L'Ecuyer, & Shen, 2016; Janssen *et al.*, 2011; van Leeuwaarden *et al.*, 2019)



## 2.4. WORKFLOW PETRI NETS (WF-NETS) AND FLOW VERIFICATION

WF-nets are specialized Petri nets with a unique source and sink place that enable formal verification of process correctness before simulation. The soundness property entails proper completion, no dead transitions, and no residual tokens at termination - helpful for detecting structural issues in multi-stage financial workflows (e.g., KYC → Scoring → Contract) before running what-if experiments. (van der Aalst, 1998)

## 2.5. DES IN BANKING/BACK-OFFICE: BENEFITS AND LIMITATIONS

DES faithfully represents multi-stage flows, resource constraints, and routing rules, making it well suited for what-if staffing, bottleneck diagnosis, and redesign validation. Compared with pure analytics, simulation better captures inter-station dependencies and queue distributions under nonstationary arrivals and service-time variability, but it requires high-quality inputs and calibration; in practice, a combined approach is most effective - analytic rules for initial sizing, with DES for end-to-end validation of performance and SLA risk. (Defraeye & Van Nieuwenhuyse, 2016; Ding & Koole, 2022; Gans *et al.*, 2003; Green *et al.*, 2003)

## 3. METHODOLOGY

### 3.1. WORKFLOW VERIFICATION WITH WF-NETS

We model the end-to-end KYC → Scoring → Contract process as a workflow Petri net (WF-net): places encode document/status states, transitions encode activities, and tokens encode cases progressing through the flow. Using a custom script in R 4.4.x, we check reachability, boundedness, and (weak) liveness, and we require WF-net soundness before any analytics or simulation. Only a sound model proceeds to subsequent steps (van der Aalst, 1998).

### 3.2. DATA AND ARRIVAL PROCESSES

Monthly arrivals are proxied by new consumer loans in the euro area (ECB MIR), covering 2011-06 to 2025-05, from which we derive daily rates for queueing and simulation (European Central Bank, 2025). Data preparation includes robust date parsing, column normalization, series alignment, and conservative treatment of missing values (no aggressive imputation). Seasonality is present; outliers are not explicitly treated in this version. For context and auxiliary checks, we collate annual unemployment and median income series (Eurostat indicators *une\_rt\_m* and *ilc\_di03*), but the forecasting model is univariate - these variables are not used as exogenous regressors in the forecasts (Eurostat, 2025a, 2025b).

*Data sources:* ECB MIR series MIR.M.U2.B.A2B.A.B.A.2250.EUR.N (European Central Bank, 2025); Eurostat *une\_rt\_m* and *ilc\_di03* (Eurostat, 2025a, 2025b). For distributional checks only, a synthetic micro-log was constructed: monthly ECB MIR volumes were mapped to daily rates by operating days, and arrival timestamps were generated as a (non)homogeneous Poisson process (“Poissonization”). On  $N=343$  inter-arrival gaps, KS and AD tests did not reject exponentiality ( $p \approx 0.960/0.965$ ). This micro-log is used solely for distributional checks; all forecasts and scenarios are based on the monthly aggregate series.

### 3.3. FORECASTING AND UNCERTAINTY → STAFFING INPUTS

Monthly arrivals (2011-06–2025-05;  $N=168$ ) were modeled by SARIMA(1,1,2)(0,1,1)[12] with a Box–Cox transform ( $\lambda \approx 0.057$ ). Parameters were estimated via state-space maximum likelihood with bias correction, and the specification was selected by AICc (AICc  $\approx -169.90$ ; BIC  $\approx -155.08$ ). Residual diagnostics (Ljung–Box  $p \approx 0.011$ ; ARCH–LM  $p \approx 0.032$ ; Jarque–Bera  $p \approx 0.000$ ) indicated departures from Gaussian IID noise; therefore, interval-wise design rates were taken from predictive quantiles rather than point means. The upper 0.95-quantile (L95; 95th percentile) per month was extracted over the 12-month horizon (2025-06 → 2026-05) and was mapped to daily piecewise-constant rates by the number of operating days; these rates feed the analytic M/M/c sizing under a utilization cap  $\rho \leq 0.90$ , as well as the DES scenarios. This quantile-to-staffing mapping is consistent with QED (square-root safety) and is widely used in call-center/back-office workforce management (Taylor, 2012; Ibrahim, Ye, L’Ecuyer, & Shen, 2016; Borst, Mandelbaum, & Reiman, 2004; Janssen, van Leeuwen, & Zwart, 2011; van Leeuwen, Mathijssen, & Zwart, 2019; Ding & Koole, 2022).



### 3.4. QUEUEING ANALYTICS AND SCENARIO CONSTRUCTION

Each service stage is modeled as M/M/c (FIFO, infinite buffer) with Erlang C; abandonment is not modeled; analytics and DES assume an infinite FIFO queue without abandonment (M/M/c). Baseline configuration specifies per-stage service rates (per case, per calendar day) and server counts; we then construct scenarios that size capacity by interval using the Section 3.3 L95 arrival rates and a utilization cap  $\rho \leq 0.90$ . Two levers are considered: (a) increase service rate  $\mu$  at fixed  $c$  (e.g., training/automation), or (b) increase servers  $c$  at fixed  $\mu$ . Scenario S1 implements (a). All numeric sizing outcomes (e.g., required  $\mu$  or  $c$  per interval) are reported later with results; here we describe the construction rules (Halfin & Whitt, 1981; Green, Kolesar, & Soares, 2003; Borst *et al.*, 2004; Janssen *et al.*, 2011; van Leeuwen, Mathijssen, & Zwart, 2019; Ding & Koole, 2022).

### 3.5. DES SETUP, REPRODUCIBILITY, AND VALIDATION PLAN

We implement a three-stage DES in R/simmer 4.x (R 4.4.x) mirroring the verified WF-net. Arrivals are nonhomogeneous (from the Section 3.3 L95 path), service times are exponential per stage (SCV = 1), routing is FIFO without preemption, and resource calendars follow observed operating hours. For the TS  $\rightarrow$  DES experiment we run 30 stochastic paths  $\times$  12 months, with warm-up = 0 for the monthly horizon, and we report  $W_q$ ,  $W$ , end-to-end lead time  $W_{total}$ , and utilization  $\rho$  as q10/q50/q90 quantiles. To ensure reproducibility, we use a predefined seed list kept consistent across conditions.

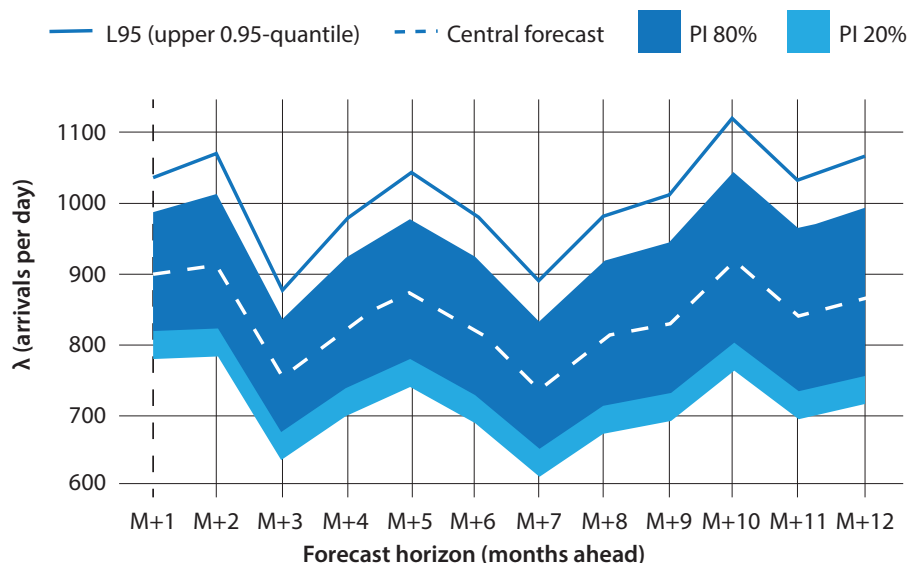
Validation plan. We (i) test inter-arrival exponentiality on a synthetic event log derived from ECB MIR, and (ii) benchmark analytic M/M/c against DES on historical slices for at least one stage. Detailed diagnostics and figures are presented in Section 4 (Results and Discussion).

## 4. RESULTS AND DISCUSSION

### 4.1. ARRIVAL FORECASTS AND UNCERTAINTY (FAN CHART; L95 PATH)

For monthly arrivals (2011-06–2025-05;  $N=168$ ), a SARIMA(1,1,2)(0,1,1)[12] model with a Box–Cox transform ( $\lambda \approx 0.057$ ) is estimated; the forecast horizon is 12 months (2025-06  $\rightarrow$  2026-05). Model selection follows AICc (AICc  $\approx -169.90$ ; BIC  $\approx -155.08$ ). Diagnostics (Ljung–Box, ARCH–LM, Jarque–Bera) indicate departures from independent, identically distributed Gaussian white noise; accordingly, capacity sizing is based on prediction quantiles. We take the upper 0.95-quantile (i.e., 95<sup>th</sup> percentile; L95) per period as the design rate. Figure 1 shows the forecast with L95 highlighted as the design rate; those interval rates feed both the queueing sizing and the DES.

Figure 1. Monthly-arrival forecast with L95 (upper 0.95-quantile) highlighted as the design rate



Source: Authors' calculations based on ECB Monetary Financial Institutions (MFI) interest rate statistics (MIR), ECB Statistical Data Warehouse, 2011–2025; forecasting via SARIMA(1,1,2)(0,1,1)[12] with Box–Cox



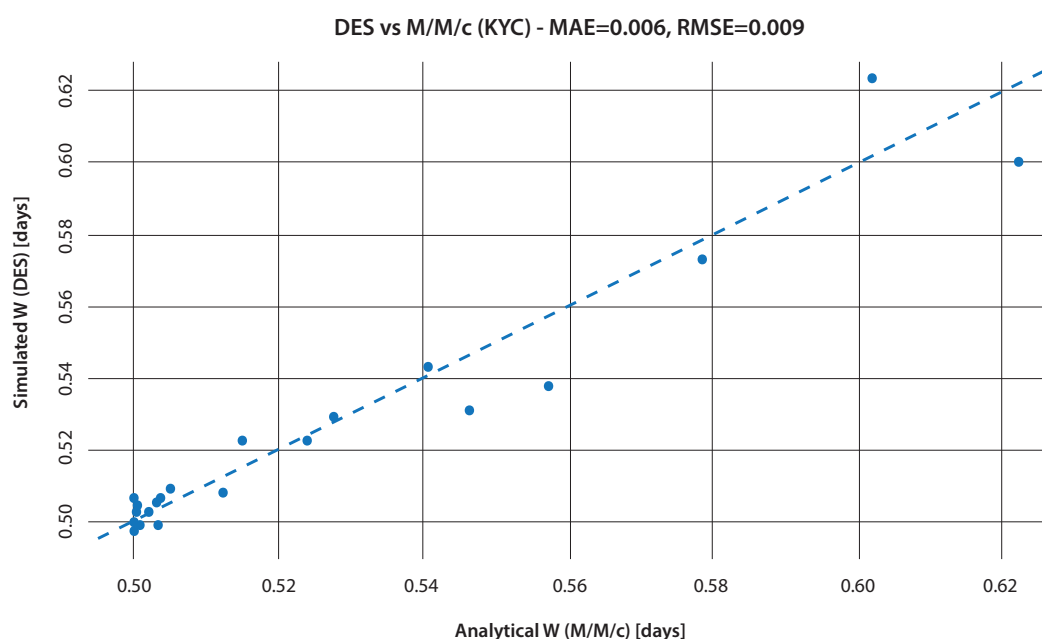
## 4.2. INPUT VALIDATION: INTER-ARRIVAL TIMES AND THE POISSON/EXPONENTIAL ASSUMPTION

Inter-arrival times were assessed on a synthetic event log derived from ECB MIR (via Poissonization), using both global and segmented tests. Globally, Kolmogorov–Smirnov (KS)  $p \approx 0.960$  and Anderson–Darling (AD)  $p \approx 0.965$  do not reject the exponential model, with estimated rate  $\lambda \approx 61.42 \text{ day}^{-1}$  (mean gap  $\bar{\Delta t} \approx 0.0163 \text{ days}$ ;  $N=343$ ). Segment tests on rescaled monthly/quarterly panels ( $N \geq 30$ ) lead to the same conclusion—no statistically significant departures from exponentiality; exponential QQ plots show no systematic deviations. Operationally, this supports modeling short-interval arrivals as a nonhomogeneous Poisson process, making M/M/c a reasonable analytic approximation for rapid sizing (Erlang C). At the same time, seasonality and peaks over longer horizons motivate interval-wise sizing based on prediction quantiles (L95) with a utilization cap  $\rho \leq 0.90$ , followed by end-to-end verification in DES.

## 4.3. ANALYTIC VS. SIMULATION RESULTS (M/M/C VERSUS DES)

On historical annual slices for the KYC stage, the analytic M/M/c model was validated against DES. For each year, the same inputs ( $\lambda, \mu, c$ ) were applied; analytic performance measures ( $W_q, W$ ) from Erlang C were compared with DES sample means under identical operational rules (FIFO, no preemption). The scatter of DES means against analytic values lies near the 45° line (see Figure 2), with MAE  $\approx 0.006$  days and RMSE  $\approx 0.009$  days (annual cuts, KYC).

**Figure 2.** DES vs. analytical M/M/c (KYC, annual slices): mean waiting time; points lie near the 45° line (MAE  $\approx 0.006$  days; RMSE  $\approx 0.009$  days)



Source: Authors' discrete-event simulation (DES) and analytical M/M/c model; synthetic micro-log via Poissonization

Within the operational utilization range ( $\rho \leq 0.90$ ) no systematic bias was detected. At very high loads ( $\rho \rightarrow 1$ ) small deviations were observed, consistent with expected Monte Carlo variability and finite-horizon aggregation effects in queueing simulations (Law, 2015); in the present setting these deviations are operationally immaterial because capacity is explicitly bounded by the  $\rho$ -cap rule. These findings support using M/M/c for rapid, interval-wise sizing, while retaining DES for end-to-end validation of the three-stage flow (KYC  $\rightarrow$  Scoring  $\rightarrow$  Contract) and for quantiles of total lead time  $W_{\text{total}}$  (median and upper tail). Section 4.4 applies the L95 +  $\rho$ -cap ( $\rho \leq 0.90$ ) and reports its impact on  $W_{\text{total}}$ .

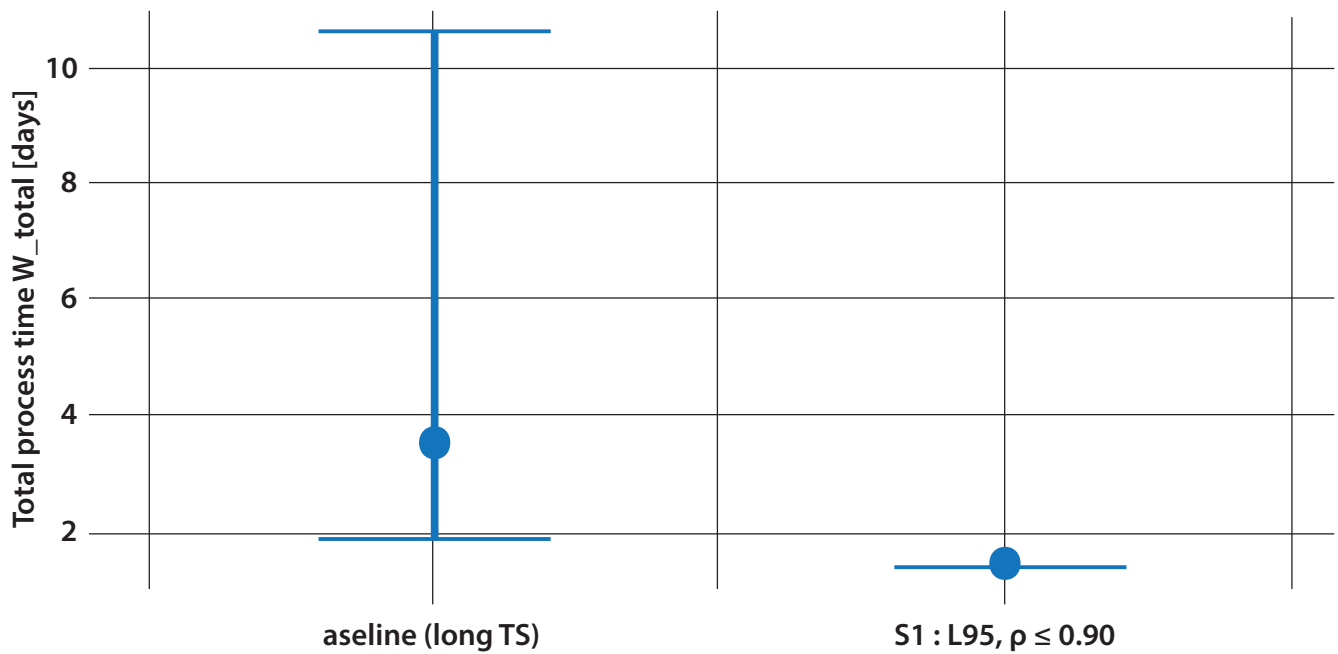
## 4.4. EFFECTS OF THE L95 + UTILIZATION CAP RULE (SCENARIO S1): END-TO-END LEAD TIME

Scenario S1 applies the sizing rule based on the L95 arrival path together with a utilization cap of  $\rho \leq 0.90$  at each stage. As the operational mechanism, service rates were increased while server counts were held fixed: approximately  $\mu_{\text{KYC}} \approx 2.516/\text{day}$ ,  $\mu_{\text{SCO}} \approx 3.019/\text{day}$ , and  $\mu_{\text{CON}} \approx 2.588/\text{day}$ . (An alternative with higher server counts yields similar performance; variant (a) is retained here because it requires fewer organizational changes.)





**Figure 3.** Baseline vs. S1 ( $L95 + \rho$ -cap) — total lead time  $W_{total}$ : quantile bars ( $q_{10}$ – $q_{90}$ ) and medians (dots). Scenario S1 sharply lowers both the median and the upper tail relative to baseline



Source: Authors' DES scenarios for the KYC workflow; arrival inputs from ECB MIR and service-rate calibration per Section 4)

Across 30 stochastic paths  $\times$  12 months, the following quantiles of the total end-to-end lead time  $W_{total}$  were obtained:

- Baseline:  $q_{10} / q_{50} / q_{90} \approx 1.915 / 3.612 / 10.589$  days.
- S1 ( $L95, \rho \leq 0.90$ ):  $q_{10} / q_{50} / q_{90} \approx 1.402 / 1.415 / 1.441$  days.

Relative to baseline, the median of  $W_{total}$  decreases by  $\approx 61\%$ , while the upper (right-tail) quantile  $q_{90}$  decreases by  $\approx 86\%$ ; dispersion narrows substantially, indicating stabilized performance and the removal of peak waiting times. All stages remain within the utilization cap ( $\rho \leq 0.90$ ) throughout the horizon. This meets the operational SLA target of  $q_{90}(W_{total}) \lesssim 1.5$  days in peak months. These findings are consistent with QED/safety-staffing logic and with the analytic M/M/c calculations in Section 3; they are confirmed end-to-end in the three-stage DES.

#### 4.5. ROBUSTNESS AND SENSITIVITY: ALTERNATIVE CAPACITY-ADJUSTMENT MECHANISM AND QUANTILE CHOICE

Alternative mechanism (adding servers). In addition to S1, where capacity is provided by increasing service rates at fixed staffing, we also consider adding servers at baseline service rates. To satisfy  $L95 + \rho \leq 0.90$ , approximate requirements are KYC  $c \approx 46$ , SCO  $c \approx 37$ , CON  $c \approx 46$  (vs. baseline 36/30/35). Under both mechanisms ( $\uparrow \mu$  or  $\uparrow c$ ), all stages remain under the utilization cap, and the distribution of total lead time exhibits a similar reduction in the median and a narrowing of the upper tail relative to baseline. The differences are primarily operational: the  $\mu$ -variant entails training/automation, whereas the  $c$ -variant requires additional staffing and higher resource consumption (measured in server-days). As summarized in Table 1, the  $L95 + \rho \leq 0.90$  target can be met either by increasing service rates at fixed staffing ( $\mu_{min}$ ) or by adding servers at baseline rates ( $c_{min}$ ), with stage-level requirements shown side by side.

**Table 1.** Capacity to meet  $L95 + \rho \leq 0.90$  (per stage)

Stage	Baseline $c$	Baseline $\mu$ (/day)	$\mu_{min}$ at fixed $c$	$c_{min}$ at fixed $\mu$
KYC	36	2.000	2.516	46
SCO	30	2.500	3.019	37
CON	35	2.000	2.588	46

Source: Authors' calculations based on ECB MIR and Eurostat datasets



We report engineering requirements; comparative costs (e.g., server-days for  $\uparrow c$  versus training/automation cost for  $\uparrow \mu$ ) are left to managerial calibration and can be added as a budget-constrained extension. Sensitivity to the quantile choice. Sizing to P90 (instead of L95) reduces required capacity but increases the risk of SLA shortfalls at seasonal peaks; conversely, P97.5 strengthens the safety margin at higher capacity cost. In analytical M/M/c calculations,  $W_q$  and SLA metrics increase nonlinearly as load approaches 1, so the chosen quantile directly sets the safety distance from near-saturation regimes. Given the goal of stabilizing the upper tail while controlling risk,  $L95 + \rho \leq 0.90$  emerges as a balanced rule; P90/P97.5 may be considered where cost minimization or, conversely, additional robustness is the primary objective.

Stage-level verification. Across all S1 simulations, no violations of the utilization cap ( $\rho \leq 0.90$ ) were observed by month and stage; observed fluctuations are within expected stochastic variability, ensuring that local targets (per stage) align with the global outcome of stabilized  $W_{total}$ .

#### 4.6. LIMITATIONS AND THREATS TO VALIDITY

*Internal validity.* The analytic M/M/c model assumes exponential service times and no abandonment; if service-time variability exceeds the exponential benchmark ( $SCV > 1$ ), waiting times are typically underestimated, so sizing relies on predictive quantiles (L95) with a utilization cap, while DES (FIFO, no preemption, observed calendars) provides end-to-end confirmation; residual diagnostics depart from Gaussian IID, and formal out-of-sample/coverage tests are left for future work.

*Construct validity.* The workflow was verified as a WF-net (soundness, boundedness, liveness) and mapped to a three-stage DES; monthly forecasts were disaggregated to daily rates via a piecewise-constant approximation; Scenario S1 increases  $\mu$  at fixed  $c$ , with the alternative (increasing  $c$  at baseline  $\mu$ ) noted in the sensitivity analysis.

*External validity.* Results rely on euro-area aggregates (ECB MIR; Eurostat) and require re-calibration of arrival rates, service mix, and operating rules for other institutions or products; the bound  $\rho \leq 0.90$  is a managerial policy.

*Reproducibility.* Pre-specified seed lists (30 paths per scenario), inputs, and scripts are documented in Section 3; public datasets are cited, and code for WF-net checks, M/M/c sizing, SARIMA, and DES is documented so that runs are reproducible with the same software versions and settings.

### 5. CONCLUSION AND DIRECTIONS FOR FUTURE WORK

This study proposes an integrated capacity-planning framework for credit approval and KYC workflows that combines formal process verification (WF-nets), analytic sizing (M/M/c with Erlang C and QED logic), and DES validation, with an explicit link from time-series forecasts to operational decisions. Workflow soundness (soundness, boundedness, liveness) provides a reliable basis for what-if analyses. Inter-arrival times are consistent with the exponential assumption over short intervals, justifying a nonhomogeneous Poisson arrival model. The analytic M/M/c approximation aligns closely with DES on historical slices, supporting its use for rapid interval-wise sizing; DES supplies end-to-end assessment in the three-stage flow.

Applying the  $L95 + \rho \leq 0.90$  rule yielded large reductions in the median and upper tail of total lead time while keeping utilization within bounds at all stages. The result is stabilized service under peak demand, supported by transparent decision rules suitable for deployment.

Managerial implications. The procedure (i) translates forecast quantiles into interval-wise capacities; (ii) enables a principled choice between raising service rates and adding servers under operational constraints; and (iii) verifies SLA targets in DES prior to implementation.

Limitations and future work. Limitations include the absence of abandonment modeling (pointing to Erlang A extensions), a piecewise-constant aggregation of intra-month rates, and no preemption or multi-skill routing. Future work will (a) model abandonment and bursty, non-Poisson arrivals in both analytics and DES; (b) optimize the quantile choice and capacity mechanism under budget constraints; (c) employ formal out-of-sample forecast evaluation; and (d) generalize the framework to additional processes and institutional settings. Additionally, rework/return loops and priority classes (e.g., enhanced due diligence) will be incorporated in the WF-net and DES to capture escalations and their impact on tail delays.



## LITERATURE

- Borst, S. C., Mandelbaum, A., & Reiman, M. I. (2004). Dimensioning large call centers. *Operations Research*, 52(1), 17–34. <https://doi.org/10.1287/opre.1030.0081>
- Defraeye, M., & Van Nieuwenhuyse, I. (2016). Staffing and scheduling under nonstationary demand for service: A literature review. *Omega*, 58, 4–25. <https://doi.org/10.1016/j.omega.2015.04.002>
- Ding, F., & Koole, G. (2022). Optimal call center forecasting and staffing. *Probability in the Engineering and Informational Sciences*, 36(4), 837–858. <https://doi.org/10.1017/S0269964820000595>
- European Central Bank. (2025). *MFI interest rate statistics (MIR) — Bank business volumes: Loans to households for consumption (new business), euro area* (Series key: MIR.M.U2.B.A2B.A.B.A.2250.EUR.N) [Data set]. European Central Bank. Retrieved August 14, 2025, from <https://data.ecb.europa.eu/>
- Eurostat. (2025a). *Unemployment by sex and age — monthly data (une\_rt\_m)* [Data set]. Eurostat. Retrieved August 14, 2025, from <https://ec.europa.eu/eurostat/databrowser/>
- Eurostat. (2025b). *Mean and median income by age and sex (ilc\_di03) — MED\_E: Median equivalised net income* [Data set]. Eurostat. Retrieved August 14, 2025, from <https://ec.europa.eu/eurostat/databrowser/>
- Feldman, Z., Mandelbaum, A., Massey, W. A., & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2), 324–338. <https://doi.org/10.1287/mnsc.1070.0821>
- Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2), 79–141. <https://doi.org/10.1287/msom.5.2.79.16071>
- Garnett, O., Mandelbaum, A., & Reiman, M. I. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3), 208–227. <https://doi.org/10.1287/msom.4.3.208.7753>
- Green, L. V., Kolesar, P. J., & Soares, J. (2003). An improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management*, 12(1), 46–61. <https://doi.org/10.1111/j.1937-5956.2003.tb00197.x>
- Halfin, S., & Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3), 567–588. <https://doi.org/10.1287/opre.29.3.567>
- Ibrahim, R., Ye, H., L'Ecuyer, P., & Shen, H. (2016). Modeling and forecasting call center arrivals: A literature survey and a case study. *International Journal of Forecasting*, 32(3), 865–874. <https://doi.org/10.1016/j.ijforecast.2015.11.012>
- Janssen, A. J. E. M., van Leeuwen, J. S. H., & Zwart, B. (2011). Refining square-root safety staffing by expanding Erlang C. *Operations Research*, 59(6), 1512–1522. <https://doi.org/10.1287/opre.1110.0991>
- Law, A. M. (2015). *Simulation Modeling and Analysis* (5<sup>th</sup> ed.). McGraw–Hill Education.
- Liu, Y., & Whitt, W. (2017). Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research*, 257(2), 500–518. <https://doi.org/10.1016/j.ejor.2016.07.018>
- Taylor, J. W. (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. *Management Science*, 58(3), 534–549. <https://doi.org/10.1287/mnsc.1110.1434>
- van der Aalst, W. M. P. (1998). The application of Petri nets to workflow management. *Journal of Circuits, Systems and Computers*, 8(1), 21–66. <https://doi.org/10.1142/S0218126698000043>
- van Leeuwen, J. S. H., Mathijssen, B. W. J., & Zwart, B. (2019). Economies of scale in many-server queueing systems: Tutorial and partial review of the QED Halfin–Whitt heavy-traffic regime. *SIAM Review*, 61(3), 403–440. <https://doi.org/10.1137/17M1133944>
- Whitt, W., & Zhao, H. (2017). Many-server loss models with non-Poisson time-varying arrivals. *Naval Research Logistics*, 64(8), 667–703. <https://doi.org/10.1002/nav.21741>