

Quantitative Tightening in Segmented Interbank Markets*

Junko Oguri[†]

Cristoforo Pizzimenti[‡]

October 2026

Working paper

Abstract

How far can a central bank proceed with quantitative tightening (QT) when banks differ in their demand for reserves and interbank trading is limited? We model a central bank choosing how far to reduce reserves through QT while limiting the risk of losing control over the monetary policy rate. Two banking systems can have identical distributions of aggregate reserve demand and identical average demand by bank type, yet require different reserve supplies. This difference remains even when aggregate demand is identical at every common funding rate. Reserves can be sufficient in aggregate but concentrated among banks with lower demand. If interbank trading cannot redistribute reserves to banks with higher demand, the central bank must retain more reserves for monetary policy rate control. Our findings show that aggregate reserve demand alone may not tell the central bank how far QT can proceed in segmented interbank markets.

Keywords: Quantitative tightening, reserve demand, interbank market segmentation, funding stress.

*An earlier version of this paper circulated as a section of our companion paper, *Ample Reserves for Whom? The Role of Foreign Banks in U.S. Monetary Policy Implementation*.

[†]Kellogg School of Management, Northwestern University. junko.oguri@kellogg.northwestern.edu.

[‡]PIMCO, Client Solutions & Analytics; previously at Northwestern University. cristoforopizzimenti2025@u.northwestern.edu. The views expressed are those of the authors and do not necessarily reflect those of PIMCO, Client Solutions & Analytics.

Liquidity is not evenly distributed in a system. And there can be times when, in the aggregate, reserves are ample or even abundant, but not in every part. And those parts where they're not ample, there can be stress, and that can cause you to prematurely stop the process to avoid the stress.

—Jerome H. Powell, March 20, 2024

1 Introduction

Determining how far to reduce the large balance sheets accumulated after the Global Financial Crisis and during the COVID-19 pandemic has become a central policy question in many countries (Ramsden, 2026; Warsh, 2025). Through quantitative tightening (QT), central banks reduce their asset holdings, which can also reduce reserves—the balances banks hold at the central bank to settle payments and meet liquidity needs.¹ For central banks operating with ample reserves, the challenge is to reduce their balance sheets while retaining enough reserves to maintain control over short-term interest rates.

Assessing how far reserves can decline also requires understanding how the reserves move across banks. In the United States, bank-to-bank trading accounts for only a small part of the federal funds market. Anbil et al. (2026b) show that funding conditions in this segment differ from those in the larger segment supplied by nonbank lenders, revealing pressures that aggregate measures can obscure. Reserve availability also matters in the much larger Treasury repo market. Copeland et al. (2025) link repo funding pressures to reserves held by major intermediaries and the timing of payments they receive. These findings raise a question for reserve planning: is knowing aggregate reserve demand enough to determine how far QT can proceed?

Our paper shows that knowing aggregate reserve demand need not be enough to determine the capacity for QT when banks differ and interbank trading is limited. Our main result is that two banking systems can have identical distributions of aggregate reserve demand and identical average demand by bank type, yet require different reserve supplies. For example, in a given state, higher demand by foreign banks can be offset by lower demand from domestic banks. Aggregate demand is unchanged, but foreign banks may be short of reserves relative to their initial balances. If interbank trading cannot cover that shortfall, the central bank needs a larger total supply to meet the same tolerance for funding stress. From identical feasible reserve supplies before QT, this leaves less capacity for reserve reduction.

¹We interpret reserve reduction as balance-sheet contraction holding other central-bank liabilities and capital fixed. Anderson et al. (2026) discuss the policy choices involved in pursuing this objective in the United States.

We develop a model with two groups of banks that differ in their liquidity needs and the costs of holding reserves.² The central bank chooses the smallest reserve supply that keeps the probability of funding stress within a given tolerance. This choice is motivated by rate-control objectives, such as keeping the effective federal funds rate within the Federal Reserve’s target range. In the model, funding stress occurs when either group’s funding spread—its funding rate minus the interest rate on reserves—exceeds a policy ceiling for that spread. The central bank chooses supply with limited information about future reserve demand, using a predictive distribution of banks’ needs. Reserves are initially distributed in proportion to bank assets. Banks then observe demand conditions and trade. We compare a frictionless interbank market with a segmented market in which transfers between the groups are limited.

We characterize when aggregate reserve demand is sufficient to assess the capacity for QT and when information about individual bank groups is also needed. With frictionless trading, the distribution of aggregate demand at the tolerated spread determines minimum reserve supply even when banks differ. With segmentation, each group must also obtain enough reserves through its initial balances and feasible interbank borrowing. We derive an explicit reserve requirement combining these aggregate and local conditions. Under uncertainty, its upper quantile corresponding to the central bank’s stress tolerance determines minimum supply.

The comparison is stronger than holding aggregate demand fixed at a single rate. The two systems can have identical total reserve demand at every common funding rate, in every demand state. This curve adds up what banks would demand if they all faced that rate; actual funding rates can differ across groups when trading is limited. Average demand by bank type also remains identical conditional on aggregate demand. Holding the policy tolerance, administered rates, initial reserve shares, and transfer capacity fixed, the required supply can still differ. The additional requirement comes from local demand mismatches that aggregate demand and average demand by type do not reveal.

Section 2 reviews the related literature. Section 3 presents the bank problem and the central bank’s reserve-supply decision. Section 4 derives minimum reserve supply in frictionless and segmented interbank markets. Section 5 compares systems with identical aggregate demand. Section 6 discusses implications for further QT, and Section 7 concludes. Proofs are in Appendix A, and Appendix E provides institutional background.

²We refer to the groups as domestic and foreign banks, motivated by the differences documented in our companion paper, [Oguri and Pizzimenti \(2026\)](#). That paper documents foreign banks’ disproportionately large reserve holdings and relates their demand to differences in balance-sheet costs and global dollar liquidity needs. The same reasoning applies to other groups whose ability to trade with one another is limited. Appendix E describes the institutional setting in more detail.

2 Related literature

Our paper is closest to research on how differences across banks interact with frictions in the redistribution of reserves. Models of search and bilateral bargaining make the distribution of reserve holdings and access to counterparties central to interbank trading and funding rates (Afonso and Lagos, 2015a,b). Differences in access to reserve remuneration and balance-sheet costs also shape trading between banks and nonbank lenders (Bech and Klee, 2011; Armenter and Lester, 2017; Afonso et al., 2019). Bianchi and Bigio (2022) connect these trading frictions to banks’ liquidity risk, reserve buffers, and lending decisions.

Participation and trading frictions are especially relevant when central banks reduce reserves. Kim et al. (2020) show that balance-sheet costs can prevent reserve reductions from reviving bank-to-bank trading. Closest to our focus on segmentation, Drehmann and Fu (2026) show how banks’ participation decisions and inactive banks’ payment flows can change the reserves available to active banks even when total reserves are unchanged. Lagos and Navarro (2025) combine heterogeneous banks with search and bargaining, estimate aggregate reserve demand, and identify reserve quantities consistent with probabilistic rate control. They also show that transaction rates can conceal differences in banks’ shadow funding costs. These papers establish why market participation and the distribution of reserves matter for monetary policy implementation.

A few recent empirical studies highlight the importance of reserve distribution. Anbil et al. (2026b) show that bank-to-bank lending can reveal funding pressures obscured by the larger segment of the federal funds market supplied by nonbank lenders. Ennis and Wolman (2026) distinguish where reserve changes initially occur from how reserves subsequently move across bank groups. These studies motivate our focus on initial reserve balances and the capacity for redistribution. Regulatory costs and liquidity requirements also help explain differences in reserve holdings and borrowing rates (Banegas and Tase, 2020; Tase, 2022; Anderson and Tase, 2023).

The reserve-supply problem also builds on research on reserve demand and uncertainty. Poole (1968) models reserve choices under uncertain payments. In Afonso et al. (2020), banks can have different reserve needs, but competitive interbank trading combines their demands at a common rate. Our frictionless benchmark has the same implication: heterogeneity alone does not prevent aggregate demand from determining minimum supply. Afonso et al. (2023) characterize optimal supply under demand uncertainty and examine standing lending facilities. Empirical work estimates aggregate demand and its changes to assess reserve amplex and the capacity for QT (Lopez-Salido and Vissing-Jorgensen, 2025; Afonso et al., 2022b). Acharya and Rajan (2024) show how deposit creation and reserve hoarding can limit the liquidity provided

by central-bank balance-sheet expansion. Our focus is on the additional reserve requirement that remains when aggregate demand is held fixed but redistribution is limited.

Connections across funding markets provide further evidence on the importance of access to liquidity. Afonso et al. (2022a) use wholesale payments to assess reserve conditions, while Copeland et al. (2025) link repo funding pressure to reserves and incoming payments at major intermediaries. d’Avernas et al. (2025) study intraday constraints on intermediation between federal funds and repo markets. Anbil et al. (2026a) show how money market funds’ liquidity supply relative to Treasury financing demand can constrain repo capacity and the minimum central-bank balance sheet. Our transfer constraint instead concerns the redistribution of available reserves between bank groups.

Our contribution is an explicit link from heterogeneous demand, the initial distribution of reserves, and finite redistribution capacity to the minimum supply satisfying a given probability bound on local funding stress. The supply requirement combines aggregate and local needs, and its upper quantile determines the central bank’s choice. We use this characterization to compare systems with identical aggregate demand at every common funding rate in every demand state and identical average demand by bank type, including conditional on aggregate demand. Holding the policy tolerance, administered rates, initial reserve shares, and transfer capacity fixed, local demand mismatches can still require different reserve supplies. This comparison isolates why knowing aggregate reserve demand need not be enough to determine the capacity for QT.

3 The reserve-supply problem

We set the stage for comparing frictionless and segmented interbank markets by describing banks’ reserve demand and the central bank’s reserve-supply decision. The two cases share the same bank problem and initial distribution of reserves, but differ in how freely banks can redistribute reserves through trading.

3.1 Banks and initial distribution of reserves

There are two groups of banks, domestic banks (d) and foreign branches and agencies (f). Each group contains a continuum of banks that take funding rates as given and have the same costs and distribution of liquidity needs. Let W_j denote group j ’s predetermined assets, $W = W_d + W_f$, and $w_j = W_j/W$, with $w_j > 0$ and $w_d + w_f = 1$. The central bank supplies a total amount of reserves S , and group j holds X_j after trading. We measure reserves relative to assets:

$$x_j = \frac{X_j}{W_j}, \quad s = \frac{S}{W}, \quad \frac{X_j}{W} = w_j x_j. \quad (1)$$

Thus x_j measures reserves relative to the group's own assets, while $w_j x_j$ measures its holdings relative to system assets. Bank size is held fixed throughout the supply comparison.

We assume that reserves are distributed in proportion to bank assets before interbank trading.³ Both groups therefore begin with the same reserve-to-asset ratio:

$$e_d(s) = e_f(s) = s, \quad \sum_j w_j e_j(s) = s. \quad (2)$$

Throughout the paper, “initial” means before interbank trading. The central bank chooses total reserve supply s , taking these initial shares as given. The shares remain fixed across demand states and supply levels, although subsequent interbank trading can change each group's share of reserves. The same initial distribution applies in the frictionless and segmented interbank markets.

Reserves earn interest i^R . Let i_j be the funding rate available to group j , and write $r_j = i_j - i^R$. Both groups face the same funding rate in a frictionless interbank market, but their rates can differ when the market is segmented. All rates and costs are measured over the same horizon.

3.2 Timing and information

The central bank chooses reserve supply before it knows the demand conditions that banks will face. Banks then trade before their liquidity needs are realized. We set these decisions into four stages, using the same timing in the frictionless and segmented interbank markets.

1. **Reserve planning and initial distribution of reserves.** The central bank observes information $\mathcal{I}$, including asset weights, the initial distribution of reserves, the trading arrangement, and a predictive distribution of demand conditions. It chooses total supply s before the demand state θ is known. Initial balances are then given by (2). The central bank commits to this reserve supply for the period, taking the initial reserve shares as given. Neither supply nor these shares change after demand conditions become known.
2. **Demand conditions become known.** The state θ is realized and becomes common knowledge among market participants. It determines banks' reserve-holding costs and the conditional distributions of their later liquidity needs. A bank knows the distribution of its liquidity need, denoted by P_j for type j , but not its realization. For example, if banks in one group expect to make larger payments, they may choose to hold more reserves in advance. The initial balances $e_j(s)$ remain fixed as these demand conditions change.

³The baseline sets domestic banks' initial reserve share $\alpha = w_d$. Appendix D allows a general share $\alpha \in (0, 1)$ that can differ from the asset share, while retaining the same bank problem and asset normalization.

3. **Interbank trading and market clearing.** Given s and θ , banks choose net borrowing and finalized reserve holdings. They assess the expected cost of a later shortfall using the information revealed in the second stage. Borrowing and lending redistribute the supplied reserves, and funding rates adjust to clear the market. Banks face a common rate in the frictionless market; in the segmented market, limits on transfers can leave the groups facing different rates. This stage determines the funding spreads that enter the central bank's rate-control objective.
4. **Late liquidity needs and payoffs.** After interbank trading closes, liquidity needs P_j are realized. A bank has the reserves x_j held at the close of trading to cover its need. Any uncovered amount $(P_j - x_j)_+$ incurs the shortfall cost specified in the bank's problem below. Banks cannot reopen interbank trading after observing this realization. Anticipating this cost gives them a reason to hold reserves when they trade.

Banks choose reserve holdings after observing the demand state θ , but before their liquidity needs P_j are realized. They therefore account for the expected cost of a later shortfall. The central bank chooses supply earlier, before θ is known, so it also faces uncertainty about banks' demand and the resulting funding rates. Its probability constraint limits the risk that either group's funding spread exceeds the ceiling when banks trade.

The central bank thus plans reserve supply in advance, and banks then redistribute reserves in response to demand conditions. We hold the predictive distribution of θ fixed when comparing different supply choices. Neither the initial distribution of reserves nor intermediation capacity adjusts to the realized state.

3.3 The central bank's objective

Let $\bar{r}$ be the highest acceptable funding spread and $\kappa \in (0, 1)$ the tolerated probability of exceeding it. The central bank chooses

$$\min_{s \geq 0} s \quad \text{subject to} \quad \Pr(R(s, \theta) > \bar{r} \mid \mathcal{I}) \leq \kappa, \quad R(s, \theta) = \max\{r_d(s, \theta), r_f(s, \theta)\}. \quad (3)$$

The same choice minimizes any strictly increasing cost of reserve supply under this constraint. In a frictionless interbank market, R is the common funding spread. In a segmented market, the constraint limits the probability that either group's spread exceeds the ceiling. It therefore addresses funding pressure within each group, which may not be visible in an average transaction rate.

Capacity for quantitative tightening. Let $s_0 = S_0/W$ denote reserve supply before QT, measured relative to the same predetermined banking-system assets. The results below give a minimum required reserve supply s^* , with all $s \geq s^*$ satisfying (3). Assume $s_0 \geq s^*$. Choosing the smallest remaining reserve supply is therefore equivalent to choosing the largest reserve reduction consistent with the stress tolerance:

$$\Delta s^*(s_0) = s_0 - s^*, \quad s_0 \geq s^*. \quad (4)$$

When comparing two systems, we use the same s_0 , high enough to satisfy the policy constraint in both. The shares defining the initial distribution of reserves in (2) remain fixed at each remaining reserve supply s . A reserve reduction therefore initially falls on each group according to its asset share. Interbank trading then redistributes the remaining balances.

We hold administered rates, facility arrangements, other central-bank liabilities, and central-bank capital fixed. With other liabilities and capital fixed, a reduction in central-bank assets corresponds one-for-one to a reduction in reserves. Thus $W\Delta s^*$ also gives the permitted balance-sheet contraction.⁴

3.4 Bank demand

Conditional on θ , a type- j bank faces a nonnegative liquidity need P_j , measured relative to its own assets. Its distribution $F_j(\cdot \mid \theta)$ is continuous, satisfies $F_j(0 \mid \theta) = 0$, and has a finite mean. Holding reserves avoids a cost $\chi_j > 0$ per unit of unmet liquidity need. The parameter $c_j \geq 0$ captures linear balance-sheet costs, while $\phi_j > 0$ captures the increase in marginal costs as reserve holdings grow.

Let b_j denote net borrowing relative to group assets. Finalized holdings satisfy

$$x_j = e_j(s) + b_j, \quad b_j \geq -e_j(s). \quad (5)$$

Positive b_j means borrowing and negative b_j means lending. Banks borrow and lend at the same local rate i_j , with no additional individual financing constraint. Costs depend on finalized reserve holdings, rather than separately on borrowing or initial wealth. The interest payoff is $i^R x_j - i_j b_j = -r_j x_j + i_j e_j(s)$. The bank takes the funding rate as given, so the last term does

⁴Changes in other central-bank liabilities or capital can alter how much securities runoff reduces reserves. The central bank plans reserve supply for one period under the specified demand forecast.

not depend on its choice. Choosing net borrowing is therefore equivalent to choosing finalized holdings to solve

$$\max_{x_j \geq 0} \left\{ -(r_j + c_j)x_j - \frac{\phi_j}{2}x_j^2 - \chi_j \mathbb{E}[(P_j - x_j)_+ | \theta] \right\}. \quad (6)$$

The first term is interest income net of funding and linear holding costs. The second captures the increasing cost of holding more reserves. The last is the expected cost of a liquidity shortfall. Liquidity needs can arise from settlement obligations, deposit withdrawals, or credit-line drawdowns. They affect banks' desired holdings; they are not an additional drain on reserves in the model.

Strict concavity gives a unique demand $D_j(r_j; \theta)$. For positive holdings, taking the first-order condition of (6) with respect to x_j and rearranging gives its inverse:

$$r_j = \nu_j(x_j; \theta) = \chi_j[1 - F_j(x_j | \theta)] - c_j - \phi_j x_j. \quad (7)$$

At a higher funding spread, the bank holds fewer reserves. Where a density exists, $-\partial \nu_j / \partial x_j = \phi_j + \chi_j f_j(x_j | \theta) > 0$. Rates rise more sharply as reserves decline when the marginal value of reserves changes more with holdings. The supply results below use decreasing demand and allow any distribution for P_j satisfying the stated assumptions. Costs and liquidity needs can differ across groups and states, so either group can have higher demand.

We assume $\bar{r} < \chi_j - c_j$ in every relevant state, so both groups demand positive reserves at the policy ceiling. The minimum reserve supply needed in each state and the quantile corresponding to the central bank's tolerance are assumed finite. These restrictions keep the focus on reserve supply when both groups demand reserves at the tolerated rate.

4 Optimal reserve supply

We compare two interbank market structures, keeping the bank problem and initial distribution of reserves unchanged. In the frictionless market, banks trade freely across groups, and a common funding spread clears the market. In the segmented market, banks trade competitively within each group, but transfers between groups are limited, so the two markets can clear at different funding spreads.⁵

⁵In the U.S. federal funds market, FHLBs lend mainly to foreign banks, while domestic banks borrow from both nonbank and bank lenders (Anbil et al., 2026b). The link in our model can be interpreted as a common lender, such as an FHLB, redirecting a fixed amount of lending between the groups, or a foreign bank lending directly to domestic banks.

4.1 Frictionless interbank market

In a frictionless interbank market, the distribution of aggregate reserve demand at the tolerated funding spread determines optimal reserve supply, even when banks differ in their costs and liquidity needs.

Starting from the initial distribution of reserves in (2), banks borrow and lend across groups at a common funding spread r , with no limit on transfers. Group j 's net borrowing is $b_j = D_j(r; \theta) - e_j(s)$. Market clearing requires total borrowing to equal total lending:

$$w_d b_d + w_f b_f = 0 \iff D(r; \theta) \equiv \sum_j w_j D_j(r; \theta) = s. \quad (8)$$

The two conditions are equivalent because $\sum_j w_j e_j(s) = s$. Initial balances affect who borrows and lends, while total supply and aggregate demand determine the common funding spread. Decreasing demand gives a unique clearing spread $r^I(s, \theta)$ for every $s > 0$, as shown in Appendix A.

Define each group's demand at the tolerated spread by

$$t_j(\theta) = D_j(\bar{r}; \theta), \quad A(\theta) = w_d t_d(\theta) + w_f t_f(\theta). \quad (9)$$

Here t_j is measured relative to group assets, while A is aggregate demand at the policy ceiling, measured relative to system assets.

Proposition 1 (Reserve supply in a frictionless interbank market). *The common funding spread exceeds $\bar{r}$ if and only if $s < A(\theta)$. With known demand, minimum supply is $A(\theta)$. Under uncertainty, the solution to (3) is*

$$s_I^* = Q_{1-\kappa}(A \mid \mathcal{I}), \quad (10)$$

where $Q_p(Y \mid \mathcal{I}) = \inf\{y : \Pr(Y \leq y \mid \mathcal{I}) \geq p\}$.

If supply is below demand at the ceiling, the rate must rise above that ceiling to induce banks to hold fewer reserves. If supply covers that demand, the clearing rate is at or below the ceiling. The requirement $A(\theta)$ applies to a particular demand state; the central bank chooses s_I^* before that state is known. With a 5 percent tolerance, it needs enough reserves to cover A with at least 95 percent probability: the 95th percentile of A . Given reserve supply before QT $s_0 \geq s_I^*$, the capacity for reserve reduction is $s_0 - s_I^*$.

Differences across banks therefore matter through their contribution to aggregate demand. Once the distribution of A is known, the central bank has enough information to determine

minimum supply in the frictionless market. With limited interbank trading, however, reserves can be sufficient in aggregate while one group cannot obtain what it needs at the tolerated spread. The next subsection derives these local supply requirements.

4.2 Segmented interbank market

We now consider a segmented interbank market with the same initial distribution of reserves and bank problem. Banks trade competitively within each group, but transfers between groups pass through an intermediation channel with finite capacity. Limits on counterparty access, intermediation, or the time available to rearrange funding can give rise to this constraint. Banks cannot transfer reserves between groups outside this channel. We hold its capacity fixed over the reserve-planning horizon.

Let Q be the net transfer from foreign to domestic banks and $K \geq 0$ its capacity. Normalize both by total banking assets: $q = Q/W$ and $k = K/W$. A positive transfer finances domestic net borrowing with foreign net lending. Local market clearing requires

$$w_d b_d = q, \quad w_f b_f = -q, \quad |q| \leq k. \quad (11)$$

Net borrowing is measured relative to each group's assets, whereas q is measured relative to system assets. Substituting $b_j = D_j(r_j; \theta) - e_j(s)$ gives the corresponding final holdings:

$$\begin{aligned} w_d D_d(r_d; \theta) &= w_d s + q, \\ w_f D_f(r_f; \theta) &= w_f s - q. \end{aligned}$$

As in the frictionless market, $w_d b_d + w_f b_f = 0$ and final holdings sum to s . The additional restriction is that net borrowing across groups cannot exceed the transfer capacity.

Intermediaries move reserves toward the market with the higher funding rate. An intermediary takes local rates as given and chooses q to maximize $(r_d - r_f)q$ subject to $|q| \leq k$. For $k > 0$ and positive holdings, equilibrium satisfies

$$\begin{cases} r_d = r_f, & -k < q < k, \\ r_d \geq r_f, & q = k, \\ r_d \leq r_f, & q = -k. \end{cases} \quad (12)$$

A domestic rate above the foreign rate draws reserves toward domestic banks until the rates equalize or transfers reach the capacity limit. Reserves move in the opposite direction when the foreign rate is higher. If rates still differ at the capacity limit, the intermediary earns the spread

because its capacity is scarce. When $k = 0$, no transfer is possible and each rate clears its local market.

Equations (11) and (12), together with bank demand, determine the equilibrium. Equilibrium reserve holdings are unique because desired holdings fall with the local rate and the rate gap falls as reserves move toward domestic banks. Appendix A proves existence and covers cases in which a group holds no reserves. The central bank chooses total reserve supply before demand is known. Banks then trade within the capacity limit after learning demand conditions.

Aggregate and local supply requirements. To keep both funding spreads at or below $\bar{r}$, each group must obtain at least $t_j(\theta)$ reserves per dollar of its own assets. Three conditions are necessary:

$$s \geq A, \quad w_d s + k \geq w_d t_d, \quad w_f s + k \geq w_f t_f. \quad (13)$$

The first condition requires enough reserves in total. The second requires that domestic banks' initial balances plus the largest possible transfer cover their demand at the ceiling. The third imposes the same requirement for foreign banks. The smallest reserve supply satisfying all three conditions is

$$C(\theta; k) = \max \left\{ A(\theta), t_d(\theta) - \frac{k}{w_d}, t_f(\theta) - \frac{k}{w_f} \right\}. \quad (14)$$

Proposition 2 (Reserve supply in a segmented interbank market). *In competitive equilibrium, neither local spread exceeds $\bar{r}$ if and only if $s \geq C(\theta; k)$. With known demand, minimum supply is C . Under uncertainty, the solution to (3) is*

$$s_K^* = Q_{1-\kappa}(C \mid \mathcal{I}). \quad (15)$$

Holding the predictive demand distribution, initial distribution of reserves, and asset weights fixed, greater transfer capacity weakly reduces s_K^* . Moreover, $s_K^* \geq s_I^*$ under the same predictive distribution.

Interbank trading also makes these three conditions sufficient. Suppose total supply covers A , but domestic banks hold less than their demand at the ceiling. Foreign banks must then hold more than their demand at the ceiling, so the domestic funding rate is above the ceiling while the foreign rate is below it. The rate difference draws reserves toward domestic banks. The domestic condition in (13) ensures that transfers can eliminate the shortfall without exceeding capacity. The same argument applies when foreign banks face a shortfall.

Segmentation can therefore require the central bank to leave more reserves in the system. Even with enough reserves in total, one group may be unable to borrow what it needs. Given the same reserve supply before QT $s_0 \geq s_K^*$ in both markets, each additional unit required reduces the capacity for QT by one unit, giving a difference of $s_K^* - s_I^*$ relative to the frictionless benchmark. The difference is zero when transfers cover the relevant mismatches and $C = A$. A capacity limit can still restrict QT even if it rarely binds in normal conditions, because supply depends on the states that determine the relevant quantile. Section 5 compares two systems with the same capacity and aggregate demand to isolate the role of local mismatches.

The conditions for each group also explain why an increase in one group's demand can require a larger increase in total supply. When the domestic requirement alone binds, a marginal increase in domestic demand measured relative to system assets raises the required reserve supply in that state at rate $1/w_d$, holding foreign demand and the initial distribution of reserves fixed. Only the share w_d of each additional unit initially reaches domestic banks, and transfers are already at capacity. The corresponding multiplier for a constrained foreign group is $1/w_f$. Allowing a different initial distribution of reserves, as in the appendix, could change this calculation.

5 Identical aggregate demand, different room for QT

Section 4 derives the reserve supply needed under each trading arrangement. We now examine what the central bank needs to know about reserve demand to determine how far QT can proceed. In the frictionless market, the distribution of aggregate demand at the tolerated spread is sufficient. With limited trading, two systems with identical aggregate demand distributions and identical average demand by bank type can require different reserve supplies, even under the same policy settings, initial distribution of reserves, and transfer capacity. The difference arises from where reserves are needed.

For a given demand state, consider reserve supply $s = A$, the amount sufficient in the frictionless market. At this supply, each group's initial balances may differ from its demand at the ceiling. The domestic demand mismatch U measures the net transfer toward domestic banks needed to meet both groups' demands, relative to system assets:

$$U = w_d(t_d - A), \quad w_d t_d = w_d A + U, \quad w_f t_f = w_f A - U. \quad (16)$$

A positive U calls for a transfer toward domestic banks; a negative U calls for a transfer toward foreign banks. Substituting into (14) separates the requirement in each demand state into aggregate demand and the extra reserves needed when trading cannot cover the mismatch:

$$C = A + L_k(U), \quad L_k(U) = \max \left\{ 0, \frac{U - k}{w_d}, \frac{-U - k}{w_f} \right\}. \quad (17)$$

If $|U| \leq k$, interbank trading covers the mismatch at supply A , so $C = A$. Otherwise, the extra requirement $L_k(U)$ is positive, even when demand is known. Under uncertainty, both aggregate demand and the mismatches that trading cannot cover affect the distribution of required supply.

5.1 Comparing systems with the same aggregate demand

The following comparison preserves the aggregate demand distribution and each type's mean demand conditional on A , while changing the local mismatch.

Proposition 3 (Same aggregate demand, different reserve requirements). *Let $w_d = w_f = 1/2$ and hold $k, \bar{r}, \kappa$, and administered rates fixed across two systems. They have the same aggregate threshold demand $A \geq A_{\min} > 0$ and $\mathbb{E}[A \mid \mathcal{I}] < \infty$. In the first, own-asset demands satisfy $t_d = t_f = A$. In the second,*

$$t_d = A + Z, \quad t_f = A - Z, \quad \Pr(Z = u \mid \mathcal{I}) = \Pr(Z = -u \mid \mathcal{I}) = \frac{1}{2}, \quad (18)$$

where Z is independent of A conditional on $\mathcal{I}$ and $A_{\min} > u > 2k$. Both systems have the same law of aggregate threshold demand and the same mean demand of each type conditional on A and $\mathcal{I}$. Their minimum reserve supplies in each state and optimal supplies under uncertainty satisfy

$$C_0 = A, \quad C_1 = A + u - 2k, \quad s_{K,1}^* - s_{K,0}^* = u - 2k > 0. \quad (19)$$

In the first system, the two groups have equal demand and receive equal shares of reserves before trading. Supplying A therefore meets both groups' demands at the tolerated spread without a transfer. In the second system, one group demands more reserves while the other demands less. These changes cancel in aggregate. Each group is equally likely to have the higher demand, so average demand by type is also unchanged.

At supply A , however, the group with higher demand needs to borrow more than the trading limit allows. Its funding spread therefore exceeds the ceiling even though there are enough reserves in total. Keeping both spreads at or below the ceiling requires more reserves. A larger supply increases the initial balances of the group that is short, without requiring further

borrowing. Each group’s initial share is fixed at one-half, so only half of each additional unit reaches the group that is short.

The required supply rises by $u - 2k$ in every demand state. Minimum supply under uncertainty therefore rises by the same amount. If reserve supply before QT s_0 satisfies the policy constraint in both systems, equation (4) leaves the second system with $u - 2k$ less room for reserve reduction.

Appendix B extends this comparison to the full demand curve. At every common funding spread, the two systems have identical aggregate demand in each state and identical average demand by type conditional on A . Each type’s full demand distribution still differs. These curves describe how many reserves banks would choose to hold if both groups faced the same spread. In the segmented market, limited transfers can leave the groups facing different rates and holding different amounts of reserves.

5.2 Funding stress missed by an aggregate assessment

At a proposed reserve supply s , a forecast using only aggregate demand misses the states in which total reserves are sufficient but one group cannot obtain enough through trading. Since $C \geq A$,

$$\Pr(C > s \mid \mathcal{I}) - \Pr(A > s \mid \mathcal{I}) = \Pr(A \leq s < C \mid \mathcal{I}) \geq 0. \quad (20)$$

In Proposition 3, supplying the amount that is optimal for the first system leaves the second system with a probability of stress above the central bank’s tolerance. In other settings, a rule based on aggregate demand can still meet the tolerance if the reserve supply it recommends is sufficiently high. The distribution of A alone, however, does not reveal how likely these additional stress states are.

To assess funding stress, the central bank therefore needs the joint distribution of aggregate demand A and mismatch U , together with the initial distribution of reserves and transfer capacity. More detailed information can raise the recommended supply if it reveals large mismatches, or lower it if it rules them out. Greater demand variance does not necessarily raise the relevant upper quantile; the effect depends on how the distribution of the required reserve supply changes.

5.3 An illustration with uncertain local demand

Suppose the central bank knows aggregate demand $A = a$ and the two groups have equal asset shares. Let $U \sim \text{Uniform}[-h, h]$, with $0 < h < a/2$. The groups' demands relative to system assets are

$$w_d t_d = \frac{a}{2} + U, \quad w_f t_f = \frac{a}{2} - U.$$

Aggregate demand is a in every state, and each group's average demand is unchanged as h varies. Appendix B shows how the bank's problem generates these demands.

Since $|U|$ is uniform on $[0, h]$, the minimum supply in each state and the optimal supply under uncertainty are

$$\begin{aligned} C &= a + 2 \max\{0, |U| - k\}, \\ s_K^*(h, k) &= a + 2 \max\{0, (1 - \kappa)h - k\}. \end{aligned} \tag{21}$$

If $(1 - \kappa)h \leq k$, interbank trading covers the mismatch often enough to meet the policy tolerance with supply a . Beyond this point, a wider range of local demand requires more reserves. Greater transfer capacity can reduce this requirement.

Figure 1 compares an aligned system, with $U = 0$, and a system with uncertain local demand, with $h = 0.01$. Both have aggregate demand $a = 0.10$ and transfer capacity $k = 0.005$. With a stress tolerance of $\kappa = 0.05$, minimum supply is where the 95th percentile of the maximum local funding spread reaches the ceiling. The aligned system requires reserves equal to 10% of system assets, while the second requires 10.9%. The additional requirement reduces capacity for QT by 0.9 percentage points of system assets.

6 Policy implications

Further QT and the functioning of the interbank market should be considered together. When reserves cannot move readily between banks, the central bank may need to maintain a larger supply to accommodate local funding needs. Improving redistribution can reduce this additional requirement and increase the capacity for QT, even when aggregate reserve demand is unchanged. The relevant question is therefore both how many reserves banks need and whether those reserves can reach the banks that need them.

This connection makes banks' incentives to lend relevant to balance-sheet normalization. [Anderson et al. \(2026\)](#) discuss changes in the remuneration of reserves, including tiering, that could encourage interbank trading and support a smaller central-bank balance sheet. Our results identify a complementary channel: greater capacity to transfer reserves between groups

can lower minimum reserve supply while holding demand and the tolerance for funding stress fixed. Measures that encourage lending are particularly relevant when they help overcome the constraints separating banks with available reserves from banks with greater needs.

The regulation of major money-market participants is also relevant for further QT. For instance, FHLBs supply funding in the federal funds market as part of managing their own liquidity. Their liquidity requirements and the regulatory treatment of alternative investments influence how they distribute funds across markets and counterparties. Greater scope to invest in interest-bearing bank deposits could shift funding away from federal funds lending while providing liquidity to the banks receiving those deposits.⁶ The implications for QT therefore depend on whether banks facing local funding shortages can obtain liquidity through the available channels.

For policy, the key issue is whether redistribution remains effective when local demand is high. More trading in ordinary conditions need not reduce the reserves required to meet the central bank’s stress tolerance. Improvements matter for the QT limit when they ease the constraints in the demand states that determine minimum supply. Assessing further QT therefore requires examining where reserve reductions leave funding gaps and whether banks can cover those gaps through interbank borrowing.

7 Conclusion

How far a central bank can proceed with QT depends on banks’ demand for reserves and their ability to borrow from one another. We derive the smallest reserve supply that keeps the probability of excessive funding pressure in either group within a given tolerance. When reserve supply before QT satisfies the policy constraint, subtracting this minimum gives the amount of reserves the central bank can withdraw. With other central-bank liabilities and capital fixed, this also determines how much the central bank can reduce its balance sheet.

Two banking systems can have identical aggregate demand curves at a common funding spread in every state and identical average demand by bank type, yet allow different amounts of QT. One group’s higher demand can be offset by another’s lower demand in the aggregate. If interbank trading cannot move enough reserves to the group that needs them, the central bank must maintain a larger reserve supply. This difference arises under the same policy tolerance, administered rates, initial distribution of reserves, and transfer capacity.

⁶Gissler et al. (2025) explain that federal funds lending helps FHLBs meet their liquidity requirements and that expanded access to interest-bearing deposit accounts could encourage substitution toward these accounts. Such deposits offer greater intraday flexibility because FHLBs can withdraw funds while Fedwire is open. The Federal Housing Finance Agency (FHFA) proposed extending the higher counterparty exposure limit applicable to overnight federal funds to these deposits in 2024, but subsequently withdrew the proposal in October 2025.

To assess further QT, the central bank needs to know which banks demand reserves, which groups initially bear reserve reductions, and whether interbank trading can meet their needs in the states that determine minimum supply.

References

- Viral V. Acharya and Raghuram Rajan. Liquidity, liquidity everywhere, not a drop to use: Why flooding banks with central bank reserves may not expand liquidity. *Journal of Finance*, 79(5):2943–2991, 2024. doi: 10.1111/jofi.13370. URL <https://doi.org/10.1111/jofi.13370>.
- Gara Afonso and Ricardo Lagos. The over-the-counter theory of the fed funds market: A primer. *Journal of Money, Credit and Banking*, 47(S2):127–154, 2015a. doi: 10.1111/jmcb.12216. URL <https://doi.org/10.1111/jmcb.12216>.
- Gara Afonso and Ricardo Lagos. Trade dynamics in the market for federal funds. *Econometrica*, 83(1):263–313, 2015b. doi: 10.3982/ECTA10586. URL <https://doi.org/10.3982/ECTA10586>.
- Gara Afonso, Roc Armenter, and Benjamin Lester. A model of the federal funds market: Yesterday, today, and tomorrow. *Review of Economic Dynamics*, 33:177–204, 2019. doi: 10.1016/j.red.2019.04.004.
- Gara Afonso, Kyungmin Kim, Antoine Martin, Ed Nosal, Simon Potter, and Sam Schulhofer-Wohl. Monetary policy implementation with ample reserves. Staff Reports 910, Federal Reserve Bank of New York, 2020. URL https://www.newyorkfed.org/research/staff_reports/sr910. Revised September 2026.
- Gara Afonso, Darrell Duffie, Lorenzo Rigon, and Hyun Song Shin. How abundant are reserves? evidence from the wholesale payment system. Staff Reports 1040, Federal Reserve Bank of New York, 2022a. URL https://www.newyorkfed.org/research/staff_reports/sr1040.
- Gara Afonso, Domenico Giannone, Gabriele La Spada, and John C. Williams. Scarce, abundant, or ample? a time-varying model of the reserve demand curve. Staff Reports 1019, Federal Reserve Bank of New York, 2022b. URL https://www.newyorkfed.org/research/staff_reports/sr1019.html. Revised November 2025.
- Gara Afonso, Gabriele La Spada, Thomas M. Mertens, and John C. Williams. The optimal supply of central bank reserves under uncertainty. Staff Reports 1077, Federal Reserve Bank of New York, 2023. URL https://www.newyorkfed.org/research/staff_reports/sr1077. Revised December 2025.
- Sriya Anbil and Zeynep Senyuz. How has monetary and regulatory policy affected trading relationships in the U.S. repo market? *International Journal of Central Banking*, 18(4):1–52, 2022. URL <https://www.ijcb.org/journal/v18n4>.
- Sriya Anbil, Alyssa Anderson, and Zeynep Senyuz. What happened in money markets in September 2019? FEDS Notes, Board of Governors of the Federal Reserve System, February 2020. URL <https://www.federalreserve.gov/econres/notes/feds-notes/what-Happened-in-Money-Markets-in-September-2019-20200227.htm>. February 27.

- Sriya Anbil, Alyssa Anderson, Ethan Cohen, and Romina Ruprecht. Beyond reserves: The federal reserve’s balance sheet and the repo market. Finance and Economics Discussion Series 2026-041, Board of Governors of the Federal Reserve System, June 2026a. URL <https://www.federalreserve.gov/econres/feds/beyond-reserves-the-federal-reserves-balance-sheet-and-the-repo-market.htm>.
- Sriya Anbil, Sebastian Infante, and Zeynep Senyuz. A tale of demand and supply for central bank reserves. Finance and Economics Discussion Series 2026-028, Board of Governors of the Federal Reserve System, 2026b. URL <https://www.federalreserve.gov/econres/feds/a-tale-of-demand-and-supply-for-central-bank-reserves.htm>.
- Alyssa Anderson and Manjola Tase. LCR premium in the federal funds market. Finance and Economics Discussion Series 2023-071, Board of Governors of the Federal Reserve System, 2023. URL <https://www.federalreserve.gov/econres/feds/lcr-premium-in-the-federal-funds-market.htm>. Forthcoming in the International Journal of Central Banking.
- Alyssa G. Anderson, Alessandro Barbarino, Anthony M. Diercks, and Stephen Miran. A user’s guide to reducing the Federal Reserve’s balance sheet. Finance and Economics Discussion Series 2026-019, Board of Governors of the Federal Reserve System, March 2026. URL <https://www.federalreserve.gov/econres/feds/files/2026019pap.pdf>.
- Roc Armenter and Benjamin Lester. Excess reserves and monetary policy implementation. *Review of Economic Dynamics*, 23:212–235, 2017. doi: 10.1016/j.red.2016.11.002. URL <https://doi.org/10.1016/j.red.2016.11.002>.
- Ayelen Banegas and Manjola Tase. Reserve balances, the federal funds market and arbitrage in the new regulatory framework. *Journal of Banking & Finance*, 118:105893, 2020. doi: 10.1016/j.jbankfin.2020.105893. URL <https://doi.org/10.1016/j.jbankfin.2020.105893>.
- Morten L. Bech and Elizabeth Klee. The mechanics of a graceful exit: Interest on reserves and segmentation in the federal funds market. *Journal of Monetary Economics*, 58(5):415–431, 2011. doi: 10.1016/j.jmoneco.2011.05.007. URL <https://doi.org/10.1016/j.jmoneco.2011.05.007>.
- Javier Bianchi and Saki Bigio. Banks, liquidity management, and monetary policy. *Econometrica*, 90(1):391–454, 2022. doi: 10.3982/ECTA16599. URL <https://doi.org/10.3982/ECTA16599>.
- Adam Copeland, Darrell Duffie, and Yilin Yang. Reserves were not so ample after all. *Quarterly Journal of Economics*, 140(1):239–281, 2025. doi: 10.1093/qje/qjae034. URL <https://doi.org/10.1093/qje/qjae034>.
- Adrien d’Avernas, Baiyang Han, and Quentin Vandeweyer. Intraday liquidity and money market dislocations. *Management Science*, 71(12):10740–10752, 2025. doi: 10.1287/mnsc.2023.04037.
- Mathias Drehmann and Xuewen Fu. Hard to shift, easy to reshape: Central bank reserve demand and frictions. BIS Working Papers 1372, Bank for International Settlements, August 2026. URL <https://www.bis.org/publ/work1372.htm>.

- Huberto M. Ennis and Alexander L. Wolman. The diffusion of bank reserves in the U.S. Working Paper 26-12, Federal Reserve Bank of Richmond, September 2026. URL https://www.richmondfed.org/publications/research/working_papers/2026/wp_26-12. Manuscript dated August 31, 2026.
- Stefan Gissler, Sam Hempel, R. Jay Kahn, Patrick McCabe, and Borghan Narajabad. Monitoring reserve scarcity through nonbank cash lenders. Feds notes, Board of Governors of the Federal Reserve System, March 2025. URL <https://www.federalreserve.gov/econres/notes/feds-notes/monitoring-reserve-scarcity-through-nonbank-cash-lenders-20250328.html>. March 28.
- R. Jay Kahn, Matthew McCormick, Vy Nguyen, Mark Paddrik, and H. Peyton Young. Anatomy of the repo rate spikes in September 2019. Working Paper 23-04, Office of Financial Research, April 2023. URL https://www.financialresearch.gov/working-papers/files/OFRwp-23-04_anatomy-of-the-repo-rate-spikes-in-september-2019.pdf. April 25.
- Kyungmin Kim, Antoine Martin, and Ed Nosal. Can the U.S. interbank market be revived? *Journal of Money, Credit and Banking*, 52(7):1645–1689, 2020. doi: 10.1111/jmcb.12693. URL <https://doi.org/10.1111/jmcb.12693>.
- Ricardo Lagos and Gastón Navarro. Monetary policy operations: Theory, evidence, and tools for quantitative analysis. Forthcoming in *Econometrica*; author draft dated October 29, 2025, 2025. URL https://www.gastonnavarro.com/_files/ugd/25a3ea_7e70ee6912014e9abf9dd5b31592cb92.pdf.
- David Lopez-Salido and Annette Vissing-Jorgensen. Reserve demand, interest rate control, and quantitative tightening. Working paper, May 1, 2025. URL https://drive.google.com/file/d/1CAF3X4pgSH0xcAVGfyYaBmm8TgCB_tEY/view.
- Junko Oguri and Cristoforo Pizzimenti. Ample reserves for whom? the role of foreign banks in U.S. monetary policy implementation. Working paper, September, 2026.
- William Poole. Commercial bank reserve management in a stochastic model: Implications for monetary policy. *Journal of Finance*, 23(5):769–791, 1968. doi: 10.1111/j.1540-6261.1968.tb00316.x. URL <https://doi.org/10.1111/j.1540-6261.1968.tb00316.x>.
- Dave Ramsden. Quantitative tightening: The next chapter. Bank of England speech, September 2026. URL <https://www.bankofengland.co.uk/speech/2026/september/dave-ramsden-speech-on-quantitative-tightening>. September 28.
- Manjola Tase. Demand segmentation in the federal funds market. Finance and Economics Discussion Series 2022-071, Board of Governors of the Federal Reserve System, 2022. URL <https://www.federalreserve.gov/econres/feds/demand-segmentation-in-the-federal-funds-market.htm>.

Kevin Warsh. Inflation is a choice: Kevin warsh on fixing the federal reserve. Interview with Peter Robinson, Uncommon Knowledge, Hoover Institution, 2025. URL <https://www.hoover.org/plus/research/inflation-choice-kevin-warsh-fixing-federal-reserve>. Recorded May 28.

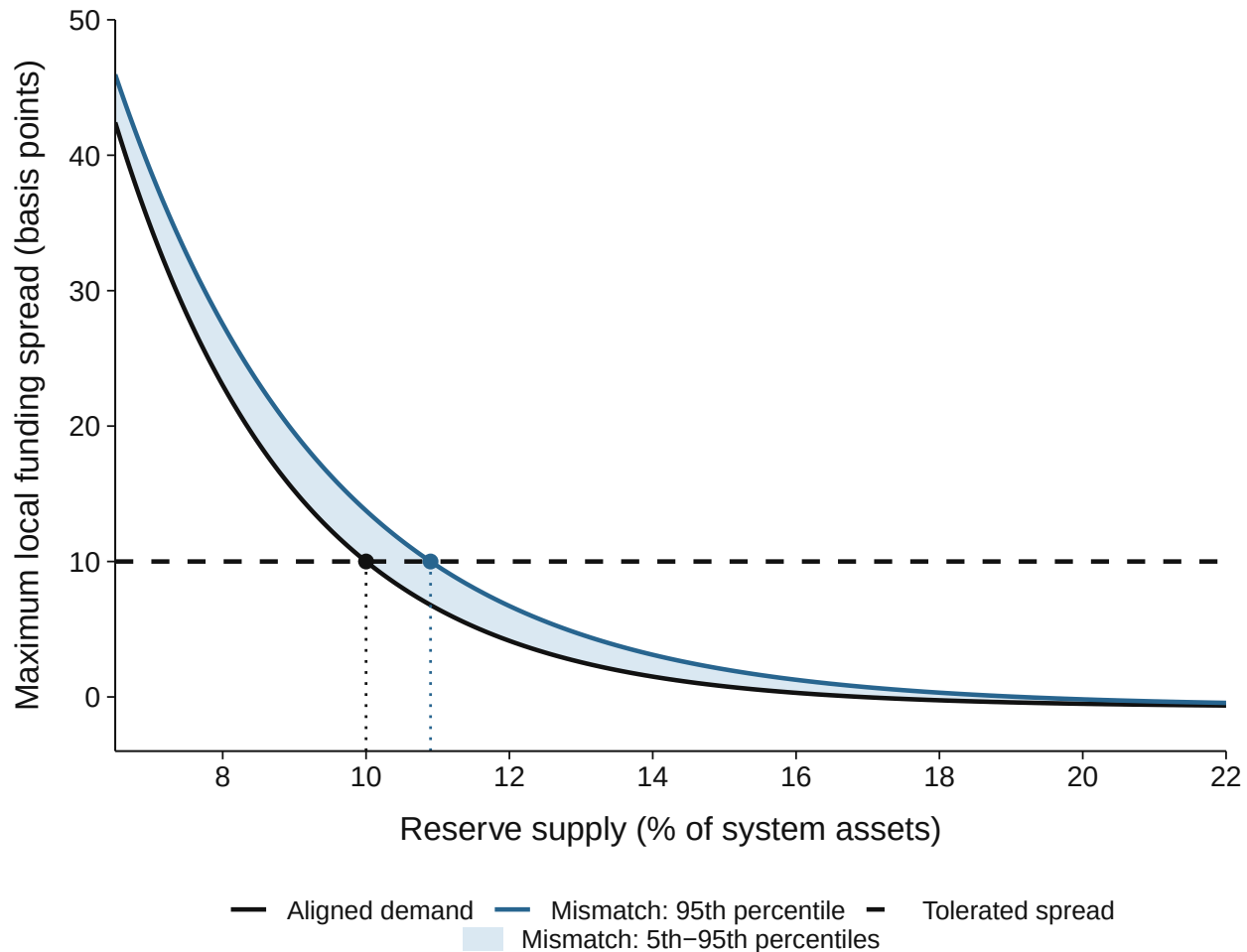


Figure 1: Identical aggregate demand, different reserve requirements

Notes: The black curve shows the funding spread with aligned demand ($U = 0$). The blue curve shows the 95th percentile of the maximum local funding spread when $U \sim \text{Uniform}[-0.01, 0.01]$; shading spans the 5th–95th percentiles. Their intersections with the tolerated spread of 10 basis points determine minimum reserve supplies of 10% and 10.9% of system assets under a 5% stress tolerance. Both systems have identical aggregate common-rate demand and average demand by type, equal initial reserve shares, and transfer capacity $k = 0.005$. Curves use Appendix B’s demand construction with $a = 0.10$, $\sigma = 0.025$, $c = 0.00005$, and $\phi = 0.0001$. Parameters are illustrative, not calibrated.

A Demand, equilibrium, and proofs

A.1 Bank demand and frictionless interbank trading

Fix a demand state. Because P_j has a continuous distribution and finite mean,

$$\mathbb{E}[(P_j - x)_+ | \theta] = \int_x^\infty [1 - F_j(p | \theta)] dp.$$

Its derivative is $-[1 - F_j(x | \theta)]$. The quadratic holding cost makes the bank's objective strictly concave. Since $F_j(0 | \theta) = 0$, demand is zero if $r \geq \chi_j - c_j$. Otherwise it is the unique positive solution to

$$r = \nu_j(x; \theta) \equiv \chi_j[1 - F_j(x | \theta)] - c_j - \phi_j x. \quad (22)$$

The inverse demand is continuous, strictly decreasing, and tends to minus infinity as x grows. These properties do not require a particular distribution for liquidity needs.

Using $b_j = D_j(r; \theta) - e_j(s)$ and $\sum_j w_j e_j(s) = s$, aggregate net borrowing is $\sum_j w_j b_j = \sum_j w_j D_j(r; \theta) - s$. Aggregate demand is continuous, strictly decreasing wherever positive, unbounded as $r \rightarrow -\infty$, and zero at $r \geq \max_j(\chi_j - c_j)$. For each $s > 0$, there is therefore a unique rate at which aggregate net borrowing is zero. This rate also determines final holdings and net trades uniquely. At $s = 0$, choose the common rate $\max_j(\chi_j - c_j)$, the limit of the positive-supply equilibrium. Since $t_j = D_j(\bar{r}; \theta) > 0$, this rate exceeds $\bar{r}$.

Proof of Proposition 1. Monotonicity and market clearing give

$$r^I(s, \theta) > \bar{r} \iff s < \sum_j w_j D_j(\bar{r}; \theta) = A(\theta).$$

The equivalence also holds at zero supply. Thus the policy constraint is $F_A(s | \mathcal{I}) \geq 1 - \kappa$. For any nonnegative random requirement Y with finite relevant quantile, the feasible set of this constraint is

$$[Q_{1-\kappa}(Y | \mathcal{I}), \infty).$$

Right-continuity of the distribution function ensures that its lower endpoint is feasible. Minimizing supply therefore selects the generalized quantile, including when the predictive distribution has atoms. Applying this argument to $Y = A$ proves the result. $\square$

A.2 Equilibrium in the segmented interbank market

For $s > 0$, market clearing gives $b_d = q/w_d$ and $b_f = -q/w_f$. The constraints $b_j \geq -e_j(s)$ and $|q| \leq k$ therefore give the feasible interval

$$\mathcal{Q}(s) = [\max\{-k, -w_d s\}, \min\{k, w_f s\}].$$

The equilibrium transfer is the unique maximizer over this interval of

$$\mathcal{V}(q) = w_d \int_0^{s+q/w_d} \nu_d(x; \theta) dx + w_f \int_0^{s-q/w_f} \nu_f(x; \theta) dx. \quad (23)$$

We use asset weights because demand is measured relative to each group's own assets, whereas q is measured relative to total assets. The objective is strictly concave in q . Its derivative is $\nu_d(x_d) - \nu_f(x_f)$, the local rate gap when both groups hold positive reserves. Its first-order conditions therefore imply equal rates at an interior transfer and the appropriate rate inequality at either capacity limit. This maximization problem describes competitive market clearing. The central bank chooses supply through the separate rate-control problem.

If a group holds no reserves, its rate must be at least its choke rate $\chi_j - c_j$. The one-sided optimality condition for (23) implies that the other group's rate is at least this choke rate. Assigning the same rate to the group holding no reserves supports both bank demand and the intermediary's profit maximization over $[-k, k]$. These prices support the reserve holdings without treating the nonnegativity of reserve holdings as an additional limit on intermediation. Other supporting rates can exist, but a group holding no reserves always faces a spread above $\bar{r}$ because its demand at the ceiling is positive. At $s = 0$, zero holdings and a common rate at or above both choke rates support equilibrium, and stress is certain. Rates may therefore be indeterminate at these corners, but this does not affect whether the policy ceiling is exceeded.

Proof of Proposition 2. For both local rates to be at most $\bar{r}$, equilibrium holdings must satisfy $x_j \geq t_j$. Adding these inequalities with asset weights and applying the transfer bound yields

$$s \geq A, \quad w_d s + k \geq w_d t_d, \quad w_f s + k \geq w_f t_f.$$

These inequalities are equivalent to $s \geq C$.

Conversely, suppose $s \geq C$ but $x_d < t_d$. Because aggregate reserves are sufficient, $x_f > t_f$, so $r_d > \bar{r} > r_f$. The competitive intermediary must then choose $q = k$. This contradicts $w_d s + k \geq w_d t_d$. The argument is symmetric if foreign banks are short. Consequently,

$$\max\{r_d, r_f\} > \bar{r} \iff s < C.$$

The quantile argument above, now applied to $Y = C$, proves the optimal supply rule. Since $C \geq A$ state by state, segmented-market supply is weakly larger than supply in the frictionless market under the same predictive distribution. Increasing k weakly lowers C in every state and hence weakly lowers its quantile. $\square$

Proof of Proposition 3. With $w_d = w_f = 1/2$ and own-asset demands $t_d = A + Z$, $t_f = A - Z$,

$$C = \max\{A, A + Z - 2k, A - Z - 2k\} = A + (|Z| - 2k)_+.$$

Conditional on $\mathcal{I}$, the first system has $Z = 0$, while the second has $Z = \pm u$ with equal probability, independently of A . Both have the same aggregate threshold demand A and the same conditional mean demand by type $\mathbb{E}[t_j | A, \mathcal{I}] = A$. The finite conditional mean of A ensures that their mean type demands given $\mathcal{I}$ also coincide. Positivity follows from $A \geq A_{\min} > u$. Since $u > 2k$, the minimum reserve supplies in each state are $C_0 = A$ and $C_1 = A + u - 2k$. Adding a constant shifts every quantile by that constant, so the optimal reserve supplies differ by $u - 2k > 0$.

Let $a^* = Q_{1-\kappa}(A | \mathcal{I})$. By the definition of the generalized quantile, $F_A(a^* - (u - 2k) | \mathcal{I}) < 1 - \kappa$. Using a^* in the second system therefore violates the rate-control tolerance, even when the distribution has atoms. More generally, $C \geq A$ implies the exact identity

$$\Pr(C > s | \mathcal{I}) - \Pr(A > s | \mathcal{I}) = \Pr(A \leq s < C | \mathcal{I}).$$

This follows by dividing $\{C > s\}$ into states in which aggregate reserves are sufficient and states in which they are not. $\square$

B A smooth bank-demand construction

The bank's optimization problem can generate the demand schedules used in Proposition 3. Choose constants $\sigma, \phi > 0$ and $c > \max\{0, -\bar{r}\}$. For each $A > 0$, define

$$\chi(A) = (\bar{r} + c + \phi A)e^{A/\sigma}, \quad \nu^0(x; A) = \chi(A)e^{-x/\sigma} - c - \phi x.$$

Let $H(r; A)$ be the associated nonnegative demand. By construction, $H(\bar{r}; A) = A$. For $Z = 0$ or $Z = \pm u$, set

$$\lambda_d = 1 + Z/A, \quad \lambda_f = 1 - Z/A, \quad D_j(r; A, Z) = \lambda_j H(r; A).$$

Both scaling factors are positive. These demands, measured relative to each group's own assets, follow from

$$\nu_j(x; A, Z) = \chi(A)e^{-x/(\lambda_j \sigma)} - c - \frac{\phi}{\lambda_j} x, \tag{24}$$

which follows from the bank's problem when liquidity needs are exponentially distributed with mean $\lambda_j \sigma$ and the quadratic cost coefficient is ϕ/λ_j . Since $w_d = w_f = 1/2$,

$$w_d D_d(r; A, Z) + w_f D_f(r; A, Z) = H(r; A)$$

at every common rate, including when demand is zero. At $\bar{r}$, demands relative to each group's own assets are $A + Z$ and $A - Z$. Each type's average demand conditional on A is also identical across the two systems at every common rate. The construction therefore preserves the entire aggregate demand curve while allowing local demand and the reserve supply required under segmentation to differ. It changes both liquidity needs and holding costs. Arbitrary offsetting shifts in nonlinear demand curves need not leave aggregate demand unchanged at every rate.

For Section 5.3, the same construction applies with $A = a$ and $Z = 2U$, where $U \sim \text{Uniform}[-h, h]$ and $0 < h < a/2$. The scaling factors remain positive and $\mathbb{E}[Z \mid \mathcal{I}] = 0$, so varying h preserves the aggregate common-rate demand curve and each type's average demand at every common rate.

C Reserve holdings and marginal adjustment

In the frictionless market, define $m_j = -\partial D_j / \partial r$ at a differentiable equilibrium and $M = \sum_j w_j m_j > 0$. Differentiating market clearing yields

$$\frac{\partial r^I}{\partial s} = -\frac{1}{M}, \quad a_j \equiv \frac{\partial(w_j x_j)}{\partial s} = \frac{w_j m_j}{M}. \quad (25)$$

The marginal adjustment share a_j generally differs from group j 's share of total reserve holdings $w_j x_j / s$. For positive holdings and a density f_j ,

$$m_j = \frac{1}{\phi_j + \chi_j f_j(x_j \mid \theta)}.$$

The slopes of banks' demand curves determine how rates respond to a reserve reduction. The supply rules depend on demand at the tolerated rate. Holding the joint distribution of demand at that rate fixed, a change in slopes elsewhere leaves A , C , and the quantiles determining supply unchanged. It can still change how far rates rise when reserves fall short.

D A general initial distribution of reserves

The baseline distributes reserves in proportion to bank assets before interbank trading. Here we allow domestic banks' initial reserve share $\alpha \in (0, 1)$ to differ from their asset share w_d . The

share is known when the central bank chooses supply and remains fixed across demand states and supply comparisons. Initial holdings remain proportional to assets within each group, so

$$e_d^\alpha(s) = \frac{\alpha s}{w_d}, \quad e_f^\alpha(s) = \frac{(1-\alpha)s}{w_f}, \quad \sum_j w_j e_j^\alpha(s) = s. \quad (26)$$

All other assumptions, including asset normalization and the bank's problem, are unchanged. The central bank chooses total supply and takes these shares as given.

Market clearing. In the frictionless market, aggregate net borrowing is still $\sum_j w_j D_j(r; \theta) - s$. The initial share affects who borrows and lends, but not the clearing rate or final holdings. Proposition 1 therefore continues to apply.

In the segmented market, net borrowing satisfies $w_d b_d = q$, $w_f b_f = -q$, and $|q| \leq k$. Final holdings satisfy

$$w_d D_d(r_d; \theta) = \alpha s + q, \quad w_f D_f(r_f; \theta) = (1-\alpha)s - q. \quad (27)$$

The intermediary faces the same profit-maximization problem and rate conditions in (12). For $s > 0$, the feasible transfer interval is

$$\mathcal{Q}_\alpha(s) = [\max\{-k, -\alpha s\}, \min\{k, (1-\alpha)s\}].$$

Replacing the upper integration limits in (23) gives

$$\mathcal{V}_\alpha(q) = w_d \int_0^{(\alpha s + q)/w_d} \nu_d(x; \theta) dx + w_f \int_0^{((1-\alpha)s - q)/w_f} \nu_f(x; \theta) dx.$$

This objective is strictly concave, with derivative equal to the local rate gap when both groups hold positive reserves. Its unique maximizer over $\mathcal{Q}_\alpha(s)$ determines equilibrium reserve holdings. The argument in Appendix A also applies when a group holds no reserves and at zero supply.

Required reserve supply. Both groups can meet their demand at the ceiling if and only if

$$s \geq A, \quad \alpha s + k \geq w_d t_d, \quad (1-\alpha)s + k \geq w_f t_f.$$

The minimum reserve supply in each state is therefore

$$C_\alpha(\theta; k) = \max \left\{ A(\theta), \frac{w_d t_d(\theta) - k}{\alpha}, \frac{w_f t_f(\theta) - k}{1-\alpha} \right\}. \quad (28)$$

The necessity of these conditions follows from market clearing. For sufficiency, suppose they hold but domestic banks have less than their demand at the ceiling. Foreign banks must then

have more, so the rate gap requires $q = k$. Domestic holdings, measured relative to system assets, would equal $\alpha s + k \geq w_d t_d$, a contradiction. The argument is symmetric for a foreign shortfall.

Let R_α denote the maximum local funding spread under this initial distribution. The same quantile argument as in Proposition 2 gives

$$R_\alpha(s, \theta) > \bar{r} \iff s < C_\alpha(\theta; k), \quad s_{K,\alpha}^* = Q_{1-\kappa}(C_\alpha \mid \mathcal{I}). \quad (29)$$

Holding demand, asset weights, and α fixed, greater transfer capacity weakly reduces this supply. Since $C_\alpha \geq A$ in every state, $s_{K,\alpha}^* \geq s_I^*$. From a feasible reserve supply before QT s_0 , the permitted reserve reduction is $s_0 - s_{K,\alpha}^*$.

Demand mismatches and the initial distribution. Define $U_\alpha = w_d t_d - \alpha A$. The required supply can be written as

$$C_\alpha = A + L_{\alpha,k}(U_\alpha), \quad L_{\alpha,k}(U) = \max \left\{ 0, \frac{U - k}{\alpha}, \frac{-U - k}{1 - \alpha} \right\}. \quad (30)$$

Setting $\alpha = w_d$ recovers the baseline initial holdings, market clearing, and supply formulas. Proposition 3 uses this case with equal asset weights, so its comparison remains valid within the general model.

When the domestic requirement alone binds, a marginal increase in domestic demand measured relative to system assets raises required supply at rate $1/\alpha$, holding foreign demand, asset weights, and the initial shares fixed. Only the share α of each additional unit initially reaches domestic banks, and transfers are already at capacity. The corresponding multiplier for a constrained foreign group is $1/(1 - \alpha)$. Changing the initial distribution can ease one group's constraint while tightening the other's.

E Institutional details of reserve and interbank markets

The U.S. illustrates how banks differ in their demand for reserves and their access to funding. Reserves are balances banks hold at the Federal Reserve to settle payments and meet liquidity needs. The interest paid on these balances helps the Federal Reserve control short-term rates. Banks adjust their holdings through unsecured and secured funding markets, but differ in their access to funding and willingness to lend. Assessing whether reserves are sufficient therefore requires knowing whether banks can obtain them through these markets.

Heterogeneity in reserve demand. Banks differ in payment activity, funding structure, and balance-sheet costs. Domestic banks rely more heavily on U.S. deposits, while foreign branches and agencies draw more on wholesale funding and transfers from related foreign offices. Their

reserve holdings can therefore reflect the dollar liquidity needs of the global banking organization (Oguri and Pizzimenti, 2026). Uninsured foreign branches also face different deposit-insurance assessment costs from insured domestic banks, although parent-level constraints can still limit their balance-sheet expansion (Banegas and Tase, 2020). These differences can lead banks to demand different amounts of reserves even when trading is frictionless.

Segmentation in the federal funds market. Federal funds are unsecured, typically overnight loans. In the post-crisis market, Federal Home Loan Banks (FHLBs) are the principal lenders and foreign branches are prominent borrowers. FHLBs hold accounts at the Federal Reserve but cannot earn interest on those balances. Eligible banks can borrow below the interest rate on reserves and earn the spread, net of balance-sheet costs. Limits on access to counterparties and bilateral bargaining constrain this arbitrage and help explain the spread (Bech and Klee, 2011; Banegas and Tase, 2020). Banks also borrow to meet their operating liquidity needs (Tase, 2022). The effective federal funds rate summarizes these transactions, but may not reveal funding pressure in every part of the market (Gissler et al., 2025).

Segmentation in the repo market. In repo markets, banks and dealers borrow against securities from cash lenders, including money market funds and other nonbanks. These lenders cannot necessarily lend directly to every borrower. Access to funding depends on collateral requirements, trading relationships, and dealers' balance-sheet constraints (Anbil and Senyuz, 2022). Institutions that trade in both federal funds and repo markets connect the two markets. Differences in counterparties and liquidity needs, however, mean that borrowers cannot always substitute one source of funding for the other (Gissler et al., 2025). Cash may therefore be available in one part of the market while borrowers elsewhere face funding pressure.

The September 2019 disruption illustrates the difficulty of assessing reserve adequacy across these markets. Repo rates rose sharply, and the effective federal funds rate briefly exceeded its target range. Corporate tax payments and Treasury securities settlements reduced reserve balances, while the settlements also increased funding demand. The Federal Reserve responded with repo operations that supplied additional reserves (Kahn et al., 2023; Anbil et al., 2020).⁷

Our model distinguishes differences in banks' demand from limits on interbank trading. Costs and liquidity needs determine how many reserves each group wants to hold. A limit on transfers determines how much banks can redistribute between groups. We use domestic and foreign banks to illustrate these groups; nationality itself is not a barrier to trading, but institutional differences (e.g. regulation) might. This setup allows us to study how demand and trading constraints affect the reserves needed for further QT.

⁷Securities runoff had ended in August 2019, but reserves continued to decline as other Federal Reserve liabilities increased (Anbil et al., 2020).