How to Shrink a Central Bank: The Optimal Pace of Quantitative Tightening*

C. Christopher Hyland† Seung Joo Lee‡

July 2026

Abstract

How quickly should a central bank shrink its balance sheet after quantitative easing? In our model, the supply of reserves affects financial conditions even when the administered interest rate is unchanged. Reserve reduction poses a timing tradeoff: moving quickly can strain payments and financial intermediation, while moving slowly extends the period over which declining reserves tighten financial conditions. Private spending responds to expected future financial conditions, so the announced path of quantitative tightening affects demand from the outset. We characterize both the reserve path and its completion date. Without an output-stabilization motive, the optimal path starts quickly and tapers to zero. Stabilization can instead bring completion forward when the cost of prolonged reserve scarcity is large, or slow the transition when short-run payment strains dominate. In a counterfactual U.S. exercise, output stabilization shortens the benchmark transition from 14.00 to 12.21 quarters. Stronger payment-flow effects produce a longer, smoother path. The dollar magnitudes are illustrative, but the timing tradeoff is general: the pace of quantitative tightening depends on the cost of moving quickly and the cost of taking too long.

*We thank Ricardo Reis for helpful comments and suggestions. All errors are ours.
†University of Oxford, Saïd Business School, Exeter College, and Oxford-Man Institute. Email: christopher.hyland.dphil@said.oxford.edu.
‡Saïd Business School, Oxford University, Park End Street, Oxford, United Kingdom, OX1 1HP. Email: seung.lee@sbs.ox.ac.uk. Corresponding Author.

1 Introduction

Large-scale asset purchases have generated a substantial literature on when and how central banks should expand their balance sheets. Their unwinding poses a less studied question: once a central bank has chosen its long-run operating regime, how quickly should it get there? In a floor system, the interest rate on reserves (IOR) and the quantity of reserves are separate policy margins. The central bank can reduce reserve balances without changing IOR, but the transition is not neutral. Lower reserve balances may constrain banks' ability to support liquid deposits and settle payments. Rapid drainage may also strain payment flows and market functioning. The pace of quantitative tightening (QT) is therefore a monetary-policy choice, not merely an implementation detail.

We study this transition by taking the terminal reserve stock and a prescribed completion-contingent administered-rate rule as given, then solving jointly for the completion date and the path of reserve drainage.[1] Existing work does not establish how a central bank should approach a prescribed reserve regime when the stock of reserves and the speed of reduction independently affect financial conditions. The two margins pull the completion date in opposite directions. Reserve scarcity makes delay costly, while payment-flow risk makes rapid drainage costly. When the stock channel is strong enough, output stabilization shortens the transition relative to the benchmark without stabilization.

The mechanism begins with a shadow-rate wedge between IOR and the market rate that governs household saving. We derive this wedge from a banking block in the spirit of Piazzesi, Rogers, and Schneider (2022), augmented with the settlement frictions in Bianchi and Bigio (2022). Households hold deposits rather than reserves, and banks issue those deposits against a collateral pool. Reserves receive the highest collateral weight because they settle payment flows. When bank balance-sheet capacity binds, another dollar of reserves relaxes the constraint

[1] Gross asset runoff need not translate one-for-one into reserve drainage because movements in currency, the Treasury General Account, overnight reverse repurchase balances, and other central-bank liabilities can offset or amplify its effect on reserves. Appendix B derives this mapping.

and earns a scarcity value beyond its administered return. This mechanism accords with the nonlinear reserve-demand curve and the rise in the federal funds rate–IOR spread as reserves leave the ample region (Afonso et al., 2022; López-Salido and Vissing-Jorgensen, 2023). Lower reserves then reduce banks' capacity to support liquid deposits, widen deposit and money-market spreads, and raise the shadow rate faced by households even if IOR is unchanged.

Drainage speed also matters independently. When reserves leave the banking system quickly, intermediaries need more intraday liquidity to settle denser transaction flows (Copeland, Duffie, and Yang, 2025; Duffie, 2026). This payment pressure tightens effective collateral capacity at a given reserve stock and temporarily widens the federal funds rate–IOR spread. The stock effect depends on how far reserves have fallen, whereas the speed effect depends on how quickly they are falling. Both raise the shadow rate, but they have opposite implications for the duration of QT.

Reserve announcements affect aggregate demand even before the stock of reserves begins to adjust. Households respond to the expected path of the shadow rate, so anticipated scarcity and drainage speed influence current expenditure. QT therefore acts as balance-sheet forward guidance even when IOR remains unchanged.

We begin without an output-stabilization motive. The central bank balances the flow cost of maintaining an oversized balance sheet against the direct implementation and market-functioning costs of rapid drainage. Optimal drainage is front-loaded and tapers linearly to zero: an early reduction avoids balance-sheet costs for longer, and that benefit declines as the completion date approaches.

Output stabilization adds two opposing forces. A stronger speed effect makes rapid drainage more contractionary and favors a longer transition. A stronger stock effect makes delay more contractionary because households anticipate reserve scarcity for longer. If the stock effect is strong enough, stabilization advances the completion date rather than producing a more gradual QT path.

We quantify these forces in a counterfactual U.S. transition path starting at the September 2021 post-QE reserve peak. Reserve balances fall from \$4.2 trillion to a conditional terminal benchmark of \$2.9 trillion, which is an estimate, based on conditions in late 2025 and early 2026, of the lower edge of the ample-reserves range under the prevailing operating framework; it is not a technological floor (Duffie, 2026; Reis, 2026; Anderson et al., 2026). The required decline is thus \$1.3 trillion. We set the direct cost of speed so that the flexible-speed transition lasts 3.5 years without output stabilization. Given the cumulative decline, this normalization implies a reference average drainage rate of \$30.95 billion per month. The exercise does not independently identify an optimal dollar pace. Instead, it measures how the two wedges and operational constraints alter the duration and shape of the transition relative to this benchmark.

The benchmark has a 7.4-basis-point stock wedge over the full reserve decline and a 5-basis-point speed wedge at the reference drainage rate. Output stabilization shortens the transition from 14.00 to 12.21 quarters and raises average reserve drainage from 31.0 to 35.5 billion dollars per month. The path remains front-loaded: drainage begins near 63.5 billion dollars per month and falls to about 10.0 billion immediately before exit. Despite the faster average pace, the announcement-date output gap becomes less negative, moving from -0.41 to -0.38 percent, because the shorter horizon limits cumulative exposure to the stock wedge. A reserve-scarcity stress advances completion further, while a payment-flow stress lengthens and smooths the transition. A \$40 billion monthly cap extends the benchmark horizon to 13.52 quarters. These conditional experiments quantify the timing tradeoff and the shape of the optimal path; they do not identify a unique historical optimum.

Related literature. The paper is closest to Benigno and Benigno (2026) and Harrison (2024). Benigno and Benigno (2026) study the joint use of reserves and the policy rate (IOR) in a New Keynesian model with a liquidity channel. Their optimal exit from a liquidity trap begins quantitative tightening before rate liftoff and withdraws liquidity slowly relative to the rate path; their emphasis is instrument coordination and the optimal long-run supply of liquidity. We instead condition on a prescribed terminal reserve stock and a completion-contingent IOR and solve for the entire drainage path and its stopping date. This formulation separates the timing incentives created by the stock of reserves and the speed of reduction. Harrison (2024) studies quantitative easing and tightening through long-bond portfolio frictions and finds that optimal balance-sheet reduction is gradual. Karadi and Nakov (2021) also obtain gradual exit because quantitative easing flattens the yield curve and slows bank recapitalization. We abstract from duration risk and bank equity dynamics and treat quantitative tightening as a reserve-drainage problem in a floor system. Gradualism obtains when the speed channel dominates but can be overturned when the stock channel makes delay costly. Our contribution is a continuous-time characterization of the optimal drainage path and stopping date, including front-loaded tapers, stock–speed comparative statics, and operational caps, conditional on the terminal regime.

A second literature studies the operating framework rather than the transition. Cúrdia and Woodford (2011) treat the central-bank balance sheet and interest on reserves as instruments distinct from the policy rate and argue that IOR should track the short-rate target in a floor-like system. Arce et al. (2020) compare large-balance-sheet floor systems with lean corridor systems. In their framework, floor regimes create policy space by pushing market rates to IOR, while corridor regimes can achieve similar stabilization if temporary QE is used promptly at the effective lower bound. Those papers discipline which steady-state reserve regime to operate. We take a floor system and a terminal ample-reserves benchmark as given, consistent with the post-crisis U.S. framework, and ask how quickly an oversized post-QE reserve stock should converge to that benchmark when both the level and pace of drainage matter.

A third strand provides the empirical and institutional motivation for our two margins. Afonso et al. (2022) and López-Salido and Vissing-Jorgensen (2023) document a nonlinear reserve-demand curve: once reserves leave the ample region, money-market rates rise relative to IOR.[2] Copeland, Duffie, and Yang (2025) show that payment delays at reserve-constrained

[2] Bigio, D'Erasmo, and Yépez (2026) estimate a downward-sloping, regime-dependent demand for U.S. reserves over 2003–2023, using a three-stage strategy that instruments for endogenous Federal Reserve operations and

intermediaries intensified as reserves declined after 2018, while Duffie (2026) argues that payment-system design itself places a floor under the Federal Reserve's balance sheet. Reis (2026) and Anderson et al. (2026) emphasize how payment needs, dealer capacity, and the operating framework shape the feasible terminal stock and the risks of shrinking too far. The cross-country evidence in Du, Forbes, and Luzzetti (2024) likewise indicates that pace and communication matter. We use this evidence to distinguish stock scarcity from drainage speed and formalize the resulting timing tradeoff.

Finally, our banking block connects the evidence on reserves and payments to models of convenience yields and liquid bank liabilities. Bansal and Coleman (1996), Krishnamurthy and Vissing-Jorgensen (2012), and Nagel (2016) study liquidity premia on safe and near-money assets. Lenel, Piazzesi, and Schneider (2019) and Piazzesi, Rogers, and Schneider (2022) embed money and banking in monetary economies in which deposits and reserves jointly determine spreads, while Drechsler, Savov, and Schnabl (2017) emphasize banks' market power in deposit markets. In Bianchi and Bigio (2022), deposit transfers settle in reserves, so reserve scarcity raises liquidity risk and affects credit conditions. We adapt this collateral-and-payments mechanism to a quantitative-tightening transition and derive separate stock and speed components of the household shadow rate. The settlement motive in Bianchi and Bigio (2022) provides the microeconomic counterpart to the payment-flow risk behind the speed effect. The announcement effect follows the New Keynesian principle that expected future intertemporal prices affect current expenditure (Woodford, 2003; Werning, 2012; Dordal i Carreras and Lee, 2024). Here that principle applies to the announced reserve path: quantitative tightening provides forward guidance through the balance sheet.

implementation-regime shifts. Their results reinforce the reserve-scarcity channel that disciplines our stock wedge.

2 Environment

Time is continuous and the economy is deterministic. Suppressing uncertainty isolates the intertemporal problem created by an announced QT path. Private agents know the future path of reserves at the announcement date, so expenditure can respond before any reserves are withdrawn. The central bank begins with a post-QE reserve stock and takes the terminal regime as given. It chooses the drainage path and its completion date. The household block maps the anticipated shadow-rate path into expenditure, while the banking block ties that rate to the reserve stock and drainage speed.

2.1 Balance-sheet normalization

Let b_t denote reserve balances. The central bank begins with b_0 reserves and takes the terminal stock b^* as given. The latter is a boundary condition for the transition problem, not a policy variable chosen within it. Define

$$b_0 > b^*, \qquad \Delta \equiv b_0 - b^* > 0, \tag{1}$$

where Δ is the total quantity of reserves to be drained.

The central bank chooses the completion date $T > 0$ and the drainage path

$$q_t \equiv -\dot{b}_t \geq 0, \tag{2}$$

so a larger q_t represents faster drainage. The sign convention makes the drainage rate positive when reserves are falling. Given a drainage path, reserve balances satisfy

$$b_t = b_0 - \int_0^t q_s \,\mathrm{d}s.$$

Feasibility requires

$$b(0) = b_0, \qquad b(T) = b^*, \tag{3}$$

and therefore

$$\int_0^T q_t \, \mathrm{d}t = \Delta. \tag{4}$$

This accounting constraint links the pace of QT to its duration. For any completion date, average drainage must equal Δ/T. An earlier finish therefore requires faster drainage on average, though the central bank can vary the rate within the transition.

At and after T, reserve balances remain at b^* and $q_t = 0$. The reserve stock is continuous at completion, but the drainage rate need not satisfy a smooth-pasting condition. It may taper to zero or drop discretely to zero when reserves reach b^*. The pair $(T, \{q_t\}_{t\in[0,T]})$ determines the duration and shape of a transition that drains the fixed quantity Δ. This terminal convention separates QT from the stationary reserve regime that follows.

2.2 Household problem

A representative household enters date t with real wealth W_t and chooses consumption C_t and real deposit balances d_t. Deposits pay the nominal rate i_t^D and provide liquidity services. The remaining wealth, $W_t - d_t$, earns the nominal rate i_t^S, which governs the household's intertemporal margin. This two-asset formulation separates the return on saving from the liquidity services used in transactions. Because households do not hold central-bank reserves directly, reserve policy reaches them through bank-intermediated rates and spreads.

Given initial wealth W_0, the household's value is

$$V_0(W_0) = \max_{\{C_t, d_t\}_{t\geq 0}} \int_0^\infty e^{-\rho t} \left[u(C_t) + v(d_t) \right] \mathrm{d}t, \qquad \rho > 0, \tag{5}$$

where u and v are increasing and strictly concave. The household takes interest rates, real income Y_t, lump-sum taxes τ_t, and inflation $\pi_t \equiv \dot{P}_t / P_t$ as given.

The choices are subject to accumulation of wealth

$$\dot{W}_t = (i_t^S - \pi_t)(W_t - d_t) + (i_t^D - \pi_t)d_t + Y_t - C_t - \tau_t, \tag{6}$$

with $W(0) = W_0$ and the usual no-Ponzi condition. For an interior solution, the first-order condition for deposits is

$$\frac{v'(d_t)}{u'(C_t)} = i_t^S - i_t^D. \tag{7}$$

The spread $i_t^S - i_t^D$ is the convenience yield on deposits: liquidity services compensate households for their lower pecuniary return. A wider spread raises the opportunity cost of deposits and, other things equal, reduces desired balances.

With CRRA consumption utility $u'(C_t) = C_t^{-1/\sigma}$, the intertemporal first-order condition is

$$\frac{\dot{C}_t}{C_t} = \sigma(i_t^S - \pi_t - \rho). \tag{8}$$

The rate relevant for intertemporal expenditure is therefore i_t^S, not the deposit or administered rate. The banking sector links reserve policy to this Euler-relevant rate.

2.3 Commercial banks

Commercial banks issue deposits backed by reserves and less-liquid assets. Reserves settle payments at par, so the two asset classes are not perfect substitutes in supporting deposit creation. When bank balance-sheet capacity binds, fewer reserves mean less deposit capacity and a higher marginal value of reserves. The scarcity premium widens the deposit spread $i_t^S - i_t^D$ and the reserve spread $i_t^S - i_t^M$, where i_t^M is IOR. Empirically, the latter is the federal funds rate–IOR gap.

Appendix B derives this spread from household deposit demand, bank collateral capacity, and payment-flow risk. We can write the Euler-relevant rate as

$$i_t^S = i_t^M + \nu(b_t) + \Theta(q_t), \qquad \Theta(0) = 0, \tag{9}$$

where $\nu(b_t)$ captures reserve scarcity and $\Theta(q_t)$ captures the contemporaneous pressure created by reserve drainage. We assume

$$\nu'(b) < 0, \qquad \nu''(b) \geq 0, \tag{10}$$

$$\Theta'(q) \geq 0, \qquad \Theta''(q) \geq 0. \tag{11}$$

The stock component rises as reserves fall, and its marginal effect weakly increases at lower reserve levels. The speed component rises with drainage and disappears when drainage stops because $\Theta(0) = 0$. Gradual drainage limits the speed component but extends exposure to reserve scarcity. The optimal transition balances these stock and flow effects.

2.4 Aggregate demand

Substituting the banking relation (9) into the household Euler equation (8) links reserve policy to expenditure:

$$\frac{\dot{C}_t}{C_t} = \sigma\left[i_t^M + \nu(b_t) + \Theta(q_t) - \pi_t - \rho\right]. \tag{12}$$

To express consumption relative to its undistorted path, let C_t^n denote natural consumption and let r_t^n denote the natural real rate, so that

$$\frac{\dot{C}_t^n}{C_t^n} = \sigma(r_t^n - \rho). \tag{13}$$

Define the output gap as

$$x_t \equiv \log C_t - \log C_t^n. \tag{14}$$

With goods-market clearing, the consumption gap coincides with the output gap. Subtracting (13) from (12) gives

$$\dot{x}_t = \sigma \left[i_t^M + \nu(b_t) + \Theta(q_t) - \pi_t - r_t^n \right]. \tag{15}$$

Equation (15) is the continuous-time counterpart of the standard New Keynesian IS relation (Werning, 2012; Galí, 2015). QT adds the stock and speed components to the short rate in the household's Euler equation. A positive wedge raises consumption growth relative to its natural path. Given a terminal condition, consumption adjusts through a lower current level. The output gap therefore depends on the expected reserve path, not just current reserves or drainage.

3 QT, aggregate demand, and the policy problem

We first derive the output gap for a given announced reserve path, holding the conventional-policy rule fixed. The central bank then chooses the path and its completion date.

3.1 Administered-rate normalization and forward guidance

To isolate the balance-sheet channel, we impose perfectly sticky prices, $\pi_t = 0$, and condition on an administered-rate rule that does not offset QT during the transition. The rule aligns the household shadow rate with the natural rate before drainage and again after reserves reach their terminal stock. During the transition, let

$$i_t^M = r_t^n - \nu(b_0), \qquad 0 \leq t < T. \tag{16}$$

When $b_t = b_0$ and $q_t = 0$, this rule implies $i_t^S = r_t^n$. As QT proceeds, the administered rate continues to offset only the initial stock wedge, allowing changes in reserves and drainage to pass through to i_t^S. For $t \geq T$, we set $i_t^M = r_t^n - \nu(b^*)$ and impose $x_T = 0$, so the terminal regime is again neutral.

Under this rule, substituting (16) into (15) and solving backward from $x_T = 0$ gives

$$x_t = -\sigma \int_t^T [\nu(b_s) - \nu(b_0) + \Theta(q_s)] \, \mathrm{d}s, \qquad x_T = 0. \tag{17}$$

Because $\nu'(b) < 0$, $\Theta(q) \geq 0$, and $b_s \leq b_0$, the integrand is nonnegative and $x_t \leq 0$ throughout the transition. Equation (17) makes the forward-guidance channel explicit. The current output gap reflects the cumulative wedge along the announced reserve path. A slower transition prolongs reserve scarcity; a faster one raises the contribution from drainage speed. Demand can therefore respond to QT at announcement, before reserves decline.

Offsetting QT. The contraction in (17) is conditional on conventional policy. If the administered rate adjusted at each date according to $i_t^M = r_t^n - \nu(b_t) - \Theta(q_t)$, then $i_t^S = r_t^n$ and $x_t = 0$. We rule out that response and optimize the reserve path conditional on (16). This keeps the balance-sheet channel distinct from a contemporaneous rate offset.

3.2 The central bank's problem

The central bank chooses the completion date T and drainage path $\{q_t\}_{t\in[0,T]}$ by solving

$$\min_{T,\{q_t\}_{t\in[0,T]}} \int_0^T \left[c_B(b_t - b^*) + \frac{\chi}{2} q_t^2 + \frac{\lambda}{2} x_t^2 \right] \mathrm{d}t \tag{18}$$

subject to

$$\dot{b}_t = -q_t, \qquad b(0) = b_0, \qquad b(T) = b^*, \qquad q_t \geq 0, \tag{19}$$

$$\dot{x}_t = \sigma \left[\nu(b_t) - \nu(b_0) + \Theta(q_t) \right], \qquad x_T = 0. \tag{20}$$

The objective contains three policy costs. Reserves above b^* incur the flow cost $c_B(b_t - b^*)$. Rapid drainage incurs the direct implementation and market-functioning cost $\chi q_t^2/2$, while deviations of output from its natural level incur the stabilization loss $\lambda x_t^2/2$. The direct drainage cost is separate from the speed wedge $\Theta(q_t)$, which affects expenditure through the household Euler equation. Reserves evolve forward from b_0, but the terminal condition $x_T = 0$ pins down the output gap. Since x_0 is free, the output-gap equation is solved backward and current demand reflects the full future path of QT wedges.

For the analytical results, we specialize the two wedges to the local linear-quadratic form

$$\nu(b_t) - \nu(b_0) = \kappa(b_0 - b_t), \qquad \Theta(q_t) = \theta q_t^2, \qquad \kappa > 0,\ \theta \geq 0. \tag{21}$$

The coefficient κ measures how much the stock wedge rises per unit decline in reserves; θ governs the convex speed wedge. Appendix B derives both coefficients and allows output to feed back into deposit demand, a channel omitted from the analytical baseline.

4 Optimal quantitative tightening

We solve the policy problem first under constant drainage, reducing policy to a horizon choice and isolating the forces that determine the completion date. We then allow the central bank to choose the entire drainage path. Appendix A contains the derivations and proofs.

4.1 Constant-speed benchmark

Suppose reserve drainage is constant over the transition. Completing the fixed reduction $\Delta = b_0 - b^*$ in T periods requires

$$q_t = q = \frac{\Delta}{T}, \qquad b_t = b_0 - \frac{\Delta}{T}t, \qquad b_t - b^* = \Delta\left(1 - \frac{t}{T}\right). \tag{22}$$

The completion date is then the only policy choice. A longer horizon mechanically lowers the drainage rate and produces a linear reserve path.

Substituting (22) into (17) decomposes the announcement-date output gap into stock and speed components:

$$x_0 = -\sigma \left[\frac{\kappa \Delta T}{2} + \frac{\theta \Delta^2}{T} \right]. \tag{23}$$

The stock contribution to $|x_0|$ rises with T because a longer transition extends exposure to reserve scarcity. The speed contribution falls with T because a lower drainage rate reduces the payment-flow wedge. The two channels therefore pull the completion date in opposite directions.

Proposition 1 (Optimal constant-speed horizon)**.** *The constant-speed problem has a unique optimal horizon $T^{CS,*} > 0$. The horizon decreases with the flow cost of maintaining an oversized balance sheet, c_B, and increases with the direct cost of rapid drainage, χ. When $\lambda > 0$, it also decreases with the stock sensitivity κ. The effect of the speed sensitivity θ is generally ambiguous when both channels are present.*

The single-channel cases resolve the source of this ambiguity. With only a speed wedge ($\kappa = 0$), output stabilization favors a longer transition because gradual drainage reduces the wedge. With only a stock wedge ($\theta = 0$), stabilization favors an earlier exit because delay extends reserve scarcity. When both channels operate, a change in θ shifts their relative importance, so the optimal horizon need not vary monotonically. Appendix A gives the exact solution, the single-channel formulas, and the comparative-static threshold.

4.2 The optimal reserve-drainage path

Relaxing the constant-speed restriction in (18)–(20), the central bank allocates the fixed reduction Δ over time and chooses when to finish. Early drainage reduces balance-sheet costs for longer, but a high drainage rate carries direct and aggregate-demand costs.

The optimal drainage rule. Let $\mu_{b,t}$ and $\mu_{x,t}$ denote the costates on reserve balances and the output gap. Under $\Theta(q_t) = \theta q_t^2$, the necessary conditions imply

$$q_t = \max\left\{0, \frac{\mu_{b,t}}{\chi + 2\sigma\theta\mu_{x,t}}\right\}. \tag{24}$$

Under the linear stock specification in (21), the costates evolve according to

$$\dot{\mu}_{b,t} = -c_B - \sigma\mu_{x,t}\nu'(b_t) = -c_B + \sigma\kappa\mu_{x,t}, \qquad \dot{\mu}_{x,t} = -\lambda x_t. \tag{25}$$

Equation (24) sets drainage equal to its shadow benefit $\mu_{b,t}$ divided by an effective marginal cost. That marginal cost combines χ with the output effect of the speed wedge. The stock sensitivity κ affects drainage through the evolution of $\mu_{b,t}$. Appendix A reports the remaining optimality conditions.

A linear taper without output stabilization. Setting $\lambda = 0$ removes the output gap from the objective and yields a closed-form drainage path.

Proposition 2 (Optimal taper without stabilization)**.** *If $\lambda = 0$, the optimal flexible-speed horizon and drainage path are*

$$T^{VS,*} = \sqrt{\frac{2\chi\Delta}{c_B}}, \qquad q_t^* = \frac{c_B}{\chi}(T^{VS,*} - t). \tag{26}$$

The taper follows from the declining value of early drainage. A unit drained near the start avoids the balance-sheet flow cost for longer than one drained near completion. The free-terminal-time condition implies $q_T^* = 0$. A larger χ lengthens the transition, while a larger c_B shortens it.

Output stabilization: speed versus stock. When $\lambda > 0$, the speed channel raises the effective marginal cost of drainage, while the stock channel changes the shadow benefit of reducing reserves.

Proposition 3 (Conditional speed effect and the optimal QT path)**.** *Suppose $\theta > 0$ and the QT wedge is nonnegative along the transition. Then $x_t \leq 0$ and $\mu_{x,t} \geq 0$. At any date with positive drainage, the speed channel weakly lowers the optimal drainage rate for a given shadow benefit of reducing reserves:*

$$q_t \leq \frac{\mu_{b,t}}{\chi}. \tag{27}$$

This conditional result does not imply that output stabilization lengthens the QT transition. The stock channel changes the shadow benefit $\mu_{b,t}$ and may make earlier completion optimal. The effect on the fully optimized drainage path and transition horizon is therefore not signed in general.

QT can change both the intensity and the duration of output losses. Rapid drainage creates a larger speed wedge while reserves are being removed. Holding the benefit of lower reserves fixed, the central bank would spread drainage over a longer period, as (27) shows. Yet a slower path also delays the return to the neutral terminal regime. Earlier completion shortens the period over which the stock wedge depresses demand. When this duration effect is large enough, the central bank drains faster and finishes earlier even though the speed channel alone favors gradualism. Proposition 3 therefore gives no general sign for the effect of output stabilization on the pace or duration of QT. Appendix A proves Proposition 3, and Appendix B extends the conditional speed effect to endogenous deposit demand.

4.3 Reserve-drainage constraints

The central bank may face an upper bound on how quickly reserves can be withdrawn without impairing market functioning or placing excessive demands on private balance sheets. We represent these considerations by the reduced-form constraint

$$0 \leq q_t \leq \bar{q}, \tag{28}$$

where $\bar{q}$ is the maximum feasible drainage rate. Under constant speed, feasibility requires $T \geq \Delta/\bar{q}$. The constrained optimum is therefore

$$T_{\bar{q}}^{CS,*} = \max\left\{T^{CS,*}, \frac{\Delta}{\bar{q}}\right\}. \tag{29}$$

The cap does not bind when the unconstrained plan is already gradual enough. Otherwise, it imposes the minimum horizon $\Delta/\bar{q}$ and lengthens the transition. Under flexible speed, it also changes the shape of the path. Since unconstrained drainage is front-loaded, a binding cap holds q_t at $\bar{q}$ initially. The policy joins the interior path once the desired drainage rate falls below the ceiling. Section 5 quantifies the delay.

5 Quantitative exercise

We apply the model to a counterfactual U.S. balance-sheet normalization beginning at the September 2021 post-QE peak in reserve balances (Duffie, 2026). The exercise traces the model-implied drainage path as the stock and speed wedges, the output-stabilization weight, and operational limits change. In mapping the model to the data, we interpret the stock wedge $\nu(b_t)$ broadly. It includes reserve scarcity and limits on private intermediation as Treasury securities move onto private balance sheets (Anderson et al., 2026).

5.1 Calibration

Calibration strategy. The reserve endpoints (b_0, b^*) determine total drainage, while targets for the two shadow-rate wedges discipline (κ, θ). The ratios χ/c_B and λ/c_B then govern the baseline implementation horizon and the weight on output stabilization. We use c_B as the objective numeraire.

Reserve endpoints. The reserve endpoints are

$$b_0 = 4.2, \qquad b^* = 2.9, \qquad \Delta = 1.3, \tag{30}$$

where quantities are in trillions of dollars. The initial stock is observed. Based on conditions in late 2025 and early 2026, we set b^* to a benchmark for the lower edge of the ample-reserves range under the prevailing operating framework. It is not a technological floor.[3] The boundary remains uncertain because it depends on reserve demand and the elasticity of central-bank liquidity supply (Afonso et al., 2022; López-Salido and Vissing-Jorgensen, 2023; Duffie, 2026; Reis, 2026). The two endpoints imply the required decline Δ.

The benchmark pace of normalization. Only the relative scale of c_B and χ matters, so we set $c_B = 1$. We choose χ so that the flexible-speed transition lasts 3.5 years without an output-stabilization motive:

$$T_0 \equiv T^{VS,*}\big|_{\lambda=0} = \sqrt{\frac{2\chi\Delta}{c_B}} = 3.5 \quad \text{years.} \tag{31}$$

This normalization implies $\chi/c_B = 4.711538$ and the reference drainage rate

$$q_{\text{ref}} \equiv \frac{\Delta}{T_0} = 0.371429 \quad \text{trillion dollars per year} = 30.95 \quad \text{billion dollars per month.} \tag{32}$$

The reference rate is the average pace required to drain Δ over T_0. It is neither an additional calibration target nor a policy cap. The choice of χ/c_B normalizes the direct cost of imple-

[3] Anderson et al. (2026) associate a roughly 2-basis-point EFFR–IORB spread with a less-ample operating point near \$2.65 trillion but emphasize considerable uncertainty. We report that estimate as context rather than use it as the terminal condition.

mentation but does not determine the payment-flow wedge, which depends separately on θ. Appendix Table 3 varies the duration normalization from 3 to 4.5 years.

Stock and speed wedges. We set the annualized stock wedge over the full reserve decline to 7.4 basis points and the annualized speed wedge at q_{ref} to 5 basis points:

$$\kappa\Delta = 0.00074, \tag{33}$$

$$\theta q_{\text{ref}}^2 = 0.0005. \tag{34}$$

We discipline the stock target with the semi-log reserve-demand estimate in López-Salido and Vissing-Jorgensen (2023).[4] In annual decimal-rate units, their coefficient implies

$$0.002 \log\left(\frac{4.2}{2.9}\right) = 0.0007407, \tag{35}$$

or approximately 7.4 basis points over the counterfactual reserve decline. We match this endpoint change with a linear stock wedge, which preserves the analytical model's linear-quadratic structure and implies $\kappa = 0.00056923$ per trillion dollars. The appendix also solves the model with the semi-log schedule itself. These reserve-demand estimates discipline the reduced-form stock wedge, but they do not separately identify the household Euler wedge or the pass-through of wholesale funding conditions.

The speed channel is less tightly disciplined. Evidence on payment delays and dealer balance sheets documents temporary funding pressure when reserve conditions are tight (Copeland, Duffie, and Yang, 2025; Duffie, 2026), but it does not identify the mapping from aggregate monthly reserve drainage to θ. We set the speed wedge to 5 basis points at q_{ref} and consider a payment-flow stress with a substantially stronger channel. This target normalizes the wedge at the reference rate; it does not cap the wedge along the optimized path. Gross SOMA runoff

[4]Note that Bigio, D'Erasmo, and Yépez (2026) find a similar semi-log reserve demand elasticity estimate.

provides institutional context but is not a one-for-one measure of reserve drainage. The target implies $\theta = 0.00362426$.

The two targets set the scale of the shadow-rate channels. They are not jointly estimated structural parameters.

Aggregate demand and policy loss. We set $\sigma = 1$ and $\lambda = 10,000$. With the output gap measured in decimals, a one-percentage-point gap sustained for one year produces a loss of 0.5, half the cost of holding \$1 trillion of reserves above b^* for one year under the normalization $c_B = 1$. We treat λ/c_B as a policy-loss normalization rather than an estimated welfare weight. Appendix Table 3 varies λ over a broad range of output-stabilization priorities.

All paths are conditional policy scenarios for a fixed terminal benchmark and a completion-contingent administered-rate rule. They show how the composition of the QT wedge changes the timing and shape of reserve drainage. Table 1 summarizes the calibration. We solve the flexible-speed problem by direct transcription. Appendix B maps (κ, θ) into banking primitives and separates the direct drainage cost χ from the effects contained in Θ.

Table 1: Benchmark calibration: inputs, targets, and normalizations

Object	Benchmark value	Role	Empirical status
b_0	\$4.2 trillion	Initial state	September 2021 post-QE peak
b^*	\$2.9 trillion	Conditional terminal benchmark	Status-quo lower edge; varied in sensitivity
c_B	1	Objective numeraire	Normalization
T_0	3.5 years	Pins down χ/c_B	Implementation normalization
q_{ref}	\$30.95bn/month	Evaluates speed target	Implied by Δ/T_0
Stock target	7.4 bp	Pins down κ	Benchmark terminal target
Speed target	5 bp at q_{ref}	Pins down θ	Benchmark scenario; weakly identified
σ	1	Intertemporal elasticity	Standard benchmark
λ	10,000	Output-loss weight	Policy-loss normalization

5.2 Results

Table 2 compares the benchmark policy with the exact no-stabilization taper and two capped alternatives. Without output stabilization, the transition lasts 14.00 quarters. Giving output losses their benchmark weight reduces the horizon to 12.21 quarters. Monthly drainage starts at 63.5 billion dollars and falls to 10.0 billion immediately before completion. The stock channel is therefore strong enough in the benchmark for output stabilization to accelerate completion. A monthly cap of $60 billion has little effect, while a $40 billion cap extends the horizon to 13.52 quarters.

Figure 1 traces the mechanism. Relative to the no-stabilization taper, the benchmark reaches b^* earlier and has a smaller initial output gap. Faster average drainage raises the speed wedge, but the shorter exposure to the stock wedge reduces the cumulative demand effect. Binding caps limit the initial front-loading of QT and postpone completion.

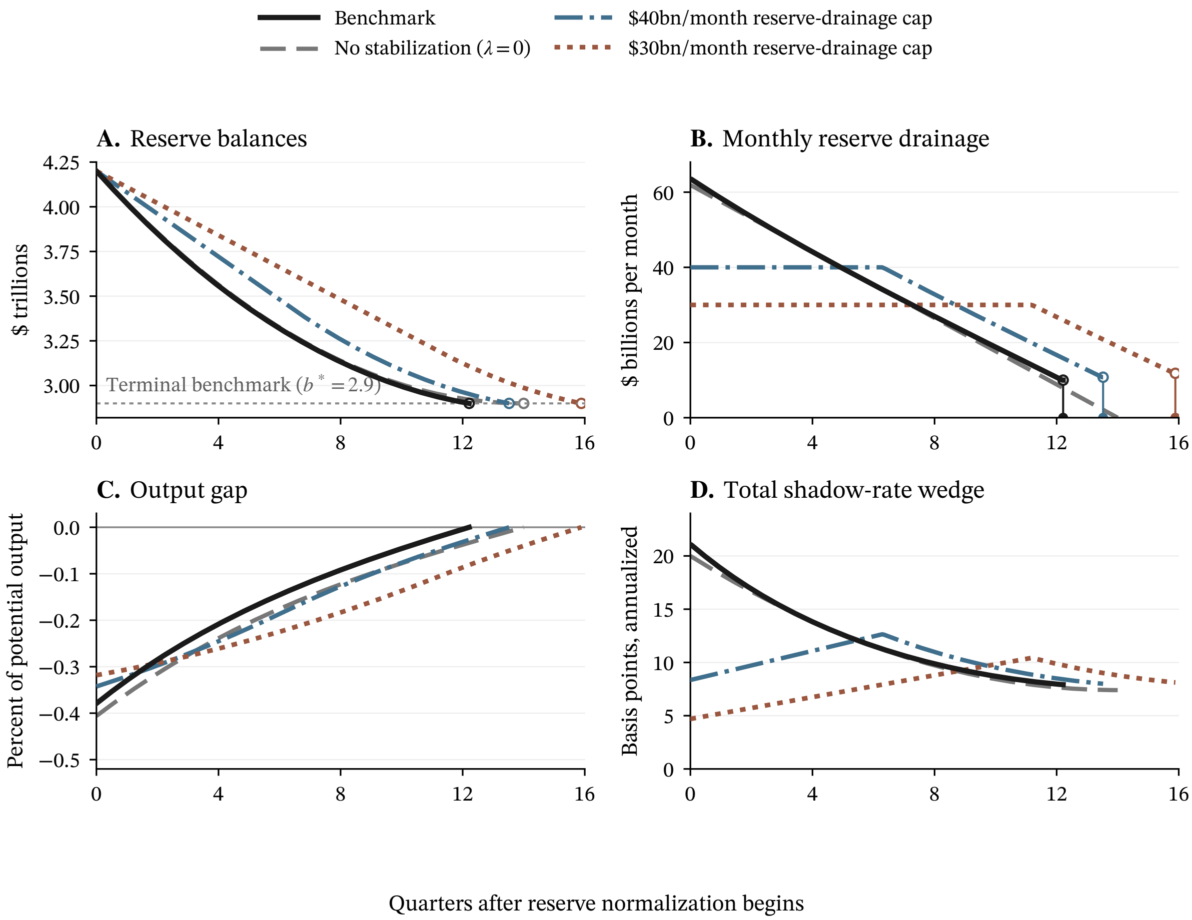

Figure 1: Optimal paths under the benchmark. Panels A–C compare the benchmark, the exact no-stabilization taper, and two reserve-drainage caps; panel D reports the total shadow-rate wedge. The benchmark uses a 7.4-basis-point stock target over the full reserve decline and a 5-basis-point speed target at the reference drainage rate. Open markers denote positive pre-exit values for policies that jump to zero at completion; the no-stabilization taper reaches zero continuously.

Table 2: Benchmark and reserve-drainage caps

Case	Horizon (quarters)	Average	Initial	Pre-exit	x_0 (percent)
		Reserve drainage ($bn/month)			
$\lambda = 0$ benchmark	14.00	31.0	61.9	0.0	-0.41
Benchmark	12.21	35.5	63.5	10.0	-0.38
Benchmark, $60bn/month cap	12.23	35.4	60.0	10.0	-0.38
Benchmark, $40bn/month cap	13.52	32.0	40.0	10.8	-0.34
Benchmark, $30bn/month cap	15.89	27.3	30.0	11.8	-0.32

Notes: The counterfactual begins at the September 2021 post-QE reserve peak. $b^* = 2.9$ trillion is a conditional terminal benchmark under the prevailing framework. Monthly quantities are declines in reserve balances generated by q_t, not gross SOMA runoff. The speed target is 5 basis points at $q_{\text{ref}} = 30.95$ billion per month; the benchmark optimum's initial speed wedge is about 21 basis points. The 3.5-year duration and $\lambda = 10{,}000$ are implementation and policy-loss normalizations. The no-stabilization row uses the exact continuous-time taper.

Figure 2 varies the composition of the QT wedge by comparing four cases: a mild reserve-scarcity scenario (terminal stock/speed wedges of 5/0 annualised basis points), the benchmark specification, a reserve-scarcity stress scenario (25/5 basis points), and a payment-flow stress scenario (0/20 basis points). A larger stock wedge brings completion forward. A stronger speed wedge produces a longer, smoother transition. See Table 4 in Appendix A.3 for corresponding values.

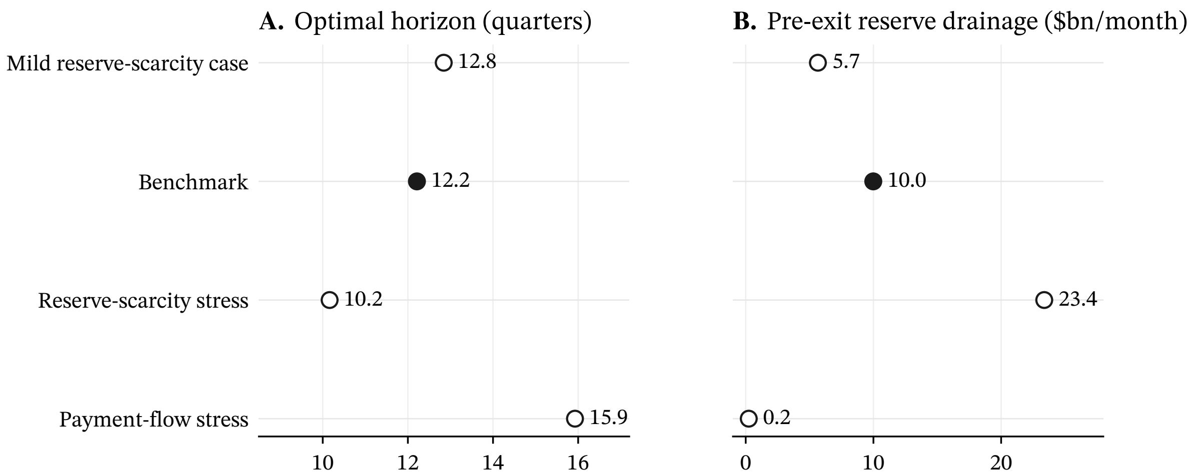

Figure 2: Wedge composition and policy timing. The corresponding terminal-stock/reference-speed targets are 5/0, 7.4/5, 25/5, and 0/20 annualized basis points. The filled marker denotes the benchmark.

Figure 3 shows when each channel matters. The speed wedge is largest at the outset, when drainage is most rapid. The stock wedge rises as reserves fall and matters increasingly near completion. Because the 5-basis-point target is evaluated at q_{ref} and optimized initial drainage exceeds that rate, the initial speed wedge is larger than the calibration target.

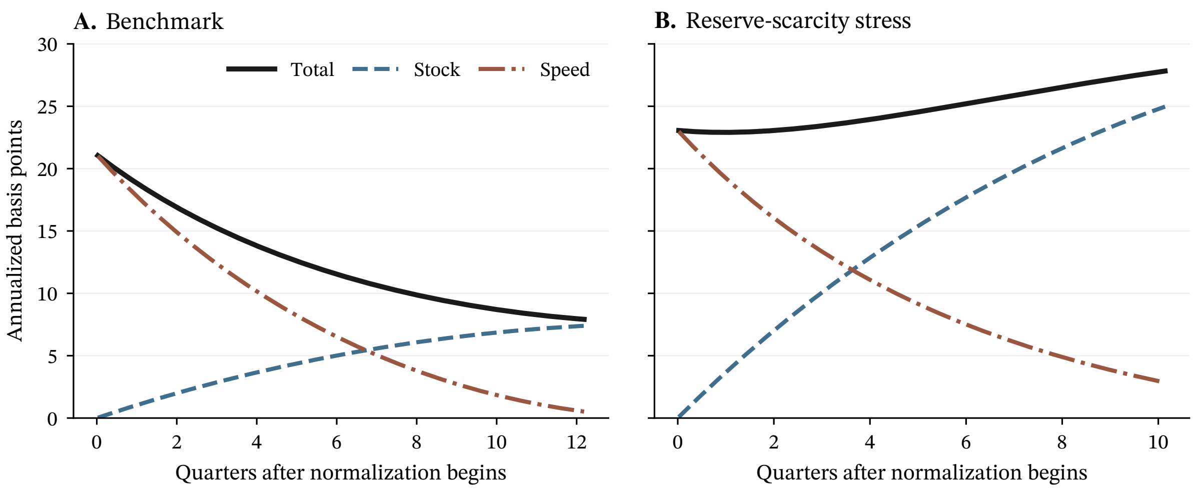

Figure 3: Stock and speed components of the shadow-rate wedge. The speed target is evaluated at the reference drainage rate rather than at the optimized initial drainage rate.

6 Conclusion

The transition to a given long-run operating regime can tighten financial conditions even when IOR is neutral at both the initial and terminal reserve stocks. As reserves decline, their scarcity value rises and private intermediation may become more constrained. Rapid drainage adds a temporary payment-flow wedge. Both wedges enter the shadow rate relevant for household expenditure, and current aggregate demand depends on their entire expected path. Designing QT is therefore a monetary-policy problem as well as an implementation problem.

The source of output losses determines the optimal transition. Without output stabilization, the central bank trades the cost of maintaining a larger balance sheet against the direct cost of rapid drainage. The solution is front-loaded and tapers linearly to zero. Once output stabilization matters, gradualism is no longer a general result. A stronger speed channel favors a longer, smoother transition. A stronger stock channel raises the cost of delay and can bring completion forward. Binding drainage caps limit front-loading and lengthen the transition.

In the counterfactual U.S. exercise, output stabilization shortens the benchmark transition from 14.00 to 12.21 quarters. The output gap at the announcement date becomes less negative, moving from -0.41 to -0.38 percent, because earlier completion reduces cumulative exposure to the stock wedge enough to offset faster average drainage. A reserve-scarcity stress advances completion further, while a payment-flow stress produces a longer, smoother path. Although the numerical results depend on the terminal reserve benchmark and the policy-loss normalization, the timing criterion is unchanged. Choosing the pace of QT requires weighing the cost of rapid drainage against the cost of prolonging the transition.

A Proofs and derivations

This appendix derives the analytical results in Sections 4.1–4.2 and reports additional quantitative sensitivity exercises. It solves the constant-speed problem and its comparative statics before deriving the necessary conditions for the unrestricted path.

A.1 Constant-speed benchmark

A.1.1 Output path and loss

Under the constant-speed path (22) and the linear-quadratic specification (21), the stock and speed wedges are

$$\nu(b_t) - \nu(b_0) = \kappa \frac{\Delta}{T} t, \qquad \Theta(q_t) = \theta \frac{\Delta^2}{T^2}. \tag{36}$$

Substituting these expressions into the output-gap equation gives

$$\dot{x}_t = \sigma \left[\kappa \frac{\Delta}{T} t + \theta \frac{\Delta^2}{T^2} \right], \qquad x_T = 0. \tag{37}$$

Integrating backward from $x_T = 0$ gives

$$\begin{aligned} x_t &= -\sigma \int_t^T \left[\kappa \frac{\Delta}{T} s + \theta \frac{\Delta^2}{T^2} \right] \mathrm{d}s \\ &= -\sigma \left[\frac{\kappa \Delta}{2T} (T^2 - t^2) + \theta \frac{\Delta^2}{T^2} (T - t) \right]. \end{aligned} \tag{38}$$

Setting $t = 0$ yields (23). The balance-sheet and direct drainage components of the objective are

$$\int_0^T c_B (b_t - b^*) \mathrm{d}t = \frac{c_B \Delta}{2} T, \qquad \int_0^T \frac{\chi}{2} q^2 \mathrm{d}t = \frac{\chi \Delta^2}{2T}. \tag{39}$$

The stabilization component follows from

$$\int_0^T x_t^2 \mathrm{d}t = \sigma^2 \int_0^T \left[\frac{\kappa\Delta}{2T}(T^2 - t^2) + \theta\frac{\Delta^2}{T^2}(T - t)\right]^2 \mathrm{d}t. \tag{40}$$

Expanding the square and using

$$\int_0^T (T^2 - t^2)^2 \mathrm{d}t = \frac{8}{15}T^5, \quad \int_0^T (T^2 - t^2)(T - t)\mathrm{d}t = \frac{5}{12}T^4, \quad \int_0^T (T - t)^2 \mathrm{d}t = \frac{1}{3}T^3, \tag{41}$$

gives

$$\int_0^T x_t^2 \mathrm{d}t = \sigma^2 \left[\frac{2\kappa^2\Delta^2}{15}T^3 + \frac{5\kappa\theta\Delta^3}{12}T + \frac{\theta^2\Delta^4}{3T}\right]. \tag{42}$$

Combining the three components yields the constant-speed objective

$$J^{CS}(T) = \frac{c_B\Delta}{2}T + \frac{\chi\Delta^2}{2T} + \lambda\sigma^2 \left[\frac{\kappa^2\Delta^2}{15}T^3 + \frac{5\kappa\theta\Delta^3}{24}T + \frac{\theta^2\Delta^4}{6T}\right]. \tag{43}$$

A.1.2 Optimal horizon and comparative statics

Differentiating (43) gives the first-order condition

$$\frac{c_B\Delta}{2} - \frac{\chi\Delta^2}{2T^2} + \lambda\sigma^2 \left[\frac{\kappa^2\Delta^2}{5}T^2 + \frac{5\kappa\theta\Delta^3}{24} - \frac{\theta^2\Delta^4}{6T^2}\right] = 0. \tag{44}$$

Multiplying (44) by T^2 gives $BT^4 + AT^2 - C = 0$, where

$$A \equiv \frac{c_B\Delta}{2} + \lambda\sigma^2\frac{5\kappa\theta\Delta^3}{24}, \qquad B \equiv \lambda\sigma^2\frac{\kappa^2\Delta^2}{5}, \qquad C \equiv \frac{\chi\Delta^2}{2} + \lambda\sigma^2\frac{\theta^2\Delta^4}{6}. \tag{45}$$

Proof of Proposition 1. Let $y = T^2$. The first-order condition becomes $By^2 + Ay - C = 0$. Since $A > 0$, $C > 0$, and $B \geq 0$, this equation has exactly one positive root. For $B > 0$,

$$T^{CS,*} = \left(\frac{-A + \sqrt{A^2 + 4BC}}{2B}\right)^{1/2}, \tag{46}$$

whereas $T^{CS,*} = \sqrt{C/A}$ when $B = 0$. The objective diverges as $T \downarrow 0$ because $\chi > 0$ and as $T \to \infty$ because $c_B > 0$. The unique stationary point is therefore the global minimizer.

For the comparative statics, let $F(T)$ denote the left-hand side of (44). Its derivative with respect to T is strictly positive:

$$F_T = \frac{\chi\Delta^2}{T^3} + \lambda\sigma^2 \left[\frac{2\kappa^2\Delta^2}{5}T + \frac{\theta^2\Delta^4}{3T^3}\right] > 0.$$

Direct differentiation gives

$$F_{c_B} = \frac{\Delta}{2} > 0, \qquad F_\chi = -\frac{\Delta^2}{2T^2} < 0,$$
$$F_\kappa = \lambda\sigma^2 \left[\frac{2\kappa\Delta^2}{5}T^2 + \frac{5\theta\Delta^3}{24}\right], \qquad F_\theta = \lambda\sigma^2 \left[\frac{5\kappa\Delta^3}{24} - \frac{\theta\Delta^4}{3T^2}\right].$$

The implicit-function theorem gives the signs stated in the proposition. When $\lambda > 0$, a larger θ lengthens the horizon if

$$8\theta\Delta > 5\kappa(T^{CS,*})^2$$

and shortens it under the reverse inequality. When $\lambda = 0$, neither κ nor θ affects the constant-speed horizon. □

A.1.3 Single-channel cases

The single-channel cases isolate the two timing effects. When $\kappa = 0$, the speed wedge is the only macroeconomic channel, and (44) gives

$$T^{CS,*} = \sqrt{\frac{\chi\Delta + \frac{1}{3}\lambda\sigma^2\theta^2\Delta^3}{c_B}}. \tag{47}$$

When $\lambda > 0$, a larger speed wedge therefore lengthens the transition. If instead $\theta = 0$, the constant-speed objective becomes

$$J^{CS}(T) = \frac{c_B \Delta}{2} T + \frac{\chi \Delta^2}{2T} + \lambda \sigma^2 \frac{\kappa^2 \Delta^2}{15} T^3, \tag{48}$$

so stabilization penalizes longer exposure to reserve scarcity. With both channels present, the derivative of the macroeconomic component of the objective is

$$\frac{\partial J^{CS}_{\text{macro}}}{\partial T} = \lambda \sigma^2 \left[\frac{\kappa^2 \Delta^2}{5} T^2 + \frac{5 \kappa \theta \Delta^3}{24} - \frac{\theta^2 \Delta^4}{6T^2} \right]. \tag{49}$$

The stock and interaction terms increase the marginal cost of a longer horizon, whereas the speed term reduces it.

A.2 Flexible-speed policy

A.2.1 Necessary conditions

Let $\mu_{b,t}$ and $\mu_{x,t}$ be the current-value costates on reserve balances and the output gap. The Hamiltonian is

$$\begin{aligned} \mathcal{H}_t = c_B (b_t - b^*) + \frac{\chi}{2} q_t^2 + \frac{\lambda}{2} x_t^2 - \mu_{b,t} q_t \\ + \mu_{x,t} \sigma \left[\nu(b_t) - \nu(b_0) + \Theta(q_t) \right]. \end{aligned} \tag{50}$$

The nonnegativity constraint on drainage gives the Kuhn–Tucker conditions

$$q_t \geq 0, \qquad \chi q_t - \mu_{b,t} + \sigma \mu_{x,t} \Theta'(q_t) \geq 0, \qquad q_t \left[\chi q_t - \mu_{b,t} + \sigma \mu_{x,t} \Theta'(q_t) \right] = 0.$$

On an interior arc, they reduce to

$$\mu_{b,t} = \chi q_t + \sigma \mu_{x,t} \Theta'(q_t). \tag{51}$$

Differentiating the Hamiltonian with respect to the states gives (25). The reserve stock is fixed at both endpoints, while the output gap is fixed only at T. Because x_0 is free, its natural boundary condition is

$$\mu_{x,0} = 0. \tag{52}$$

Because b_T and x_T are fixed, their terminal costates are unrestricted. The free terminal date adds the condition $\mathcal{H}_{T^-} = 0$, where terminal controls and costates are understood as their pre-exit left limits. Substituting $\Theta'(q) = 2\theta q$ into the Kuhn–Tucker conditions yields (24).

A.2.2 Proof of Proposition 2

Proof. When $\lambda = 0$, the output-gap block does not affect the objective and can be omitted. The problem reduces to

$$\min_{T,\{q_t\}} \int_0^T \left[c_B(b_t - b^*) + \frac{\chi}{2} q_t^2 \right] \mathrm{d}t \tag{53}$$

subject to $\dot{b}_t = -q_t$, $b(0) = b_0$, and $b(T) = b^*$. The Hamiltonian is

$$\mathcal{H}_t^0 = c_B(b_t - b^*) + \frac{\chi}{2} q_t^2 - \mu_t q_t. \tag{54}$$

The first-order condition and costate equation are

$$\chi q_t = \mu_t, \qquad \dot{\mu}_t = -c_B. \tag{55}$$

Since the terminal date is free and $b_T = b^*$ is fixed, $\mathcal{H}^0_{T^-} = 0$. Evaluating this condition at the terminal reserve stock gives

$$0 = \mathcal{H}^0_{T^-} = \frac{\chi}{2} q^2_{T^-} - \mu_T q_{T^-}. \tag{56}$$

Together with the first-order condition $\mu_T = \chi q_{T^-}$, this implies

$$q_{T^-} = 0, \qquad \mu_T = 0. \tag{57}$$

The costate and control paths are therefore

$$\mu_t = c_B(T - t), \qquad q_t = \frac{c_B}{\chi}(T - t). \tag{58}$$

Imposing total drainage,

$$\Delta = \int_0^T q_t \mathrm{d}t = \frac{c_B}{2\chi} T^2. \tag{59}$$

Solving this equation for T and substituting into (58) yields (26). □

A.2.3 Proof of Proposition 3

Proof. By (17), a nonnegative QT wedge implies $x_t \leq 0$ for $t < T$. The output-gap costate therefore satisfies $\dot{\mu}_{x,t} = -\lambda x_t \geq 0$. Since $\mu_{x,0} = 0$, it follows that $\mu_{x,t} \geq 0$. Equation (24) then establishes (27) on any interior arc.

Using $\nu'(b_t) < 0$, the reserve costate can be written as

$$\dot{\mu}_{b,t} = -c_B + \sigma \mu_{x,t} |\nu'(b_t)|. \tag{60}$$

For a given terminal value $\mu_{b,T}$, integration gives

$$\mu_{b,t} = \mu_{b,T} + \int_t^T \left[c_B - \sigma \mu_{x,s} |\nu'(b_s)| \right] \mathrm{d}s. \tag{61}$$

Thus, stabilization also changes the shadow benefit in the numerator of the drainage rule. Since both the terminal costate and the terminal date are endogenous, the overall response of q_t and T cannot be signed in general. □

Allowing deposit demand to respond to output leaves the conditional speed effect unchanged. Appendix B derives the modified output-gap costate and shows that $x_t \leq 0$ still implies $\mu_{x,t} \geq 0$.

A.2.4 Terminal speed

Let $\Delta\nu = \nu(b^*) - \nu(b_0) > 0$, and suppose the pre-exit drainage rate q_{T^-} is positive and unconstrained. Since $b_{T^-} = b^*$ and $x_T = 0$, the free-terminal-time condition is

$$0 = \frac{\chi}{2} q_{T^-}^2 - \mu_{b,T} q_{T^-} + \sigma \mu_{x,T} \left(\Delta\nu + \theta q_{T^-}^2 \right).$$

The interior first-order condition gives

$$\mu_{b,T} = \left(\chi + 2\sigma\theta\mu_{x,T} \right) q_{T^-}.$$

Substitution into the free-terminal-time condition yields

$$q_{T^-}^2 = \frac{\sigma \mu_{x,T} \Delta\nu}{\chi/2 + \sigma\theta\mu_{x,T}}. \tag{62}$$

For $\theta \geq 0$, (62) implies a positive pre-exit drainage rate whenever $\mu_{x,T} > 0$. Although x_T is fixed at zero, its terminal costate is unrestricted and can be positive. The control then jumps to zero when the terminal regime begins. A smooth-landing constraint would rule out the jump and replace the terminal condition above.

A.3 Quantitative sensitivity

Table 3 varies the no-stabilization horizon T_0, which determines χ/c_B, and the output-loss weight λ. A larger output-loss weight shortens the benchmark transition across these implementation normalizations because the stock channel dominates the speed channel. The effect is stronger under the reserve-scarcity stress. Table 4 reports the stock and speed scenarios plotted in Figure 2.

Table 3: Sensitivity to implementation and output-loss weights

Scenario	T_0 (years)	λ	Horizon (quarters)	Average drainage ($bn/month)	x_0 (percent)
Benchmark	3.0	0	12.00	36.1	-0.35
	3.0	2,500	11.21	38.7	-0.33
	3.0	10,000	10.63	40.8	-0.33
	3.0	40,000	9.82	44.1	-0.32
	3.5	0	14.00	31.0	-0.41
	3.5	2,500	12.95	33.5	-0.39
	3.5	10,000	12.21	35.5	-0.38
	3.5	40,000	11.22	38.6	-0.37
	4.0	0	16.00	27.1	-0.46
	4.0	2,500	14.66	29.5	-0.44
	4.0	10,000	13.76	31.5	-0.43
	4.0	40,000	12.58	34.4	-0.43
	4.5	0	18.00	24.1	-0.52
	4.5	2,500	16.35	26.5	-0.50
	4.5	10,000	15.27	28.4	-0.48
	4.5	40,000	13.91	31.2	-0.48
Scarcity stress	3.0	0	12.00	36.1	-0.70
	3.0	2,500	10.10	42.9	-0.60
	3.0	10,000	8.99	48.2	-0.55
	3.0	40,000	7.63	56.8	-0.52
	3.5	0	14.00	31.0	-0.82
	3.5	2,500	11.52	37.6	-0.68
	3.5	10,000	10.16	42.6	-0.63
	3.5	40,000	8.53	50.8	-0.60
	4.0	0	16.00	27.1	-0.93
	4.0	2,500	12.90	33.6	-0.77
	4.0	10,000	11.27	38.4	-0.71
	4.0	40,000	9.36	46.3	-0.68
	4.5	0	18.00	24.1	-1.05
	4.5	2,500	14.23	30.4	-0.85
	4.5	10,000	12.33	35.1	-0.79
	4.5	40,000	10.15	42.7	-0.77

Notes: T_0 changes $\chi/c_B = T_0^2/(2\Delta)$. These are policy-loss and implementation sensitivities, not estimated preference parameters.

Table 4: Illustrative wedge compositions and optimal transition paths

Scenario	Stock Target (bp)	Speed	Horizon (quarters)	Average Drainage ($bn/month)	Pre-exit	x_0 (percent)	Initial speed Wedge (bp)	Maximum total
Mild reserve-scarcity case	5	0	12.84	33.7	5.7	-0.10	0.0	5.0
Benchmark	7.4	5	12.21	35.5	10.0	-0.38	21.0	21.1
Reserve-scarcity stress	25	5	10.16	42.6	23.4	-0.63	23.0	27.9
Payment-flow stress	0	20	15.93	27.2	0.2	-0.89	102.8	102.8

Notes: Stock targets are evaluated over the full reserve decline. Speed targets are evaluated at the reference reserve-drainage rate of 30.95 billion per month. The scenarios are conditional comparisons rather than jointly estimated structural calibrations.

Replacing the linear 7.4-basis-point stock wedge with the estimated semi-log schedule changes the optimal horizon from 12.21 to 12.24 quarters and moves the announcement-date output gap from -0.38 to -0.37 percent. The small changes show that the linear approximation does not drive the benchmark result.

Table 5: Robustness to the empirical semi-log stock schedule

Stock schedule	Horizon (quarters)	Average Reserve drainage ($bn/month)	Initial	Pre-exit	x_0 (percent)
Linear 7.4-bp benchmark	12.21	35.5	63.5	10.0	-0.38
Empirical semi-log schedule	12.24	35.4	63.4	9.9	-0.37

Notes: The semi-log stock wedge is $0.002 \log(b_0/b_t)$ in annual decimal-rate units. Both rows use the same 5-basis-point reference-speed target, reserve endpoints, and policy-loss normalizations.

B A banking foundation for the QT wedge

This appendix derives a banking foundation for the local specification

$$\nu(b_t) - \nu(b_0) = \kappa(b_0 - b_t), \qquad \Theta(q_t) = \theta q_t^2,$$

used in the main text. The mechanism combines deposit demand with a collateral constraint in a floor-system banking model. Households value deposits for their liquidity services, and banks issue deposits against a quality-weighted collateral pool in which reserves receive the highest weight. A binding constraint transmits reserve scarcity to deposit and money-market spreads.

In Bianchi and Bigio (2022), deposit transfers ultimately clear in reserves. That settlement motive explains why the constraint depends on both the reserve stock and payment-flow risk.

The wedge is a theoretical counterpart to the nonlinear reserve-demand evidence in Afonso et al. (2022) and López-Salido and Vissing-Jorgensen (2023). A local expansion around the pre-transition regime delivers the stock coefficient κ and speed coefficient θ. We also distinguish reserve drainage from gross asset runoff and derive the output feedback omitted from the analytical baseline.

Layered payments. Households use bank deposits for payments, while banks settle in central-bank reserves. Banks hold reserves b_t and other assets a_t against deposits d_t. Reserves therefore affect household choices through the supply of bank-issued money, not through direct household holdings. Reserve drainage can tighten household financial conditions when banks' settlement capacity binds even if it leaves household portfolios unchanged (Piazzesi, Rogers, and Schneider, 2022; Bianchi and Bigio, 2022).

Household liquidity demand. Let ξ_t denote the current-value marginal utility of wealth. The Hamiltonian associated with (5) and (6) is

$$\mathcal{H}_t^H = u(C_t) + v(d_t) + \xi_t \left[(i_t^S - \pi_t)(W_t - d_t) + (i_t^D - \pi_t)d_t + Y_t - C_t - \tau_t \right]. \tag{63}$$

An interior solution satisfies

$$u'(C_t) = \xi_t, \qquad v'(d_t) = \xi_t(i_t^S - i_t^D), \tag{64}$$

and

$$\frac{\dot{\xi}_t}{\xi_t} = \rho - (i_t^S - \pi_t), \tag{65}$$

$$\lim_{t\to\infty} e^{-\rho t}\xi_t W_t = 0. \tag{66}$$

These conditions yield the deposit-demand and Euler equations in Section 2.2.

In particular, the liquidity spread on deposits satisfies

$$\omega_t \equiv i_t^S - i_t^D = \frac{v'(d_t)}{u'(C_t)}. \tag{67}$$

Let $u'(C_t) = C_t^{-1/\sigma}$ and $v'(d_t) = \Omega^{1/\eta_m} d_t^{-1/\eta_m}$. Equation (67) then yields

$$d_t = \Omega\, C_t^{\eta_m/\sigma}\, \omega_t^{-\eta_m}, \qquad \eta_m > 0, \qquad \text{equivalently} \qquad \omega_t = \left(\frac{\Omega C_t^{\eta_m/\sigma}}{d_t}\right)^{1/\eta_m}. \tag{68}$$

When $\eta_m = \sigma$, deposit demand is proportional to consumption. More generally, the separable specification preserves the Euler equation used in the main text while allowing the money-demand and intertemporal elasticities to differ. Piazzesi, Rogers, and Schneider (2022) obtain related liquidity demand from a richer CES aggregator; Nagel (2016) and Lenel, Piazzesi, and Schneider (2019) study convenience yields on near-money claims. When bank balance-sheet capacity contracts, deposits become scarce relative to consumption and ω_t rises. This is the deposit-market counterpart to the spreads emphasized by Drechsler, Savov, and Schnabl (2017).

Bank collateral capacity. Banks support deposits with reserves b_t and other assets a_t. We normalize the collateral weight on reserves to one and give other assets a weight $\rho_A \in (0, 1)$ because their settlement and liquidity properties are weaker. Banks must retain enough capacity to meet a stressed deposit outflow. Let $\Lambda_t \geq 0$ denote payment-flow risk and $\ell \in (0, 1]$ denote leverage capacity. Deposit issuance must satisfy

$$[1 + \Lambda_t(1 - \ell)]\, d_t \le \ell\,(b_t + \rho_A a_t)\,, \tag{69}$$

and when the constraint binds,

$$d_t = \mathcal{L}_t\,(b_t + \rho_A a_t)\,, \qquad \mathcal{L}_t \equiv \frac{\ell}{1 + \Lambda_t(1 - \ell)}. \tag{70}$$

Deposit capacity rises with ℓ and ρ_A and falls with Λ_t, strictly so for the latter when $\ell < 1$. A decline in reserves also reduces the quantity of deposits that banks can support. This stock channel tightens (69) and raises the shadow value of another unit of b_t.

Drainage speed can tighten collateral capacity through settlement pressure, intraday liquidity needs, and greater reliance on money-market intermediation. These forces are documented by Copeland, Duffie, and Yang (2025) and emphasized by Duffie (2026). We model them as

$$\Lambda_t = \Lambda_0 + \omega_\Lambda q_t^2, \qquad \omega_\Lambda \ge 0. \tag{71}$$

Rapid drainage may also reduce the effective collateral quality of other assets: $\rho_A(q_t) = \rho_A - \omega_\rho q_t^2$, with $\omega_\rho \ge 0$. This local specification requires $\rho_A - \omega_\rho q_t^2 \ge 0$ along the path. The proposition below allows both mechanisms. Either one raises the interest-rate wedge for a given reserve stock and generates the speed term $\Theta(q)$ used in the main text.

The reserve-collateral spread. We summarize the relation between the deposit liquidity spread and the reserve-collateral spread as

$$\nu_t \equiv i_t^S - i_t^M = \psi\, \omega_t, \qquad \psi > 0. \tag{72}$$

Note that this relation is not causal, as both spreads inherit the shadow value of the binding collateral constraint (Lenel, Piazzesi, and Schneider, 2019). The results require only positive pass-through, and ψ is absorbed into the baseline wedge ν_0. Combining (68)–(72) gives

$$\nu_t = \psi \left(\frac{\Omega C_t^{\eta_m/\sigma}}{\mathcal{L}_t \left(b_t + \rho_A a_t \right)} \right)^{1/\eta_m} . \tag{73}$$

The wedge rises as collateral-supported deposit capacity falls relative to consumption. Because drainage speed affects both $\mathcal{L}_t$ and effective collateral quality, (73) is generally not separable in b_t and q_t. The additive stock-and-speed specification in the main text is a local approximation around the pre-QT regime.

Local approximation. Holding consumption, nonreserve assets, and payment risk at their initial values isolates the stock channel. Equation (73) then becomes

$$\nu(b_t) = \nu_0 \left(\frac{b_0 + \rho_A a_0}{b_t + \rho_A a_0} \right)^{1/\eta_m}, \qquad \nu'(b) < 0, \qquad \nu''(b) > 0, \tag{74}$$

which satisfies (10) and has the convexity found in empirical reserve-demand schedules such as Afonso et al. (2022). The next proposition expresses the local coefficients (κ, θ) as functions of primitive parameters in the banking block, and gives the output-feedback term ϕx_t omitted from the analytical baseline.

Proposition 4 (Banking foundation for the QT wedge)**.** *Consider the floor-system banking block in which households hold deposits d_t and banks hold reserves b_t and other collateral a_t. Hold the stock of other collateral fixed at $a_t = a_0$ locally, apart from its speed-dependent collateral weight, and normalize natural consumption to be locally constant so that $\widehat{C}_t = x_t$. Deposits satisfy the binding collateral constraint*

$$d_t = \frac{\ell}{1 + (\Lambda_0 + \omega_\Lambda q_t^2)(1 - \ell)} \left[b_t + (\rho_A - \omega_\rho q_t^2) a_t \right],$$

with $a_0 \geq 0$, $0 < \ell \leq 1$, $0 < \rho_A < 1$, $\omega_\Lambda \geq 0$, *and* $\omega_\rho \geq 0$. *Household money demand satisfies* (68) *and pass-through satisfies* (72). *Then, around* $(C_0, b_0, a_0, q = 0)$,

$$\nu_t - \nu_0 = \phi x_t + \kappa(b_0 - b_t) + \theta q_t^2 + o\left(|x_t| + |b_t - b_0| + q_t^2\right),$$

where

$$\phi = \frac{\nu_0}{\sigma} > 0, \qquad \kappa = \frac{\nu_0}{\eta_m} \frac{1}{b_0 + \rho_A a_0} > 0$$

if other collateral is held fixed, and

$$\theta = \frac{\nu_0}{\eta_m} \left[\frac{(1-\ell)\omega_\Lambda}{1 + \Lambda_0(1-\ell)} + \frac{\omega_\rho a_0}{b_0 + \rho_A a_0} \right] \geq 0$$

with strict inequality if and only if $(1-\ell)\omega_\Lambda > 0$ *or* $a_0\omega_\rho > 0$. *Reserve drainage therefore raises the Euler-relevant shadow rate through a stock-scarcity channel and, under either strict-positivity condition, a payment-flow or collateral-quality speed channel. The baseline model sets* $\phi = 0$ *to isolate the reserve-supply channel.*

Proof. Let

$$Z_t(q_t) = b_t + (\rho_A - \omega_\rho q_t^2) a_t, \qquad \mathcal{L}_t(q_t) = \frac{\ell}{1 + (\Lambda_0 + \omega_\Lambda q_t^2)(1-\ell)},$$

so that the binding constraint gives $d_t = \mathcal{L}_t(q_t) Z_t(q_t)$. Money demand (68) and pass-through (72) then imply

$$\nu_t = \psi \left(\frac{\Omega C_t^{\eta_m/\sigma}}{\mathcal{L}_t(q_t) Z_t(q_t)} \right)^{1/\eta_m}.$$

A first-order expansion in logs around $(C_0, b_0, a_0, q = 0)$ gives

$$\frac{\nu_t - \nu_0}{\nu_0} = \frac{1}{\sigma}\widehat{C}_t - \frac{1}{\eta_m}\left[\widehat{\mathcal{L}}_t + \widehat{Z}_t\right] + o\left(|\widehat{C}_t| + |\widehat{Z}_t| + q_t^2\right),$$

where hats denote log deviations from the baseline. Under the local normalization that natural consumption is constant, $\widehat{C}_t = x_t$. More generally, $\widehat{C}_t = x_t + \widehat{C}^n_t$, which adds an exogenous natural-consumption term. Direct differentiation yields

$$-\widehat{Z}_t = \frac{b_0 - b_t}{b_0 + \rho_A a_0} + \frac{\omega_\rho a_0}{b_0 + \rho_A a_0} q_t^2 + o\left(|b_t - b_0| + q_t^2\right),$$

and

$$-\widehat{\mathcal{L}}_t = \frac{(1-\ell)\omega_\Lambda}{1 + \Lambda_0(1-\ell)} q_t^2 + o\left(q_t^2\right).$$

Substituting these expressions and multiplying by v_0 yields the stated coefficients. □

Asset runoff, reserve drainage, and collateral incidence. The policy object in this paper is reserve drainage q_t, but gross asset runoff need not reduce reserves one-for-one. Changes in the overnight reverse-repurchase facility, the Treasury General Account, currency, and other liabilities can absorb or amplify the reserve effect of QT (Anderson et al., 2026; Du, Forbes, and Luzzetti, 2024). Let A_t^{CB} denote central-bank assets and let o_t collect all liabilities other than reserves:

$$o_t \equiv \mathrm{ONRRP}_t + \mathrm{TGA}_t + \mathrm{Currency}_t + \mathrm{Other}_t. \tag{75}$$

Total central-bank liabilities are therefore $b_t + o_t$. Holding central-bank capital $\bar{k}^{CB}$ fixed, the balance-sheet identity is

$$A_t^{CB} = b_t + o_t + \bar{k}^{CB}. \tag{76}$$

Define gross asset runoff and reserve drainage, respectively, as

$$Q_t \equiv -\dot{A}_t^{CB}, \qquad q_t \equiv -\dot{b}_t.$$

Differentiating (76) gives

$$q_t = Q_t + \dot{o}_t. \tag{77}$$

Gross asset runoff drains reserves one-for-one if nonreserve liabilities are unchanged. If these liabilities decline, they absorb part of the balance-sheet contraction and $q_t < Q_t$. For example, money-market funds may finance Treasury purchases by reducing ON RRP balances, so ON RRP rather than reserves bears part of the adjustment. An increase in the Treasury's account or another nonreserve liability can instead make reserve drainage exceed contemporaneous asset runoff.

When $Q_t > 0$ and $0 \leq q_t \leq Q_t$, define

$$\beta_t \equiv \frac{q_t}{Q_t} \in [0, 1], \qquad q_t = \beta_t Q_t. \tag{78}$$

Then equation (77) implies $\beta_t = 1 + \dot{o}_t / Q_t$. Hence $\beta_t = 1$ when other liabilities are unchanged and $\beta_t < 1$ when they decline. The representation $q_t = \beta_t Q_t$ is a convenient special case, not a general accounting identity: higher nonreserve liabilities can make reserve drainage exceed gross asset runoff.

This identity concerns the liability side of the central-bank balance sheet, whereas asset incidence raises a separate question: who ultimately holds the securities leaving the central bank? Conditional on a reserve decline, let a_t^B denote bank-held nonreserve collateral, with initial value a_0, and let $\alpha \in [0, 1]$ denote the share of the associated securities entering the bank collateral pool. For constant α,

$$\mathrm{d}a_t^B = -\alpha \, \mathrm{d}b_t, \qquad a_t^B = a_0 + \alpha(b_0 - b_t).$$

The case of $\alpha = 1$ corresponds to full bank absorption, while $\alpha = 0$ corresponds to nonbank absorption: banks lose reserves without acquiring replacement collateral. Intermediate values describe mixed absorption.

Writing $Z_0 = b_0 + \rho_A a_0$, quality-weighted bank collateral is

$$Z_t = b_t + \rho_A a_t^B = Z_0 + (1 - \alpha\rho_A)(b_t - b_0),$$

and the local stock coefficient becomes

$$\kappa_\alpha = \frac{\nu_0}{\eta_m} \frac{1 - \alpha\rho_A}{b_0 + \rho_A a_0}.$$

Since $\alpha\rho_A < 1$ under the maintained parameter restrictions, the stock channel remains contractionary. Bank absorption replaces some lost reserve collateral but cannot replace all of it because nonreserve assets carry the lower weight $\rho_A < 1$.

For the corresponding stock-only schedule,

$$Z_t^\alpha = Z_0 + (1 - \alpha\rho_A)(b_t - b_0), \qquad \nu^\alpha(b_t) = \nu_0 \left(\frac{Z_0}{Z_t^\alpha}\right)^{1/\eta_m}. \tag{79}$$

Its local expansion around b_0 yields κ_α. The parameters β_t and α describe different incidence margins. The first determines how much gross asset runoff passes through to reserves. The second determines how much replacement collateral reaches bank balance sheets conditional on that reserve decline. Treasury purchases financed by lower ON RRP balances imply a relatively low β_t, while purchases by nonbanks financed from bank deposits imply substantial reserve drainage but a low α. Dealer and repo constraints may further reduce intermediation capacity as Treasuries leave the central bank's portfolio (Anderson et al., 2026); in the reduced-form model, these effects appear as a stronger stock wedge. The main analysis conditions on q_t and leaves both incidence margins in the background. Money-market-fund portfolio adjustment may also affect deposit demand, but a full treatment lies outside the present banking block.

Output feedback. Money demand creates an output feedback that the analytical baseline sets to zero. With locally constant natural consumption, a negative output gap reduces deposit demand and dampens the reserve-collateral spread through ϕx_t, where $\phi = \nu_0/\sigma > 0$. Define

$$h_t \equiv \kappa(b_0 - b_t) + \theta q_t^2.$$

The full local output-gap equation and its terminal condition are

$$\dot{x}_t = \sigma(\phi x_t + h_t), \qquad x_T = 0.$$

Multiplying the differential equation by $e^{-\sigma\phi t}$ and integrating backward from T gives

$$x_t = -\sigma \int_t^T e^{-\sigma\phi(s-t)} h_s \, \mathrm{d}s. \tag{80}$$

The feedback discounts more distant wedges. Since the kernel is positive, $h_s \geq 0$ on $[t, T]$ still implies $x_t \leq 0$.

For the flexible-speed problem, the full Hamiltonian is

$$\mathcal{H}_t^{\phi} = c_B(b_t - b^*) + \frac{\chi}{2} q_t^2 + \frac{\lambda}{2} x_t^2 + \mu_{b,t}(-q_t) + \mu_{x,t}\sigma\left[\phi x_t + h_t\right].$$

The output-gap costate therefore satisfies

$$\dot{\mu}_{x,t} = -\frac{\partial \mathcal{H}_t^{\phi}}{\partial x_t} = -\lambda x_t - \sigma\phi\mu_{x,t}.$$

Because x_0 remains free, its natural boundary condition is $\mu_{x,0} = 0$. Solving the costate equation forward gives

$$\mu_{x,t} = \lambda \int_0^t e^{-\sigma\phi(t-s)}(-x_s) \, \mathrm{d}s.$$

Hence $x_s \leq 0$ implies $\mu_{x,t} \geq 0$. Since ϕx_t does not depend directly on the control, the constrained drainage rule remains

$$q_t = \max\left\{0, \frac{\mu_{b,t}}{\chi + 2\sigma\theta\mu_{x,t}}\right\}.$$

The denominator is at least χ, so the conditional speed effect in Proposition 3 is unchanged. If natural consumption varies, the local expansion also contains the exogenous term $\phi\widehat{C}_t^n$.

The baseline sets $\phi = 0$ to isolate the reserve-supply channel and retain the polynomial closed forms in the analytical sections. The optimal-control problem remains well defined for $\phi > 0$. Numerically, this case replaces the linear within-interval recursion with the exact exponential backward map in (80), after which Gaussian quadrature evaluates the squared output-gap path.

A conditional terminal benchmark. The collateral constraint supplies one determinant of the terminal target b^*. It complements the payment-system floor in Duffie (2026) and the operating-framework considerations in Reis (2026) and Anderson et al. (2026). Let d^* denote the quantity of inside money the banking system must support in the terminal regime under normal payment-flow risk Λ_0, and let a^* denote other collateral in that regime. Conditional on these primitives, the minimum nonnegative reserve stock capable of supporting d^* is

$$b^* = \max\left\{0, \frac{1 + \Lambda_0(1-\ell)}{\ell}\, d^* - \rho_A a^*\right\}. \tag{81}$$

When the expression inside the maximum is positive, the collateral constraint binds. In this region, b^* rises with required deposit services d^* and, for $\ell < 1$, with payment-flow risk Λ_0. It falls with leverage capacity ℓ, with collateral quality ρ_A when $a^* > 0$, and with the stock of nonreserve collateral a^* when $\rho_A > 0$. Payment technology, regulation, supervision, liquidity facilities, and the choice between floor and corridor systems can shift these objects materially (Arce et al., 2020; Cúrdia and Woodford, 2011). We therefore treat \$2.9 trillion as a conditional benchmark under the prevailing framework, not as a unique technological reserve requirement.

Direct speed and balance-sheet costs. The model contains two distinct costs of rapid drainage. Derived from the banking block above, $\Theta(q_t)$ affects expenditure through the household shadow rate. The residual term $\chi q_t^2/2$ captures implementation and financial-stability costs not already reflected in that wedge, including operational frictions or market-functioning strains that do not pass one-for-one into $i_t^S - i_t^M$. The term $c_B(b_t - b^*)$ is likewise a reduced-form fiscal, political, or institutional cost of maintaining a larger balance sheet rather than an implication of the banking block. We interpret χ net of the effects in Θ to avoid double counting.

References

Afonso, Gara, Domenico Giannone, Gabriele La Spada, and John C. Williams. 2022. "Scarce, Abundant, or Ample? A Time-Varying Model of the Reserve Demand Curve." Federal Reserve Bank of New York Staff Report No. 1019, revised November 2025.

Anderson, Alyssa G., Alessandro Barbarino, Anthony M. Diercks, and Stephen Miran. 2026. "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.

Arce, Óscar, Galo Nuño, Dominik Thaler, and Carlos Thomas. 2020. "A Large Central Bank Balance Sheet? Floor vs Corridor Systems in a New Keynesian Environment." *Journal of Monetary Economics* 114: 350–367.

Bansal, Ravi, and Wilbur John Coleman. 1996. "A Monetary Explanation of the Equity Premium, Term Premium, and Risk-Free Rate Puzzles." *Journal of Political Economy* 104(6): 1135–1171.

Benigno, Gianluca, and Pierpaolo Benigno. 2026. "Managing Monetary Policy Normalization." *Economic Journal*, advance article, ueag085.

Bianchi, Javier, and Saki Bigio. 2022. "Banks, Liquidity Management, and Monetary Policy." *Econometrica* 90(1): 391–454.

Bigio, Saki, Pablo D'Erasmo, and Luis Yépez. 2026. "Monetary Policy Implementation Regimes and Reserve Demand in the US." Working paper.

Board of Governors of the Federal Reserve System. 2026. "Factors Affecting Reserve Balances of Depository Institutions and Condition Statement of Federal Reserve Banks." H.4.1 statistical release.

Copeland, Adam, Darrell Duffie, and Yilin (David) Yang. 2025. "Reserves Were Not So Ample After All." *Quarterly Journal of Economics* 140(1): 239–281.

Cúrdia, Vasco, and Michael Woodford. 2011. "The Central-Bank Balance Sheet as an Instrument of Monetary Policy." *Journal of Monetary Economics* 58(1): 54–79.

Dordal i Carreras, Marc, and Seung Joo Lee. 2024. "Higher-Order Forward Guidance." Working paper.

Drechsler, Itamar, Alexi Savov, and Philipp Schnabl. 2017. "The Deposits Channel of Monetary Policy." *Quarterly Journal of Economics* 132(4): 1819–1876.

Du, Wenxin, Kristin Forbes, and Matthew N. Luzzetti. 2024. "Quantitative Tightening Around the Globe: What Have We Learned?" NBER Working Paper No. 32321.

Duffie, Darrell. 2026. "The Payment System Puts a Floor on the Fed's Balance Sheet." *Brookings Papers on Economic Activity*, Spring 2026 Conference Draft.

Galí, Jordi. 2015. *Monetary Policy, Inflation, and the Business Cycle*. Second edition. Princeton University Press.

Harrison, Richard. 2024. "Optimal Quantitative Easing and Tightening." Bank of England Staff Working Paper No. 1063.

Karadi, Peter, and Anton Nakov. 2021. "Effectiveness and Addictiveness of Quantitative Easing." *Journal of Monetary Economics* 117: 1096–1117.

Krishnamurthy, Arvind, and Annette Vissing-Jorgensen. 2012. "The Aggregate Demand for Treasury Debt." *Journal of Political Economy* 120(2): 233–267.

Lenel, Moritz, Monika Piazzesi, and Martin Schneider. 2019. "The Short Rate Disconnect in a Monetary Economy." Working paper.

López-Salido, David, and Annette Vissing-Jorgensen. 2023. "Reserve Demand, Interest Rate Control, and Quantitative Tightening." Working paper.

Nagel, Stefan. 2016. "The Liquidity Premium of Near-Money Assets." *Quarterly Journal of Economics* 131(4): 1927–1971.

Piazzesi, Monika, Ciaran Rogers, and Martin Schneider. 2022. "Money and Banking in a New Keynesian Model." Working paper.

Reis, Ricardo. 2026. "Payment Needs and the Size of the Federal Reserve's Balance Sheet." Discussion of Duffie (2026), *Brookings Papers on Economic Activity*, Spring 2026.

Werning, Iván. 2012. "Managing a Liquidity Trap: Monetary and Fiscal Policy." Working paper.

Woodford, Michael. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press.