1-2015

Mechanism Design for Near Real-Time Retail Payment and Settlement Systems

Zhiling GUO
Singapore Management University, ZHILINGGUO@smu.edu.sg

Robert John KAUFFMAN
Singapore Management University, rkauffman@smu.edu.sg

Mei LIN
Singapore Management University, mlin@smu.edu.sg

Dan MA
Singapore Management University, madan@smu.edu.sg

Mechanism Design for Near Real-Time Retail Payment and Settlement Systems

Zhiling Guo, Robert J. Kauffman, Mei Lin and Dan Ma
Singapore Management University
{zhilingguo, rkauffman, mlin, madan}@smu.edu.sg

Abstract

Rapid expansion of e-commerce, along with rising domestic and cross-border payments, has fueled the demand among financial institutions for cost-effective means to achieve real-time settlement of retail payments. Traditionally, retail payments have made extensive use of interbank netting systems, in which payments are accumulated for end-of-day settlement. This approach, known as deferred net settlement (DNS), reduces the liquidity needs of a payment system, but bears inherent operational risks. As large dollar volumes of retail payments accumulate swiftly, real-time gross settlement (RTGS) is an attractive option. It permits immediate settlement of transactions during the day, but it brings up other risks that require consideration. We propose a hybrid payment management system involving elements of both DNS and RTGS. We explore several hybrid system mechanism designs to allow payment prioritization, reduce payment delays, enhance liquidity by pooling payments from banks, and optimize settlement. We provide a modeling framework and experimental set-up to evaluate the proposed approach. Our results shed new light about cost-effective and value-maximizing mechanisms to quickly settle increasingly large volumes of retail payments.

> "Only at the end of the day were the inter-bank claims settled, on a net basis. [The] system worked well: so long as [the] payments could be made then settlement went ahead. [I]f just one bank could not make its payments, the whole inter-bank settlement process would have been disrupted. This exposed a severe vulnerability – potentially threatening the stability of the entire financial system – if a bank failed."
>
> Tim Hampton [21], Economist, Financial Markets Department, Reserve Bank of New Zealand, 1999.

1. Introduction

The growth of global e-commerce, domestic payments, and cross-border payments has created new demand for real-time gross settlement (RTGS) for payments [8]. Compared with large-value payments that rely on central bank-operated settlement systems [5], retail payments make extensive use of interbank netting systems [9]. As retail payment flows accumulate with high velocity, large dollar volumes for settlement build up and the inherent operational risks in a netting system increase. These days, there is increasing interest worldwide to bring the capabilities of real-time processing and settlement to retail payments [27]. Financial innovation and market globalization have pushed banks to embrace cost-effective RTGS [1, 12].

Historically, interbank payments have been settled by using *deferred net settlement* (DNS) mechanisms, such as clearing houses and netting systems, where payments are accumulated and settlement is delayed [24]. Netting is an efficient way to reduce the liquidity needs of a payment system. The delays in settlement create vulnerabilities for the financial system though. The risk of liquidity raises concerns for retail payments, as well as wholesale settlement [17] and other types of high-value payments. In the 2011 UBS scandal, for example, a trader exploited the delay of exchange-traded funds (ETF) transactions by creating fake hedges that later caused more than US$2 billion dollars in losses for UBS [33].

RTGS payments are processed individually and settlement occurs with finality in the full amount immediately [26]. Although this may reduce operational risk by avoiding short-term debt between participants, it is higher in operational cost and creates intraday liquidity needs to smooth payment flows that are not synchronized. Central banks provide intraday liquidity for a fee or require it to be backed by collateral to control risk. By giving intraday credit, central banks assume risk.

RTGSs have been implemented among central banks for large-value payments since 1990 [2, 25, 39].[1] More recently there have been cross-border RTGS systems initiatives. An example is TARGET2 RTGS by the European Central Bank [15, 16] for large-value funds transfers between banks in Europe. RTGS is less of a reality for retail, though there have been initiatives indicative of future trends in the market.[2,3]

[1] Well-known systems include China's National Advanced Payment System (CNAPS) [13], Hong Kong's Clearing House Automated Transfer System (CHATS) [13], Mexico's Sistema de Pagos Electrónicos Interbancarios (SPEI) [11], the Monetary Authority of Singapore's Electronic Payment System Plus (MEPS+) [11], and the U.S. Federal Reserve Wire Network (Fedwire) [13, 31].

[2] They include the Faster Payments Service (www.fasterpayments.org.uk) implemented in the U.K. since the mid-2000s, the mobile

[3] The Society for Worldwide Interbank Financial Telecommunications (SWIFT) is at the epicenter of these developments, since its

1530-1605/15 $31.00 © 2015 IEEE
DOI 10.1109/HICSS.2015.573

Hybrid centralized payment management systems that combine various functions of RTGS, DNS, and payment priority queuing represent another possible solution [10, 46]. Queue-augmented real-time systems queue payments using a centralized or internal queue managed by individual banks, as these payments enter the system [35, 36]. Such hybrid systems will have less delay than end-of-day netting systems, and will lower liquidity needs to a greater extent than RTGS does. Efficiency gains and cost savings can be achieved by consolidating payment streams into a central platform.

The transformation of payment and settlement systems raises mechanism design issues.[4] One is the *participation incentive* for banks [3].[5] Since each bank has its own unique liquidity needs, it is not clear whether every bank will be interested in participating in a real-time or hybrid settlement system. Other important questions related to banks' participation also arise. For example, how will the central bank's credit policy affect the banks' intraday borrowing? Will all of the participating banks be better off from the liquidity pooling benefits of centralized payment management?

The second challenge is *incentive compatibility* [38]. Since individual banks have private information and make decisions about when to place payments into the central queuing system, do they have an incentive to delay submission? If so, what is the economic explanation? Also, will decentralized submission of payments and uncoordinated decision-making make centralized payment management system less valuable for the banks? Delayed and asynchronous submission of payments will adversely affect a payment settlement system's ability to match the payments it received with cash for final settlement. So a market mechanism should take into consideration how the banks will release their payments for settlement and coordinate them to synchronize their actions to mitigate the possible failure of an RTGS-based hybrid mechanism.

The third issue is *liquidity* [7, 24, 28, 29]. After payments are submitted to the central payment management system, having an effective payment settlement rule is crucial for market liquidity. It directly affects how payments in the queue from different banks get settled. In addition to the liquidity created by payment pooling from participating banks, the centralized system (possibly managed by a digital intermediary representing the central bank) may extend credit that allows the system operators to economize on liquidity.

Successful development and implementation of the hybrid system requires a deep understanding of the economic incentives and business value that arise from the adoption of a multi-sided technology platform like a centralized payment management system. Retail and corporate customers, merchants and banks, and government regulators all have a stake in achieving effective outcomes. A systematic evaluation of hybrid system performance should consider the banks' participation incentives, and central bank's credit and liquidity provision decision policies.

We address key mechanism design issues for an effective hybrid payment settlement system from the infrastructure design, participation incentives, and market coordination perspectives. These cover retail and financial services in the economy, information in the payment process, technology as a solution for digital intermediation, and economics as a theoretical lens through which to view and resolve some of the issues.

We ask: What constitutes an efficient design for a hybrid centralized payment management system that will support fully-automated straight-through (FAST) processing of payment settlements at low cost? How does a central payment management system solution alter the banks' economic incentives and payment submission tactics, and reduce their operational costs while controlling credit risks? And how are the technological developments, bank behavior, and regulations likely to drive market adoption for RTGS innovations?

2. Literature

RTGS, DNS, and hybrid settlement systems all have been discussed in the literature since the 1990s. Recently, Johnson et al. [24] proposed a deferred settlement mechanism based on the settlement of queued payments related to incoming payment value and not the account balance. Their mechanism reduces intraday credit extensions while modestly delaying the average time of payment settlement. They showed that the preference for RTGS or a hybrid system depends on how credit risk and liquidity efficiency trade-off.

Bech and Garratt [2] analyzed bank behavior under three credit regimes: *free intraday*, *collateralized*, and *priced credit*. Among these three, free credit is not a viable option for most central banks due to risk and moral hazard. They reported that collateralized credit is

payment mechanism design and technology staff have been involved in planning the implementation of large-value payment settlement systems around the world for over 25 years. SWIFT has experience in 20-plus small-value payment systems too [32].

[4] The design of a hybrid system is not trivial, if key economic considerations are made. Various operational and liquidity costs must be spread across many transactions when delayed net settlement is used, and the operational costs of handling individual transactions in real-time gross settlement is high. A central design challenge is to achieve real-time speed at low cost, while ensuring liquidity for "anytime" settlement, so banks avoid unnecessary reserves.

[5] This problem arises whether participation is accomplished through *settlement tiering* or *piggybacking*. This occurs with foreign banks in the U.S. that are not members of the Clearing House Interbank Payments Systems (CHIPS). Australia's Reserve Bank Information and Transfer Systems (RITS) impose minimum requirements for banks to participate, as opposed to avoid participation through settlement tiering relationships with other banks.

the prevalent option in Europe, while priced credit dominates in the U.S. They showed that payment delays emerge under various intraday credit policy regimes. They counter-intuitively concluded that it may be socially efficient for banks to delay payments.[6] There are benefits to synchronizing payments under priced credit. A related issue is how banks should coordinate.

Prior research also has assessed the benefit of payment settlement systems with simulation [24, 29] and agent-based methods [18], and examined the network topology of payments, and how payments to and from banks shift in the presence of market shocks [40]. A major limitation in this line of work is that bank behavior is typically viewed as exogenous to the system. In reality, banks will initiate actions, based on their own underlying decision-making motivations, such as which payments will be submitted to the system, what time submission will occur, and whether it is based on the payer, the nature of the transaction, etc.

Thus, we expect that the banks' actions should be endogenous and will largely depend on the payment system design. Guo et al. [19, 20] provided a theoretical and experimental market design framework to model order submissions, trade matching, and market-clearing dynamics in a distributed system based on economic considerations of value. Similar to their approach, we focus on the hybrid system's mechanism design by considering the participating banks' actions, the central bank's liquidity-related credit policy, and how the central payment management policy is structured. Various types of hybrid systems are possible based on the mechanism design ideas that we will discuss. We evaluate the performance of the proposed mechanisms using experimental methods.

3. Mechanism, Model and Management

We next outline a simplified hybrid settlement mechanism that implements the central payment management approach, and combines DNS and RTGS. Our model includes payment timing, credit risk, delay cost and settlement emphasizing the banks' viewpoint.

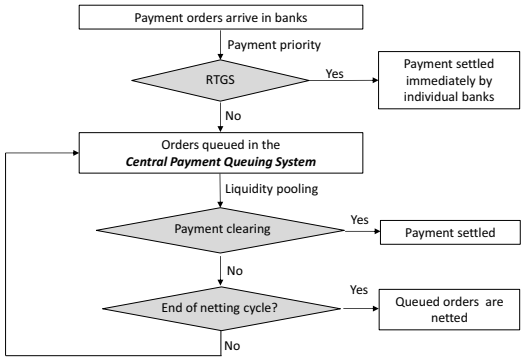
3.1. A Payment Settlement Mechanism Design

We propose a hybrid retail payment settlement mechanism that also provides liquidity management functions. These include setting payment priorities, maintaining payment inventories in reserve, pooling liquidity from participating banks, and optimizing payment settlement. Our proposed approach is different from other RTGSs in which banks are able to borrow money from the central bank to strengthen their intraday liquidity to meet transactions demand. The system we propose pools liquidity from all of the participating banks through the use of a centralized payment queue management system that we call a *central payment queuing system* (CPQS). It offsets a bank's payment orders using pooled receipts from other banks and the central bank's inventory, which makes settlement short and effective.

We also combine other aspects of netting and queuing in payment settlement. In CPQS, the amount of payments received in a one-minute period (or some other flexibly-determined period of time) provides the available liquidity for other payments to be released from the queue. Payments can be settled when CPQS receives sufficient incoming funds. Payments subject to deferral will be held in queue if they have not been offset by other incoming payments during that minute. Queuing enables the banks to automatically synchronize their outgoing payments with their incoming payments. CPQS will pool liquidity and be more effective in offsetting payments as more banks participate, a *network effect* [1], though *competitive externalities* may arise in the process [2].

The proposed process. As shown in Figure 1, first, payment orders arrive at the banks, for example, an ATM transaction, a payment to a merchant, or a mobile payment. Each bank will set its own payment priority. Orders with high priority will be settled in real-time at individual banks, while orders of low priority will enter CPQS and queue with other banks' payment requests, with or without a priority rank. CPQS pools all payment orders from all of the participating banks.

Figure 1. A Hybrid Payment Settlement Mechanism with a Central Payment Queuing System (CPQS)



In each one-minute interval, the system will check whether the total value of the payment orders received is greater than the total value of the queued payment. If this is the case, then all of the queued payment orders will be settled – in essence, *full near real-time settlement* – and the queue will be cleared. If the total value

[6] For a discussion of how banks handle synchronization of payments inflows and outflows, and how this affects the timing of their payment order submissions into a settlement system, see McAndrews and Rajan [31], who describe this for Fedwire in the U.S.

of the payments received plus the central queue's inventory (if any is present) is less than the value of all of the queued payment, then *partial near real-time settlement* will occur. If the central payment queue is rank-ordered, then the top-ranked orders will be cleared first. If the central queue is not rank-ordered, then the payment orders to be settled will be determined by the CPQS settlement mechanism. All unsettled payment orders will remain in the central queue, either to wait for incoming payments in the next minute, or be netted if it is the end-of-day netting time.

3.2. Model

Since different settlement rules in CPQS imply different payment priorities, the economic behavior of banks will be influenced by the rules of the payment system. In addition to the settlement rules, we consider several key factors influencing banks' decisions: payment timing, credit risk, and delay cost. We focus on quantifying the effects of different settlement approaches on banks' behavior and performance, and demonstrating how transaction pooling-related benefits arise in our hybrid payment settlement system. Through our economic analysis, we are also able to gain insights into infrastructure mechanism design for settlement systems.

Assumptions. We assume payment orders arrive according to a Poisson process. The time between arrivals of payment orders has an exponential distribution with parameter λ. The bank determines payment time-criticality and assigns different priorities to payment orders. *High-priority payment orders* go to the real-time queue for immediate settlement. Low-priority payment orders–*regular payment orders* in the bank's business–go to CPQS. These regular payment orders will be queued for pooled settlement with a possible time delay that is viewed as an economic trade-off.

Suppose there are I banks in total, each operating T minutes during a day. We define these variables as:

- p_{ij}^t: the dollar amount of the payment order made at time t that go from bank i to bank j;
- s_{ij}^t: the dollar amount of the payment order settled in real time at time t from bank i to bank j;
- $q_{ij}^{t,r}$: a payment order from bank i to bank j that enters the central queue at time t, with priority rank r in the queue, and q_{ij}^t if no rank is assigned;
- Q_i^t: bank i's set of released payments from the central queue at time t.
- B_i^t: bank i's available account balance at time t.

We further assume that no two payment orders that go from bank i to bank j will arrive at the same time. Also, once a payment order is settled, it will be removed from the queue immediately.[7]

The bank's decision. Banks have heterogeneous preferences regarding settlement delays and the credit risks they may have to bear related to the customers for whom they handle payments. Compared to RTGS, other alternative designs typically trade off credit risk with payment settlement delays [30, 37, 46].

Banks asynchronously receive heterogeneous payment requests. In a near real-time system, the bank will make two decisions at the end of each minute to prioritize its payment orders: the number of high priority orders p_{ij}^m that will get settled immediately and the number of regular payment orders q_{ij}^m that will be submitted to CPQS. Since real-time settlement has no delay cost, the bank's decision is to minimize the total cost of its payment order delay up to time m in a day, plus a credit penalty with a per unit cost δ, when the bank's account balance is negative.

At the end of each minute m, bank i's problem is to assign order priority by determining $\{p_{ij}^m, q_{ij}^m\}$: what orders should settled immediately or be submitted for CPQS queuing, subject to a balance constraint:

$$\min_{p_{ij}^m, q_{ij}^m} \sum_{t=1}^{m} \sum_{j \neq i, q_{ij}^t \notin Q_i^m} q_{ij}^t \,(m-t) - \delta \min\,(B_i^m, 0)$$

$$\text{s.t. } B_i^m = B_i^{m-1} - \sum_{m-1<t\leq m} \sum_{j \neq i} p_{ij}^t + \sum_{m-1<t\leq m} \sum_{j \neq i} s_{ji}^t - \sum_{j \neq i, q_{ij}^t \in Q_i^m} q_{ij}^t \text{, for } m-1 < t \leq m.$$

In bank i's objective function, the term $q_{ij}^t(m-t)$ measures its delay cost related to bank j. We can think of this as an agreed upon penalty for missing a value date that requires compensation for unavailable funds, or a diminution in goodwill between the banks. The summation over j is its total delay costs related to all other banks. The summation up to time m is the cumulative delay cost for all its unsettled payments in the central queue. The term $\delta \min\,(B_i^m, 0)$ is bank i's credit penalty, which is incurred only if $B_i^m < 0$, when bank i faces a negative account balance at the end of minute m. Since B_i^m is the cumulative measure of bank i's available funds, it does not have the summation sign. The objective function reflects a bank's trade-off between the costs of obtaining liquidity from the central bank and delaying payments by choosing which ones to submit to the central queue.

The balance constraint specifies bank i's account balance at the end of minute m, which should be equal to its account balance at the end of the previous minute, B_i^{m-1}, minus the payment orders settled in real

[7] If they arrive together, two payment orders can be combined into one order. This is a technical assumption to ease our exposition.

time from bank i to other banks, between time m - 1 and m ($\sum_{m-1<t\leq m}\sum_{j\neq i} p_{ij}^t$), plus the receipt of payments settled in real time from other banks to bank i, during time m - 1 and m ($\sum_{m-1<t\leq m}\sum_{j\neq i} s_{ji}^t$), and minus regular payment orders from bank i to other banks that are settled by the central queue within the one-minute interval ($\sum_{j\neq i, q_{ij}^t\in Q_i^m} q_{ij}^t$). The set of released payments Q_i^m is identified by the CPQS rule.

3.3. Management

The CPQS design involves rules for order entry, queuing and settlement that define the management process for hybrid payment settlement. For example, orders may enter the queue using a first-in, first-out (FIFO) rule. Payments entering earlier will be given higher priority. Priorities for payments can also be based on other criteria. One is the personal payment history of a customer [45]. Another is payment value, as in the Clearing House Automated Payment System (CHAPS) in the U.K. There are many other ways to set the priority. Such criteria assign a priority to payments entering the queue.

Settlement criteria define the rules to release queued payments. Settlement at the end of each minute allows multiple payments to settle simultaneously, if offsetting funds to match are found. By the end of each minute, certain payment orders can be released from the queue as soon as sufficient receipts arrive to cover the outgoing funds. Also at the beginning of each minute, queued payments are reset to reflect new orders and settlements. At the end of the day, all payment orders in the queue will be netted. Of course, the end-of-day netting (*late netting*) can be adjusted to reflect other design considerations, such as hourly (*frequent netting*) and noon netting (*early netting*).

Ranked queue clearing. The system assigns a rank to currently active payment orders in the queue. Assume that at time m a total of n payment orders are in the queue. $q_{ij}^{t,r}$ is the payment amount that enters the central queue at time t, which goes from bank i to bank j, and is currently ranked in the rth position. The payment order-matching problem is:

$$\min_k \sum_{t=1}^{m}\sum_{r=k+1}^{n} q_{ij}^{t,r}\,(m-t)$$
$$s.t.\ \ \sum_{r=1}^{k} q_{ij}^{t,r} \leq \sum_{m-1<t\leq m}\sum_{i=1}^{I}\sum_{j\neq i} s_{ji}^t,$$
$$k \leq n, \text{and } k \text{ is } integer.$$

The objective is to choose the top k-ranked payment orders for settlement so the total delay cost up to time m for orders still in queue is minimized. CPQS settlement constrains that the first k queued payment orders are released at the end of minute m when the value of the total pooled receipts within that minute are greater than or equal to the value of the k ranked payments. This constraint assumes the central bank does not provide any liquidity to CPQS. If the central bank offers liquidity level c through credit, the constraint can be modified to $\sum_{r=1}^{k} q_{ij}^{t,r} - \sum_{m-1<t\leq m}\sum_{j\neq i} s_{ji}^t \leq c$. After settlement, the rank will be adjusted.

Non-ranked queue clearing. q_{ij}^t is the payment amount from bank i to bank j at time t that enters the central queue. Assume the system maintains a liquidity level c. The maximum liquidity level that the central bank provides will not be larger than the total credit cost the central bank has to bear in a decentralized system. Otherwise, it would be inefficient to use a centralized system. The related order-matching problem is:

$$\min_{y_{ij}^t} \sum_{t=1}^{m}\sum_{i=1}^{I}\sum_{j\neq i} q_{ij}^t\,(m-t)\,(1-y_{ij}^t)$$
$$s.t. \sum_{i=1}^{I}\sum_{j\neq i} q_{ij}^t\,y_{ij}^t - \sum_{m-1<t\leq m}\sum_{i=1}^{I}\sum_{j\neq i} s_{ji}^t \leq c$$
$$y_{ij}^t = 1\ or\ 0.$$

Here, y_{ij}^t is a binary decision variable. When $y_{ij}^t = 0$, the payment order from bank i to bank j that enters the system at time t is not chosen to be matched, so it imposes a delay cost of $q_{ij}^t(m-t)$ on the system. When $y_{ij}^t = 1$, the payment order will be matched. As a result, the order is removed from the queue and there is no delay related to this order. The objective is to minimize the total payment delay cost by selecting payments from the CPQS queue to settle without violating the settlement constraint. The constraint ensures the net settlement amount, with outgoing payments removed, will not exceed system liquidity c.

Note that here we have not specified how participating banks finally get compensated by contribution to the central payment inventory pool. A number of possibilities for this may affect the model that we have proposed. One is a per transaction fee when payment inventory is drawn down. The second possibility is similar, only any fees would be based on the dollar amount of the payments that are covered. Pricing in this manner is likely to be ineffective though, since it will only diminish the attractiveness of the mechanism that we have proposed. Another possibility is to assess which banks are net liquidity providers or net deficit consumers of the funds in the central payment inventory pool, and then to charge them *post hoc* fees. The last is similar to Vickrey-Clarke-Groves payment rules used in auctions, where the system charges each bank the social cost of its payment withdrawn from CPQS incurred by the rest of the banks.

4. Illustration of the Hybrid Mechanism

To illustrate the cost savings that can be achieved

by pooling payment orders, we offer an example for two banks' incoming and outgoing payment flows within a one-minute interval. See Figures 2 and 3.

Figure 2. Settling Banks' Account Balances

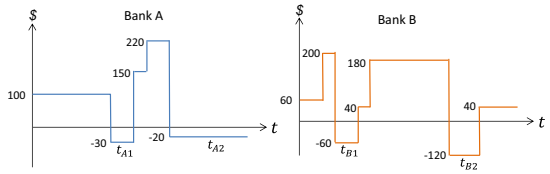
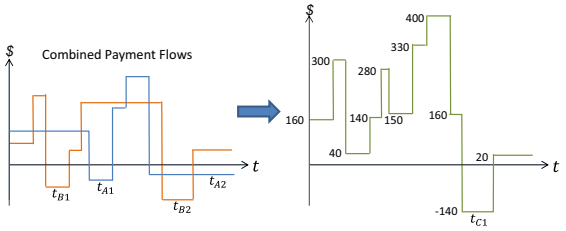


Figure 3. Pooled Payments in a Centralized System



Suppose that Bank A has an initial account balance of \$100. The settlement of an outgoing real-time payment of \$130 reduces the bank's account balance to -\$30. After t_{A1} time, suppose that the bank receives an incoming payment of \$180, which brings its account balance back to \$150. The next incoming payment of \$70 builds up the account balance to \$220. Then another payment of \$240 that is settled in real time will result in a negative balance of -\$20. Since there are no other payments, the negative balance will persist through time t_{A2}. Further assume the liquidity cost is linear in time. The total delay cost for Bank A thus will be \$30 × t_{A1}+ \$20 × t_{A2}. Bank B's order flows can be interpreted in a similar manner, and its liquidity cost is \$60 × t_{B1} + 120 × t_{B2}.

Now, assume the two banks submit their payments to a central settlement management system. By pooling their payments, the two banks will operate as if they are one integrated bank. The combined payments are shown in Figure 3. The total liquidity cost of the integrated bank is 140 × t_{C1}= 140 × t_{B2}. For the combined payments, the overdraft occurs when Bank B makes a large payment at the beginning of t_{B2} and ends when Bank B receives a new payment at the end of t_{B2}.

Further note that $t_{B2} < t_{A2}$. This implies that the total liquidity cost for Banks A and B operating separately will be $30t_{A1}$+ $20t_{A2}$+ $60t_{B1}$+ $120t_{B2}$ > $140t_{B2}$. As a result, the integrated bank that is made possible by the pooled payments inventory incurs a smaller total liquidity cost. The gain can be shared between the two banks so the operational cost of each is reduced.

The real challenge, however, is that individual banks may not have an incentive to submit their payment orders in real time, as this is their private information. Since incoming payments are a substitute for costly borrowing from the central bank, the banks have an incentive to delay submitting payments into the system if the delay cost is smaller than the liquidity cost. This is the incentive compatibility concern that arises in this payments platform setting.

To illustrate this, assume the cost of per unit liquidity is \$10 in each minute. If Bank A can delay the first payment of \$130 until it receives the incoming payment \$180, then Bank A will incur a delay cost of \$130 t_{A1}, but avoids a liquidity cost of \$30 × 10 = \$300. Thus, Bank A will have an incentive to delay the submission of the \$130 payment if t_{A1} is not too long.

Private information is unlikely to be reported truthfully by the banks without appropriate incentives. They may wish to manipulate their payment timing to avoid credit costs for central bank borrowing. The proposed system may alleviate such concerns because combined payment settlement enhances liquidity.

Our professional experience with bank payments management and consulting, and knowledge from prior research suggest that there are underlying reasons for why payment orders for settlement may cluster.[8] The related agency problems can be alleviated by allowing the central bank to price-discriminate against the banks in the provision of intraday liquidity. For example, the central bank may consider to price intraday liquidity and settlement differently at different times during the day to give the banks incentives to settle at the desired times for different kinds of payments. Whether it does so is a policy decision that will be assessed based on the quality of the mechanism used for settlements.

5. A Mechanism Design Experiment

We will simulate and compare the performance of different settlement systems designs. A RTGS mechanism is useful as a measurement benchmark against which we can compare alternative queuing and DNS performance. The different designs that we will assess are based on the *frequency of DNS*, similar to Willison's [46] modeling and numerical simulation approach. It will also be based on the *payment settlement priority queuing rules* that are imposed. Our factorial experiment and treatment set-ups are discussed below.

5.1. Experimental Set-Up

To understand the effects of key mechanism design factors on performance, we propose a controlled experiment with the design description in Table 1.

[8] See Hong Kong Monetary Authority [22] on its 1996 implementation of a real-time gross settlement system; McAndrews and Rajan [31] on Fedwire's demand for settlement at different times of the day; and Willison [46], on intraday morning and afternoon offsets.

Table 1. Mechanism Design Treatment

MECHANISM DESIGN	EXPERIMENTAL CONDITIONS	EXPERIMENTAL TREATMENTS
Central bank	Liquidity credit	Uniform price Price discrimination
Banks	Liquidity needs	Low High
Central Order Management System Queue	DNS system netting times	Hourly netting Morning netting Afternoon netting
	Information transparency	Queue visible Queue not visible
	Queue entry	Random entry Threshold rule-based entry Value-based entry
	Queue settlement	Arrival time-based priority Payment value-based priority
	Liquidity	Credit available No credit available

Credit to provide liquidity. The first experimental treatment that we will explore is related to the central bank's decision problem. We assume there is a per unit cost δ when a bank requires credit from the central bank to have sufficient liquidity to settle the payments that it is handling. This is an implementation of a *uniform pricing rule* for payments settlement. An alternative design is to allow the central bank to charge different prices for credit at different times of the day. This is a *price discrimination rule*. For example, there might be different hourly, morning or afternoon prices, etc.

The need for liquidity. A bank's need for liquidity largely depends on the time at which its payment orders arrive. This is the focus of our second experimental treatment. We can separately model the payment order arrivals between different banks with different Poisson arrival rates. In a specific time interval, if a higher dollar volume of outgoing payment orders arrives, then a bank's credit needs will increase, creating pressure for it to borrow from the central bank so it will be able to settle funds on an intraday basis.

Delayed netting settlement. Willison [46] modeled different delayed net settlement approaches, including one-hour, morning, and afternoon netting. The one-hour netting mechanism has a high frequency of payment settlement: net settlement of queued payments occurs every hour. The net amounts settle immediately thereafter in real time. It flushes the payment order queue at the end of the hour. The morning and afternoon netting approaches are intended to manipulate the amount of time delay that is introduced into settlement.

Information transparency. *Information transparency* determines to what degree a bank can observe the payments that other banks have submitted to the central queue. Payment orders to pay out funds can only be offset if there are payment receipts in the central queue, from other banks or the central bank taking the offsetting position. The level of information transparency may affect a bank's order submission tactics.

Queue entry. There are a number of different ways that the payment queue can be handled in terms of orders to pay out funds and receipts of funds. Our experimental treatment for the queue entry condition consists of: (1) a *random percentage* (say 25%, 50%, 75%) of payments are selected by the bank to enter the queue; (2) payments *smaller than some threshold value* are selected by the bank to enter the queue; and (3) payments that are *not rank-ordered by priority* enter the payment queue. The first two cases do not involve strategic submission decision-making on the part of the bank. A bank can automate queue entry based on prespecified criteria. Only in the last case will the bank strategically assign the priority it places on payments that it wishes to see enter the queue. We call this *value-based queue entry.*

Queue settlement. The next experimental treatment that we will assess involves the impacts of priority-ranked payment queue settlement based on payment order arrival time and dollar value. In contrast, we will assess payments that are not priority-ranked based on the offsetting funds that are received in the system.

Liquidity provision. A final experimental treatment is related to the provision of liquidity. The central bank can choose to provide its own inventory of received payments to banks involved in payments settlement, as a way to supplying liquidity to them. If it does this, then the speed of settlement will increase when RTGS is used. The central bank will have to bear the exposure of credit risk to the banks involved though. If the central bank does not supply liquidity, the speed of payment settlement will slow down, introducing operational risks. This is the trade-off between liquidity and operational risks.

5.2. Mechanism Design Evaluation

Meeting the demand for liquidity to cover payment orders that require the outflow of funds from a bank, and the speed with which payments are settled are important criteria for evaluating the performance of a payment settlement mechanism. We define several measures that will be useful for this purpose. We plan to use these measures to evaluate the treatments of the experiments.

Delays. The *settlement system delay* is the difference between the time the payment is received by the bank and the time it is settled. *Total settlement system delay* is given by this *normalized delay index*: $1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{q_{ij}^t \in Q_i^m} q_{ij}^t (T-t)}{\sum_{t=1}^{T}\sum_{i=1}^{I} q_{ij}^t (T-t)}$. The denominator of the index is the value of the queued payment orders multiplied by the time that payments would have been queued, had their settlement been delayed until the

time the queue closes. The numerator is the savings that arise from the delay: the value of settled payments multiplied by the time savings (between when they exit the queue and when it closes). The index takes a value between 0 and 1. If all queued payments are immediately settled in real time, then the index will equal 0. For end-of-day delayed net settlement, the set Q_i^m will be empty for $t < T$, and so the index will equal 1.

Overdrafts. An *overdraft* will occur when a bank's account balance falls below $0. The Federal Reserve Bank measures overdraft positions at banks in the U.S. at the end of each minute of the day, and from this it is able to compute a bank's *average daily daylight overdraft*. This is defined as the sum of all the overdraft minutes of the day for the member bank divided by the number of operating minutes of the day. The Fed charges banks fees based on their average overdrafts for credit use. Let $OD_i^t = \max(-B_i^t, 0)$ be bank i's overdraft at time t. If the balance is negative, then the bank will have an overdraft equal to the absolute value of the balance; if the balance is positive, no overdraft will occur, and so $OD_i^t = 0$. The average overdraft is $\overline{OD} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{I} OD_i^t}{T \times I}$. Coordination among multiple banks can reduce their overdrafts.

Average funds transfer. Another way to evaluate the performance of a payment settlement mechanism is to assess how much the account balances fluctuate for the banks involved in the settlement of payments. For this purpose, we define the *average funds transfer* amount for all banks as $\frac{\sum_{t=2}^{T}\sum_{i=1}^{I}|B_i^t - B_i^{t-1}|}{(T-1)I}$. This measures the variation in account balances to settle payments, and is also useful as a measure of liquidity. In addition, the *average maximum funds transfer* measures the average maximum funds that must be transferred from an individual account on a per minute basis, across the minutes of the day, to complete all payments. We also will measure the *average absolute change in balances* that occurs per minute for each bank in our experimental simulation. This is the dollar amount of funds that a bank has to move for any given minute of the day, in or out of its account. We also can apply variance measures for how the balance fluctuates as banks settle their payments.

6. Discussion

In the second quarter of 2014, The SWIFT Institute [44], the London-based research arm of SWIFT, called for research proposals on the "Transformation in the Payment and Settlement System Infrastructure." There is great interest on the part of commercial banks, central banks, merchants, and banking customers – and third-party service providers like SWIFT – around the world to move to a more economical and less risky infrastructure for retail payments. It is especially interesting as we enter 2015 to consider the role of technological innovation in payment clearing and settlement, as the technologies and industrial organization of payments change around us.

The year of 2014, based on the assessment of many observers, was a breakout year for the development and diffusion of mobile payments [41, 42], where there is going to be significant pressure to create new support for settlement. As mobile payments further spur the growth of e-commerce, the volume of retail payments in the settlement system will increase significantly. The system we proposed is scalable by design. In fact, our model illustrates that a high number of payment requests benefit the banks and the system as a whole by improving liquidity. Moreover, Mobile payments are introducing new roles into the financial system. Digital players such as PayPal, Square, Google, Adyen, and many others are more equipped than banks for innovation with mobile technologies [14]. Thus, the settlement system needs to consider a high degree of heterogeneity among the participants. Our experiments aim to test different system configurations for the effects of different participants on system performance. The trend of mobile payments also suggests the need for other kinds of payment settlement systems refinements that deal with low-value Internet payments, utility services payments, and point-of-sale payments. So it will be important to differentiate the increasingly wide variety of payments. Our model takes into account such payment heterogeneity.

We see similar developments related to low-value transactions and payments in international trade in support of supply chain management. For example, bank payment obligations, a relatively new means of payment solutions in international trade finance, represent "an irrevocable undertaking given by a bank to another bank that payment will be made on a specified date after successful electronic matching of data according to an industry-wide set of ICC rules" [23]. They create the basis for increases in open account terms in trade, diminishing the role of documentary credits and collections as the primary vehicles of trade finance. This is possible since supply chain management and procurement-related trade documentation can now be tracked more effectively around the world.

Examples of new document and payment platform-related providers in this area that are changing the financial side of supply chain management include: Orbian (www.orbian.com), GT Nexus (www.gtnexus.com) and Bolero (www.bolero.net). These developments will make low-value trade transactions more economical, and support trade and exchange between small enterprises in different parts of the world with lower transaction costs. This is another reason why the kinds

of approaches that we have discussed related to low-value domestic payments are likely to be extended to handle multi-currency cross-border payments as well.

These developments suggest a number of changes that will need to occur in the operational environment in which payment settlement. For new forms of low-value retail payments, for example, the role that the central banks play will need to be revisited. Central banks will increasingly need to provide the technical infrastructure for low-value payments, just like an electrical utility must handle the delivery of electricity services to all sorts of customers.

A possible alternative is that the *geometry of payment platforms* in financial services will shift to favor other third-party service providers. An example of this in the electronic bill presentment segment of financial services in the U.S. is NACHA (www.nacha. org), The Electronic Payments Association (previously known as the National Automated Clearing House Association). It is operated, funded and governed by the financial institutions that are its members. Just as we have modeled the central bank as the primary provider of liquidity, so too is it possible for third-party service providers to explore new roles, including the role of providing liquidity in exchange for compensation due to the risks that are undertaken.

Other areas of settlement coverage will need to be extended to remote and cross-border payments. The settlement system should then be ready to take into account country-specific regulatory policies and resolve any issues related to volatility in currency rates. The system in this work is designed to handle shocks related to the changes in the number of participants, which regulations may generate.

In addition, their systems will need to change to support different kinds of services from what we see today. Financial services suffer from a problem that is known in the industry as the *reference data problem*, limiting the capacity that firms have to integrate their data, software and systems, largely due to insufficiently compatible data scheme and ontologies. The ongoing reference data revolution in financial services is driving toward more effective cross-functional data standards, and greater systems integration across different business areas. The expected impacts are especially high in financial markets, retail banking and lending, among other commercial banking activities.

The acceptance of a proposed settlement system is determined by the system capability in serving the many participants that interact in the payments ecosystem. Our system handles the demand of requests from a large number of banks, and it also addresses the problems of the central bank by appropriately prioritizing requests and moderating risk. Meanwhile, the system has the potential to be tailored for new roles that emerge in the continuing growth of the payment industry. Thus, both new and existing participants in the settlement system may have an incentive to adopt the proposed model. Moreover, from a social welfare perspective, our system offers other substantial benefits. To more closely analyze and quantify the benefits of the system, we have planned a number of experiments to show the payoffs of different participants under various system configurations. This will illustrate the appropriateness of the proposed model.

7. Conclusion

In this study, we explore various mechanism designs of a hybrid payment management and settlement system that involves elements of both RTGS and DNS. The different mechanisms that we presented trade off the liquidity costs incurred when a bank borrows funds under a line of credit from the central bank, and the delayed payment costs that arise from using the central payment management system that involves different frequencies of periodic netting. We examined a central payment queuing system design that incorporates payment priority queuing rules for settlement with and without priority rankings. We will present these findings at HICSS 2015.

There are two limitations of our approach that will affect the external validity of our experimental findings. First, we assume only one payment channel, rather than several channels – card payments, Internet payments, mobile payments, etc. – as is the case in real-world payment services. The limitation that this imposes is the likely heterogeneity of risk across the channels, the self-selection of users in different channels, and the underlying differences in automation and straight-through processes in the channels. Second, there are other proposals that have been discussed among central bankers that are not represented in our current experimental set-up. As time passes, and we gain additional experience with our modeling approach, we will extend it to ensure fuller coverage of existing approaches, while assessing the value of our innovations.

In spite of these issues, we nevertheless expect to use this research to encourage discussion in the minitrack on experimental approaches to business problems involving digital intermediation, and to engender a fuller understanding of the issues that low-value payments are creating for clearing and settlement.

References

[1] Allsop, P., Summers, B., Veale, J. 2009. The evolution of real-time gross settlement: access, liquidity and credit, and pricing. The World Bank, Washington, DC.

[2] Angelini, P. 1998. An analysis of competitive externali-

ties in gross settlement systems. *J. Bkg. Fin*., 22, 1-18.
[3] Arculus, R., Hancock, J., Moran, G. 2012. The impact of payment system design on tiering incentives. Working paper, Res. Bank of Australia, Sydney, Australia.
[4] Bech, M.L., Garratt, R. 2003. The intraday liquidity management game. *J. Econ. Theory*, 109, 198-219.
[5] Bech, M.L., Hobijn, B. 2006. Technology diffusion within central banking: the case of real-time gross settlement. Federal Reserve Bank of New York.
[6] Bertilsson, C., Hult, F. 2013. Future payment solutions in Sweden: critical success factors and scenarios from a stakeholder perspective. Lund U., Stockholm, Sweden.
[7] Cirasino, M. 2003. The role of the central banks in supervising the financial system: the case of the oversight of payment systems. World Bank, Wash., DC.
[8] Cirasino, M., Garcia, J.A. 2008. Measuring payment system development. World Bank, Washington, DC.
[9] CPSS. 1993. Central bank payment and services with respect to cross-border and multi-currency transactions. Bank for Intl. Settlements, Basel, Switzerland.
[10] CPSS. 2005. New developments in large-value payment systems. Bank for Intl. Sett., Basle.
[11] CPSS. 2011. Payment, clearing and settlement systems in the CPSS countries. Vol. 1, Bank for Intl. Sett., Basle.
[12] CPSS. 2012a. Innovations in retail payments. Bank for Intl. Sett., Basel.
[13] CPSS. 2012b. Payment, clearing and settlement systems in the CPSS countries. Vol. 2, Bank for Intl. Sett., Basle.
[14] Denecker, O, Gulati, S., Niederkorn, M. 2014. The digital battle that banks must win. *McKinsey & Co. Insights and Publications*, New York, NY, August.
[15] European Central Bank. 2004. Future developments in the TARGET system. *ECB Mnthly. Bull.*, April, 59-65.
[16] European Central Bank. 2013. TARGET2. Frankfurt, Germany, September.
[17] FFIEC. 2004. Wholesale payment systems. ISACA.
[18] Galbiati, M, Soramäki, K. 2008. An agent-based model of payment systems. Bank of England, London, UK
[19] Guo, Z., Koehler, G.J., Whinston, A.B. 2007. A market-based optimization algorithm for distributed systems. *Mgmt. Sci.*, 53(8), 1345-1358.
[20] Guo, Z., Koehler, G., Whinston, A. 2012. A computational analysis of bundle trading markets design for distributed resource allocation. *Info. Sys. Res.*, 23, 823-843.
[21] Hampton, T. 1999. Intra-day liquidity: 18 months on. *Bulletin*, Reserve Bank of New Zealand, 62(4), 34-46.
[22] Hong Kong Monetary Authority. 1997. Hong Kong's real-time gross settlement system. Monetary Policy and Markets Department, Hong Kong.
[23] International Chamber of Commerce. 2013. Bank payment obligation. Paris, France.
[24] Johnson, K., McAndrews, J.J., Soramäki. 2004. Economizing on liquidity with deferred settlement mechanisms. *Econ. Pol. Rev.*, Fed. Res. Bk. NY, 10(3), 56-72.
[25] Kahn, C., Roberds, W. 2001. Real-time gross settlement and costs of immediacy. *J. Mon. Econ*. 299-319.
[26] Kahn, C., Roberds, W. 2009. Why pay? an introduction to payments economics. *J. Fin. Inter.*, 18(1), 1-23.
[27] KPMG. 2012. The great payments transformation: insights into the payments ecosystem. New York, NY.
[28] Leinonen, H., Soramäki, K. 1999. Optimizing liquidity usage and settlement speed in payment systems. Paper 16, Bank of Finland, Helsinki.
[29] Leinonen, H., Soramäki, K. 2003. Simulating interbank payment and securities settlement mechanisms with the BoF-PSS2 simulator. Bank of Finland, Helsinki.
[30] Manning, M., Nier, E., Schanz, J. 2009. *The Economics of Large-Value Payments and Settlement: Theory and Policy Issues for Central Banks*. Oxford U. Press, NY.
[31] McAndrews, J.J., Rajan, S. 2000. The timing and funding of Fedwire funds transfers. *Econ. Pol. Rev.*, Federal Reserve Bank of New York, New York, NY, 17-32.
[32] Moon, M. 2014. Global developments in retail real-time payments. *Australia Bus. Rev.*, March 9.
[33] Murphy, M. 2011. UBS trader Adoboli held over US$2bn loss. *Financial Times*, September 15.
[34] *Payments, Cards & Mobile.* 2013. Swish mobile payments, December 6.
[35] Peñaloza, R.A. 2009. A duality theory of payment systems. *J. Math. Econ.*, 45(9-10), 679-692.
[36] Peñaloza, R.A. 2011. Implementation of optimal settlement functions in real-time gross settlement systems. Working paper, Dept. Econ., Univ. of Brazil.
[37] Schulz, C. 2011. Liquidity requirements and payment delays: participant type and dependent preferences. Paper 1291, European Central Bank, Frankfurt, Germany.
[38] Selgin, G. 2004. Wholesale payments: questioning the market failure hypothesis. *Intl. Rev. Law Econ.*, 2(3), 333-350.
[39] Shen, P. 1997. Settlement risk in large-value payment systems. *Econ. Rev*., Federal Reserve Bank of Kansas City, Second Quarter, 46-62.
[40] Soramäki, K., Bech, M.L., Arnold, J., Glass, R.J., Beyeler, W.E. *Physica A*, 6(1), 317-333.
[41] Stone, B., Kharif, O. 2013. Easy mobile payments are almost here. *Bloomberg BusinessWeek*, November 14.
[42] Taulli, T. 2014. Ka-ching: mobile payments predictions for 2014. *Forbes*, January 1.
[43] Teague, S. 2014. Singapore closes in on real-time payments as Australia prepares for mandate. *Euromoney*, January 30.
[44] The SWIFT Institute. 2014. Call for proposals: transformation in t
[45] he payment and settlement system infrastructure. London, UK.
[46] Wallace, N. 2000. Knowledge of individual histories and optimal payment arrangements. *Federal Reserve Bank of Minneapolis Quarterly Review*, 24(3), 11-21.
[47] Willison, M. 2005. Real-time gross settlement and hybrid payment systems: a comparison. Working paper, The Bank of England, London, UK.