SWIFT INSTITUTE

SWIFT INSTITUTE WORKING PAPER NO. 2014-004

NEAR REAL-TIME RETAIL PAYMENT AND SETTLEMENT SYSTEMS MECHANISM DESIGN

ZHILING GUO

ROB KAUFFMAN

MEI LIN

DAN MA

PUBLICATION DATE: 08 SEPTEMBER 2015

NEAR REAL-TIME RETAIL PAYMENT
AND SETTLEMENT SYSTEMS MECHANISM DESIGN

Zhiling Guo, Robert J. Kauffman, Mei Lin, Dan Ma

School of Information Systems, Singapore Management University
80 Stamford Road, Singapore 178902
{zhilingguo, rkauffman, mlin, madan}@smu.edu.sg

Last revised: September 4, 2015

ABSTRACT

Rapid expansion of e-commerce, along with rising domestic and cross-border payments, has fueled the demand among financial institutions for a cost-effective means to expedite clearing and settlement of low-value retail payments. These are called *faster payments settlement systems.* 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 the payment system, but bears inherent operational and financial risks for unsettled intraday payments. As large dollar volumes of retail payments accumulate, *real-time gross settlement* (RTGS) has been recognized as an attractive option, especially for high-value payments. It permits immediate settlement of transactions during the day, but it brings up other risks and incentives issues that require consideration. This research proposes a *hybrid faster payments settlement system* involving elements of both DNS and RTGS. We explore a mechanism that supports centralized queuing, permits payment prioritization, reduces payment delays, enhances liquidity, and optimizes the settlement process. We offer a modeling framework and experimental simulations to evaluate the proposed approach. Our evaluation is based on a set of *system performance conjectures* that support operational performance evaluation and mechanism design-related policy insights. Our results point out the qualities of a cost-effective and value-maximizing mechanism to quickly settle increasingly large volumes of retail payments, while ensuring that the *incentives for payment system stakeholders* are given careful consideration.

Keywords. Clearing, faster payments, hybrid mechanism, management science, mechanism design, payment, policy analytics, payment, settlement, systems design

Acknowledgments. Different aspects of this research have been presented in an early form at the January 2015 Hawaii International Conference on System Sciences (HICSS), in the Society, Information, Technology, and Economics Mini-Track, as well as at the April 2015 SWIFT Business Forum London. We also benefited from discussions with attendees at the FINEXTRA Future of Money Conference in London in April 2015. We would like to thank the following individuals for their helpful guidance and contributions to our work: Hugh Watson, Eric Clemons, Thomas Weber and Rajiv Dewan at HICSS; Terrence Goh and Nelson Yeon Lee Chua at the Monetary Authority of Singapore; Carlos Palmers, Marie-Christine Diaz, Brett Lancaster, Elie Lasker, Chermaine Lee, Nancy Murphy, Peter Ware and anonymous reviewer at SWIFT; and Enoch Ch'ng, Tin Seong Kam and Alan Megargel at Singapore Management University. This research was conducted with grant support from the SWIFT Institute, under its Call for Proposals entitled "Transformation in the Payment and Settlement System Infrastructure." All opinions, errors, and omissions are the sole responsibility of the authors.

> *"Many payment systems around the globe are undergoing fundamental changes to reflect the new realities of digital commerce, especially when it comes to the speed in which payment requests are processed. The electronic payment systems that were designed in previous decades can no longer meet all of the expectations of a society where devices with enormous computing power are literally in the hands of most adults and instantaneous response has become the norm, even when it isn't necessarily required for business reasons."*
>
> \- Clear2Pay, a payment modernization company (McIntosh et al. 2014).

1. INTRODUCTION

In the digital economy, everything wants to be faster – faster product design, faster sales capabilities, and faster delivery of goods and services. In the middle of it all though, faster payments practices need to be given more attention. This is because they enable the exchange of funds at nearly the same time that economic transactions occur, more effectively supporting the related consumer, business and social objectives. Making payments "faster" – especially faster settlement of funds – has not been easy due to various business process and technological reasons, as well as operational risk and liquidity management concerns in large, diverse and complex banking systems (Todd 2015).[1] In addition, firm-level and public payment infrastructure costs and benefits have been difficult to measure (Greene et al. 2015).

1.1. Technological Innovation and Faster Payments

Technological innovation has led to the rise of the new digital economy, which is driving the need for faster settlement of payments, as the above quotation so clearly states. Important by-products of advancing technology are fundamental changes in the patterns of public demand and consumption for products and services, new business models for the retail and wholesale sectors, and a changed competitive landscape in the financial services (Cognizant 2014). In addition, the patterns of demand for cash in transaction-making have also been changing, as the use of credit and debit cards, mobile phone-based payments, and the Internet channel have grown. These forces also have led to the need on the part of participants in

[1] SWIFT is an emerging leader in the global marketplace for retail-level real-time domestic payments, most recently with its involvement in the development of Australia's New Payments Platform (NPP) (SWIFT 2014a). We were fortunate to talk with the people involved in that effort at SWIFT, and to participate in the recent SWIFT London Business Forum where these issues were debated and discussed in a keynote panel on faster payments.

the retail banking ecosystem around the world to handle a much larger number of low-value payments, including both domestic (BNY Mellon 2014) and cross-border payments (Aite Group 2015). And there has been a simultaneous call by consumers, industry observers, government regulators, and banking sector innovators to expedite the settlement of these payments so customers receive funds as soon as possible (Groenfeldt 2014, Summers 2015).

The businesses created by electronic commerce have been especially influential in shifting consumption patterns and unleashing the new wave of low-value payments. Information technology (IT) allows online businesses to reach far beyond the traditional approach in which consumers mostly buy from large retailers. Many online marketplaces now procure and sell goods, and facilitate transactions between many small and medium enterprises (SMEs) or individual merchants and online consumers (International Trade Center 2009). The result has been an explosion in the variety of transacted goods. Another related development has been a new "long tail" of the distribution of consumption due to the increased variety of goods available on the Internet (Anderson 2008). Combined with consumers' ability to purchase anytime and anywhere, there is a much larger volume of low-value payments in the retail sector (Boston Consulting Group 2014).

The digital economy also has important implications for banks, which now face an ecosystem with demand for different levels and kinds of services, as well as new competitors. In addition, in recent years due to new IT, the market for financial services has become a newly-vulnerable market (Clemons et al. 2002, Granados et al. 2008). Future-focused technology innovation and entrepreneurs who are devoted to changing how financial services business processes work view the sector as newly-easy to enter, attractive to attack, and difficult for incumbent firms to defend their positions against innovative "fintech" startups (Economist 2015). In fact, other solutions, such as PayPal, Secure Vault Payment (SVP), and P2P payment options, have emerged in response to the need for faster payments in general (Daly 2013), though they do not directly address the issue of settlement. Moreover, with an increasing number of SMEs and small merchants, delayed settlement of their funds is increasingly undesirable. These merchants need to receive funds as soon as possible to avoid the borrowing costs of funding their supply acquisition, inventory management, and operations.

The same is true for large corporations, for which banks will have credit risk when funds are made available to their customers in advance of the related payments settlement. Thus, improved timeliness in settling payments is critical for both low-value and high-value payments, though the issues and complications of the mechanisms that will work are very different.[2] To sustain their competitiveness, banks need to create infrastructure solutions that transform their capabilities in the transition to "future money." [3]

1.2. Faster Payments Mechanisms

Historically, interbank payments have been settled using *deferred net settlement* (DNS) *mechanisms*, such as clearinghouses and netting systems, where payments are accumulated and settlement is delayed (Alexander et al. 2006). Netting is an efficient way to reduce the overall liquidity needs of a payment system; however, the delays in settlement are undesirable and create vulnerabilities for the financial system. The goals of payment systems innovation have been to make retail payment transfers more timely and services more accessible, and to lay the foundation for electronic, mobile, and block chain-based payments that have been rapidly gaining traction. *Real-time gross settlement* (RTGS) *mechanisms* have been implemented among central banks for large-value payments, where the residual risks of unsettled transactions may be severe (Angelini 1998, Kahn and Roberds 2001, Shen 1997).[4] It has generally been the case that DNS settlement has been more often used with lower-value payments, while RTGS is more focused on higher-value payments, where gross settlement has been viewed as less risky.[5]

[2] An anonymous reviewer suggested that only with larger-value payments, due to the randomness and variance of the number and amount of individual payments, will there be great enough pressure on the liquidity in the system to warrant the settlement of payments with a hybrid approach.

[3] This was the essential theme of the recent FINEXTRA Future of Money conference, held in London in April 2015 (www.finextrafuturemoney.com). It was accompanied by SWIFT's InnoTribe tech entrepreneurship presentations, many of which focused on new financial technologies-related business models as well.

[4] Well-known systems include China's National Advanced Payment System (CNAPS), Hong Kong's Clearing House Automated Transfer System (CHATS), the Monetary Authority of Singapore's Electronic Payment System Plus (MEPS+), and the U.S. Federal Reserve Wire Network (Fedwire) (Committee on Payment and Settlement Systems 2011, 2012a; McAndrews and Rajan 2000).

[5] An anonymous reviewer pointed out to us that the typical range of dollar values for payments settled via DNS is on the order of $2,000 to $5,000, while for RTGS, it is about $2 million to $5 million, a difference of 1000 times.

In contrast, payment system innovations in the retail sector have not always focused on real-time settlement, although faster clearing has enabled banks to make unsettled funds available to their customers. According to Clear2Pay (McIntosh et al. 2014), *faster payments systems* are “[d]omestic, interbank, purely electronic payment systems in which irrevocable funds are transferred from one bank account to another and where confirmation back to the originator and receiver of the payment is available in one minute or less.” Various implementations of such systems have been in operation in countries in Asia, Europe, South America, Australia, and Africa, from the start of the 21st century up to the present.

1.3. The Research Conducted in This Study

In this research, we propose a *hybrid payment settlement system* that combines various functions of RTGS, DNS, and payment priority queuing (Wallace 2000, Committee on Payment and Settlement Systems 2005, Willison 2005).[6] This design recognizes the issues that banks face, and their incentives to adopt such a system. *Queue-augmented systems* sequence payments using a centralized management approach by a central bank or a third-party organization, or internal queue-based, payment-specific decentralized management approach by individual banks, as payments enter their systems (Peñaloza 2009, 2011). Such hybrid systems will have less delay for payments settlement than end-of-day netting systems, and will lower the liquidity needs of participating banks to a greater extent than RTGS does. Furthermore, in the centrally-managed queue, efficiency gains and cost savings can be achieved by consolidating the individual banks’ payment streams on a central platform. In association with the banks’ reserve accounts at a central bank, a centralized solution will have the capability to provide additional intertemporal liquidity, in the event that settlement demand causes a bank participant to be transiently short of funds, as well as if the participant fails.[7] It also is likely to diminish any bank’s incentive to delay its own payments to

[6] Willison’s (2005) work has been especially useful for the research we have done, since his article compares the performance of RTGS and hybrid payments settlement systems using simulation methods.

[7] A recent public disclosure by the Faster Payments Scheme Limited (2014) provides additional details on the essence of how liquidity issues can be effectively managed in advance of the times when they arise. “*The consequences of a significant central processing outage or of a failure of the settlement process (including member failure to meet settlement obligations) are severe enough to be afforded a range of controls to prevent the risk occurrence. These include service levels and monitoring, secure messaging, secure dual site processing and strict change control. To mitigate settlement risk, Faster Payments Members’ net settlement positions are limited using hard debit*

the system, and to not delay payments to other banks in a self-interested way. This reinforces the benefits associated with a centralized system.

Our proposed mechanism addresses issues for effective payment settlement system design from several different viewpoints: the technology infrastructure that supports the business process; the participation incentives of banks that handle payments; and the capability to implement market coordination actions for settlement decision-making to create business value for the banks and social value for consumers and the economy. The modeling approach that we propose considers: retail and wholesale payments, and related financial services in the economy; information that becomes available in the payment process; the role of IT as a solution for the digital intermediation of payments; and management science as a methodology that we can use to resolve some key economic issues in this problem space. We aim to answer the following questions: (1) What constitutes an efficient design for a hybrid payment settlement system that will support faster settlement of payments on average at low cost? (2) How does a hybrid payment settlement solution align the banks' incentives and payment submissions, and reduce their operational costs, while controlling the related credit and settlement risks? (3) How can such a proposed system be evaluated, so that it is possible to draw attention to the resolution of incentive problems that appear to have stalled some aspects of the adoption of faster payments systems?

2. THE VARIED DESIGNS OF FASTER PAYMENTS SYSTEMS AROUND THE WORLD

2.1. A First Premise: Understanding Incentive Issues Affecting Faster Payments System Adoption

In recent years, the accumulating pressure for the transformation of payment and settlement systems has pointed to several core issues that need to be identified and explained to understand what is going on.[8]

caps. The caps are partially collateralised as a requirement of the Scheme's Liquidity and Loss Share Agreement (LLSA). If a Member institution fails to settle, the LLSA also requires surviving Members to provide liquidity to meet any shortfall in the settlement obligations of the failed Member (up to the value of the largest Member). Surviving Members are subsequently partially refunded through liquidation of the failed Member's collateral. It is in the Scheme's 2014 Operating Plan that all collateral will be fully prefunded in cash by the end of 2014 eliminating any credit risk of default." This essentially represents a strong set of foundational agreements on liquidity management to ensure there are not overly-complex liquidity-related incentives problems.

[8] The design of a hybrid system is complex. Various operational and liquidity costs must be spread across many transactions when DNS is used, and the operational costs of handling individual transactions via RTGS is high. A design challenge is to achieve real-time speed at low cost, while ensuring liquidity for "anytime" settlement, so

Our view is that the *first premise* for understanding the drive toward faster payments is incentives.[9] For example, one is the *participation incentive* for banks (Arculus et al. 2012). Since every bank has its own unique operational requirements, customer base, and strategic approach to its market, it is unclear whether every bank will be interested in an RTGS or a DNS-RTGS hybrid settlement system. Although we will not model these aspects in such detail in this exploratory research, we nevertheless acknowledge their importance in practical terms to further increase the value of this research.

Other important questions related to banks' participation also arise. For example, how will the central bank's credit policy affect a bank's costs for intraday borrowing to fund liquidity shortfalls that may arise from the real-time settlement of payments? Will all of the participating banks be better off contributing some level of reserves to support liquidity pooling to add resilience to a central system? Should they immunize one another by contributing post-payment bailouts when liquidity shortfalls arise? And should there be any penalties or pricing for individual banks' use of liquidity in the central system that takes advantage of the contributions of member banks? Considerations like these are critical to identify the extent to which banks will have appropriate incentives to participate.

Another challenge is *incentive compatibility* (Selgin 2004). Since individual banks have private information about their payments, their customers and the payment risks involved, they will need to either make decisions about when to send their payments for settlement to a central queuing system, or to give the decision rights to some extent for managing their payments flow to the settlement intermediary. Under what circumstances will banks or the settlement intermediary have an incentive to delay payment submission or payment settlement? If so, what is the explanation for the behavior we may observe? Also, will decentralized submission of payments and uncoordinated decision-making cause a centralized payment

banks avoid having to tie up their funds by contributing funds to their central bank reserve accounts (Payments Systems Studies Staff 2000).

[9] This problem arises whether participation is accomplished through *settlement tiering* or *piggybacking* (Kahn and Roberts 2009). 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) has imposed minimum requirements for banks to participate, as opposed to avoiding participation through settlement tiering relationships with other banks.

management system to be 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 receives for final settlement. So an effective mechanism should take into consideration how banks will release payments for settlement and coordinate to synchronize their actions to mitigate possible failure.

The third issue is the management of *liquidity* (Cirasino and Garcia 2008, Johnson et al. 2004, Leinonen and Soramäki 1999, 2003). After payments are submitted, having an effective payment settlement rule is crucial for market liquidity, since it will affect how payments from different banks get settled. With liquidity created by payment pooling from participating banks, a centralized system, possibly managed by a digital intermediary representing the central bank, may provide funds that allow the system participants to economize on liquidity.

Successful development and implementation of an effective faster payments system requires a deep understanding of the economic incentives and business value that arise. A centralized payment management system, for example, will require the adoption of a *multi-sided platform* (Hagiu 2014). Retail and wholesale customers, merchants and banks, and government regulators all have a stake in achieving effective outcomes. So a systematic evaluation should consider the banks' participation incentives, and the decision rules that are implemented to structure how liquidity formation and the funding of liquidity shortfalls will be handled.

2.2. Learning from the Implementation of Faster Payments Approaches around the World

Preliminary observations. In a recent SWIFT Business Forum London, held in April 2015, in a brief public update on our research, we offered a perspective to suggest that the spectrum of past and ongoing implementations of the faster payments approach across different countries and over 20-plus years is helpful to understand both the incentive issues, and the characteristics of the different approaches that have been implemented. In related public discussion at FINEXTRA 2015 in London, a key question that the participants grappled with was: "Who will invest and build the technological infrastructure and update their legacy systems to make real-time payments possible?"

A couple of observations are useful. First, for clearing, the chief concern is the credit risk for the receiving bank to extend funds to customers when the payer's bank has not yet settled funds with individual customers. For settlement, liquidity risk is more important, especially when a large dollar value of unsettled funds may be subject to systematic risks in the market, and the possible failure of the payers' bank. Second, on top of these risks, it's never obvious in a "public goods" setting, such as faster payments systems, who can monetize technology infrastructure investments and systems updates. There is never short-term ROI, and the issue is about "play or pass," and "hook up or lose out."

The different implementations of faster payments approaches reflect the underlying incentives issues that the banks, the central bank and their respective stakeholders are dealing with in their countries. As a result, in terms of *settlement immediacy*, we see some countries that operate faster payments systems with settlement a few times per day (Brazil, Denmark and South Korea), whereas other countries operate in near real-time (Mexico, Sweden, and Poland). Another observable difference is the systems' *operating hours.* Some operate 24 x 7 x 365, while others support fewer hours and do not operate on weekends or holidays. Some systems prioritize payments by differentiating them based on value, some do so based on payment type, and some do not implement priorities for payments at all. The other contrast is the reliance on DNS versus RTGS, with more or less frequency netting. We will explore these issues in depth shortly.

Even in countries where the legacy payment systems have not yet been upgraded, discussions are occurring, policy-making exchanges between central banks and commercial bankers are in process, and infrastructure and bank-level systems renewal investments are being made (Kauffman 2015). Bankers are confronting the real threat of technology disruption and non-bank solutions in the digital business model space (PayPal, Bitcoin, etc.). Australia, for example, is in the midst of developing the New Payment Platform (NPP), which is projected to be operational in 2017 (Bolt et al. 2014). The U.S., albeit without a fully-developed plan, has also initiated a formal discussion (The Federal Reserve Banks 2013). The National Automated Clearinghouse Association (NACHA) in the U.S. has proposed a same-day automated clearinghouse (ACH) mechanism, which will be a stepping-stone toward a faster settlement system for retail payments (Oliver 2015). The U.S. has lagged in making this move, but the country is large, the institutions that need to agree upon the arrangements are many, the technology investment and process rework will surely be expensive (McKinsey & Company 2014), and the Federal Reserve has had other pressing matters to manage since the financial crisis of 2007 and 2008.

Global diversity in faster payments settlement system. The recent global efforts for transforming payment systems for faster payments settlement show diversity in the system design choices that have been made (McIntosh et al. 2014). Consider the fast settlement systems for retail payments, most of which went live since 2000. (See Table A1 in Appendix A.)

Some of these systems have increased the *DNS frequency* to multiple times a day. The frequency of intraday settlement ranges from every few minutes to several times a day. Some systems follow a blended approach of settling certain kinds of payments, such as low-value payments using RTGS and high-value payments using RTGS, or a blend of both. Other countries' systems are real-time, including Mexico's Interbank Electronic Payment System (SPEI, Sistema de Pagos Electronicos Interbancario), Sweden's Payments in Real-Time (BiR, Betalningar i Realtid), and Switzerland's Swiss Interbank Clearing (SIC) system. Substantiating our earlier comments, we observe operating hours for some systems that are restricted to business hours, but extend to 24 hours a day in other cases.

Variety of implementation factors for a faster payments system. The factors that play a role in the differences among faster payments systems around the global are related to the motivation to implement them, country-specific economy factors, and the legacy systems that are present. This final factor weighs heavily on the observed outcomes, reflecting the limits of financial resources available (The Federal Reserve System 2015).[10] In many cases, migration to a faster payments system is mandated by the regulators, the central banks and monetary authorities of the nations (McIntosh et al. 2014), but not all central

[10] An example of the recognition of the limited resources that are available in the financial services community for building new public payments infrastructure is exemplified in the recent Federal Reserve System (2015) report on the various options that are at its disposal to bring faster payments clearing and settlement to fruition in the U.S. The *American Banker / Bank Technology News* comments on the options, including: "*evolving the existing PIN debit infrastructure, which is currently used in retail stores and at ATMs, to enable real-time payments; using common*

banks have the authority to do so (Grover 2015).

Payments innovation is critical for the health of the financial system and economy in a country. In many cases, the regulators have mandated migration to a faster payments system (McIntosh et al. 2014), but not all central banks have the authority to do so (Grover 2015). However, banks often do not have the financial incentives or resources available to make investments in such a large-scale implementation, whose business value will be hard to appropriate (VocaLink / PricewaterhouseCoopers 2009). In most of the cases listed in Appendix A Table A1, the faster payments system was a response to a regulatory mandate. Even in these cases though, the motivation of the banks and regulators have differed. For example, Sweden's BiR became operational in 2012 under competitive pressure from non-bank payment providers in the industry. The threat of non-bank competition for the banks has been more salient in recent years, especially the start-up fintech innovators that have flooded the markets. Faster payments systems that have resulted from competition are often equipped with more advanced features, such as more timely settlement, support for e-commerce and mobile payments, 24 x 7 operations, and other capabilities.

The economic conditions in some countries may also create the impetus for implementing a faster payments system. In both South Africa and India, for example, the large populations that did not have access to banking services acted as obstacles for economic growth. *Financial inclusion* was a main goal in these countries to expedite payment settlement (Committee on Payment and Settlement Systems 2012b). According to the U.K.-based Payments Council (2015), the faster payments systems that are implemented also have needed to facilitate payment transfers via mobile technology, which addresses access for the unbanked population. Support for low-value payments has to be emphasized too.

Depending on the state of infrastructure at the time of implementation, faster payments systems have been achieved in several stages. South Korea rolled out HOFINET in 2001, and Brazil's SITRAF went live in 2002, both of which preceded other more recent implementations by a number of years. Neither

protocols and standards to facilitate the clearing of transactions over the Internet; building a new payments infrastructure that would build on existing technology and only have limited uses; or building a new payments infrastructure that would process a wider range of transactions" (Wack 2015).

system was built with the ISO 20022 standard in mind though. Nevertheless, in 2010 Brazil announced the plan to migrate its system to meet the standard. Switzerland, whose SIC was built much earlier in 1987, also only recently implemented the ISO 20022 standard. Given the basis of their domestic payments systems, South Africa and China chose not to support ISO 20022 in their faster payments systems, although they have committed to future migration toward this standard.

The uniqueness of priority queuing: Mexico SPEI and Switzerland SIC. These countries' payments clearing and settlement systems are unique in their implementations of *priority queuing*, a feature also incorporated in the hybrid system proposed in this research. In both SPEI and SIC, prioritized payments are settled on arrival at the banks if possible, whereas other payments are queued on the systems of a settlement intermediary. Settlement is conditional on the balance available in the participating banks' settlement accounts. If the balance is insufficient for settlement, the payments will be delayed to the next settlement period. Furthermore, the exact settlement sequence of payments may depend on other factors, and is determined by complex algorithms. In our proposed mechanism, we will take into account several key factors – such as payment priority, payment delay and payment value, as well as the availability of pooled liquidity – in determining whether payment settlement should be immediate or delayed. We also will analyze the payment-related decisions of the participating banks and the settlement intermediary.

Ongoing efforts in Australia and the United States. Recent efforts that are underway include the development of faster payments services in Australia, and increasingly close study with the payments system stakeholders in the U.S. In Australia, the initiative for building the New Payments Platform (NPP) received funding from a consortium of Australian financial institutions in December 2014, who were encouraged by the Reserve Bank of Australia (2012). They have been exploring innovation in the payments system, to move forward with faster payments. The motivation for the NPP has been driven by concerns about payments relative to the public's welfare, the global trend in the direction of faster payments, and the growing maturity of the underlying technologies. The NPP is designed to address critical issues in a faster payments system, including real-time settlement, 24 x 7 operation, richer messaging, and easier addressing (Bolt et al. 2014). The support also will include the basic infrastructure for an "Overlay Service" that will enable financial industry firms to offer specially-designed services to their customers if they wish (Australian Payments Clearing Association 2015). It will also provide a Reserve Bank of Australia-supported "Fast Settlement Service" (FSS), as well as an "Initial Convenience Service" (ICS). The latter includes a fundamental message set that makes it possible to provide availability for near real-time funds (Hume-Cook 2015).

In the U.S., policy-makers and industry participants are engaging in active discussions. In September 2013, the Federal Reserve identified major gaps in the current payment environment and the opportunity for near real-time payments system (The Federal Reserve Banks 2013). The U.S. has also taken initial steps towards expediting payments in a recent proposal by NACHA, which laid out the needs and the plan to enable same-day automated clearinghouse operations. Richard Oliver (2015), a former Federal Reserve executive vice president, commented on this proposal, stating that "*the changes necessary to implement same-day ACH are eminently doable in a reasonably short time frame, and that same-day ACH will create, if nothing else, a nice bridge to the ultimate nirvana of a real-time retail payments environment*."

3. ISSUES WITH THE DESIGN OF FAST PAYMENTS SETTLEMENT SYSTEMS

In different regions around the world, regardless of whether they are in place, regulators as well as practitioners have reached a consensus that faster payments are crucial for economic growth and the future health of a country's financial system (McIntosh et al. 2014). The World Bank carried out a study that involved 142 countries to examine the development level of their payments systems, such as the speed of payments settlement (Cirasino and Garcia 2008). Taking a global perspective, KPMG (2012) opined on how payment innovation spurred by economic growth in Asia may affect western countries, highlighting the new trend of accelerated payment systems. On the country level in the U.S., aside from the fruitful progress worldwide to speed up payments, the Federal Reserve Banks (2013, 2015) also have identified the need for faster payments and released the strategies for implementing technological innovations. (See Table A1 in Appendix A.) Meanwhile, the private sector has become more motivated to speed up payments in the U.S. (Pelegero 2013). The literature is helpful for understanding the design of settlement mechanisms in greater detail.

3.1. Strategic Thinking about Faster Payments Implementation

Successful implementation of a faster payments system requires the participation, coordination, and harmonious collaboration of banks. An example is the initiation of Australia's NPP implementation. It was a result of support by a consortium of 12 leading Australian financial institutions with the appointment of SWIFT as the vendor to build and operate the NPP (Australian Payments Clearing Association 2014, SWIFT 2014b). To identify the stakeholders and evaluate the potential for industry support, the U.S. Fed has researched the business case for profit contribution of faster payments for participating banks (The Federal Reserve Banks 2014). The conclusions are that, although implementing a faster payments system may be only profit-neutral to banks, its strategic implications are strong, and that engaging stakeholders is instrumental. In most countries where faster payments systems have become operational, regulatory mandates have been necessary to incentivize banks' participation (McIntosh et al. 2014). But this does not suggest that the banks must view such investment as a negative NPV project.

From the banks' perspective, it is important to recognize that the opportunity from participating in implementing the faster payments systems lies in their readiness for future payment innovations. It is not about near-term profit. By adapting to the faster payments trend, banks can be innovators, influence the emerging industry solutions, and leverage their IT capabilities to create new products for their customers (Pelegero 2013). The alternative is unattractive: falling behind with legacy payment systems puts banks at a disadvantage, as new competition heats up with new, non-bank market entrants. An understanding of the potential business value for different stakeholders is related to the adoption and implementation of faster payments systems in most countries.

3.2. Trade-Offs Involving Credit, Risk, and Business Value

We earlier documented a number of different cases of faster payments implementations around the world. Not all faster payments systems use immediate settlement because of the various trade-offs between RTGS and conventional DNS systems involving risk and business value. Through RTGS, payments are processed individually, and settlement occurs immediately with finality for the full amount of the transaction (Kahn and Roberds 2009). Although RTGS may reduce credit risk by avoiding short-term debt between the participants, there are higher operational risks involved, and RTGS creates the possibility of intraday liquidity needs not being met to smooth payment flows that are not synchronized. Also, in RTGS systems, either the banks themselves may provide intraday liquidity based on the reserves they put on deposit, or central banks may provide it for a fee, possibly via an intermediary. In both cases, the banks typically are required to back up the RTGS systems' liquidity risks by providing collateral to mitigate risk. When they give intraday credit to the bank participants, central banks assume some of the risk.

With DNS, payments are accumulated and settlement is delayed (Johnson et al. 2004). Netting is an efficient way to reduce the liquidity needs of a payment system, and diminishes some – but not all – of the risks from the banks' point of view. The delays in settlement create vulnerabilities for the financial system. It raises concerns for retail payments, as well as wholesale and other types of high-value payments settlement (Federal Financial Institutions and Examination Council 2004).

Bech and Garratt (2003) analyzed bank behavior under three credit regimes related to payments settlement: *free intraday credit*, *collateralized credit*, and *priced credit*. Among these three, free credit is not a viable option for most central banks in payment settlement systems due to risk and moral hazard. They reported that collateralized credit is 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, and concluded that it sometimes may be socially efficient for banks to delay payments. [11] There are benefits to synchronizing payments under priced credit. A related issue is how banks should coordinate their actions with one another. We adopt the priced credit approach for the present research.

3.3. Hybrid Payment Settlement Systems with Payment Queuing

[11] For a discussion of how banks handle the synchronization of payments inflows and outflows, and how this affects the timing of their payment submissions into a settlement system, see McAndrews and Rajan (2000), who describe this for Fedwire in the U.S.

Hybrid payment settlement systems that combine the various functions of RTGS, DNS, and payment priority queuing represent a possible solution to address the various trade-offs. They have been discussed in the literature since the 1990s. Johnson et al. (2004) proposed a DNS mechanism based on the settlement of queued payments related to incoming payment value rather than the account balance. Their mechanism reduces intraday credit extensions while modestly delaying the average time of settlement. They showed that a bank's preference for an RTGS or a hybrid system depends on how credit risk and liquidity efficiency trade off with one another.

In contrast, we consider the central queue as an extra source of liquidity. We evaluate the trade-offs among total settlement delays, the density of payment transaction processing demand, and the cost of borrowing from the settlement intermediary to address liquidity issues that may arise when settling funds.

3.4. Methods for the Study of Faster Payments Mechanisms

Prior research has assessed the benefit of payment settlement systems with simulation (Leinonen and Soramäki 2003, Johnson et al. 2004) and agent-based methods (Galbiati and Soramäki 2008). It also has examined the network topology of payments, and how payments to and from banks shift in the presence of market shocks (Soramäki et al. 2007). A major limitation in this line of work is that bank behavior is typically viewed as *exogenous to the system*. In reality though, banks 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.

In contrast, we believe that the banks' actions should be viewed as *endogenous to the system*, and will largely depend on the payment system design that has been implemented. Guo et al. (2007, 2012) 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 will focus on the hybrid system's mechanism design by considering the participating banks' actions, the volume and frequency of payment transactions to be processed, and how the central queue settlement rules are structured. We will also consider the secondary issue of the central bank's liquidity-related credit policy, while recognizing that the cost of funds has been relatively low during the past few years. We note that in other times gone by – and possibly in the future – the cost of credit may be much more important to how things actually work. Finally, we note that various implementations of hybrid settlement systems are possible, based on the design ideas that we will discuss. We will evaluate the performance of our proposed approach using *experimental simulation* – a technique that we have experience with in prior published research (Kauffman et al. 2008, Li and Kauffman 2012, Li et al. 2014, Shang et al. 2012)

4. MECHANISM AND MODEL DESCRIPTION

We next propose a *hybrid payment settlement mechanism* that combines the functions of DNS and RTGS, and implements a *central payment management system* (CPMS) that leverage computational power for minimizing settlement system-wide cost. The mechanism serves multiple functions, including setting payment priorities, optimizing payment settlement, and providing intertemporal liquidity in the presence of an imbalance of funds. The model design is grounded on our knowledge of settlement systems in practice, central bank and university-based research, and contemporary coverage of the business press and leading payments organizations, such as the SWIFT Institute and its business partners.

4.1. Key Elements in a Model for Faster Payments Mechanism Design Evaluation

Based on our review of faster payments settlement systems in the different countries around the world, we identified the following characteristics as key elements in our model.

Settlement frequency. We have observed different choices for *settlement frequency*, which typically is defined in terms of how often DNS-based netting occurs. This is a fundamental issue that needs to be addressed when designing the settlement mechanism. For example, Australia's New Payments Platform (NPP) is intended to provide dedicated services to facilitate settlement in real-time. Denmark's RealTime24/7 conducts netting six times a day; its newest development, StraksClearing, is targeted toward real-time settlement. In contrast, India's Immediate Payment Service (IMPS) is a DNS with three settlement periods a day, and Singapore's Fast and Secure Transfers (FAST) is a DNS that settles twice each day. Our hybrid system differs from RTGS, in which settlement occurs continuously as payments enter the system; our hybrid system also differs from DNS, in which the *netting frequency* defines the settlement frequency. Because central queue-based settlements occur periodically, we use the number of settlement periods in a day as a measure of *settlement frequency*. Some payments either are settled in real-time by the bank, or can be settled immediately by the settlement intermediary, without being delayed in the intermediated queue. (See Table B1 in Appendix B for the model's variables and their definitions.)

The payments clearing and settlement intermediary. The settlement systems in the countries that we reviewed typically are owned and operated by a *settlement intermediary*. In most cases, the intermediary is the country's national central bank, such as in Switzerland, or a consortium of financial institutions, such as in India and Australia. The role of the intermediary in such interbank settlement systems is critical: it builds the required platform infrastructure (and may fund its construction); it designs the rules for payments settlement; it manages and delivers settlement services; and it monitors the risks involved in the process. We include a settlement intermediary, though we choose not to specify its identity. It could be the central bank of the country, or a bank association, or even a third-party platform provider.

Settlement systems design. Payments settlement mechanisms for RTGS typically trade off the liquidity risk and settlement delays that banks experience (Manning et al. 2009, Schulz 2011, Willison 2005). Among them, Switzerland's SIC system utilizes a queuing-based settlement mechanism. A bank participant queues its payments and waits for them to be settled when the settlement period ends. In this mechanism, participants can manage the queuing sequence of their payments by assigning different priorities to each of them. Another example is Mexico's SPEI system, in which banks assign higher priorities to some payments and settle them immediately, and send lower-priority payments to the queue. The system periodically settles the DNS queue payments based on funds availability. Payments that cannot be settled due to unavailable liquidity are pushed to the next settlement period.

Our proposed mechanism includes payment queuing prioritization to some extent. Participating banks can differentiate the payments to process immediately and to delay; the latter type of payments then enter the settlement intermediary's system. We consider the *internal queue* that is managed by the individual bank and the *central queue* that is managed by the intermediary. For the central queue, it will be the intermediary's choice of which payments to settle first and which to queue until offsetting funds are available, while considering the priority assignment by the related bank.

Availability of reserves to fund liquidity shortfalls. Reserve funds that are available for a participating bank to draw upon to cover its intraday liquidity shortfalls will affect its incentive to delay or accelerate the settlement of payments. Big banks typically have large reserve funds available in their central bank accounts, so the settlement of low-value payments will be somewhat less of a concern compared to high-value payments are settled. And so long as they have healthy financial activities in other sectors of their operations, they should be able to support faster settlement of payments due to relatively lower pressure from liquidity shortfalls in their retail and low-value payments activities. Small banks, in contrast, will be more vulnerable to liquidity shortfalls that may require them to pay intraday overdraft penalties or borrow funds from the settlement intermediary in a faster payments settlement system. Our modeling approach takes into account the availability of reserve funds that can be drawn upon for each participating bank via the settlement intermediary, as well as the ability to pool such liquidity across the banks.

Number of participating banks. In recent years, banks' participation in most countries' faster payments clearing and settlement systems has increased. For example, India's IMPS has 51 member banks, up from its 4 founding banks, which expanded to 9 banks within a year, and 84 participants by 2015. The number of participating banks represents the size of the payments settlement system network; thus, more participants create greater network effects for all parties involved. With more banks, the settlement intermediary's liquidity pooling capability will be stronger, and will be able to support a faster payments settlement without undue fears about liquidity shortfalls (Allsop et al. 2009). Our model will examine how the number of member banks affects the performance of the proposed mechanism.

Demand for payments services. Demand for services to process a high volume of low-value payments transactions, or a low volume of high-value payment transactions may affect the performance of different payments settlement mechanism designs. Our model includes a variable that describes the density of payments services demand, which is proxied by the likelihood of a participating bank receiving payment requests to pay other banks in a specified time interval. We model the value of payments as a random variable from a pre-specified distribution, without creating an order of magnitude difference in payment values. This characterizes industry DNS and RTGS systems based on the evidences provided earlier.

Relevant costs in payments settlement. There also are various costs in the process that we view as exogenous variables. They are beyond the control of the participating banks. The first is *liquidity costs*, which occur when participating banks borrow money from the settlement intermediary. Related to this is *overdraft penalty costs*, which a participating bank incurs when its funds in the reserve account drop below zero. Finally, there are *delay costs* that apply to payments requests that are not processed immediately. They are increasing in the duration that payments have been queued for settlement.

4.2. How the Model Works

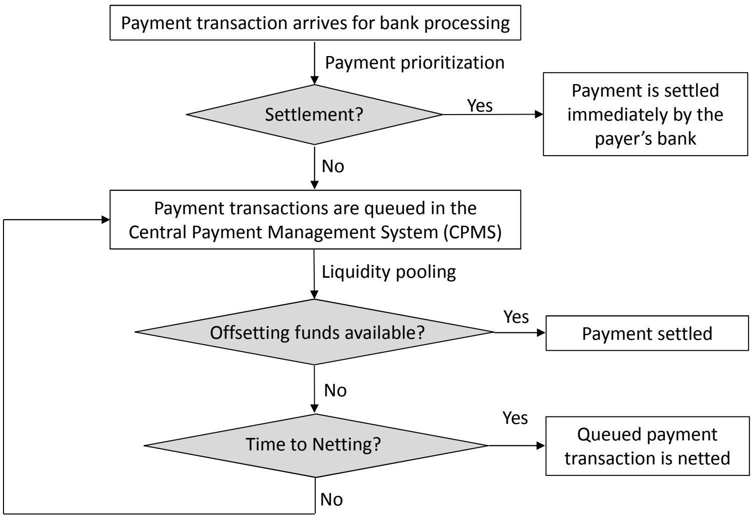
Liquidity provision is an interesting and important feature in our proposed faster payments settlement system. On the one hand, participating banks can borrow money from the settlement intermediary to strengthen their ability to meet payments settlement demand. On the other hand, the system also employs *pooled liquidity* from all of the participating banks through the intermediary's payment prioritization queue. Participating banks' payments are offset using pooled funds from all of the banks, some of which will be directed to the settlement intermediary to create its own liquidity.

Figure 1 shows the proposed *settlement process* for a payment transaction received by a bank.[12] We consider the actions and decisions of participating banks and the intermediary in each settlement period during a day. First, payment transactions arrive at the participating banks. Examples include a debit card transaction, a payment to a merchant, or a mobile payment – each involving the payment or receipt of funds. Each bank will need to set its own priority for making payments to settle different kinds of payment transactions. Payments identified as high priority will be settled in real-time if possible, while those

[12] We should emphasize that the banks will settle the payments immediately when they have the funds to do so, while other payments will be forwarded to the central queue. Similarly, the intermediary will settle high-priority payments immediately if there is available liquidity, or hold off on settling payment via the DNS queue.

of low priority will enter the intermediary's queue, and be handled with other banks' payments.

Figure 1. A Hybrid Payments Settlement Mechanism with Central Payment Queuing


Payment transaction arrives for bank processing
Payment prioritization
Settlement?
Yes
Payment is settled immediately by the payer's bank
No
Payment transactions are queued in the Central Payment Management System (CPMS)
Liquidity pooling
Offsetting funds available?
Yes
Payment settled
No
Time to Netting?
Yes
Queued payment transaction is netted
No

In each period, the current monetary value of payments received by the intermediary, together with reserve funds that are accessible to the intermediary, will supply the available liquidity for delayed payments that are queued. The liquidity pooling capability of our approach makes payments settlement faster on average, especially when many banks are participating.

Multiple payments can be settled simultaneously, as long as the central payment management system receives sufficient incoming funds. The settlement system determines which payments to be settled, with consideration given to the priority that the banks assign. For example, if the central payment queue is rank-ordered, then the top-ranked payments will be settled first. If the central queue is not rank-ordered, then the payment transactions to be settled will be determined differently. This can be by time or by value, or some other desirable criterion. All unsettled payment transactions will remain in the queue, either to wait for offsetting funds in the next settlement period, or to be processed when the final period DNS netting occurs.

We now present the optimization problem for the settlement intermediary and the participating bank, and explain the constraints that are involved.

4.3. The Optimization Process of the Settlement Intermediary

The intermediary's objective in each settlement period is to minimize the total costs of the settlement system, which is based on the delay costs for all payments and the total cost of funds to cover the liquidity shortfalls of all participating banks. It is important to track two key variables: (1) the *net settlement amount for the bank*, and (2) the *delay cost per payment.* The delay cost arises when the intermediary delays making these payments in a settlement period, while the total cost of funds comes from borrowing liquidity to support the net settlement value in the period.

In each settlement period, the intermediary makes decisions based on a set of binary decision variables that represent choices for each payment transaction in the queue: 1 for settlement in the current period, and 0 for delayoing settlement to the next period. All payments assigned "1" will be processed and removed from the queue, with no delay costs imposed. These payments will be used to calculate each participating bank's *net settlement amount* in the period:

Bank's Net Settlement Amount = Total Payments It Makes to All Other Participating Banks

– Total Payments It Obtains from All Other Participating Banks

If the bank's net settlement amount is positive, then it will need to borrow from the intermediary to cover this liquidity shortfall, so it will incur a liquidity cost. If its net settlement amount is negative, the bank will have sufficient funds to cover the payment transaction, and does not need to borrow.

Payment transactions with a "0" assigned will stay in the queue and the related will incur a *delay cost for settlement*, as follows:

Delay Cost for Settlement = Unit Delay Cost · Time Delay · Payment Amount

The reader should think of the delay cost as a factor that will affect the bank's reputation for the speed of settlement of payments relative to other banks.[13] The *unit delay cost* measures the sensitivity of delay of a payment. The higher the payment amount, the longer the time delay of the unsettled payment, the

[13] It is probably not useful to think of this in terms of the time value of money, but rather as a delay in making funds available to a bank's customers that will have some effect on the funds they have available to use. We have not tried to model this additional detail. We are currently exploring other ways to handle this aspect of our model, since the assignment of delay costs will be quite arbitrary in practice. This is a useful first step though.

higher the total delay cost associated with the payment.

One constraint applies to all banks simultaneously: the total amount borrowed by all banks cannot exceed the total reserve funds that are available at a given point in time for lending by the settlement intermediary. (For the full mathematical details of the model, see Equation 3 in Appendix C.)

4.4. The Decision-Making Process of Participating Banks

Individual banks either will handle payments via their own internal payment queue for settlement, or they will submit them with some prioritization to the intermediary's queue, called the central queue. Here, we will focus on decision-making related to the central queue. (Refer to Equation 1 in Appendix C for decision-making for payments in the banks' internal queues.)

In each settlement period, each participating bank may receive a request to pay another participating bank a random *payment amount*. When payment requests arrive, a bank will examine then and assign different priorities. In our model, we include an exogenous *threshold-value priority rule* for banks. If an incoming *payment request amount* is greater than the threshold value, then it will be treated as a high-priority payment, and will be settled immediately by the participating bank regardless of the bank's reserve balance. If the incoming *payment request amount* is lower than the threshold value, then it will be treated as a low-priority payment, and the bank will determine when to settle the payment based on availability of funds.

The bank makes two decisions. The first, as we have just explained, will determine payments that should be settled immediately, and the payments that should be submitted to the central queue. As a result, each participating bank will have two sets of payments: a set of *immediate payments* and a set of *delayed payments*. Only the latter will incur delay costs; the former will not. The second decision is how much the bank will pay the intermediary back in the period, if funds have been borrowed to cover the liquidity shortfall for settling payments. As a result, the bank's optimization problem in each period will be:

Minimize: Total Delay Cost of Unsettled Payments + Total Cost of Funds

where

Total Cost of Funds = Unit Overdraft Penalty Cost × Bank Overdrafts

+ Unit Liquidity Cost × Debt

The first term is the bank's total payment delay costs. This involves a summation of the delay costs over all unsettled payments that are identified as being of low priority and submitted to the central queue.

The second term is the bank's total cost of funds, which consists of two parts. The first part is the funding cost for overdrafts, which are incurred only if a participating bank's reserve account balance drops below 0. The bank's reserve fund balance at the end of the period is equal to the bank's reserve account balance from the last period minus the high-priority payments that are settled by the bank itself immediately and the low-priority payments that are settled by the intermediary in this period, and also less the amount paid back this period. The second part is the bank's borrowing cost from the central queue to fund payment liquidity shortfalls. This is calculated based on the difference between the bank's debt to the intermediary from last period and the amount that is paid back this period.

There is one constraint in this optimization problem: a bank will never pay back to the intermediary more than the amount it owes. As a result, the amount paid back in the current period will be less than or equal to its debt to the intermediary in the prior period. (For details, see Equation 2 in Appendix C.)

4.5. System Performance Measures

Meeting the demand for liquidity to cover payments, and the speed they are settled are important criteria for evaluating payment settlement performance. We define several measures for this purpose.

Normalized delay index (*NDI*). *Settlement delay* for a payment is the difference between the time the payment it is received by a bank and the time it is settled. To measure the system-wide *settlement delay*, which covers all payments and all participating banks, we propose a *normalized delay index* (*NDI*), a ratio between 0 and 1. Its numerator is the total delay cost incurred for all delayed payments in the proposed settlement mechanism, and the denominator is the total delay cost for using an end-of-day DNS system rather than what we propose. For RTGS systems, where payments are immediately settled, *NDI* is 0, while for end-of-day DNS systems, *NDI* is equal to 1. This enables us to set up a comparison.

Average funds required (*AFR*). An *overdraft* occurs when a bank's reserve account balance falls below zero, when payments are settled. The Federal Reserve Bank measures overdraft positions at banks in the U.S. at the end of each minute of the day, and with this information, it is able to compute a bank's *average daily daylight overdraft* (Governors of the Federal Reserve Bank System 2012). Two sources of funds cover liquidity shortfalls in our proposed mechanism: the central bank reserves overdrafts and the liquidity funds from the settlement intermediary. We compute the average funds required across both sources. Other approaches are possible.

Average funds transfer (*AFT*). Another way to evaluate the performance of a payment settlement mechanism is to assess how much the participating banks' reserve account balances fluctuate. For this purpose, we define the *average funds transfer amount* (*AFT*) for all banks. For each bank, we calculate the difference between its fund transfers in two adjacent periods, and then sum over all periods and all banks. This amount is averaged based on number of banks in the network and the number of periods that are under study. This measurement is a useful measure of system-wide variation in account balances. (Details of the measures and the implementation of the algorithm are provided in Appendices D and E.)

5. CONJECTURES AND EXPERIMENTAL SIMULATIONS

A modern payments settlement system should be more than just faster. It should be more flexible in handling payments, more cost-effective, and lower in operational risk exposure. We developed a set of conjectures that are intended to act as leading questions for the evaluation of our proposed mechanism. We also developed experimental simulations to see whether our conjectures can be validated.

5.1. Conjectures

We explored various design issues to understand how high performance with the proposed settlement system can be achieved. This includes understanding the conditions under which the central queue creates large benefits, and the appropriate pricing for liquidity to align the incentives of participating banks. We then present a broader view of the simulated performance for different payment networks.

Design options. There are several options for payment transaction queue design. One choice is that banks manage and settle all the payments via their own internal queues rather than submitting their payment transactions to a central queue. When this is the case, the potential benefits of liquidity pooling with other banks and the central queue's liquidity provision will be lost. This makes our first conjecture worthwhile to evaluate:

- **Conjecture 1 (Queue Design Options).** *A central queue will perform better than an internal queue for faster payments settlement.*

A key benefit of the intermediated settlement is that it supplies liquidity to reduce interbank settlement pressure when funds received and funds to be paid are not balanced. When demand for payment transaction processing is high and a bank's reserve account is low in funds, the bank will experience pressure if it wishes to settle payments quickly. The benefit will be greater when the level of the liquidity that intermediary can provide is higher. In addition, as more banks join, the system's liquidity pooling capability will be stronger – a positive network effect, which also should occur when larger banks join. We expect that more payments can be settled faster on average in such a system. With this in mind, we offer:

- **Conjecture 2 (Intermediary Performance and Liquidity Provision**). *A hybrid payments settlement intermediary will settle a higher proportion of total payments in number and volume when payment demand is higher, the banks' reserves are lower, and the number of banks is larger.*

Effective operation of the system relies on well thought out incentives to effectively coordinate the participating banks' actions. There are trade-offs though. When it is less costly to borrow from the settlement intermediary than to endure overdrafts to settle payments, individual banks will not have an incentive to pay the funds back.[14] As a result, the intermediary will run out of liquidity, since funds will not be replenished fast enough. On the other hand, if the intermediary's liquidity costs were too high, it will not be justified to rely on its liquidity provision to settle delayed payments with the trade-off between delay cost and the cost of funds, as individual banks make decisions. This leads us to make a conjecture about the pricing of funds to help banks with liquidity to support faster payments.

- **Conjecture 3 (Intermediary's Pricing of Liquidity Funding).** *For liquidity funding provided to*

[14] Too high a price for central queue-based payment settlement services will discourage bank participants to queue their payments with the settlement intermediary. Too low a price will discourage them from paying back the funds they have been granted through overdraft credit. The relationship between the overdraft penalty cost and the settlement intermediary's liquidity cost will cause the banks to either increase or decrease its central bank reserves.

the banks by the intermediary, a cost per unit of funds that is slightly higher than the penalty cost per unit of the overdraft will help to achieve faster settlement on average.

System performance. We observe that countries such as Switzerland and Mexico have adopted priority queuing systems to support near real-time settlement of retail payments. In contrast, the U.S. has been late to implement faster payments. We conjecture that the concentration of payments transaction processing services among the banks will play a role in system performance. Because our proposed system requires harmonious coordination among the banks to improve central queue performance, the economic gain of the proposed system will be higher when there are a smaller number of banks with relatively higher concentration of demand for services. Thus, we offer:

- **Conjecture 4 (Concentration of Demand for Payments Transaction Processing).** *Other things being equal, intermediated payments settlement will perform better in a payment network that has a small number of banks with more concentrated payment services demand than in a network with a large number of banks and less concentrated payment services demand.*

As we noted earlier, there is consensus in the payments industry that real-time settlement can be an expensive endeavor (McKinsey & Company 2014). As the volume of payments increases, a country may benefit from simultaneously increasing the number of intraday settlement windows to control systemic risk while creating improved funds release protocols for payees in payments transactions. It is natural to expect that increasing settlement frequency will reduce settlement delays in the payments system on average at the expense of higher funding costs for the banks. However, to what extent will more frequent settlement be justified in terms of the relevant costs and benefits? Will faster payments settlement occur via the central queue on average, or will individual banks take the lead? For this, we will explore:

- **Conjecture 5 (Settlement Frequency).** *Other things being equal, higher settlement frequency reduces settlement delays and increases the total payments, but the required funds to cover liquidity shortfalls increase, and more settlements will be handled by individual banks.*

5.2. Experimental Simulations

We propose two sets of *experimental simulations* to understand the effects of key design factors on settlement system performance. The first set is a *controlled experiment* (see Table 1). The second set utilizes three specially-designed networks that will allow us to evaluate the performance of the central queue approach in those networks. (The detailed experimental set-up is provided in Appendix D.)

Table 1. Experimental Treatments

MECHANISM DESIGN	EXPERIMENTAL CONDITIONS	EXPERIMENTAL TREATMENTS
Payment System	Number of banks	I = {Small, Large)
	Number of settlement periods	T = {Short, Long}
Participating Banks	Payment density	P = {Low, Medium, High}
	Reserve account funds	B = {Low, High}
Central Bank	Cost of funds per unit of overdrafts	δ from 0 to 0.08, step size 0.02
Settlement Intermediary	Cost of funds per unit of liquidity	ω from 0 to 0.08, step size 0.02
	Liquidity provision	c = {Low, High}

A *payment system* is defined in terms of the *number of participating banks* (I) and the *number of settlement periods* (T). Each bank receives *payment transaction processing demand* during a settlement period. Payments arrive stochastically – in other words, in random fashion according to an identifiable probability distribution. We operationalize system-wide payment demand in terms of bank-level *payment density* (P). This refers to the likelihood that a bank will send a payment transaction to each of the other banks in a given settlement period. A higher payment density implies higher demand for payment services from each bank. As Table 1 shows, our experiment consists of 12 (= 2 × 2 × 3) main configurations based on three variables: the number of participating banks, the number of settlement periods, and payment density. The different instantiations of the three variables define the different payment network configurations.

Central banks require their member banks to hold reserve account balances. An individual bank may incur an overdraft penalty cost for the central bank to settle its high-priority payments if it does not have sufficient funds available in its reserve account. Our proposed settlement mechanism offers an alternative: to provide liquidity to offset delayed payments in the central queue. By offering funds to enable earlier release of payments on average in the queue, the speed of settlement will increase, which ought to trigger the further release of other payments involving other banks. Because the two sources of liquidity are substitutable, the cost and the level of each will affect the banks' incentives for faster payments.

To focus on the trade-off between the central bank's credit extension and settlement intermediary's liquidity provision, we tuned the experimental simulation so that the cost of funds per unit of the overdrafts δ and the cost of funds per unit of liquidity ω vary in the interval [0,0.08] with step size 0.02.

There are two levels of treatment to assess the effects of bank borrowing to mitigate the effects of liquidity shortfalls: each bank's central bank reserve account balance *B*, and the intermediary's liquidity provision *c*. We will perform an experimental simulation with low and high levels for both parameters.

We define a *concentrated payment network* as one in which a large number of payment transactions occurs among a small number of banks with high probability. We ask: Will a concentrated payment network have an advantage for faster payments performance over one in which demand for a similar number of payment transactions is spread over more banks? We start with a *baseline network* to explore this:

- **Network 1 (Low Payment-Concentration, Low Settlement-Frequency Network).** *This network has a larger number of banks with lower payment density and lower settlement frequency, and provides a baseline for making comparisons.*

We will compare the baseline network to two other contrasting networks with the same expected total payment demand to test our conjectures about payment concentration and frequency:

- **Network 2 (High Payment-Concentration, Low Settlement-Frequency Network).** *This network has a smaller number of banks with higher payment density, and the same settlement frequency as Network 1.*

- **Network 3 (Low Payment-Concentration, High Settlement-Frequency Network).** *This network has the same number of banks as Network 1, with half its payment density and twice its settlement frequency.*

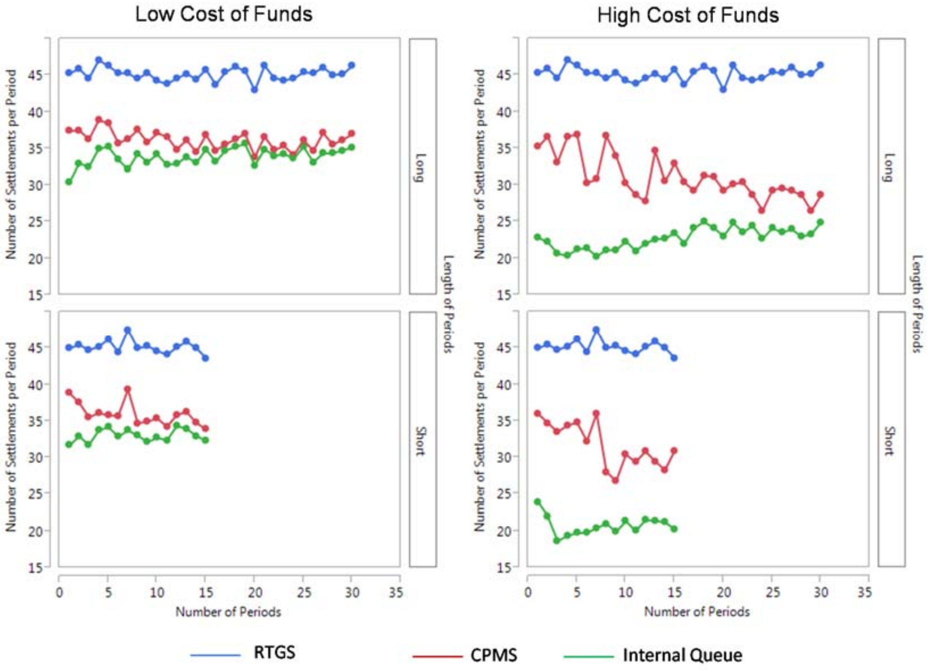
6. EXPERIMENTAL SIMULATION RESULTS

In this section, we present some key experimental simulation results to verify our conjectures, and offer the related insights. More details about the experiment set-up, the implementation procedure and analysis of other results are provided in Appendices B to G.

6.1. Results on System Design

When all payments that arrive in a period get settled without being delayed to the next period, this is a form of near real-time settlement that resembles an RTGS system. The RTGS system serves as a useful benchmark based on which we demonstrate our queue-based system performance. Figure 2 illustrates the settlement process for the internal queue and the central queue using RTGS as a benchmark.

Figure 2. An Illustration of the Settlement Process with Three Different Mechanisms

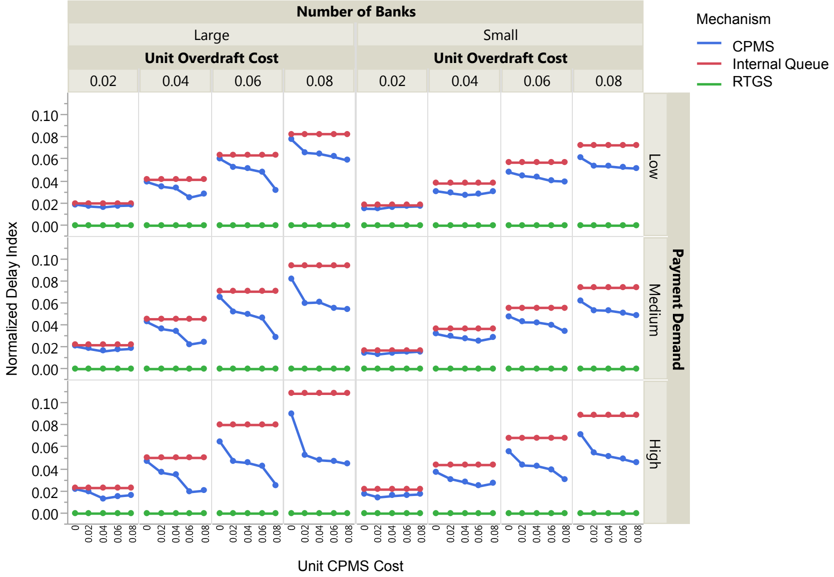


We focus on the case when the banks have low reserve balances on deposit, so they will have some borrowing needs to fund liquidity shortfalls. The *cost of funds* includes a bank's *overdraft penalty cost* due to its reserve account overdrafts. The penalties are based on how long the overdraft occurs and its magnitude. Another is the *liquidity cost* that the intermediary levies for lending money to address a bank's liquidity shortfall. For comparison, we set these costs at the same rates so settlement via the intermediary does not impose additional costs on the banks, but provides an opportunity for them to acquire liquidity when they need it.

The performance of the internal queue (green line) and central queue (red line) are similar when the cost of funds is low, and both are lower than RTGS (blue line). When the cost of funds is high, the central queue is clearly preferred over the internal queue. We did not analyze the initial periods because of the initial condition effect that pertains to constructed stochastic processes. In the central queue system, initially, the intermediary settles more payments because of its provision of liquidity. The number of payments that are settled decreases due to the decreased liquidity. In contrast, the internal queue settles a smaller number of payments initially because the low-priority payments enter the queue and do not incur high cumulative delay costs yet. As time goes by, payments that entered the queue early will accumulate higher delay costs and get settled by individual banks. Thus, we observe an increasing trend early with internal queue settlement, but the two settlement processes seem to stabilize at different levels in the later periods. In general, the intermediated central queue settles more payments than the internal queue in almost every time period.

Next, we will show the differences in performance for the internal queue and central queue with variation in the values of the key variables for the different system configurations. Figure 3 again compares the performance of RTGS (which is always essentially zero delay), the central queue and the internal queue settlement mechanisms for the normalized delay index (NDI) measure.

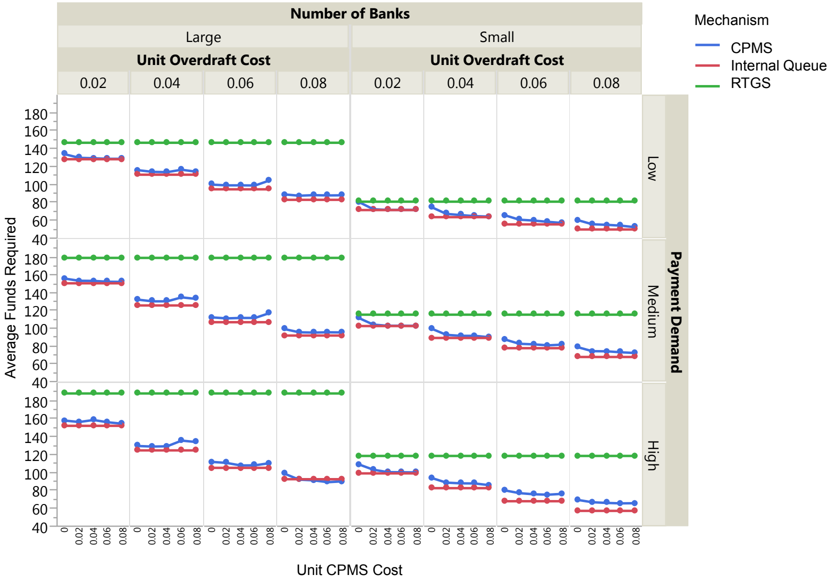
Figure 3. Comparison of the Settlement Mechanisms Based on Normalized Delay Index (NDI)



The central queue outperforms the internal queue through the reduction of payment delays. The effect is more significant when there is a larger number of participating banks, the banks have higher demand for payment transaction processing, and the cost of funds to cover overdrafts is high.

Figure 4 presents the average funds required (AFR) to cover the liquidity shortfalls under different scenarios.

Figure 4. Comparison of the Settlement Mechanisms Based on Average Funds Required (AFR)



Because the RTGS has zero delay, it causes the highest average level of overdrafts, as Figure 4 shows. In most cases, a central queue significantly reduces the average funds required in comparison to RTGS. Its performance is either comparable to or slightly exceeds the performance of the internal queue-based settlement mechanism in mitigating the occurrence of overdrafts.

Overall, our analysis confirms the Queue Design Options Conjecture (C1). With a large number of banks, each with high demand for payment processing, and a high unit cost to fund overdrafts, the intermediated central queue achieved the greater performance improvement over the internal queue. It reduced the delay for settling payments and lowered the average funding cost simultaneously.

Figure 3 also supports the Intermediary's Pricing of Liquidity Funding Conjecture (C3). We observe that, for a given level of overdraft penalty cost, the lowest delay occurs when the intermediary's liquidity cost is slightly higher (at the next experimental level) than the overdraft penalty cost. But do we observe a positive network effect? Figure 5 provides an indication.

Figure 5. The Effect of Network Size on Payments Settlement Delay

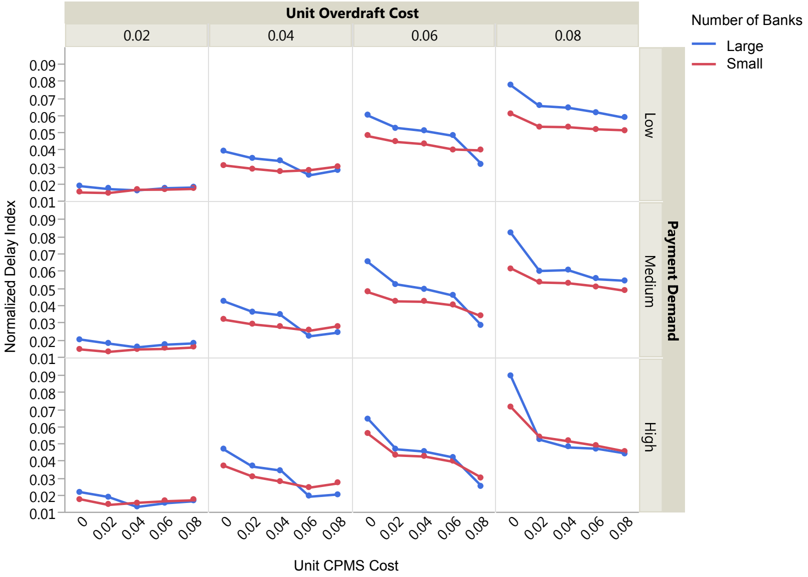
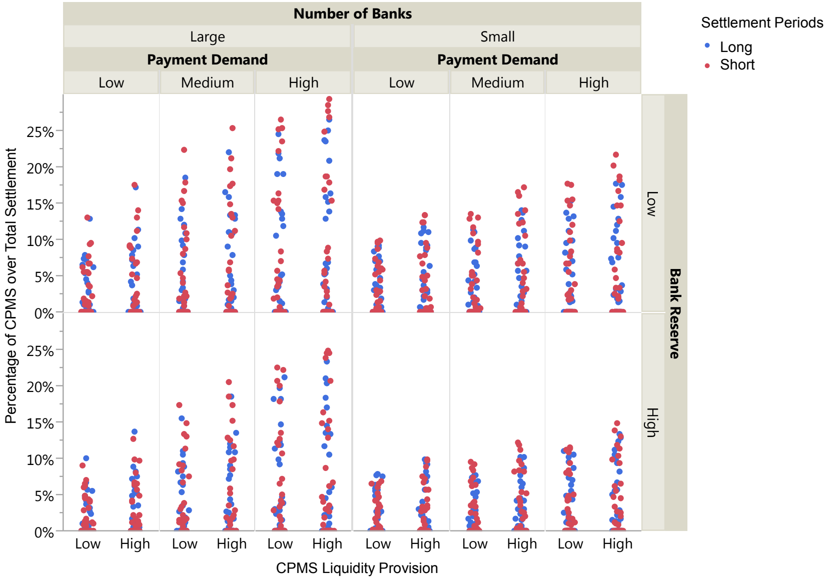


Figure 5 shows that large networks outperform small networks (red line) when the unit liquidity cost is relatively higher than the cost of funds per unit of the overdrafts that occur. Under optimal pricing, as suggested by Conjecture C3, we observe positive network effects. In general, small networks perform better than large networks in speeding up settlement when the cost of funds is relatively low. This reveals the importance of system configuration. If the settlement system is not properly designed, the system may not permit the desired business value to be appropriated by its participants. More generally, value appropriation is a difficult issue in shared system and platform investment settings.

The next issue to explore is how the central queue's performance varies with different levels of liquidity that it can provide. Figure 6 shows the ratio of intermediated payments settled to the total number of payments settled under different scenarios. Each cluster of dots represents a specific system configuration characterized by: the number of banks, the payment services demand, the bank reserve funds on account, and the ability of the intermediary to provide different levels of liquidity, up to its limit. Each dot in the cluster represents an aggregate performance measure over 30 random simulations based on different combinations of the unit overdraft penalty cost and the intermediary's liquidity cost for loaned funds to cover payment-related overdrafts. The colors of the dots are intended to distinguish between results that were obtained for shorter and longer settlement periods. We observe the same trend with respect to the volume of settlements due to the prioritization of different payments.

Figure 6. The Effect of the Intermediary's Liquidity Provision on Payments Settlement



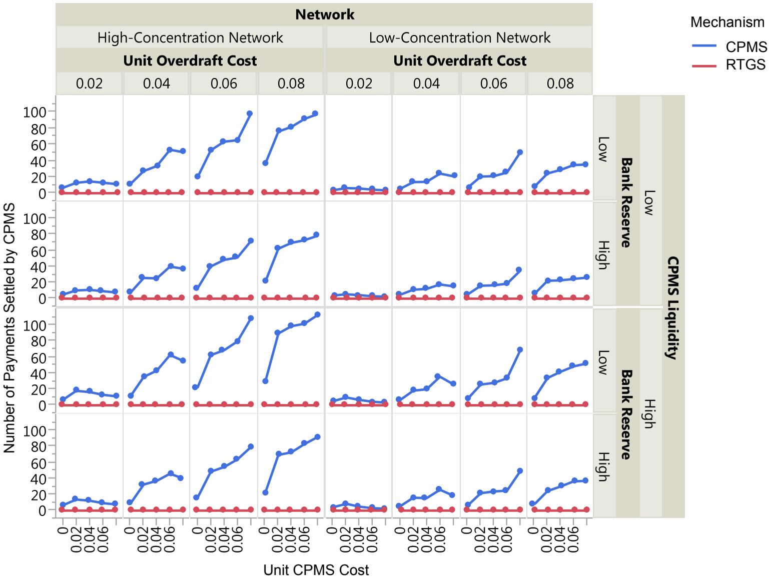
Let's now look at the details now, by comparing the first three columns with the last three columns in Figure 6. We observe that a higher percentage of settled payments is achieved via the intermediary's central-queue-based settlement process as the payment processing demand increases.

When the number of banks is large, the central queue achieves a higher proportion of payments settlement. Comparing the upper panel with the lower panel, we also observe that bank reserves play a role. When bank reserves are low, the effect of central queue-based settlement is greater. Comparing the two clusters in each sub-graph, we see that the intermediary settles more payments when its liquidity provision is high. In our simulation, it handles as much as 30% of the total payments, which is significant.

6.2. Comparative Analysis of Three Simulated Payment Networks

We designed low-concentration and high-concentration bank payment networks to assess what happens to central queue performance when the number of payments and transaction values in these two networks are compared. Figure 7 presents the number of payments settled by the intermediary.

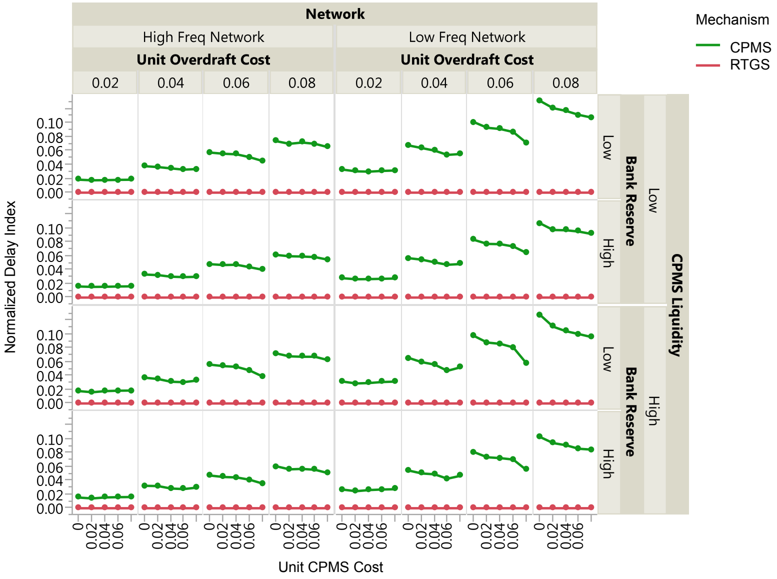
Figure 7. The Effect of Payment Concentration on Central Queue-Based Payments Settlement



The graphical presentation of the results suggests that the high-concentration payment network relies more on central queue settlement than the low-concentration payments network does. Figures F1 to F4 in Appendix F show that payment delays are reduced and average funds required to cover the liquidity shortfalls are lower in the high-concentration network, suggesting our mechanism's superior performance relative to its support for the low-concentration network. The Concentration of Demand for Payments Transaction Processing (C4) is supported.

If two payment networks have comparable numbers of payments and transaction values, it is natural to expect that higher settlement frequency will lead to lower settlement delay, as shown in Figure 8.

Figure 8. The Effect of Settlement Frequency on Payments Settlement Delay



In terms of speeding up payments settlement, our experimental simulation shows that the high-frequency network outperforms the low-frequency network under all of the scenarios that we tested. The performance improvement appears to be more significant when bank reserves are low and the cost of fund is high. We also see in Figures G1 and G2 in Appendix G that, although the high-frequency payment network settles more payments than the low-frequency payment network does, the number of payments settled by the central queue in the high-frequency network is lower there than in the low-frequency network. This suggests that individual banks will handle more payments. As a result, it appears that the banks bear more settlement pressure if settlement frequency increases, which increases the average funds required to cover the liquidity shortfalls. (See Figure G3 in Appendix G.).This may explain why banks do not necessarily prefer to settle faster. Thus, our experimental simulations also supported the Settlement Frequency Conjecture (C5).

7. DISCUSSION

Today, in many countries around the world, new economy firms and existing non-banking institutions are actively pursuing payments systems innovations, and entrepreneurs are developing fintech start-ups.

Consumers, meanwhile, increasingly see the potential for greater convenience of payments systems that are entirely electronic and do not require financial intermediation from banks for many kinds of transactions. There is rising demand, as a result, for blockchain-based crypto-currency solutions, such as Bitcoin and Ether. But the overwhelming majority of transactions of this sort still require other specialized digital intermediaries, such as Circle, Coinbase and BitPay – and even then only a relatively limited number of large companies have begun to accept crypto-currencies directly. The time is coming when the market pressures on banks and the demand from businesses and public organizations inevitably rise higher, creating a groundswell of support for change.

To respond to future challenges imposed by innovations in payments, banks in different countries worldwide need to improve their current payment systems. Faster payments settlement represents one such direction to invest in technological advances to build new infrastructures that can support faster clearing and settlement of payments. However, faster settlement speeds create issues: For example, imbalance of payments inflows and outflows, and short-term liquidity difficulties. Because smooth settlement of funds among banks and other financial institutions is so important for a stable financial system, making disruptive changes in the current ways payments are settled or moving to real-time gross settlement without carefully evaluating the consequences have been viewed as unacceptable. Our research, thus, focuses on a hybrid payments clearing and settlement mechanism design that aims to improve settlement speed without incurring high funding costs and creating operational and financial risks. We demonstrate that a central queue managed by a digital intermediary can be an effective solution to address the challenges for making faster payments work.

On the path toward faster settlement of payments, central banks need to make clear commitments to maintain stability in the system. Traditionally, central banks have been the “lenders of last resort” and guarantors that absorb shocks from money in the economy, payment systems operations and performance, and the condition of financial institutions.

Our research examined the effects of central bank reserves, overdraft credit-related cost of funds, and the settlement intermediary-based liquidity provision on the banks' settlement behavior and system stability under different market conditions and settlement mechanisms. Our results show that an intermediated central queue can play an important role in coordinating faster payments settlement on average without harming system stability. New efficiencies can only be achieved when faster settlement services appropriately price their pooled liquidity. Banks need to operate with incentives that are aligned with the system's objective of speeding up the settlement of payments in the most cost-effective way.

We remind our readers that faster payments settlement system design is more complex than what we have considered in this research – yet the ideas we have shared can produce deep and useful insights. First, we modeled the cost of overdrafts when a bank's central reserve account balance becomes negative. Central banks generally provide priced overdrafts coverage to banks, though there are various forms of how this is done, depending on the country being considered. So developing a detailed understanding of central bank operations for implementing overdraft credit are worthwhile for future research.

Second, we conceptualized payment settlement delay cost in terms of the value and time it takes for the funds related to a payment transaction to be settled. This is a simplification: there is an assumption that different banks will be able to use such observable variables as the time duration, the customer type, the payment amount, and so on to approximate a value drawn from a distribution for the variable we specified for payment priority. This specification is weak: bankers will not be easily able to implement it. So we are continuing to explore other more easily-implemented ways to determine payment priority that are realistic for practitioners. Such an approach might be to discretize the payment delay sensitivity level according to the payment channel (high, medium, low), consider the characteristics of the payee or the payer, or include other appropriate information that is available to the bank. Banks then can automate how payment settlement delay sensitivity is assigned to make the payment prioritization decisions machine-based, and not subject to the vagaries of human judgment. Furthermore, we assume random payments and delay coefficients were drawn independently in our simulation. Presumably these two variables can be either positively or negatively correlated. For example, higher value payments have higher delay cost, and so on. This limitation is largely due to our inability, at present at least, to access data on payments from organizations such as banks or a central bank in a country. If such data on payments were available – for example, via a central bank, multiple domestic banks, or via SWIFT or a national faster payments services provider – we would be able to implement a richer and more defensible model to support insights based on observed payment transactions. We plan to explore these directions in our future work.

Third, we also made each decision period discrete and assumed the banks' and the intermediary's settlements occurred after a fixed duration of time. This eased our settlement tracking and account updates in the simulation. But again, we must own up to its limitations. For example, the longer the time lag between clearing and final settlement with netting, the higher will be the potential delay cost and settlement risk associated with unsettled payments in the system. Shortening the duration of time to final settlement when no transactions can be pushed for settlement in the next period prevents the further build-up of unsettled positions, and reduces delays while mitigating risk. Alternatively, we can directly use the clock time to define each period.

As a first step to understand the factors impacting the fast payments settlement systems design, we take a controlled experimental approach that focuses on a few key design parameters, such as the number of banks, the payment density, the number of periods, etc. We only consider banks with similar sizes and payments transactions in our simulation. We do not model heterogeneous banks where few key big banks play more influential roles or being more liquid than small banks in the payments eco-system. We only focus on low-value retail payments. So catastrophical events like default are unlikely to happen. However, we note that these are realistic problems with the higher-value payments settlement. This is the direction of our future research in this area.

Future faster payments settlement system design requires smooth delivery of funds from payments transactions anywhere and anytime. Central banks play a crucial role in streamlining and stabilizing the settlement process because they provide the means to ensure that banks can achieve "finality" in the settlement of payments. Our research also suggests the need for central banks to find ways to incentivize banks in their countries to improve their payment services and facilitate completion of payments.

8. CONCLUSION

In this research, we constructed a mathematical model and performed experimental simulations to offer insights into the impacts of the key design factors on our proposed payments settlement system. This work offers value and delivers new, important messages for technology platforms and platform strategy. The literature on technology platforms (Eisenmann et al. 2006, Evans 2011) has had a strong focus on network effects. And the most recent readings by experts in key areas of tech innovation, such as the Internet of Things, make such effects even more evident (Regalado 2014). Whereas existing studies show that a larger network of participants is critical for the growth, profitability, and competitiveness of a platform, our findings further articulate the dependencies of such network effects in the payments context. For instance, a smaller network of banks can contribute to the superior performance of the payments system when the payment services demand is more concentrated. Moreover, other factors may support network effects, such as payment services demand.

Our research also provides a new angle for examining platform pricing issues, a key element of successful platform strategy (Bonchek and Choudary 2013). The literature analyzes different types of fees on platform pricing, such as entry fees, fixed fees, interchange fees, per-transaction royalties, and two-part tariffs (Rochet and Tirole 2002, Armstrong 2006). We consider financial costs such as the price of funds borrowed from the central queue and various overdraft costs. Our findings point to the importance of the cost-effective pricing of funds provided by the central queue based on banks' overdraft cost.

Our research suggests that payment networks that have a small number of banks with relatively concentrated demand will have an advantage over the networks that have a large number of banks with relatively dispersed payment demand (Conjecture 4). This is supported by the observation that, small countries such as Sweden (RTGS) and Poland (RTGS), found it much easier to leapfrog, than large countries such as China (DNS) and the U.S. (none).

While many countries have adopted the real-time payments settlement, the actual payments settlements in most countries only occur several times a day. Examples include Singapore (2 times), United

Kingdom (3 times), Chile (3 times), India (3 times), and Denmark (6 times). But why so? One of our results (Conjecture 5) revealed that higher risk, such as the increased cost of funds to cover overdrafts and more settlements handled by individual banks, may be what prevents banks from increasing the frequency of payments settlement. We view this as a bigger problem for high-value than low-value payments.

In addition, countries such as Switzerland and Mexico have implemented priority queuing systems to enable near real-time payments settlement. In terms of the priority queuing mechanism design, our research suggests that a central-queue management system outperforms the internal-queue design (Conjecture 1). A settlement intermediary has the ability to provide liquidity when payment demand is high, the participating banks' reserves are low, and the number of participating banks is large (Conjecture 2). The superior performance of the intermediary can be supported by appropriate pricing of liquidity at a level slightly higher than the overdraft penalty cost (Conjecture 3). Overall, our results show that the hybrid payments settlement system design supported by priority queuing improves system performance, and thus facilitates the achievement of faster payments settlement on average.

There are several limitations in this research. First, we have only focused on a specific set of values for the variables in our experiments. A larger-scale simulation and more thorough exploration of the impacts of the various model variables need to be done in the future. Also, if we can obtain payment and settlement data from the banking industry, we will have an excellent opportunity to demonstrate the external validity of our experimental simulation findings. We also offered ideas earlier about how to address and possibly recast some of the variables in our simulation model that banking industry practitioners may find challenging to implement based on the operational aspects of payments settlement, or where some machine-based readings will be more pragmatic and doable cheaply, with adjustments that we can make to the modeling.

Second, we haven't modeled the different risks that are associated with different participating banks and different channels for payments. For example, payments can come from different channels: card payments, Internet payments, mobile payments, etc. Not only are the payments and settlement risks across the different channels different due to the self-selection of consumers and organizations who need these services, but also the underlying technologies and software that support fully-automated straight-through processing may differ. Accounting for these kinds of things can bring more realism into our research.

Third, our concentration on the hybrid payments settlement mechanism design offers important insights into digital financial intermediation and market-based coordination. Our results suggest 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. With the support of appropriate pricing for the services that are provided, efficient delivery of high-quality services for low-value payments will become a reality.

REFERENCES

Aite Group (2015) In the faster payments frenzy, don't forget cross-border payments. Nancy H. Atkinson's blog. Boston, MA.

Alexander K, Dhumale R, Eatwell J (2006) Global governance of financial systems: the international regulation of systemic risk. Oxford University Press, New York, NY.

Allsop P, Summers B, Veale J (2009) The evolution of real-time gross settlement: access, liquidity and credit, and pricing. Payment Systems Policy and Research, Financial Infrastructure Series, The World Bank, Washington, DC.

Anderson C (2008) *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion, New York, NY.

Angelini P (1998) An analysis of competitive externalities in gross settlement systems. *Journal of Banking and Finance* 22:1-18.

Arculus R, Hancock J, Moran G (2012) The impact of payment system design on tiering incentives. Working paper, Reserve Bank of Australia, Sydney, Australia.

Armstrong M (2006) Competition in two-sided markets. *RAND Journal of Economics* 37(3): 668-691.

Australian Payments Clearing Association (2014) Australia's leading financial institutions sign up to build the New Payments Platform. Media Release, Sydney, New South Wales, Australia, December 2. Available at c.ymcdn.com/sites/www.aiia.com.au/resource/dynamic/blogs/20141204_225703_17434.pdf.

Australian Payments Clearing Association (2015) New Payments Platform: phases 1 & 2 – requirements and sourcing. Sydney, New South Wales, Australia, December 2.

Banco de México (2015) Interbanking Electronic Payment System. Mexico City, Mexico, May 3. Available at www.banxico.org.mx/sistemas-de-pago/informacion-general/sistemas-de-pago-de-alto-valor/interbanking-electronic-payme.html.

Bech ML, Garratt R (2003) The intraday liquidity management game. *Journal of Economic Theory*, 109(2): 198-219.

BNY Mellon (2014) Global payments 2020: transformation and convergence. New York, NY.

Boston Consulting Group (2014) Global payments 2014: the interactive edition. Boston, MA.

Bolt S, Emery D, Harrigan P (2014) Fast retail payment systems. *The Bulletin - Reserve Bank of Australia*. December: 43-52.

Bonchek M, Choudary SP (2013) Three elements of a successful platform strategy. *Harvard Business Review*, January 31. Available at hbr.org/2013/01/three-elements-of-a-successful-platform.

Cirasino M, Garcia JA (2008) Measuring payment system development. Payment Systems Policy and Research, Financial Infrastructure Series, World Bank, Washington, DC.

Clemons EK, Gu B, Lang KR (2002) Newly vulnerable markets in an age of pure information products: an analysis of online music and online news. *Journal of Management Information Systems* 19(3): 17-41.

Cognizant (2014) The changing face of payments 2014: a review of current infrastructures, drivers for change and implications for the future. In cooperation with VocaLink, London, UK.

Committee on Payment and Settlement Systems (CPSS) (2005) New developments in large-value payment systems. Bank for International Settlements, Basel, Switzerland.

Committee on Payment and Settlement Systems (CPSS) (2011) Payment, clearing and settlement systems in the CPSS countries. Vol. 1, Bank for International Settlements, Basel, Switzerland.

Committee on Payment and Settlement Systems (CPSS) (2012a) Payment, clearing and settlement systems in the CPSS countries. Vol. 2, Bank for International Settlement, Basel, Switzerland.

Committee on Payment and Settlement Systems (CPSS) (2012b) Innovations in retail payments. Bank for International Settlement, Basel, Switzerland.

Daly J (2013) Cover story: faster payments. *Digital Transactions: Trends in the Electronic Exchange of Value.* Boland Hill Media, Chicago, IL, June.

Economist (2015) The fintech revolution. May 9.

Eisenmann T, Parke G, Van Alstyne MW (2006). Strategies for two-sided markets. *Harvard Business Review* 84 (10): 92-101.

Evans DS (2011) Platform Economics: Essays on Multi-Sided Businesses. Competition Policy International, Boston, MA.

Faster Payments Scheme Limited (2014) CPSS-IOSCO self-assessment public disclosure for Faster Payments Scheme Limited (FPSL). London, UK, July.

Federal Financial Institutions and Examination Council (FFIEC) (2004) Wholesale payment systems, In *IT Examination Handbook*. Infomation Systems Audit and Control Association, July.

Galbiati M, Soramäki K (2008) An agent-based model of payment systems. Bank of England, London, UK.

Governors of the Federal Reserve Bank System (2012) Overview of the Federal Reserve's payment systems risk policy on intraday credit. 10th edition, Washington, DC, July.

Granados NF, Kauffman RJ, King, B (2008) How has electronic travel distribution been transformed? a test of the theory of newly vulnerable markets. *Journal of Management Information Systems* 25(2): 73-95.

Greene C, Rysman M, Schuh S, Shy Z (2015) Costs and benefits of building faster payment systems: the U.K. experience and implications for the United States. No. 14-5, Current Policy Perspectives, Federal Reserve Bank and Boston, Boston, MA, February 24.

Groenfeldt T (2014) Federal Reserve study says the U.S. needs faster payments. *Forbes*, November 17.

Grover E (2015) Fed should tread carefully in faster payments plan. *American Banker / Bank Think*, February 18.

Guo Z, Koehler GJ, Whinston, AB (2007) A market-based optimization algorithm for distributed systems. *Management Science*, 53(8): 1345-1358.

Guo Z, Koehler GJ, Whinston AB (2012) A computational analysis of bundle trading markets design for distributed resource allocation. *Information Systems Research*, 23(3, Part 1): 823-843.

Hagiu A (2014) Strategic decisions for multisided platforms. *MIT Sloan Management Review*, Winter.

Hume-Cook M (2015) Australia's New Payments Platform: opportunities for innovation. Odecee Pty. Ltd., Melbourne, Victoria, Australia, April.

International Trade Center (2009) Secrets of electronic commerce: a guide for small and medium enterprises, 2nd ed. World Trade Organization / United Nations, Geneva, Switzerland.

Johnson K, McAndrews JJ, Soramäki K (2004) Economizing on liquidity with deferred settlement mechanisms. *Economic Policy Review*, Federal Reserve Bank of New York, NY, 10(3): 56-72.

Kahn C, Roberds W (2001) Real-time gross settlement and costs of immediacy. *Journal of Monetary Economics*, 47(2): 299-319.

Kahn C, Roberds W (2009) Payments settlement: tiering in private and public systems. *Journal of Money, Credit and Banking*, 41(5): 855-884.

Kauffman RJ (2015) Real-time payment-driven digitization in banking. Interview with Becky Butcher. *Asset Servicing Times*, London, UK, May.

Kauffman RJ, Spaulding T, Wood CA (2008) Are online auction markets efficient? An empirical study of market liquidity and abnormal returns. *Decision Support Systems* 48(1): 3-13.

KPMG (2012) The great payments transformation: insight into the payment ecosystem. KPMG International, Amsterdam, Netherlands.

Leinonen H, Soramäki K (1999) Optimizing liquidity usage and settlement speed in payment systems. Paper 16, Bank of Finland, Helsinki, Finland.

Leinonen H, Soramäki K (2003) Simulating interbank payment and securities settlement mechanisms with the BoF-PSS2 simulator. Discussion paper 23/2003, Bank of Finland, Helsinki, Finland.

Li T, Kauffman RJ (2012) Adaptive learning in service operations. *Decision Support Systems* 53(2): 306-319.

Li T, Kauffman RJ, van Heck E, Vervest PHM, Dellaert B (2014) Consumer informedness and firm information strategy. *Information Systems Research* 25(2): 345-363.

Manning M, Nier E, Schanz J (eds.) (2009) *The Economics of Large-Value Payments and Settlement: Theory and Policy Issues for Central Banks*. Oxford University Press, Oxford, UK.

McAndrews JJ, Rajan S (2000) The timing and funding of Fedwire funds transfers. *Economic Policy Review*, Federal Reserve Bank of New York, New York, NY, 17-32.

McIntosh G, Whisler E, Koninckx M, Hartley M, Gardiner W (2014) Flavours of fast – a trip around the world in immediate payments. Clear2Pay, Brussels, Belgium.

McKinsey & Company (2014) Transforming national payment systems. In *McKinsey on Payments*, 20, New York, NY, 23-31.

Negrin J, Ocampo D, de los Santos A (2015) Recent innovations in inter-bank electronic payment system in Mexico: the role of regulation. *IFC Bulletin* 31: 473-494.

Oliver R (2015) Why a former Fed exec is on board with same-day ACH. *Pymnts.com*. February.

Payments Council (2015) Delivering world-leading mobile payments: how does the U.K. compare internationally? London, UK.

Payments Systems Studies Staff (2000) What is a payment system? Payment system central banking seminar, Federal Reserve Bank of New York, New York, NY, October 13.

Pelegero RM (2013) The need for real-time payments in the US. Strategy Note Series, Retail Payment Global Consulting Group, www.rpgc.com, Woodinville, WA, June.

Peñaloza RAS (2009) A duality theory of payment systems. *Journal of Mathematical Economics*, 45(9-10): 679-692.

Peñaloza RAS (2011) Implementation of optimal settlement functions in real-time gross settlement systems. Working paper, Department of Economics, University of Brazilia, Brazilia, Brazil.

Regelado A (2014) The economics of the Internet of Things. MIT Technology Review, May 20.

Reserve Bank of Australia (2012) Strategic review of innovation in the payments system: conclusions. Sydney, New South Wales, Australia, June.

Rochet JC, Tirole J (2002) Cooperation among competitors: some economics of payment card associations. *RAND Journal of Economics* 33(4): 549-570.

Schulz C (2011) Liquidity requirements and payment delays: participant type dependent preferences. Working paper no. 1291, European Central Bank / Eurosystem, Frankfurt, Germany, February.

Selgin G (2004) Wholesale payments: Questioning the market failure hypothesis. *International Review of Law and Economics*, 2(3): 333-350.

Shang R, Huang J, Yang Y, Kauffman RJ (2012) Exploring spot market users' willingness-to-pay for service-level agreements in cloud computing. In *Proceedings of the 2012 Workshop on E-Business*, Orlando, FL, December 2012.

Shen P (1997) Settlement risk in large-value payment systems. *Economic Review,* 2nd Quarter, Federal Reserve Bank of Kansas City, Kansas City, MO, 46-62.

Soramäki K, Bech ML, Arnold J, Glass RJ, Beyeler WE (2007) The topology of interbank payment flows. *Physica A: Statistical Mechanics and Its Applications*, 379(1): 317-333.

Summers BJ (2015) Facilitating consumer payment innovation through changes in clearing and settlement. Federal Reserve Bank of Kansas City, Kansas City, MO, August 6, 175-229.

SWIFT (2014a) SWIFT and the New Payments Platform. Brussels, Belgium. Available at www.swift.com/assets/swift_com/documents/news/AUNPP_Brochure.pdf.

SWIFT (2014b) SWIFT enters into real-time real-time domestic payments. Brussels, Belgium. Available www.swift.com/about_swift/shownews?param_dcr=news.data/en/swift_com/2014/PR_AU_SWIFT.xml.

The Federal Reserve Banks (2013) Payment system improvement – public consultation paper. FedPaymentsImprovements.org. September.

The Federal Reserve Banks (2014) Research results summary: Faster payments assessment summary. FedPaymentsImprovements.org. August.

The Federal Reserve Banks (2015) In pursuit of a better payment system: new on FedPayments improvements – Federal Reserve announces launch of faster and secure payments task forces. FedPaymentsImprovements.org. March.

The Federal Reserve System (2015) The strategies for improving the U.S. payment system. January.

Todd S (2015) The road to faster payments: a banker's guide. *American Banker / Bank Technology News,* March 30.

VocaLink / PricewaterhouseCoopers (2009) Tomorrow happened yesterday: how banks are building a business case for faster payments. London, UK.

Wack K (2015) Fed offers four options for speeding up payments. *American Banker / Bank Technology News*, January 26.

Wallace N (2000) Knowledge of individual histories and optimal payment arrangements. *Federal Reserve Bank of Minneapolis Quarterly Review*, 24(3): 11-21.

Willison M (2005) Real-time gross settlement and hybrid payment systems: a comparison. Working paper, The Bank of England, London, UK.

APPENDIX A. GLOBAL FASTER PAYMENTS SYSTEMS – COUNTRY COMPARISONS

Table A1. Examples of Faster Retail Payments Settlement Systems, Ordered by Year Implemented

COUNTRY	SYSTEM	YEAR	TYPE	COMMENTS
Switzerland	Swiss Interbank Clearing (SIC)	1987	Near to real-time	Mostly 24 x 7 x 365. Both high-value and retail payments. Current underway to upgrade to new standards including EuroSIC and ISO 20022. Implementation of priority queuing.
South Korea	Electronic Banking System (EBS), HOFINET	2001	DNS, next day	24 x 7 x 365 operation. Upgrade from Automatic Response Service (ARS); support e-payments from telebanking, Internet banking, and mobile banking; settled at 11:30am next day.
Brazil	Funds Transfer System (SITRAF)	2002	DNS, every 5 minutes	Not 24 x 7 x 365. Economic factors are the main driver for the establishment of this payment system.
Mexico	Interbank Electronic Payment System (SPEI)	2004	Real-time	Mostly 22 x 7 x 365 operation. Central bank ownership and operation. Implementation of priority queuing. Designed with regional characteristics in mind, to facilitate funds transfer between U.S. and Mexico [Banco de Mexico 2015, Negrin et al. 2015].
South Africa	Real-Time Clearing (RTC)	2006	DNS, approx. hourly	24 x 7 x 365 operation. A main objective is to enable payments via mobile phones for the large unbanked population. Commitment to migrate to ISO 20022 standard.
United Kingdom	Faster Payments (FPS)	2008	DNS, 3 cycles per day	24 x 7 x 365 operation. A main focus is to enable overlay of innovative solutions such as Paym, Pingit, and Zapp for mobile payments.
Chile	Transferencias en Línea (TEF)	2008	DNS, 3 cycles/day	24 x 7 x 365 operation. Private sector initiative following regulatory mandate to eliminate float.
China	Internet Banking Payment System (IBPS)	2010	DNS	24 x 7 x 365 operation. Planned upgrade to CNAPS II to connect to ISO 20022 standard.
India	Immediate Payment Service (IMPS)	2010	DNS, 3 cycles per day	24 x 7 x 365 operation. To facilitate banking services via mobile phone for the large unbanked population. Not yet adopted the ISO 20022 standard.
Sweden	Payments in Real Time (BiR)	2012	Real-time	24 x 7 x 365 operation. Integrated with Bankgirot; good fit with mobile payment service Swish. Partially driven by the competitive pressure from non-bank entities.
Poland	Express ELIXIR	2012	Real-time	24 x 7 x 365 operation. Participation of banks is limited without regulatory mandate.
Singapore	Fast and Secure Transfers (FAST)	2014	DNS, two cycles/day	24 x 7 x 365 operation. Electronic funds transfer service. 14 participating banks. Also called G3 Clear2Pay.
Denmark	RealTime24/7	2014	DNS, six cycles/day	24 x 7 x 365 operation. Three parts, SumClearing (DNS), IntradagClearing (Intra-Day), and StraksClearing (real-time), together form the retail payments infrastructure.
Australia	New Payments Platform (NPP)	2017	Real-time	To operate 24 x 7 x 365, adopt ISO 20022 standards, build business overlays on NPP to encourage innovation and entrepreneurship, and run the clearing and settlement processes both in real-time.
United States	None	After 2018	Unknown	Currently exploring approaches to implement faster payments nationally, with the Federal Reserve Bank's encouragement.

APPENDIX B. MODELING NOTATION AND DEFINITIONS

All of our notation is presented below. This will be useful to fully understand our modeling, methods, analytical work, and experimental simulation work.

Table B1. Modeling Notation and Definitions

NOTATION	DEFINITION
$t \in \{1,2,...,T\}$	There are T settlement periods, and all but the last has the possibility of delay for a payment to be settled in the next period. In period T , netting will occur.
$i \in \{1,2,...,I\}$	There are I participating banks.
$\omega \geq 0$	Liquidity cost, which is the borrowing cost per unit time for covering a bank's funding shortage for payment settlement via the intermediary-managed central queue.
$\delta \geq 0$	Overdraft penalty cost, which is a penalty cost per unit time that the bank pays to the central bank if its balance is negative.
$c^0 \geq 0$	The intermediary's initial liquidity provision.
$B_i^0 \geq 0$	Bank i 's initial reserve funds.
p_i	Probability that bank i receives a random payment request from any other bank in each period.
q_{ij}^t	Payment from bank i to bank j that arrives in period t.
$d_{ij}^t \geq 0$	Unit delay cost associated with payment q_{ij}^t
$\theta_i > 0$	Settlement priority threshold of bank i.
$y_{ij}^t = \{0,1\}$	Bank i 's decision of whether to settle in settlement period a t, where 0 means no settlement and 1 implies settlement.
$m = \{1,2,...,T\}$	The decision period m when all delayed payments with $t \leq m$ are considered for possible settlement.
P_i^m	Bank i 's immediate payment set in period m.
D_i^m	Bank i 's delayed payment set in period m.
$b_i^m \geq 0$	Amount bank i borrows from the intermediary at the beginning of settlement period m.
$z_i^m \in [0, v_i^m]$	Amount bank i pays back to the intermediary in settlement period m.
$v_i^m \geq 0$	Bank i 's net settlement amount in settlement period m at the settlement intermediary.
c^m	The intermediary's liquidity provision in settlement period m.
B_i^m	Bank i 's available reserve funds at the beginning of settlement period m.
OD_i^m	Bank i 's overdraft amount in its reserve account when payments are made at the end of settlement period m.
NDI	Normalized delay index, which measures the system-wide settlement delay.
AFR	Average funds required, which measures the average system-wide funds that are used to cover liquidity shortfalls from both the central bank and the settlement intermediary.
AFT	Average funds transfer, which measures the average system-wide variation in account balance.

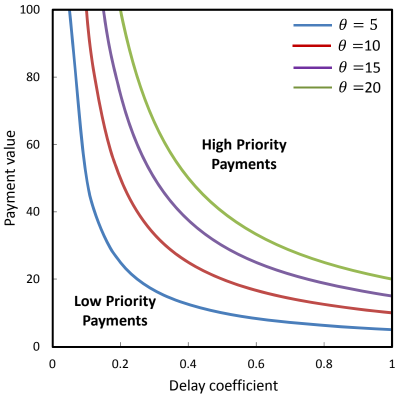
APPENDIX C. THE MODEL AND ITS ASSUMPTIONS

The banks and the settlement procedure. We analyze banks' decision-making by considering the trade-off between the total settlement delay and the total cost of funds. All delayed payments will be settled and all accounts will be reset to the reserve level when netting occurs in the last period of the process, and the system will be restarted. We set up an experiment to model the system's performance. To begin, we assume there are I banks, each is subject to T settlement periods, and final settlement via netting occurs in the last period. Suppose the system has a settlement period of 6 minutes or 10 times per hour. Then every 3 hours – which is a long enough period of time to observe system performance – there will be $T = 30$ periods in which settlement-related decision-making will occur. This frequency can be viewed as near to real-time. Similarly, when $T = 15$ periods, settlement will occur every 12 minutes, which is still a relatively high frequency of settlement within a 3-hour period of time. We can easily extend these ways of modeling time periods for settlement and final settlement to a full day of operation, without loss of generality. For example, because settlement with netting restarts the whole process, we can simulate 24 hours of system operations by aggregating 8 random instances of our simulation. The qualitative insights remain unchanged this way. So we have limited our coverage to system performance over 3 hours within a day. It is sufficient to demonstrate our essential findings by focusing on system performance across the periods in which settlement occurs up to the final period involving settlement with netting.

Banks receive payment requests that arrive in real time, but only settle periodically. In each settlement period $t \in \{1,2,...,T\}$, bank i receives random requests to pay bank j with probability p_i, which we denote as q_{ij}^t. If payment requests can be settled directly by the bank within the period, it is considered as an instance of *near real-time settlement* handled via RTGS. If not, the payment requests will either enter bank i's internal queue or the intermediary-managed central queue, depending on the system design. In the case of central queue, priorities of payments are set by the banks, but settlement is decided on by the intermediary. Some payments are settled via the intermediary's RTGS if sufficient liquidity is available, while others are handled via *delayed settlements*. We further assume that no two payments that go from bank i to bank j will arrive at the same time. If they arrive together, two payments can be combined into one payment request. Furthermore, we assume each payment request has a delay coefficient d_{ij}^t, which can be understood as a *unit delay cost* that measures the sensitivity of delay.

Payment prioritization. If the *weighted value of the payment request* (the payment amount times the delay cost parameter) is greater than a threshold value θ, then it will be treated as a *high-priority payment*; otherwise, it will be considered as a *low-priority payment*. This approach proxies for the banks' decisions, and this reflects differences in how payments will be handled. Figure C1 illustrates the related *payment prioritization rule*.

Figure C1. Payment Prioritization Rule



The bank will settle the high-priority payments as soon as possible in the period when the payment request arrives. So there will be no delay cost incurred for high-priority payments. However, this will sometimes lead to an overdraft of the bank's reserve account at the central bank, with a per unit penalty cost of overdraft δ . In contrast, low-priority payments will be handled differently based on the bank's reserve account balance. If funds are available in the account, then the request will be settled in this period; otherwise, the payment will be queued with other unsettled payments in the bank's internal queue or in the central queue, depending on the mechanism. The queued payments will be selectively settled either by the bank in subsequent periods when funds become available, or by the intermediary. The delayed low-priority payments reduce the possible overdraft cost at the expense of delayed settlement, which is an economic trade-off.

Each individual bank determines its own threshold level θ_i for $i \in \{1,2,...,I\}$. In settlement period m , denote the immediate payment set by $P_i^m = \{q_{ij}^t \mid d_{ij}^t q_{ij}^t \geq \theta_i, t = m\}$ and the delayed payment set with $D_i^m = D_i^{m-1} \cup \{q_{ij}^t \mid d_{ij}^t q_{ij}^t < \theta_i, t = m\}$. Note that if the payments arrival time t is modeled in clock time rather than with discrete periods, $t = m$ can be modified as $m - 1 < t \leq m$.

The Case of an Internal Queue. In the case of an internal queue, the banks will solve the following optimization problem in each settlement period m :

$$\begin{aligned} \min_{y_{ij}^t=\{0,1\}} \quad & \sum_{q_{ij}^t \in D_i^m} d_{ij}^t (m-t) q_{ij}^t (1 - y_{ij}^t) + \delta \max(-B_i^m, 0) \\ s.t. \quad & B_i^m = B_i^{m-1} - \sum_{q_{ij}^t \in P_i^m} q_{ij}^t - \sum_{q_{ij}^t \in D_i^m} q_{ij}^t y_{ij}^t \end{aligned} \tag{1}$$

Here, y_{ij}^t is a binary decision variable. When $y_{ij}^t = 0$, the payment request from bank i to bank j that enters the system at time t is not chosen to be settled, so it imposes a delay cost of $d_{ij}^t (m-t) q_{ij}^t$ on the system. When $y_{ij}^t = 1$, the payment request will be settled. As a result, the payment will be removed from the queue and there will be no delay related to this payment. The summation over D_i^m is bank i 's total delay costs related to its unsettled payments to all other banks. The term $\delta \max(-B_i^m, 0)$ is bank i 's overdraft cost in this settlement period, which is incurred only if $B_i^m < 0$, when bank i faces a negative account balance at the end of settlement period m . Since B_i^m is the cumulative measure of bank i 's available funds, it does not require the summation sign.

The balance constraint specifies bank i 's account balance at the end of settlement period m , which should be equal to its account balance at the end of the previous period, B_i^{m-1} , minus the high-priority payments to other banks that must be settled in this period, minus low-priority payments to other banks that are selected to be settled in this period.

The Case of a Central Queue. In the case of a central queue, the settlement intermediary may help the banks to cover the liquidity shortfall in the presence of an imbalance of funds in the system. To give banks an incentive to pay back the central queue in a timely manner, the central queue may charge a fee ω per unit for its intertemporal liquidity provision.

Bank i will make two decisions. First, it will determine what payments it will settle itself in this period and what payments should be submitted for queuing. Second, in the case that the bank has an outstanding debt with the central queue, it will decide how much to pay back to the intermediary. The bank's objective is to minimize the total cost of its payment delay costs up to time m in a day, plus the overdraft cost to the central bank and the liquidity cost to the central queue. The objective function is given by:

$$\begin{aligned} &\min_{y_{ij}^t=\{0,1\}, z_i^m \geq 0} \sum_{q_{ij}^t \in D_i^m} d_{ij}^t (m-t) q_{ij}^t (1-y_{ij}^t) + \delta \max(-B_i^m, 0) + \omega(b_i^m - z_i^m) \\ &s.t. \quad B_i^m = B_i^{m-1} - \sum_{q_{ij}^t \in P_i^m} q_{ij}^t - \sum_{q_{ij}^t \in D_i^m} q_{ij}^t y_{ij}^t - z_i^m \\ &\qquad z_i^m \leq b_i^m \end{aligned} \tag{2}$$

Compared with the decision-making model related to the internal queue, the additional term $\omega(b_i^m - z_i^m)$ is the *liquidity cost of borrowing* from the intermediary if a bank has not repaid all its debt. The objective function trades off the total payment delay cost, the overdraft penalty cost charged by central bank, and the liquidity cost of borrowing from the intermediary.

The balance constraint is interpreted similar to that for the internal queue. In addition to the immediate payments and the delayed payments settled in this period, the bank also will consider the repayment to the intermediary. The second constraint states that the amount paid back to the settlement intermediary is no more than the amount the bank owed: no banks keep their cash reserves with the intermediary.

Intermediated settlement. Settlement at the end of each period allows multiple payments to be settled simultaneously, if offsetting funds to match are available. At the end of T periods, all payments in the queue will be netted. The intermediary's objective is to minimize the total delay cost and the cost of borrowing from the central queue for all banks in period m. It solves the following optimization problem:

$$\begin{aligned} &\min_{y_{ij}^t=\{0,1\}, v_i^m \geq 0} \sum_{q_{ij}^t \in D_i^m} d_{ij}^t (m-t) q_{ij}^t (1-y_{ij}^t) + \omega \sum_{i=1}^{I} \max(0, v_i^m) \\ &s.t. \quad \sum_{q_{ij}^t \in D_i^m} q_{ij}^t y_{ij}^t - \sum_{j \neq i} \sum_{q_{ji}^t \in D_j^m} q_{ji}^t y_{ji}^t = v_i^m, \quad i = 1, \ldots, I \\ &\qquad \sum_{i=1}^{I} \max(0, v_i^m) \leq c^m \end{aligned} \tag{3}$$

The interpretation of the settlement decision variable y_{ij}^t is the same as for the internal queue. The objective is to minimize the total payment delay cost and the total borrowing cost from the intermediary for liquidity. The means of doing this is by selecting payments from the central queue to settle without violating the settlement constraints.

The first constraint is the equality that computes the net outgoing payment amount v_i^m for each bank i . It is the total net settlement of bank i, which is equal to the total payments bank i pays other banks less the total receipts bank i obtains from other banks. If $v_i^m < 0$, then bank i will have a net receipt of funds from the other banks and it will not need to borrow from the intermediary. If $v_i^m > 0$, then bank i will borrow the amount v_i^m from the intermediary. However, the second constraint ensures that the total amount borrowed by all banks does not exceed the *intermediary's liquidity pool of funds available*, *c^m^*, in settlement period m.

APPEDIX D. EXPERIMENTAL SIMULATION SET-UP

In our experiment, the major design factors are set up according to Table 1 in the main report. The specific values in the experiment were implemented as described in Table D1.

Table D1. Mechanism Design Experiment Conditions and Treatments

MECHANISM DESIGN	EXPERIMENTAL CONDITIONS	EXPERIMENTAL TREATMENTS
Payment System	Number of banks	$I = \{5,10\}$
	Number of settlement periods	$T = \{15,30\}$
Participating Banks	Payment density	$P_i = \{0.3, 0.5, 0.7\}$
	Reserve account funds	$B_i = \{50,100\}$
Central Bank	Cost of funds per unit of overdrafts	$\delta = \{0, 0.02, 0.04, 0.06, 0.08\}$
Settlement Intermediary	Cost of funds per unit liquidity in the central queue	$\omega = \{0, 0.02, 0.04, 0.06, 0.08\}$
	Liquidity provision	$c^0 = \{50,100\}$

Since the number of banks and the number of settlement periods before final netting occurs (like the end-of-day netting) are important characteristics of the system configuration, we consider two levels for each variable. $I = \{5,10\}$ is intended to represent a smaller and a larger number of banks. $T = \{15,30\}$ can be understood as less frequent settlement (15 times in 3 hours) versus more frequent settlement (30 times in 3 hours).[15]

Note that our experimental set-up is different from Willison (2004), who modeled different DNS approaches, including one-hour, morning, and afternoon netting. The one-hour netting mechanism had a higher frequency of payment settlement: net settlement of queued payments occurred every hour. His system flushed the payment queue at the end of each hour. And then, after the final netting occurred, the whole system restarted.

Payment density in our experimental design refers to the probability that a bank will send a payment to each of the other banks in any period. For example, $P_i = 0.3$ means that bank i has a 30% probability to send a payment to each of the other banks in each settlement period. We use 0.3, 0.5 and 0.7 to represent low, medium and high probabilities. Central banks typically require banks to set aside individual reserves to satisfy their individual liquidity needs, in the event that settlement problems arise. We set two levels of *reserve balances*: $B_i = \{50,100\}$, representing low and high reserve balances at the central bank for bank i.

We assume there is a per unit penalty cost δ when a bank requires credit to have sufficient liquidity to settle payments that it is handling. We experiment with five different values: $\delta = \{0, 0.02, 0.04, 0.06, 0.08\}$.

The central queue also can choose to provide its own intertemporal liquidity in offsetting delayed payments in the central queue by charging a fee per time unit ω. To compare this with the central bank's overdraft penalty cost, we experiment with five different levels: $\omega = \{0, 0.02, 0.04, 0.06, 0.08\}$. Similarly,

[15] Both will have only 1 instance of final settlement with netting, since the 15th and 30th periods both occur after the same amount of total time has gone by. It also is possible for us to have focused on the number of times that netting occurs, although we chose not to do this for the specific simulation results that we will report on. We think of these two different approaches as *settlement frequency* and *netting frequency* treatments. The latter we will leave for future research.

we use two levels for the settlement intermediary's liquidity provision: $c^0 = \{50,100\}$.

In general, q_{ij}^t can be modeled as a random variable from a distribution that can be statistically described from banks' historical transaction data. Since we do not have access to such information, we assume that $q_{ij}^t \in [0,100]$ follows a uniform distribution. For simplicity, we assume both random payment q_{ij}^t and its associated delay coefficient d_{ij}^t follow a uniform distribution over [0, 100] and [0,0.01], respectively. Therefore, the maximum delay cost considering both the delay sensitivity and the payment value is 0.1 · 100 = 10. We set $\theta_i = 5$ as the *delay cost threshold for immediate payment settlement.*

Under each level of payment density $P_i = \{0.3, 0.5, 0.7\}$, we simulate 30 instances of the payment request arrivals for each bank in each time period for every network configuration characterized by $I \times T$, where $I = \{5,10\}$ and $T = \{15,30\}$. All together, we have 360 independent random samples (= 30 · 3 · 2 · 2) for which we can test the effects of different treatments.

In addition, to test the interactive network effect by P_i , I and T , we constructed three bank payment networks as follows:

- **Network 1 (Low-Concentration, Low-Frequency Network): 10 low-payment demand banks in a network.** Each bank will send payments to another bank with probability 0.3 in each period. The expected number of outgoing payments for each bank is 0.3 · 9 = 2.7. So the total expected number of payments per period is 2.7 · 10 = 27. We chose the number of periods is 15 to simulate less frequent settlement. The total expected number for payments is 27 · 15 = 405.

- **Network 2 (High-Concentration, Low-Frequency Network): 6 high-payment demand banks in a more concentrated network.** Each bank will send payments to another bank with a high probability (P = 0.9) in each period. On average, each bank will send 0.9 · 5 = 4.5 payments in each period, which is greater than 2.7. So banks have higher payment demand in Network 2. Moreover, the number of periods is 15, and the expected number of payments in the network is 4.5 · 6 · 15 = 405, which is the same as Network 1.

- **Network 3 (Low-Concentration, High-Frequency Network): 10 low-payment demand banks with a high payment settlement frequency.** Each bank sends payments to another bank with probability 0.15 in each period. The number of periods is 30 so Network 3 settles twice as fast as Network 1. The expected number of payments is 0.15 · 9 · 10 · 30 = 405.

We propose three main performance measures.

- *Normalized delay index,* $NDI = \frac{\sum_{m=1}^{T}\sum_{i=1}^{I}\sum_{q_{ij}^t \in D_i^m} q_{ij}^t}{\sum_{m=1}^{T}\sum_{i=1}^{I} q_{ij}^t (T-t)}$. The index takes a value between 0 and 1. If all queued payments are immediately settled, then the index will equal 0. For end-of-day delayed net settlement, q_{ij}^t remains in the set D_i^m from t to T , so the index will equal 1.

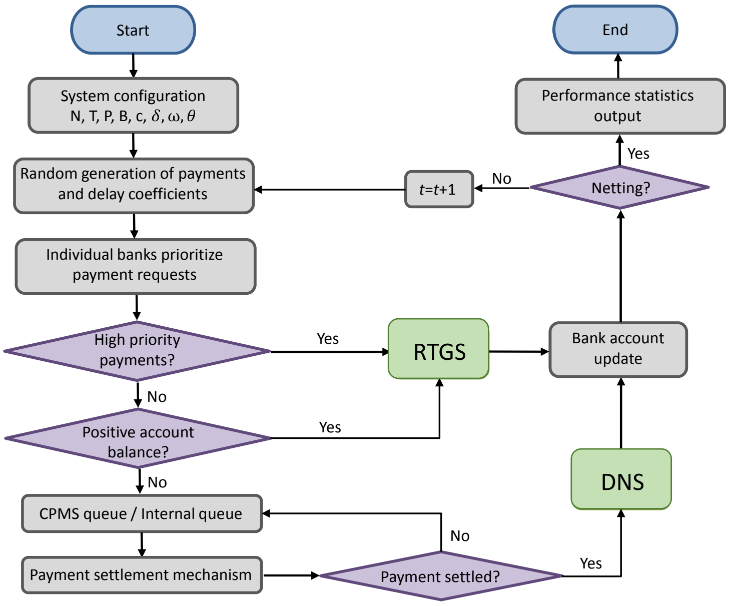
- *Average funds required,* $AFR = \frac{\sum_{m=1}^{T}\sum_{i=1}^{I} OD_i^m}{T \times I}$, where $OD_i^m = \max(-B_i^m, 0) + z_i^m$ is bank i 's overdraft penalty cost in period m to the central bank (the first term) and the borrowing amount from the central queue (the second term). If the bank's reserve account balance is negative, then the bank will have an overdraft equal to the absolute value of the balance, plus its borrowing amount from the central queue if there is any; if the balance is positive, no overdraft will occur, and so $OD_i^m = 0$.

- *Average funds transfer,* $AFT = \frac{\sum_{m=2}^{T}\sum_{i=1}^{I}|B_i^m - B_i^{m-1}|}{(T-1)\times I}$. This measures the variation in the reserve account balance to settle payments, and is also useful as a measure of liquidity.

APPEDIX E. IMPLEMENTATION: THE MECHANISM AND EXPERIMENTAL SIMULATION

The following flowchart represents our implementation of the proposed payments settlement mechanism and the related experimental simulation.

Figure E1. Flowchart


Start
System configuration
N, T, P, B, c, δ, ω, θ
Random generation of payments and delay coefficients
Individual banks prioritize payment requests
High priority payments?
Yes
No
Positive account balance?
Yes
No
CPMS queue / Internal queue
Payment settlement mechanism
Payment settled?
No
Yes
RTGS
DNS
Bank account update
Netting?
No
Yes
t=t+1
Performance statistics output
End

Note: To maintain a straightforward presentation of the ideas in this process, we suppressed one step that would generalize the flow that we have depicted for this mechanism. For example, it might be possible to have multiple 3-hour periods in a day in which final settlement occurs, with no payment requests carried forward. This suggests that there would be an *inner loop* similar to what we depicted above, and an *outer loop*, for which it is possible to instantiate the number of times per day that the inner loop runs. This will bring what we are suggesting a little closer to what will need to be done in an industry setting.

- **Step 1 (System Configuration).** Set the total number of banks I and the number of settlement periods T. Initialize the bank balance B_i^0, the priority threshold level θ_i, and payment arrival probability P_i for all banks $i = 1, 2, ...I$. Initialize the overdraft penalty cost δ, the intermediary's liquidity cost ω, and its inventory level c^0. Set the queuing mechanism design choice *CPMS* to 0 or 1.
- **Step 2 (System Initialization).** Set the delayed payment set $D_i^0 = \Phi$ and the initial borrowing from CPMS $b_i^0 = 0$. Set the time period m. Repeat Steps 3 and 4 until $m = T$.
- **Step 3 (Payments Arrival and Banks' Decision-Making).** In each settlement period m, with probability p_i, random payment requests q_{ij}^t arrive associated with a delay cost d_{ij}^t, where $t = m$. Each bank prioritizes its payments, set $P_i^m = \{q_{ij}^t \mid d_{ij}^t q_{ij}^t \geq \theta_i\}$, $D_i^m = D_i^{m-1} \cup \{q_{ij}^t \mid d_{ij}^t q_{ij}^t < \theta_i\}$, and solves the optimization problem in Equation 1 if *CPMS* = 0, or with Equation 2 if *CPMS* = 1. Assume the optimal solution is y_{ij}^{t*} and z_i^{m*}, where $z_i^{m*} = 0$ if *CPMS* = 0, and payments are settled for each bank.

Bank i receives $\sum_{j \neq i}(\sum_{q_{ji}^{t} \in P_{j}^{m}} q_{ji}^{t} + \sum_{q_{ji}^{t} \in D_{j}^{m}} q_{ji}^{t} y_{ji}^{t*})$ and the budget is updated as:

$B_{i}^{m} = B_{i}^{m-1} - \sum_{q_{ij}^{t} \in P_{i}^{m}} q_{ij}^{t} - \sum_{q_{ij}^{t} \in D_{i}^{m}} q_{ij}^{t} y_{ij}^{t*} + \sum_{j \neq i}(\sum_{q_{ji}^{t} \in P_{j}^{m}} q_{ji}^{t} + \sum_{q_{ji}^{t} \in D_{j}^{m}} q_{ji}^{t} y_{ji}^{t*}) - z_{i}^{m*}$. Moreover,

$D_{i}^{m} \leftarrow D_{i}^{m} - \{q_{ij}^{t} \mid y_{ij}^{t*} = 1, \forall\ j \neq i\}$, $b_{i}^{m} \leftarrow b_{i}^{m} - z_{i}^{m*}$, and $c^{m} \leftarrow c^{m} + \sum_{i=1}^{I} z_{i}^{m*}$.

- **Step 4 (Queuing and Settlement).** If *CPMS* = 0, go to step 3; otherwise, the intermediary solves Equation 3. Assume the optimal solution is y_{ij}^{t*} and v_{i}^{m*}, and payments are settled for each bank. The updates for B_{i}^{m} , D_{i}^{m} , b_{i}^{m} and c^{m} are as follows: $B_{i}^{m} \leftarrow B_{i}^{m} - \sum_{q_{ij}^{t} \in Q_{i}^{m}} q_{ij}^{t} y_{ij}^{t*} + \sum_{j \neq i} \sum_{q_{ji}^{t} \in Q_{j}^{m}} q_{ji}^{t} y_{ji}^{t*}$,

  $D_{i}^{m} \leftarrow D_{i}^{m} - \{q_{ij}^{t} \mid y_{ij}^{t*} = 1, \forall j \neq i\}$, $b_{i}^{m} \leftarrow b_{i}^{m} + \max(0, v_{i}^{m*})$, and $c^{m} \leftarrow c^{m} - \sum_{i=1}^{I} \max(0, v_{i}^{m*})$.
- **Step 5 (End-of-Period *T* Final Settlement with Netting).** Compute the average normalized delay index, NDI_{i} , average funds required, AFR_{i} , and average funds transfer, AFT_{i} , for each bank i . All banks settle all their set of delayed payment requests, D_{i}^{T} , and reset their reserve account balances to B_{i}^{0} .
- **Step 6 (Output Performance Measures).** Compute the average performance measure for each bank based on the average values of NDI , AFR and AFT across the T periods in which the payments settlement experimental simulation was run. Stop.

Again, please note that, in this experiment, we only focus on one set of settlement periods up to the end of period T , when final settlement with netting occurs, and the whole process restarts. So there will not be any cumulative effect if there were another set of simulated periods that operate independent of what we have simulated with the T periods.

APPENDIX F. EXPERIMENTAL RESULTS FOR PAYMENT NETWORK CONCENTRATION

Figures F1 and F2 show that the central queue in the high-concentration network has less system delay than the low-concentration network when the cost of the intermediary's liquidity provision is not zero. The total number of payments in the high-concentration network is also higher than in the low-concentration network.

Figure F1. The Effect of Payment Network Concentration on Payment Settlement Delay

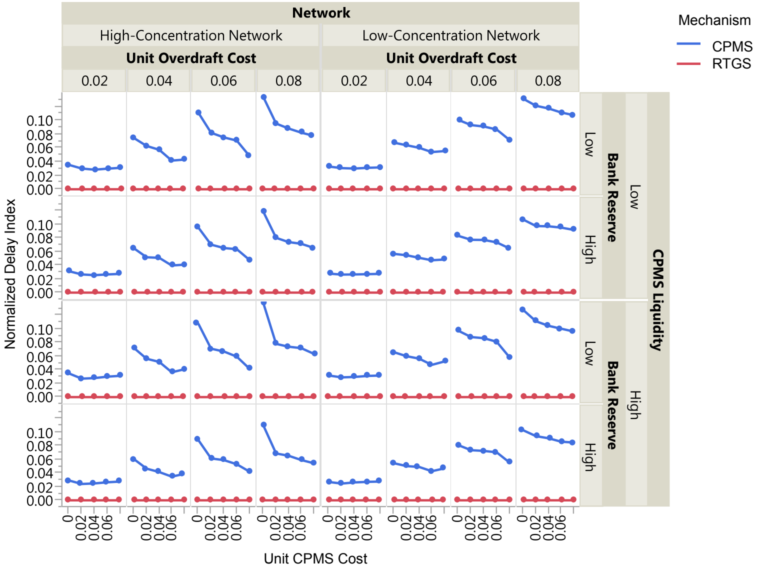


Figure F2. The Effect of Payment Network Concentration on Total Settlements

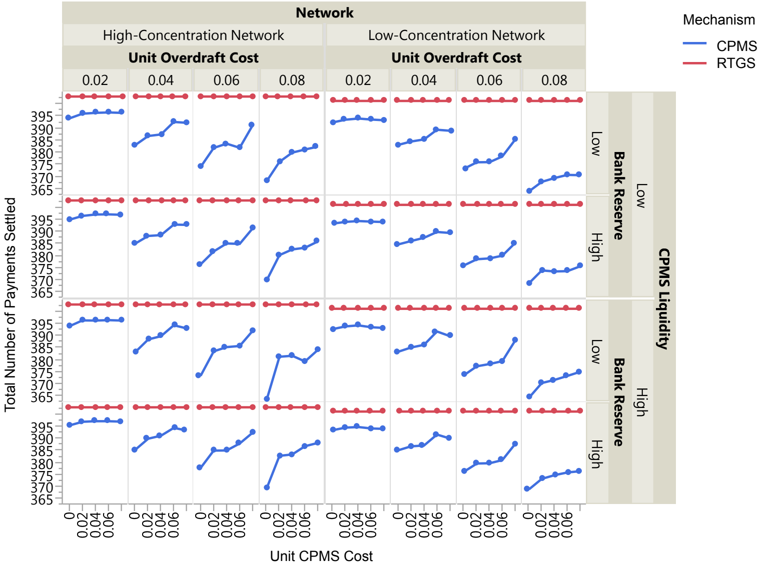


Figure F2 illustrates that the total number of settlements is the highest when both the overdraft penalty cost and the intermediary's liquidity cost are low. It also shows that free liquidity provision by the intermediary will not lead to good performance. This is because borrowing without interest costs will not create any incentive for the banks to pay back. Instead, a borrowing cost as small as 2% (0.02) can effectively increase the number of total payments settled and reduce the delays. By imposing a small fee, the banks' incentives will be better aligned with the intermediary's liquidity provision. Banks will pay back when it makes sense to do so, and the intermediary can maintain a healthy level of cash inventory to supply intertemporal liquidity for all participating banks.

Does the performance improvement in the high-concentration network come at a higher cost? Figure F3 compares the system performance in terms of the average funds required to cover the liquidity shortfalls. We see that the high-concentration network has a lower average funds required than the low-concentration network, showing a fewer funds shortages in the high-concentration network. For both networks, for a given overdraft penalty cost, the average funds required does not seem to vary much as the unit liquidity cost the intermediary changes. However, as the overdraft penalty cost becomes higher, the average funds required decreases. This is mainly because the banks settle smaller number of payments immediately, as shown by the higher normalized delay index in Figure F1. Overall, the central queue liquidity cost has a higher impact than the overdraft penalty cost on the average funds required in the system to cover liquidity shortfalls.

Figure F3. The Effect of Payment Network Concentration on the Average Funds Required

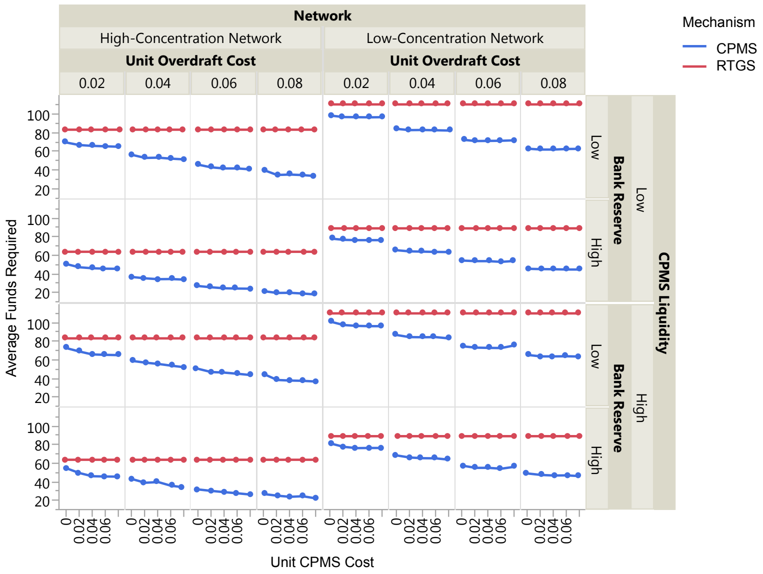
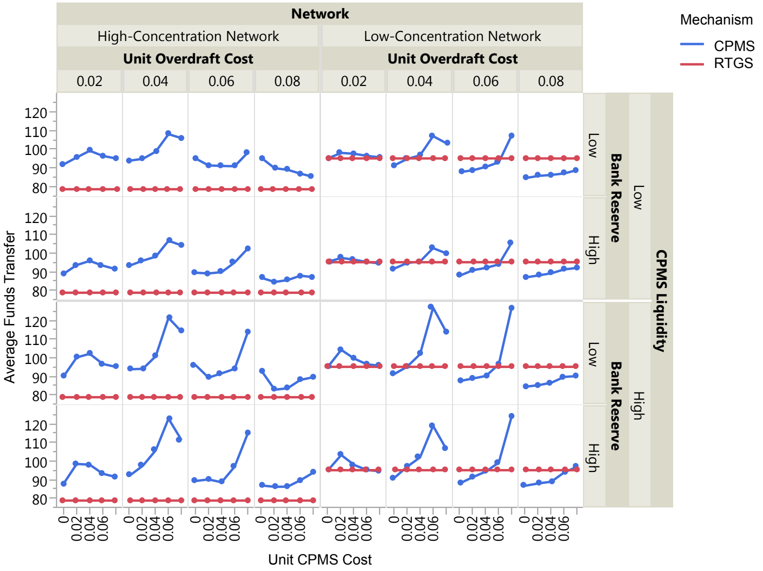


Figure F4 presents the average total funds transfer, which sheds light on the variation in the banks' reserve account balances. Similar to Figure F3, the high-concentration network has a lower average amount of funds transferred than the low-concentration network has with RTGS. Both Figures F3 and F4 provide evidence that, other things being equal, a high-concentration network is better suited to implementing an RGTS system in comparison to a low-concentration network.

Nevertheless, comparing the central queue with RTGS for the low-concentration network, we see that a properly-priced intermediated-queue service can effectively reduce the average variation of the banks' reserve account balances below the RTGS level, especially when the overdraft penalty cost is high. This shows that the intermediary can mitigate systemic risk in the low-concentration network.

Figure F4. The Effect of Payment Network Concentration on the Average Funds Transfer



Overall, the lower payment delay that we shared in Figure 7 in the main report, and the lower average overdraft cost in Figure F3 here provide strong support that payment network concentration plays a significant role in priority queuing mechanism performance. The practical implication is that the priority queuing mechanism is likely to be more effective in countries where a small number of banks handle high demand for payments than in countries where a large number of banks share the payment demand.

APPENDIX G. EXPERIMENTAL RESULTS RELATED TO SETTLEMENT FREQUENCY

Figure 8 in the main report shows that higher settlement frequency reduces settlement delay. Figure G1 (G2) further show that the total number of payments (the number of central queue settlements) settled by the high-frequency network is higher (lower) than the low-frequency network. This suggests that increasing settlement frequency shifts more settlements to individual banks.

Figure G1. The Effect of Settlement Frequency on the Central Queue Settlements

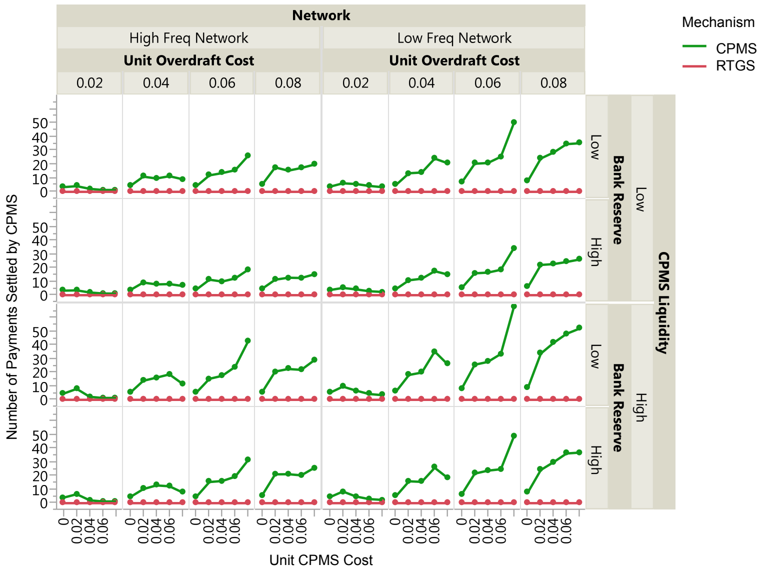


Figure G2. The Effect of Settlement Frequency on Total Settlements

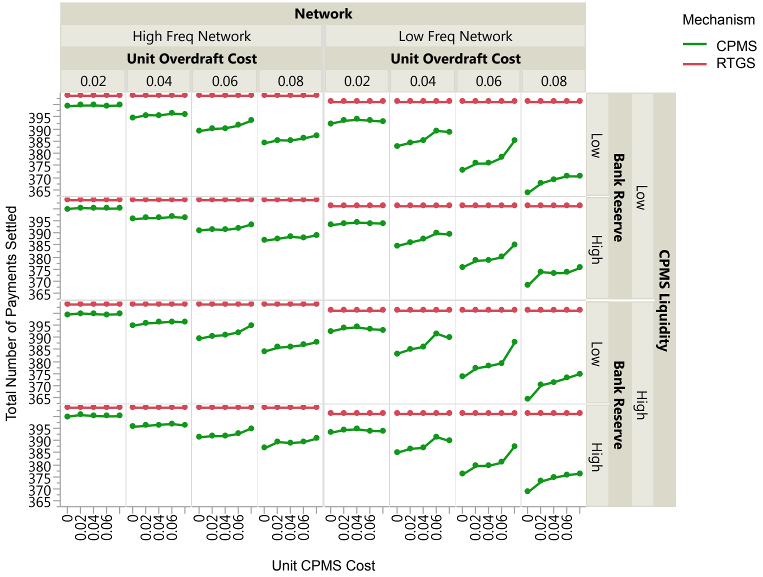


Figure G3. The Effect of Settlement Frequency on Average Funds Required

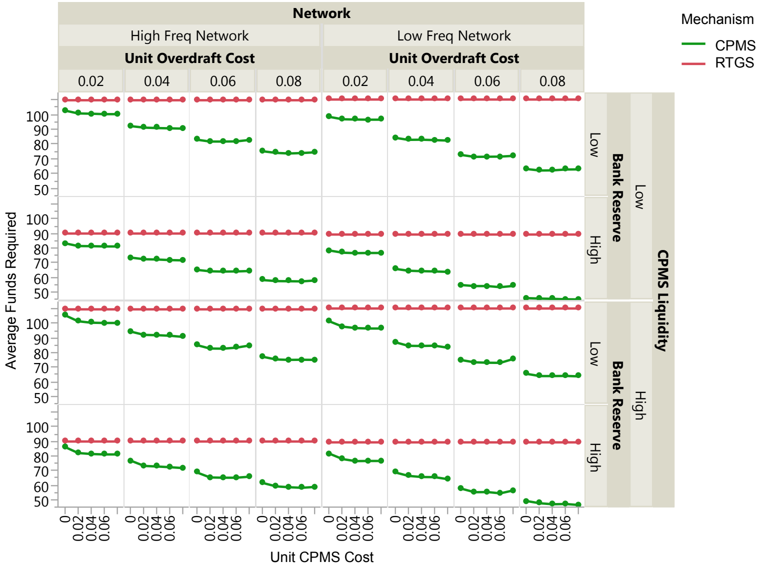


Figure G3 shows that the average funds required is higher in the high-frequency network, indicating higher liquidity cost with the central queue settlement. Moreover, the central queue liquidity cost has a higher impact than overdraft penalty cost on the average funds required to cover liquidity shortfalls.

Figure G4. The Effect of Settlement Frequency on the Number of Funds Transfers

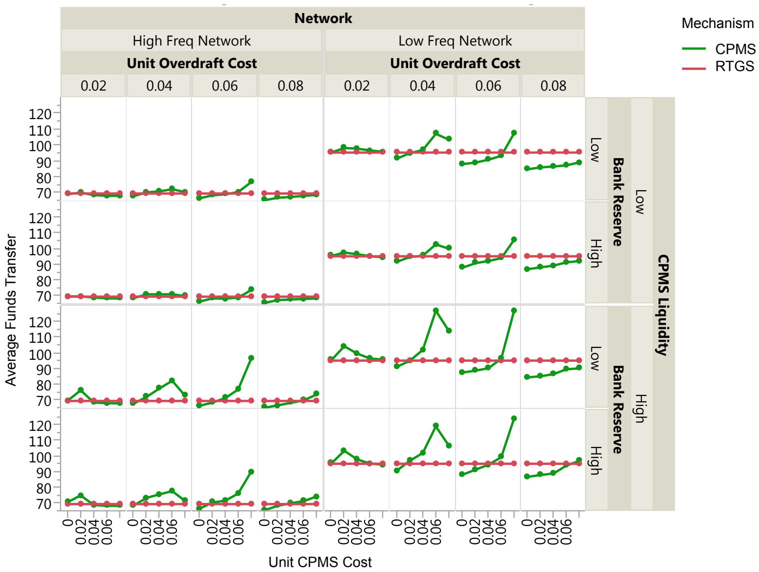


Figure G4 further shows that average amount of funds transferred is lower in the high-frequency network. This is mainly because, as the settlement frequency increases, the number of payments that need to be settled in each settlement period falls. The delayed payments are spread over more periods for settlement, leading to less variation in the banks' reserve account balances.