Chapter 2

New approaches for payment system simulation research

Kimmo Soramäki – Walter Beyeler – Morten Bech – Rober Glass

2 New approaches for payment system simulation research.....16

Abstract 16

2.1 Introduction.....16
2.2 Modelling interbank payment flows.....17
2.2.1 Network representation of payment systems.....18
2.2.2 Fedwire as an example of a complex network.....23
2.3 Modelling payment system dynamics 24
2.3.1 Network dynamics 24
2.3.2 Simple payment system model.....26
2.4 Modelling bank behaviour.....29
2.4.1 Decision-making, learning and adaptation 29
2.4.2 Multi-agent model of bank behaviour 30
2.5 Conclusion 33

References 35

The authors would like to thank Marco Galbiati (Bank of England) and Jeffrey Arnold (Federal Reserve Bank of New York) for comments and their input in the papers summarized in this article.

Soramäki: Helsinki University of Technology, Finland. Corresponding author: www.soramaki.net
Beyeler and Glass: Sandia National Laboratories, Albuquerque, NM, USA. Supported through the National Infrastructure Simulation and Analysis Center (NISAC), a program of the Department of Homeland Security's Infrastructure Protection/Risk Management Division and comprised of a core partnership of Sandia National Laboratories Los Alamos National Laboratory. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.
Bech: Federal Reserve Bank of New York, New York, NY, USA. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

2 New approaches for payment system simulation research

Abstract

This article presents new directions for simulation research in interbank payment systems that integrates network topology, network dynamics and agent-based modelling of bank behaviour. In the process it also reviews literature in the field and presents applications of the ideas presented. While the focus of the article is on systemic risk in interbank payment systems, the concepts and models presented are applicable to address questions related to other payment systems and topics such as liquidity flow efficiency as well.

2.1 Introduction

At the apex of the financial system is a network of interrelated financial markets by which domestic and international financial institutions allocate capital and manage their exposure to risk. Critical to the smooth functioning of these markets are a number of financial infrastructures that facilitate clearing and settlement. The events of 11 September 2001 underscored both the resiliency and the vulnerabilities of these financial infrastructures to wide-scale disruptions. Any interruption in the normal operations of these infrastructures may seriously impact not only the financial system but also the economy as a whole.

A growing body of policy-oriented research is available. One segment of the literature focuses on simulating the default of a major participant and evaluating the effects on other institutions in payments[1] and securities settlement systems[2]. Another segment presents detailed case studies on the responses of the US financial system to shocks such as the 1987 stock market crash and the attacks of 11 September 2001.[3] Much of the research has been conducted

[1] See Humphrey (1986), Angelini et al (1996), Kuussaari (1996), Bech et al (2002), Northcott (2002), Bech and Soramäki (2005), Bedford et al (2005) and Mazars and Woelfel (2005).
[2] See Hellqvist and Koskinen (2005) and Devriese and Mitchell (2006).
[3] See Bernanke (1990), McAndrews and Potter (2002) and Lacker (2004).

using data from real operating environments with the given payment flows and settlement rules of the respective systems. As such they are useful for assessing the operation of the particular system under disruptions, but the results are difficult to generalise to systems with other characteristics. Little research has focused on explaining the relationship between the characteristics of the system and its performance during and following disruptions. Also the behaviour of participants has been generally exogenously defined or assumed unchanged (or to change in a predetermined manner) when the policy parameters of the system change or when a bank changes its settlement behaviour as a consequence of operational or financial problems. Such assumptions are unlikely to hold in the case of real disruptions.

This article argues that three aspects are important for answering the still unanswered questions on what makes a payment system and its participants robust or fragile towards disruptions, and what are the most efficient measures to reduce the likelihood and magnitude of disturbances. First, understanding the pattern of liquidity flows among the system participants. Second, understanding how the rules of the system affect the dynamics of liquidity flows. Third, the ability to evaluate likely behavioural changes of the participants before, during and following disruptions or as a consequence of policy changes.

This article presents new approaches at answering the above questions. It is organised as follows. Section 2.2 discusses how payment system interactions can be described by means of network topology and presents empirical results for the US Fedwire system. Section 2.3 describes dynamics that can take place in interbank payment systems and presents a simple model of a payment system based on simple rules of settlement. Section 2.4 presents some possible directions for modelling participant behaviour in payment systems. Section 2.5 concludes.

2.2 Modelling interbank payment flows

A payment system can be treated as a specific example of a complex network (see eg Newman, 2003). In recent years, the physics community has made significant progress towards understanding the structure and functioning of complex networks. The literature has focused on characterising the structure of networked systems and how the properties of the observed topologies relate to stability, resiliency and efficiency in case of perturbations and disturbances.

From a technical perspective, most payment systems are star networks where all participants are linked to a central hub (the operator) via a proprietary telecommunications network. From a payment processing perspective, payment systems are generally complete networks as all nodes (participants) are linked in the sense that they can send and receive payments from each other. However, these representations do not necessarily reflect the actual behaviour of participants that controls the flow of liquidity in the system and thus the channels for contagious transmission of financial disturbances. In common with other of social networks mediated by technology (such as email or telephone calling), the networks formed by actual participant behaviour are of more interest than the network structure of the underlying communication system.

2.2.1 Network representation of payment systems

Networks have been modelled in several disciplines such as in mathematics and computer science under graph theory, in applied mathematics and physics under network theory and in sociology under social network analysis. While the terminologies and research questions in the different traditions vary, common to all is the representation of the topic under study as (at minimum) two types of elements: nodes and connections between them, ie links. The following paragraphs summarise the main concepts.

Links can be either undirected or directed. Links can have weights attached to them representing the importance of the relationship between nodes. The strength of a node can be calculated as the sum of the weights of all the links attached to it. For a directed network, strength can be defined over both the incoming and outgoing links.

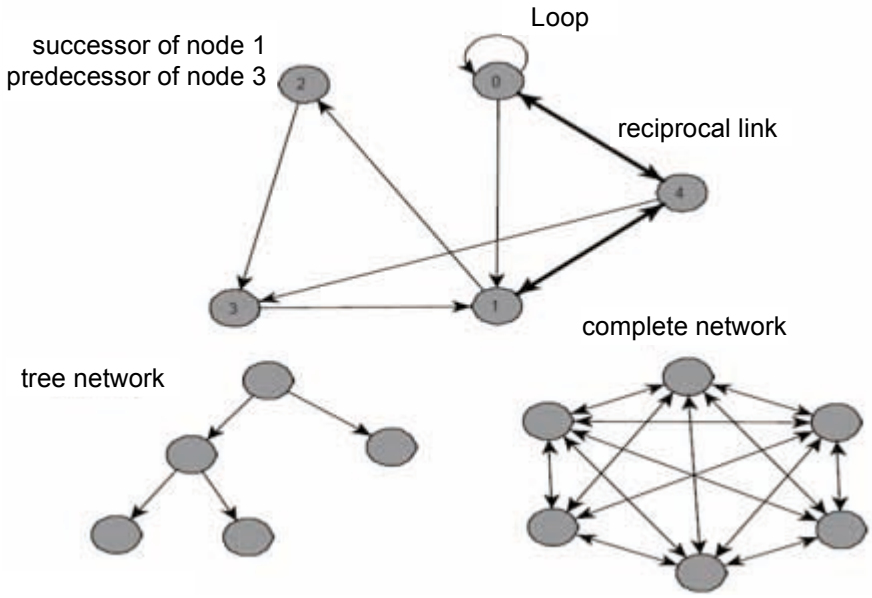
A link from a node to itself is called a loop. The neighbours of a node are all the nodes to which it has a link. The predecessors of a node are the nodes that have a link to the node and the successors are the nodes that have a link from the node. A walk is a sequence of nodes in which each node is linked to the next. A walk is a path if all its nodes are distinct. The length of a path is measured by the number of links. If the start node and the end node of a path are one and the same, then it forms a cycle.

A complete network is a network where all nodes have a link to each other. A tree is a network in which any two nodes are connected by exactly one path. A connected network is a network where any two nodes can be joined by a path while a disconnected network is made

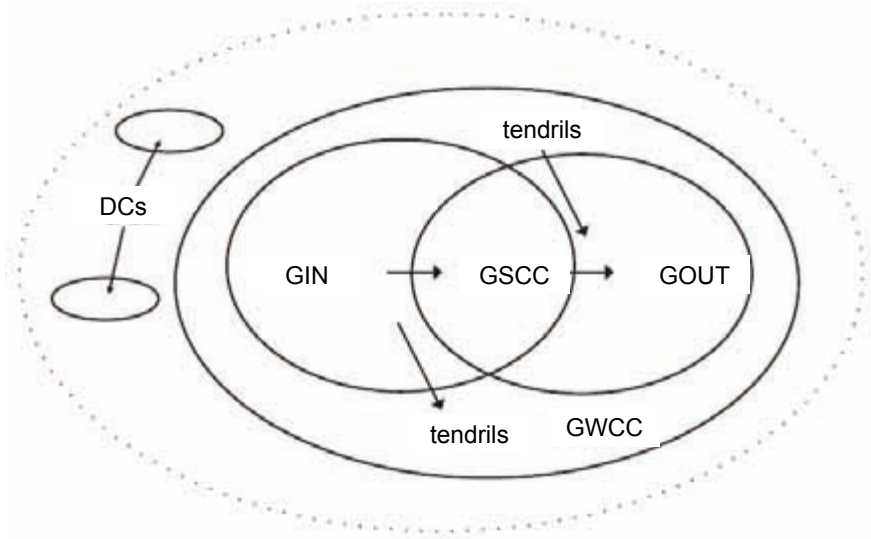
up of two or more connected components or sub-networks. These concepts are illustrated in Figure 2.1a.

Figure 2.1 **Network modelling**

a)



b)



The most basic properties of a network are the number of nodes n and the number of links m. The number of nodes defines the size of a network while the number of links relative to the number of possible links defines the *connectivity* of a network. The *degree* of the network is the average number of links for each node in the network.

A starting point for the quantitative analysis of a network is to partition the set of nodes into *components* according to how they connect with other nodes. Dorogovtsev et al (2001) divide a network into a single giant weakly connected component (GWCC) and a set of *disconnected components* (DCs). The GWCC is the largest component of the network in which all nodes connect to each other via undirected paths. The DCs are smaller components for which the same is true. The GWCC consists of a *giant strongly connected component* (GSCC), a *giant out-component* (GOUT), a *giant in-component* (GIN) and *tendrils*. The GSCC comprises all nodes that can reach each other through a directed path. A node is in the GOUT if it has a path from the GSCC but not to the GSCC. In contrast, a node is in GIN if it has a path to the GSCC but not from it. Tendrils are nodes that have no directed path to or from the GSCC. They have a path to the GOUT or a path from the GIN (see Figure 2.1b).

Application of the component analyses to liquidity flows between banks provides insights on the structure of these flows within the payment system and gives clues with respect to the relative importance and vulnerability of banks in the system in case of disruptions. As banks in GOUT only receive funds from other banks in the GSCC, a disruption by a bank in GOUT would only affect other banks in that component. Banks in GIN are affected only by disruptions in the same component, and not by banks in other components as their payment processing is not dependent on incoming liquidity from these banks. Banks outside the GSCC tend to be smaller whereas all money center banks belong to the GSCC.

Two important characteristics of a node in a directed network are the number of links that originate from the node and the number of links that terminate at the node. These two quantities are referred to as the *out-degree* and *in-degree* of a node respectively. The average degree of a node in a network is the number of links divided by the number of nodes, ie $\langle k \rangle = m/n$. Networks are often categorised by their degree distributions. The degree distribution of a classical random network (ER-network, Erdõs and Rényi, 1959) is a Poisson distribution. Many real networks have fat-tailed degree distributions and a large number have been found to follow the power law

$P(k_i = x) \sim k^{-y}$ for large-degree nodes. Networks with a power-law distribution are sometimes referred to as scale-free networks[4]. Scale-free networks have been found to remain better connected when subjected to random failures than other types of networks. Albert et al (1999) and Crucitti et al (2004) find that the connectedness of scale-free networks is robust to random failures but vulnerable to targeted attacks. However, one must be a bit careful here as the process acting on the network influences such analyses of robustness and vulnerability.

Simply put, banks that have a low in-degree and high weights for these links are likely to be more vulnerable to disturbances than other banks as the removal of one link will severely limit the amount of incoming funds. Conversely, banks with high out-degree have ceteris paribus the potential to affect more counterparties if their payment processing is disrupted. Understanding the topology of payment flows is likely to be important in assessing the resiliency of a payment system to wide-scale disruptions.

It is also common to analyse *distances* between nodes in the network. The distance from node i to node j is the length of the shortest path between the two nodes. The average distance from a node to any other node in a strongly connected network is commonly referred to as the average path length of a node. If the network is not strongly connected, paths between all nodes may not exist. In a payment network the path length may be important due to the fact that the shorter the distances between banks in the network, the easier liquidity can re-circulate among the banks. On the other hand, a payment system where liquidity flows over short paths is also likely to be more vulnerable to disruptions in these flows.

Sociologists have long studied *clustering* in social networks, ie the probability that two nodes which are the neighbours of the same node themselves share a link. This is equivalent to the observation that two people, each of whom is your friend, are likely to be friends with each other. One way of measuring the tendency to cluster is the ratio of the actual number of links between the neighbours of a node over the number of potential links among them. A tree network has a clustering coefficient of zero, and a complete network a coefficient of one. In a classical random network, the clustering coefficient is the unconditional probability of connection, ie $\langle C \rangle = p$.

[4] This is because the power law distribution is the only scale-free distribution, ie if the scale by which x is measured is increased by a factor, the shape of the distribution p(x) is unchanged, except for an overall multiplicative constant (see Newman, 2005).

In a payment network, the clustering coefficient measures the prevalence of payments between a bank's counterparties. In terms of resilience one could hypothesise that disturbances in banks with a higher clustering coefficient might have a compounding impact on their counterparties, as some of the disturbance may be passed on by the bank's neighbours to each other – in addition to the direct contagion from the source of the disruption.

There are various measures of the centrality that indicate the relative importance of nodes in a network. Four measures of centrality are commonly used in network analysis: degree, closeness, betweenness, and eigenvector centrality. The first three were described in their current form by Freeman (1979) while the last was proposed by Bonacich (1972). Degree centrality takes into account only the immediate neighbourhood of the node, ie it is simply the number of links the node has. Closeness centrality as defined by Freeman is the sum of shortest paths from all other nodes. Betweenness centrality may be defined loosely as the number of times that a node is on the shortest path between any pair of nodes. Eigenvector centrality encapsulates the idea that the centrality of a node depends also on the centrality of the nodes that it is linked by (or links to). A famous commercialisation of this centrality measure is the PageRank algorithm by Google (Brin and Page, 1995). In general, the importance of the node will depend on process taking place in the network. Borgatti (2005) provides a good overview of alternative processes in networks and centrality measures applicable for their analysis.

Finally, a key question in the study of networks is how the topologies that are seen in reality have come into being. There are two classes of network formation models some times referred to as equilibrium and non-equilibrium models (Dorogovtsev and Mendes, 2003). Equilibrium models have a fixed set of nodes with randomly chosen pairs of nodes connected by links. Erdõs and Rényi (1959) develop a basic model of a n node network, with each pair of nodes connected by a link with probability p. This type of network is commonly referred to as a classical random network. Non-equilibrium network models grow a network by successively adding nodes and setting probabilities for links forming between the new nodes and existing nodes and between already existing nodes. Many of these models, notably the Barabasi and Albert (1999) model (BA model), are based on preferential attachment. Preferential attachment assigns a probability of a link forming with a node that is increasing with the number of prior links of the node.

2.2.2 Fedwire as an example of a complex network

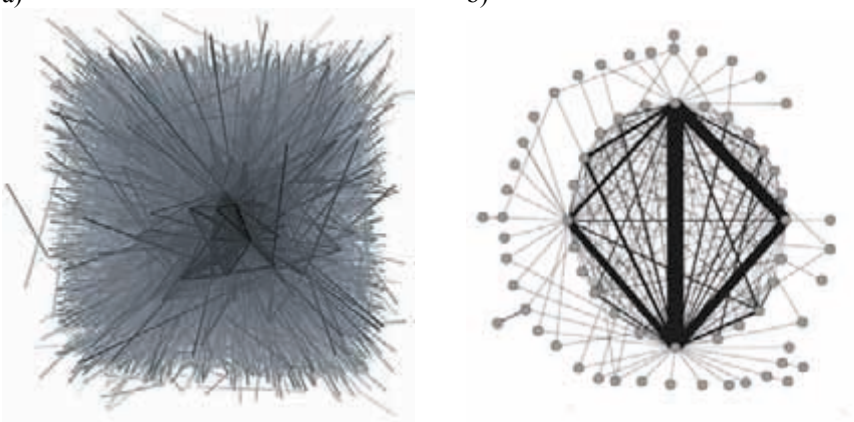
Soramäki et al (2007) analyse the topology of daily networks formed by the payment flows between commercial banks over Fedwire for a period of 62 consecutive business days. Apart from a few holidays, the statistics characterising the network were quite similar from day to day. These networks shared many characteristics with other empirical complex networks, such as a scale-free degree distribution, high clustering coefficient and the small world phenomenon (short path lengths in spite of low connectivity). Like many other technological networks, high-degree nodes tend to connect to low-degree nodes. Similar conclusions can also be reached from analysis on BoJ-NET by Inaoka et al (2005).

Moreover, Soramäki et al (2007) report that the topology of the network was significantly altered by the attacks of 11 September 2001. The number of nodes and links in the network and its connectivity was reduced, while the average path length between nodes was significantly increased. Interestingly, these alterations were of both similar magnitude and direction to those that occurred on several of the holidays contained within the period.

Figure 2.2a shows liquidity flows in Fedwire as a visual graph. The figure includes over 6,600 nodes and more than 70,000 links. Each link between two banks is shaded by the value of payments exchanged between them, with darker shades indicating higher values. Despite the appearance of a giant fur ball, the graph suggests the existence of a small group of banks connected by high value links. To gain a clearer picture of this group, a subset of the network where the focus is on high value links is displayed in Figure 2.2b. This graph shows the largest undirected links that comprise 75% of the value transferred. The network consists of only 66 nodes and 181 links. The prominent feature is a densely connected sub-graph, or clique, of 25 nodes to which the remaining nodes connect. By itself it is almost a complete graph. A small number of banks and the links between them thus dominate the value of all payments sent over the network.

Figure 2.2 **Visualisation of the liquidity flow network (Soramäki et al, 2007)**

a) b)



The analysis finds that payment networks have characteristics similar to other social and technological networks. An unanswered question is why the network has the structure it does: the network may grow over time by a logic that is very general or that is particular to payment systems, or to specific policies of a given system. This is an interesting topic for future research. The network structure has also implications for its robustness. Robustness of the network, however, also depends on the processes taking place in it. This is the topic of the next sections.

2.3 Modelling payment system dynamics

2.3.1 Network dynamics

A number of payment system simulations carried out in recent years have used actual or generated payment data. These simulations have studied the actual dynamics of payment systems, where system rules have varied from simple real-time gross settlement to complex hybrid settlement mechanisms with offsetting and multilateral settlement capabilities. The research can be summarised as trade-off questions between liquidity, speed of settlement and risks. The impact of bank behaviour has not been taken endogenously into account in these simulations. A summary of this line of research is provided in Leinonen (2005) and is not presented here.

From a network perspective, the performance of banks (nodes) is often dynamically dependent on the performance of other banks within the network and upon the structure of linkages between banks. A failure by one node in the network, for example, may hinder flows in the network and adversely impact the performance of the other nodes as the disturbance propagates in the network.

One branch of network literature has investigated the resilience of different network topologies in terms of a connectivity threshold (ie percolation threshold)[5] at which a network dissolves into several disconnected components. A well-known finding is that scale-free networks are more robust to random failures than other types of networks. However, they are very susceptible to the removal of the very few highly connected nodes. These static failure analyses may be applicable to some networks if the interest is the availability of paths between nodes in the network – but are less applicable to networks of monetary flows which contain both flows via the shortest paths as well as longer walks within the network.

Another branch of the literature has studied the impact of perturbations that cascade through the network on the basis of established theoretical or domain-specific rules[6]. In these dynamical models nodes generally have a capacity to operate at a certain load and, once the threshold is exceeded, some or all of the node's load is distributed to neighbouring nodes in the network (Bak et al, 1987). While the detailed dynamics depend on the rules applied for the cascades, generally the most connected nodes (or nodes with highest load in relation to overall capacity) are more likely than average nodes to trigger cascades. Increased heterogeneity makes the system more robust to random failures, but more susceptible to targeted attacks that may cause global cascades.

Cascade models have been applied by physicists to systems within fields ranging from geology to biology to sociology (eg Jensen, 1998). This research has demonstrated that models made of very simple agents, interacting with neighbouring agents, can yield surprising insights about system-level behaviour. In the spirit of these cascade models, Beyeler et al (2007) formulate a simple agent-based model for liquidity flows within a payment system.

---

[5] Eg Bollobas (1985), Moore and Newman (2000) and Callaway et al (2000).

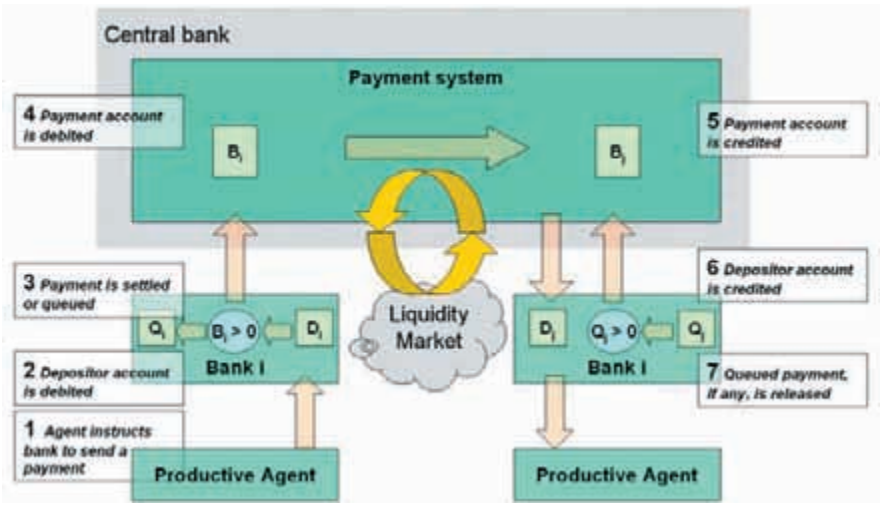
[6] Eg Watts (2002) and Crucitti et al (2004b) for random and complex networks, respectively, and Sachtjen et al (2000) and Kinney et al (2004) for power networks.

2.3.2 Simple payment system model

The model of Beyeler et al includes only the essential processes of a payment system and its accompanying liquidity market. A set of banks exchange payments through a single common payment system. All payments occur only along the links of a scale-free network – as was shown to be representative of Fedwire liquidity flows. Banks' customers randomly instruct them to make a unit payment to a neighbouring connected bank. Banks are reflexively cooperative: they submit the payment if the balance in their payment system account allows; otherwise they place the instruction on a queue for later settlement.

If the receiving bank has instructions in its queue, the payment it just received enables it to remove a queued instruction and submit a payment in turn. If the bank that receives that payment is also queuing instructions, then it can make a payment, and so on. In this way a single initial payment made by a bank can cause many payments to be released from the queues of the downstream receiving banks. This is an example of the cascade processes typically studied in other models of self-organised criticality. Statistics on these settlement cascades are an indicator of the extent of interdependence of the banks, and in the model they are a controlled by two parameters: the overall liquidity and market conductance.

Figure 2.3 **Simple payment system model (Beyeler et al, 2007)**

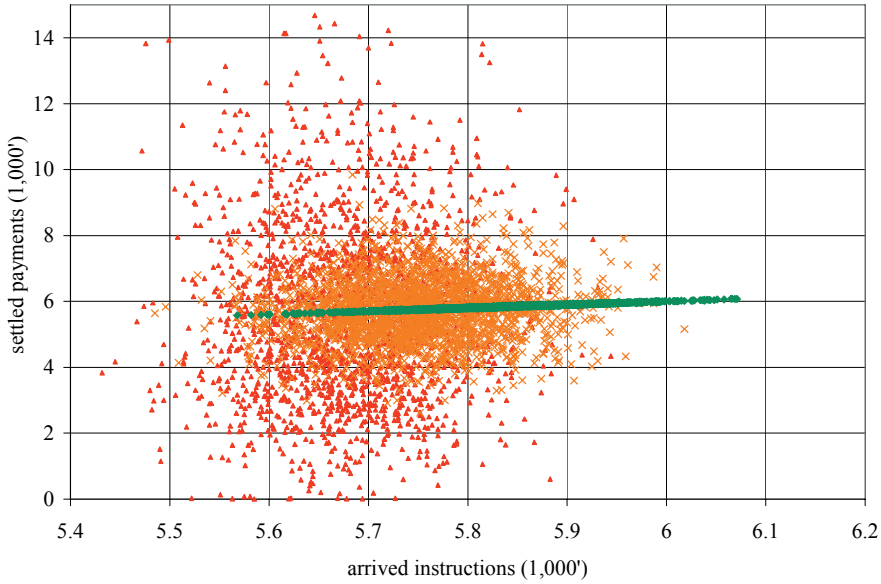


In the absence of a liquidity market, only abundant liquidity allows banks to operate independently; reducing liquidity increases the likelihood that a given bank will exhaust its balance and begin queuing payments. A bank that has exhausted its balance must wait for an incoming payment from one of its neighbours. When liquidity is low a bank's ability to process payments becomes coupled to its neighbours' ability to process. The output of the payment system as a whole is no longer determined by overall input, but instead becomes dominated by the internal dynamics of the system. Figure 2.4a shows how the correlation between arriving instructions and submitted payments degrades in the model as liquidity is reduced (1: high liquidity; 2: medium liquidity; 3: low liquidity). A settlement cascade, that is the release of queued payments as a result of a single initiating payment, can comprise hundreds of queued payments as illustrated in Figure 2.4b.

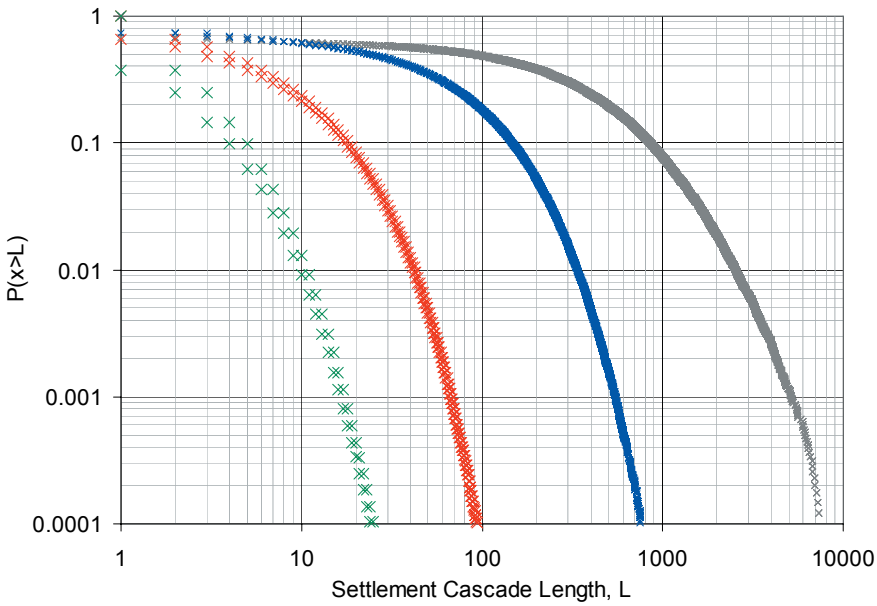
To explore how liquidity markets reduce coupling among network neighbours and thereby reduce congestion, market transactions were represented as a diffusive process where a bank's balance plays the role of a potential energy or pressure. Banks with high balances tend to contribute liquidity to the market, while banks with low balances tend to draw liquidity from the market. There is no decision-making or price-setting in this simple market model, but it reflects two essential features of a real market: liquidity flows from banks with surplus funds to banks that need funds, and liquidity can flow from any bank to any bank – flows are not confined to the links of the payment network. It creates a separate global pathway for liquidity flow. The ease of liquidity flow through the market is described by a single conductance parameter.

Figure 2.4 **Instruction and Payment Correlation (a) and Settlement Cascade Length Distribution (b).**

a)



b)



With a liquidity market included, the number of payments closely tracks the number of instructions as the coupling between banks is weakened and the size of the settlement cascades is reduced. The rate of liquidity flow through the market relative to the rate of flow through the payment system was surprisingly small. The performance of the system can be greatly improved even though less than 2% of the system through-put flows through the market.

2.4 Modelling bank behaviour

2.4.1 Decision-making, learning and adaptation

Wide-scale disruptions may not only present operational challenges for participants in the interbank payment system, but they may also induce participants to change the way they conduct business. The actions of participants have the potential to either mitigate or exacerbate adverse effects. Hence, understanding how participants interact and react when faced with operational adversity will assist operators and regulators in designing countermeasures, devising policy, and providing emergency assistance, if necessary.

The first approach to study bank behaviour in payment systems has been to use standard game theory. Angelini (1998) and Kobayakawa (1997) use a setup derived from earlier literature on precautionary demand for reserves. Angelini (1998) shows that in a RTGS system, where banks are charged for intraday liquidity, payments will tend to be delayed and that the equilibrium outcome is not socially optimal. Kobayakawa (1997) models the intraday liquidity management process as a game of uncertainty, ie a game where nature moves after the players. Kobayakawa (1997) shows that both delaying and not delaying can be equilibrium outcomes when intraday overdrafts are priced. McAndrews and Rajan (2002) study the timing and funding of transfers in the Fedwire funds transfer system. They show that banks benefit from synchronising their payment pattern over the course of the business day because it reduces the overdrafts. Bech and Garratt (2003) develop a stylised two-period-two-player model with imperfect information. They analyse the strategic incentives under different intraday credit policy regimes employed by central banks and characterise how the Nash equilibria depend on the underlying cost parameters for liquidity and delays. It turns out that two classical paradigms in game theory emerge: the Prisoner's Dilemma in the case where intraday credit is provided

against collateral and the Stag Hunt coordination game in the case where the central bank charges a fee. Hence, many policy issues can be understood in terms of well-known conflicts and dilemmas in economics.

Other approaches that have been applied to similar problems of repeated interaction among a large number of players are evolutionary game theory and reinforcement learning (such as Q-Learning by Watkins et al, 1992). Agents who learn about each others' actions through repeated strategic interaction is a leading theme in evolutionary game theory. In most of the existing literature it is customary to look at the players' asymptotic behaviour in situations where the payoffs are some known function of players' strategies. In one strand of the literature, this knowledge is a prerogative of the players, who can therefore use adaptive rules of the type 'choose a best reply to the current strategy profile'. In a second research line, the learning rules do not require knowledge of the payoff function on the part of the learners. Such rules are instead of the kind 'adopt more frequently a strategy that has given a high payoff'.

Galbiati and Soramäki (2007) use methods from reinforcement learning (Barto and Sutton, 1998) and fictitious play (Brown, 1951) to numerically solve a model with interactions among a large number of banks that settle payments on a continuous basis under imperfect information, stochastic payoffs and a finite but long sequence of settlement days. The model is summarised and discussed in more detail below.

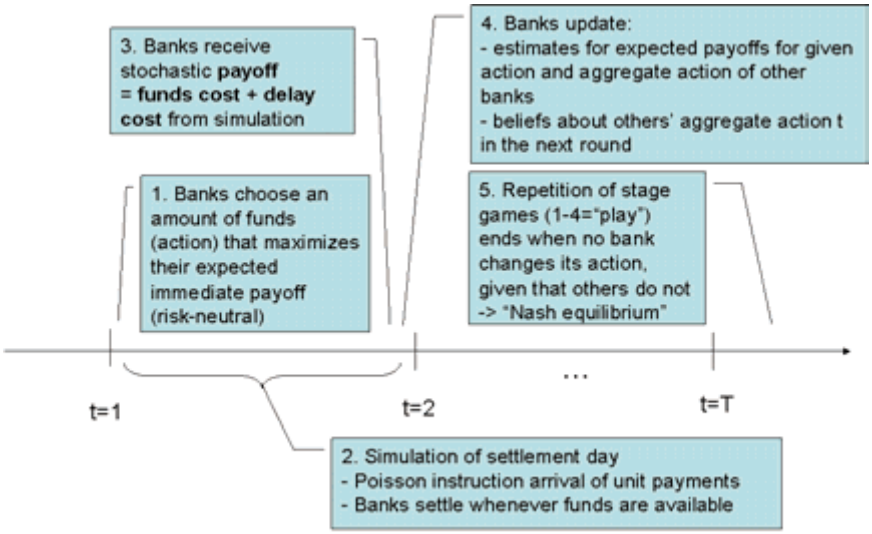
2.4.2 Multi-agent model of bank behaviour

Galbiati and Soramäki (2007) develop a dynamic multi-agent model of an interbank payment system where payments are settled on the basis of pre-committed funds. In the model banks choose their level of committed funds on the basis of private payoff maximisation.

The model consists of a sequence of settlement days. Each of these days is a simultaneous-move game, in which each bank chooses the amount of liquidity to commit for payment processing and receives a stochastic payoff. Payoffs are determined by means of simulating the settlement day with the amounts of liquidity chosen by the banks. Instructions to be settled by the banks arrive on the basis of a Poisson process and are ex-ante unknown to the banks. As shown in Section 2.3.2, the relationship between instruction arrival and payment settlement is very complex and could not so far be described analytically. Adaptation takes place through reinforcement learning

with Bayesian updating, with banks maximising immediate payoffs. Figure 2.5 shows the sequence of decisions, events and learning in the model.

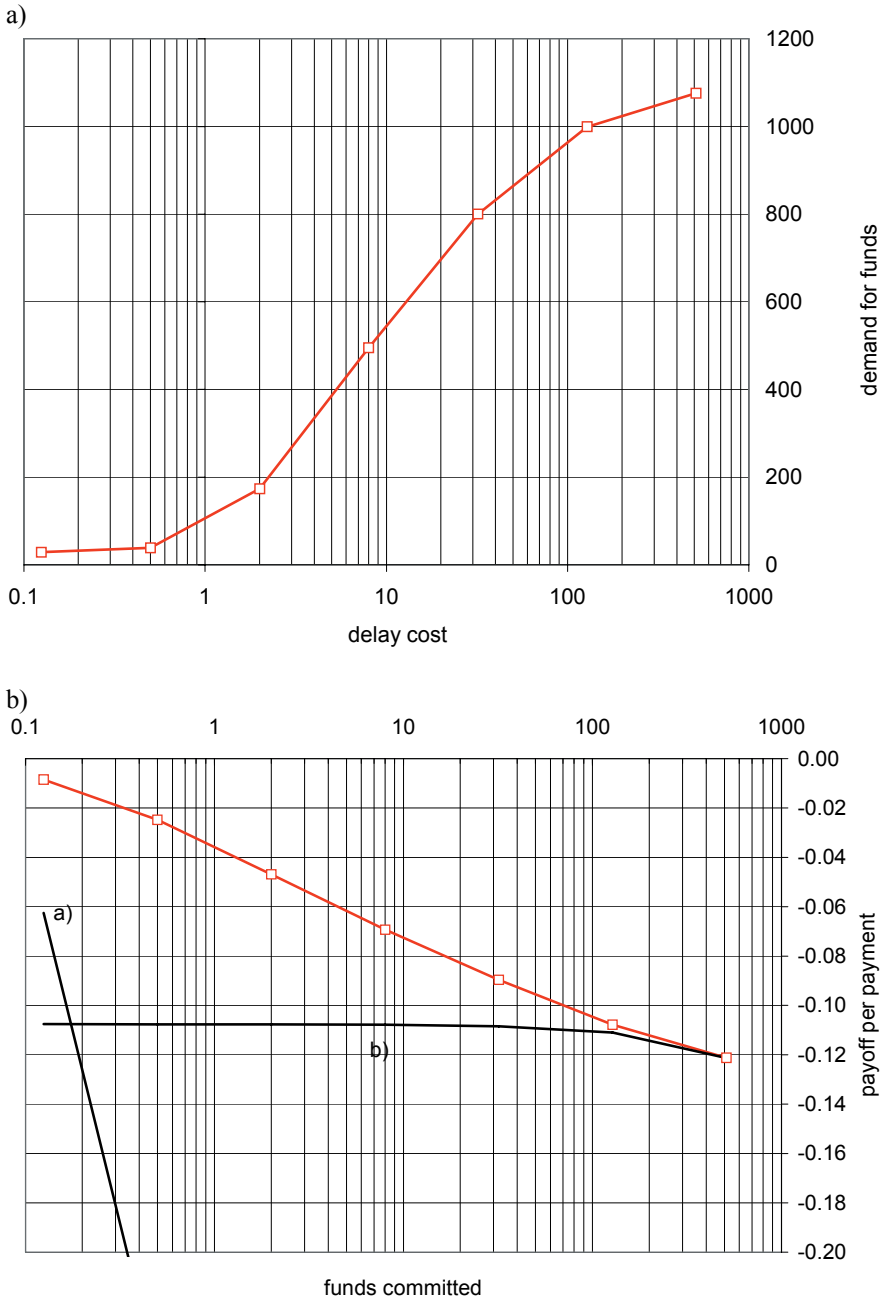
Figure 2.5 **Overview of a multi-agent learning model of a payment system (Galbiati and Soramäki, 2007)**



By the process of individual pay-off maximisation, banks adjust their demand for liquidity up (reducing delays) when delay costs increase and down (increasing delays), when they rise. It is well known that the demand for intraday credit is generated by a tradeoff between the costs associated with delaying payments and liquidity costs. Simulating the model for different parameter values, they find that the demand for intraday credit is an S-shaped function of the cost ratio between intraday credit costs and the costs associated with delaying payments[7] (see Figure 2.6a).

[7] In the model both costs are assumed to be linear.

Figure 2.6 **Demand for intraday credit (a), Payoff comparison (b)**



An interesting question is how good the performance of the banks is in absolute terms. To understand this we compare the payoffs received by the banks through adaptation with two extreme strategies:

a) delay all payments to the end of the day;
b) commit enough liquidity to be able to process all payments promptly.

The performance of these three strategies is shown in Figure 2.6b. For any level of the delay cost, the adaptive banks obtain better payoffs than either of the two extreme strategies as they manage to learn a convenient trade-off between delay and liquidity costs. On the contrary, the strategy under a) becomes quickly very expensive as delay costs increase, and the strategy under b) is exceedingly expensive when delays are not costly.

Ideally, banks should be taking into consideration the future stream of pay-offs as well. This would create a value of information to the banks as discounting expected future payoffs would create an explicit trade-off between exploitation (the use of actions that appear optimal in the light of the available information) and exploration (the use of seemingly sub-optimal actions, which might appear such because of lack of experimentation). Banks may also be risk-averse, interested not only in the expected pay-off but also its variability. These are among the topics for future research.

2.5 Conclusion

This article presented three elements of payment systems, new approaches for understanding and analysing them, and examples on how these approaches can be applied to specific research questions. It argues that performance of a payment system is a function of network topology, the 'physics' of the system and the behaviour of banks – one factor alone is not enough to evaluate efficiency or robustness.

First, the payment system can be understood as a network of liquidity flows and can be modelled as a graph. Each model of a payment system assumes some topology, be it random, complete or a topology closer to the system being modelled - such as the scale-free topology of Fedwire. Graph theory and social network analysis provide good tools for analysing the structure of interbank payment systems and their liquidity flows. Understanding how banks are connected in the payment network is important for analysing their

robustness. The concepts developed in the field can help us structurally analyse payment flows in the system (see eg Newman, 2003). Measures of average path length can tell us how quickly disturbances are likely to reach other banks in the network. More research is clearly needed to identify measures that explain the connection between system topology and its robustness. Centrality measures can help us identify banks that are not only important through their size, but also due to their position in the network and due to their linkages to other banks (see eg Borgatti, 2005). A likely fruitful area in payment system research would be to use such approaches for the identification of important (and vulnerable) banks in networks representing RTGS or netting systems.

Second, payment systems have rules, procedures and technical constraints for the processing of individual payments that may produce emergent behaviour at the system level. An example of these is the settlement cascades that take place at low levels of liquidity and low market conductance. The model of payment system dynamics exhibits a transition from independent to highly interdependent behaviour and allows the study of factors that control system-wide interdependence. Complexity theory and models developed in statistical mechanics (see eg Bak, 1987, and Sachtjen et al, 2000) can help explain how simple local rules create emergent system-level behaviour.

Third, banks react to changes in the environment – be these changes in policy or disruptions to the system's operation or changes in the behaviour of other banks. Understanding how banks might react, and the impact of simultaneous reactions at the system level, greatly helps in evaluating risks and efficiencies of payment systems. While the incentives of banks may be analysed individually in isolation or when operating in a stipulated environment, their interaction in a system of banks with their own incentives necessitates a model. In modelling bank behaviour, methodologies developed under reinforcement learning (Sutton and Barto, 1998) and learning in games (Fudenberg and Levine 1998) may prove useful. As seen by the given example, mere simple 'intelligence' by agents can produce realistic behaviour and add value to the analysis of payment systems. In the development of more realistic behaviour for banks in settling payments, an important unanswered question is whether and what kind of bank behaviour can be identified from empirical payment data.

References

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

Angelini, P – Maresca, G – Russo, D (1996) **Systemic risk in the netting system.** Journal of Banking and Finance 20, 853–868.

Albert, R – Jeong, H – Barabasi, A-L (1999) **Error and attack tolerance of complex networks.** Nature 406, 378–382.

Bak, P – Tang, C – Wiesenfeld, K (1987) **Self Organized Criticality: An explanation of 1/f Noise.** Physical Review Letters 59:4, 381–384.

Barabasi, A-L – Albert, R (1999) **Emergence of scaling in random networks.** Science 286, 509–512.

Barto, A G – Sutton, R S (1998) **Reinforcement Learning: an introduction.** MIT Press, Cambridge, Massachusetts.

Bech, M – Madsen, B – Natorp, L (2002) **Systemic Risk in the Danish Interbank Netting System.** Danmarks Nationalbank Working Paper No. 8.

Bech, M – Soramäki, K (2002) **Liquidity, gridlocks and bank failures in large value payment systems.** In E-money and Payment Systems Review. Central Banking Publications.

Bech, M – Soramäki, K (2005) **Systemic risk in a netting system revisited.** In Leinonen, H (ed.) Liquidity, risks and speed in payment and settlement systems – a simulation approach. Bank of Finland Studies E:31.

Bech, M – Garratt, R (2003) **The intraday liquidity management game.** Journal of Economic Theory, Vol. 109(2), 198–219.

Bech, M – Garratt, R (2006) **Illiquidity in the Interbank Payment System following Wide-Scale Disruptions.** Federal Reserve Bank of New York Staff Report 239.

Bedford, P – Millard, S – Yang, J (2005) **Analysing the impact of operational incidents in large-value payment systems: a simulation approach.** In Leinonen, H (ed.) Liquidity, risks and speed in payment and settlement systems – a simulation approach. Bank of Finland Studies E:31.

Bernanke, B (1990) **Clearing and Settlement during the Crash.** Review of Financial Studies, Vol. 3, Issue 1, 133–151.

Beyeler, W – Bech, M – Glass, R – Soramäki, K (2007) **Congestion and Cascades in Payment Systems.** Physica A. Forthcoming.

Bollobas, B (1985) **Random Graphs.** Academic Press, London, 123–136.

Bonacich, P (1972) **Factoring and weighting approaches to status scores and clique identification.** Journal of Mathematical Sociology 2, 113–120.

Borgatti, S P (2005) **Centrality and network flow.** Social Networks 27, 55–71.

Brin, S – Page, L (1998) **The anatomy of a large-scale hypertextual Web search engine.** Proceedings of the seventh international conference on World Wide Web 7, 107–117.

Brown, G W (1951) **Iterative Solutions of Games by Fictitious Play.** In Koopmans, T C (ed.) Activity Analysis of Production and Allocation. Wiley, New York.

Callaway, D – Newman, M – Strogatz, S – Watts, D (2000) **Network robustness and fragility: Percolation on random graphs.** Phys. Rev. Lett. 85, 5468–5471.

Crucitti, P – Latora, V – Marchiori, M – Rapisarda, A (2004a) **Error and attack tolerance of complex networks.** Physica A 340, 380–394.

Crucitti, P – Latora, V – Marchiori, M (2004b) **A model for cascading failures in complex networks.** Phys. Rev. E 69, 045104.

Devriese, J – Mitchell, J (2006) **Liquidity Risk in Securities Settlement.** Journal of Banking and Finance, Vol. 30, Issue 6, 1807–1834.

Dorogovtsev, S N – Mendes, J F F – Samukhin, A N (2001) **Giant strongly connected component of directed networks.** Phys. Rev. E 64.

Dorogovtsev, S N – Mendes, J F F (2003) **Evolution of Networks.** Oxford University Press.

Erdõs, P – Rényi, A (1959) **On random graphs.** Publicationes Mathematicae 6, 290–297.

Freeman, L C (1978) **Centrality in social networks: I. Conceptual clarification.** Social Networks 1, 215–239.

Fudenberg, D – Levine, D K (1998) **The Theory of Learning in Games.** MIT Press. Cambridge, Massachusetts.

Galbiati, M – Soramäki, K (2007) **A competitive multi-agent model of interbank payment systems.** arXiv:0705.3050v1 [cs.MA].

Hellqvist, M – Koskinen, J (2005) **Stress testing securities clearing and settlement systems using simulations.** In Leinonen, H (ed.) Liquidity, risks and speed in payment and settlement systems – a simulation approach. Bank of Finland Studies E:31.

Humphrey, D B (1986) **Payments Finality and Risk of Settlement Failure.** In Saunders, A and White, L J (eds) Technology and the Regulation of Financial Markets: Securities, Futures, and Banking. Heath: Lexington.

Inaoka, H – Ninomiya, T – Taniguchi, K – Shimizu, T – Takayasu, H (2004) **Fractal Network derived from banking transaction – An analysis of network structures formed by financial institutions.** Bank of Japan Working Papers No. 04-E-04.

Jensen, H J (1998) **Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems.** Cambridge University Press, ISBN 0-521-48371-9.

Kinney, R – Crucitti, P – Albert, R – Latora, V (2005) **Modelling cascading failures in the North American power grid.** Eur. Phys. J. B 46, 101–107.

Kobayakawa, S (1997) **The comparative analysis of settlement systems.** Centre for Economic Policy Research Discussion Paper No. 1667.

Kuussaari, H (1996) **Systemic Risk in the Finnish Payment System: An Empirical Investigation.** Bank of Finland Discussion Paper 3.

Lacker, J M (2004) **Payment system disruptions and the federal reserve following September 11, 2001.** Journal of Monetary Economics, Vol. 51, No. 5, 935–965.

Leinonen, H (ed) (2005) **Liquidity, risks and speed in payment and settlement systems – a simulation approach.** Bank of Finland Studies E:31.

Mazars, E – Woelfel, G (2005) **Analysis, by simulation, of the impact of a technical default of a payment system participant.** In Leinonen, H (ed.) Liquidity, risks and speed in payment and settlement systems – a simulation approach. Bank of Finland Studies E:31.

McAndrews, J – Potter, S M (2002) **The Liquidity Effects of the Events of September 11, 2001.** Federal Reserve Bank of New York Economic Policy Review, Vol. 8, No. 2, 59–79.

McAndrews, J – Rajan, S (2000) **The timing and funding of Fedwire funds transfers.** Federal Reserve Bank New York Econ. Pol. Rev. 6, 17–32.

Moore, C – Newman, M E J (2000) **Exact solution of site and bond percolation on small-world networks.** Physical Review E 62: 7059–7065.

Newman, M E J (2003) **The Structure and Function of Complex Networks.** SIAM Review, Vol. 45, No. 2, 167–256.

Newman, M E J (2005) **Power laws, Pareto distributions and Zipf's law.** Contemporary Physics, Vol. 46, No 5, 323–351.

Northcott, C A (2002) **Estimating Settlement Risk and the Potential Contagion in Canada's Automated Clearing Settlement System.** Working Paper 2002-41. Bank of Canada.

Sachtjen, M L – Carrerras, B A – Lynch, V E (2000) **Disturbances in a power transmission system.** Physical Review E, 61:5, 4877–4882.

Soramäki, K – Bech, M L – Glass, R J – Beyeler, W E (2007) **The Topology of Interbank Payment Flows.** Physica A. Vol. 379, 317–333.

Watkins, C – Dayan, P (1992) **Technical note: Q-learning.** Machine Learning, 8(3/4), 279–292.

Watts, D J (2002) **A simple model of global cascades on random networks.** PNAS, April 30, 2002, Vol. 99, No. 9, 5766–5771.

Chapter 3

From PNS to TARGET2: the cost of FIFO in RTGS payment systems

Fabien Renault – Jean-Baptiste Pecceu

3 From PNS to TARGET2: the cost of FIFO in RTGS payment systems 43

Abstract 43

3.1 Optimisation in RTGS.....44
3.1.1 From net to hybrid systems 44
3.1.2 Optimisation and the FIFO rule.....45
3.1.3 Formalisation of the problem 46
3.1.4 Non-FIFO features in PNS and TARGET2.....47
3.1.5 Objectives of the paper.....48
3.2 Bilateral optimisation 49
3.2.1 Bilateral optimisation in PNS and TARGET2.....49
3.2.2 The bilateral Greedy algorithm 50
3.2.3 New ideas regarding bilateral optimisation 51
3.2.4 A few basic examples.....52
3.2.5 Relative efficiency of bilateral optimisation algorithms.....54
3.2.6 Conclusion regarding bilateral optimisation.....57
3.3 Multilateral optimisation 57
3.3.1 Multilateral optimisation in PNS and TARGET2 57
3.3.2 A new concept: pre-conditioning 58
3.3.3 The multilateral Las-Vegas algorithms 59
3.3.4 Relative efficiencies of multilateral optimisation algorithms.....60
3.4 Optimisation in PNS in case of an operational failure.....63
3.4.1 The PNS system 63
3.4.2 Simulating the technical default of the largest participant in the system.....64

3.4.3 Results 65
3.4.4 Payments rejected at the end of closure 68
3.5 Conclusions 70

References 72

Appendix 1. Bilateral optimisation algorithms 73
Appendix 2. Multilateral optimisation algorithms 76
Appendix 3. The Greedy algorithm of Güntzer et al and superincreasing payment values distributions 81

3 From PNS to TARGET2: the cost of FIFO in RTGS payment systems

Abstract

Most of the recent RTGS payment systems are equipped with various optimisation algorithms that are able to increase the settlement speed by resolving fully or partially some of the gridlock situations that arise in the system. Today, most of the optimisation algorithms in use follow – at least partially– the FIFO (First In First Out) rule, meaning that they always settle the queued payments in their order of arrival. While the FIFO rule may be desirable based on some other considerations, for example legal ground, it creates an additional constraint to the optimisation problem, potentially leading to a less efficient solution in terms of settled value. The aim of this paper is to try to quantify to which extent non-FIFO optimisation algorithms can be more efficient than FIFO algorithms.

In the first part of this paper, some simulations performed on randomly generated sets of payments are used to evaluate the efficiency of several FIFO and non FIFO optimisation algorithms. This analysis is conducted both in the case of bilateral optimisation and in the case of multilateral optimisation. The results show that in those conditions, some non-FIFO algorithms are able to improve significantly on their FIFO counterparts.

In a second part, the impact of the different optimisation algorithms is investigated further by simulating the complete PNS system using real data. In the context of a liquidity crisis created by the technical failure of the largest participant of the system, the use of some non-FIFO algorithms is shown to reduce the number of rejected payments at the end of the day.

3.1 Optimisation in RTGS

3.1.1 From net to hybrid systems

The last two decades have witnessed important transformations in the field of payment systems. Pure DNS (Deferred Net Settlement) systems, in which payment orders are stored throughout the day and the resulting net balances are settled only once at the end of the day, were the predominant form of LVPS (Large Value Payment Systems) in the 1980s. Although DNS systems are extremely efficient in terms of central bank money usage, the absence of intraday finality leading to potentially large intraday exposures raised some concerns in the context of ever-increasing values exchanged. Indeed if one participant fails to meet its end-of-day payment obligations in an unprotected DNS system, some or all payments involving this participant have to be unwound, potentially leading to the default of other participants and further unwinding. This potential domino effect can have unpredictable consequences on the final cash balances of each participant and on the number of rejected payments at the end of the day and thus undermines confidence in the payment system.

For these reasons, DNS systems were progressively replaced in the 1990s by RTGS (Real Time Gross Settlement) systems, in which payments are settled one by one as soon as the payment orders enter the system (and provided sufficient liquidity is available). Compared to DNS systems, RTGS systems tremendously reduce the risks associated with exchanging large value payments, but they also require significantly higher levels of central bank money to operate.

In order to reduce the central bank money usage of their participants, RTGS systems progressively adopted several payment-offsetting features.[1] Payments that cannot be settled immediately are held in a centrally-organised queue, and more or less sophisticated optimisation algorithms are used to try and simultaneously settle groups of queued payments that can not be settled individually.

Examples of such RTGS systems with offsetting mechanisms, sometimes referred to as 'hybrid systems', include the French LVPS PNS (Paris Net Settlement) and the future pan-European system TARGET2. Besides offsetting algorithms, those two systems offer the participants the possibility to establish bilateral sending limits towards

[1] Here offsetting is to be understood as the gross execution of individual payments simultaneously within one legal and logical second. From a legal perspective, offsetting in RTGS is very different from the netting process in DNS.

their counterparties. A bilateral limit is the net amount of money a participant is willing to pay another participant before being paid back. This feature is helpful for risk management purposes and creates incentives to submit payments early into the system. Indeed, when intraday liquidity is scarce in a payment system, some participants might delay their payments in order to get a free ride on other participants' liquidity (see eg Bech and Garratt, 2003). When no bilateral sending limit feature is available, if bank A is not willing to grant bank B free intraday credit, the only solution bank A has is to retain its payments towards bank B in its own internal queue (located in its private IT infrastructure and invisible to the system and other participants). Conversely, if bank A can establish a bilateral limit towards bank B, bank A can submit payments towards bank B and let them be blocked by the RTGS system. Bank B is therefore incentivised to submit payments towards bank A. Doing so will not deplete bank B's liquidity stock because bank B's submission of payments towards bank A will trigger the release of bank A's payments towards bank B. Bilateral sending limits, together with offsetting mechanisms, thus transform intraday liquidity management from a competitive game (whoever submits his payments last wins) into a cooperative game (I will pay you at the exact time you pay me, so it is optimal for you to pay me early).

3.1.2 Optimisation and the FIFO rule

The benefits provided by offsetting algorithms in terms of lower liquidity needs in RTGS have been extensively investigated in recent years, notably thanks to the development of simulation tools for RTGS systems. Koponen and Soramäki (2005) and Leinonen and Soramäki (2005), among others, clearly showed how offsetting algorithms could for a given level of liquidity reduce the settlement delay and conversely reduce the liquidity needs for a given level of delay.

However, most of the analysis done until now relates to the use of optimisation algorithms that follow the First In First Out (FIFO) rule, meaning that payments have to be settled in the order they entered the system. While this constraint might be supported by some participants wishing to keep full control of their payment queue and might also be desirable from a legal point of view, it potentially lowers the efficiency of the optimisation algorithm in terms of settled value. Clearly, if a single very large payment is first in the queue, it might

block many later-submitted smaller payments, and a FIFO algorithm will not be able to do anything about it.

The aim of this paper is to investigate other types of offsetting algorithms which do not necessarily follow the FIFO rule and to try and quantify to what extent non-FIFO optimisation algorithms can be more efficient than FIFO algorithms. In other words, we will try to calculate the cost of the FIFO rule for RTGS systems in terms of decreased efficiency of the optimisation mechanisms.

Bech and Soramäki (2001 and 2005) formalised the problem by introducing a clear distinction between the Gridlock Resolution Problem (GRP, ie the problem of optimisation under the FIFO constraint, as defined by Bech and Soramäki) and the Bank Clearing Problem (BCP, ie the free optimisation problem, as referred to by Güntzer et al, 1998).

3.1.3 Formalisation of the problem

The notations used in this section come from Bech and Soramäki (2001). We consider n banks ($i = 1 \ldots n$) participating in a RTGS system, each characterised by its initial amount of liquidity S_i. The queue of bank i contains m_i payments waiting to be settled. The kth payment sent by bank i is characterised by its value $a_{i,k}$ and the receiving bank designated by the integer $r_{i,k} \in \{1,2,...,n\} \setminus \{i\}$.

In order to be able to characterise any subset of the queued payments, we will use the indicators $x_{i,k} \in \{0,1\}$. A value of 1 (respectively 0) attributed to $x_{i,k}$ simply means that the kth payment of bank i is included (respectively not included) in the considered subset.

Bech and Soramäki define the **Gridlock Resolution Problem** as finding the $(x_{i,k})_{\substack{i=1...n \\ k=1...m_i}}$ that maximise the total value settled $V = \sum_{i=1}^{n} \sum_{k=1}^{m_i} a_{i,k} x_{i,k}$ under the dual condition

$$\begin{cases} \forall i \in \{1...n\}, \quad S_i - \sum_{k=1}^{m_i} a_{i,k} x_{i,k} + \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{k=1}^{m_j} a_{j,k} x_{j,k} \delta_{r_{j,k}=i} \geq 0 \\ \forall i \in \{1...n\}, \quad \forall k \in \{1...m_i - 1\}, x_{i,k+1} \leq x_{i,k} \end{cases}$$

where $\delta_{r_{j,k}} = i$ is equal to 1 if $r_{j,k} = i$ and zero otherwise.

The first condition is the liquidity constraint. It simply states that a bank cannot have a negative cash balance within the considered payment system. The second condition is the sequence constraint. It simply translates the fact that bank i wants its payments settled in the chronological order in which they were received by the system.

The **Bank Clearing Problem**, as defined by Güntzer et al (1998) is similar to the Gridlock Resolution Problem with the difference that the second constraint (the sequence condition) is not present in the BCP.

3.1.4 Non-FIFO features in PNS and TARGET2

Neither the French LVPS PNS nor the future pan-European RTGS TARGET2 totally comply with the sequence constraint of the GRP problem, as explained in the previous section. Indeed, the FIFO rule is arguably breached on several occasions.

First, in both PNS and TARGET2, the FIFO principle is to be understood on a bilateral basis. A payment from bank A to bank B can be settled before a payment from bank A to bank C that entered the system earlier. Moreover, it is clear that such an exception to the FIFO rule will be present in all systems offering the participants the possibility to set bilateral limits towards their counterparties. Indeed, if the payment from bank A to bank C is queued because the bilateral limit bank A has set towards bank C has been reached, bank A will still want to be able to settle payments towards its other counterparties.

Furthermore in PNS, a low value payment (whose value is lower than EUR 1 million) from bank A to bank B will be settled directly by the entry mechanism of the system, provided bank A has the necessary funds and whether or not earlier submitted payments from bank A to bank B are present in the queue. The aim of this rule is to avoid a situation where a very large queued payment creates a blockage, unnecessarily delaying the settlement of many small payments. A similar feature exists in the entry mechanism of TARGET2: indeed when a normal priority payment[2] is submitted, '*it is not checked whether the normal [priority payments] queue is empty, because the FIFO principle can be breached for normal [priority] payments'*.[3]

[2] In TARGET2, the participants will be able to choose either normal priority or urgent priority for each payment they emit.

[3] TARGET2 User Detailed Functional Specifications, first book version 2.0, page 145.

Even when retaining a bilateral definition of the FIFO rule, PNS (for payments lower than EUR 1 million) and TARGET2 (for normal priority payments) do not comply with this rule in the entry mechanism.

Finally, another breach of the FIFO rule occurs in the multilateral optimisation algorithm of the PNS system which attempts to settle simultaneously all queued payments of all participants. In case it is impossible to settle all queued payments because one or several participants do not have sufficient liquidity, the algorithm will consider the participant having the largest net debit position and de-activate the smallest of its payments whose value exceeds the value of its net debit position (in case no payment exceeds the value of the net debit position, the biggest payment of the participant having the largest net debit position is de-activated). In this special case, the payments are then selected according to their value, and not according to the order they arrived in the system.

3.1.5 Objectives of the paper

We have just shown that the settlement process of PNS, in particular for low value payments, and of TARGET2 in the case of normal priority payments, breach the FIFO rule on several occasions. Moreover, in TARGET2 normal priority payments can by-pass other queued payments in the entry mechanism while they are treated according to a strict FIFO rule (in a bilateral-FIFO sense) in the bilateral optimisation algorithm. One can thus feel entitled to investigate the benefits non-FIFO optimisation algorithms could bring to the system.

There are several good reasons for a payment system to follow the FIFO principle: it makes the rules of the system easier and allows participants to keep full control of the order their payments are settled. For this last reason in particular, some treasurers are very fond of the FIFO principle. Moreover, FIFO optimisation algorithms are fast, simple to understand and easy to implement while efficient enough to solve many gridlock situations.

In theory the drawback associated with the lack of flexibility the FIFO principle represents is decreased settlement efficiency. All other things being equal, a pure FIFO RTGS is characterised by a higher settlement delay than a RTGS equipped with more advanced non-FIFO offsetting algorithms.

The aim of this paper is not to discuss whether or not the FIFO principle should continue to be applied in today's RTGS, as many

other considerations may influence the conclusion that could be made regarding this topic. Instead, the objective of this contribution is to try and quantify the expected increase in settlement efficiency that would allow the use of non-FIFO offsetting algorithms.

Two types of optimisation algorithms co-exist in PNS and in the future TARGET2 system: bilateral optimisation and multilateral optimisation. We will examine them successively in a theoretical framework before moving to a 'real-life case' in the PNS system.

3.2 Bilateral optimisation

In this section we focus on bilateral optimisation, ie we examine two participants A and B and consider only queued payments from A to B and from B to A. The objective of a bilateral optimisation algorithm is to settle simultaneously a set of queued payments for as high a total cumulated value as possible (the number of settled payments is also of interest as a 'secondary objective', although the settled value is usually considered more important).

One may wonder why optimisation should be performed on a bilateral basis rather than directly on a multilateral basis, ie considering all queued payments of all participants at the same time. In theory, any solution provided by a bilateral optimisation algorithm could also be found by a multilateral optimisation algorithm while the opposite is not true. In practice, bilateral optimisation takes profit from the usually relatively high level of reciprocity of payment networks in order to drastically reduce the number of variables and the complexity of the problem. Another important element is the presence of bilateral sending limits (cf 3.1.1) which create a strong linkage between the payments exchanged by a pair of participants (A will pay B if and only if B pays A). While the treatment of bilateral limits is cumbersome in a multilateral optimisation algorithm, it is very easily implemented and effective in a bilateral optimisation algorithm.

For those reasons, bilateral optimisation and multilateral optimisation can be considered as complimentary and are both used in PNS and in TARGET2.

3.2.1 Bilateral optimisation in PNS and TARGET2

The two systems, PNS and TARGET2, rely on the same bilateral optimisation algorithm. This algorithm follows the FIFO rule in a

bilateral sense. First, the algorithm tries to settle all payments queued between the two banks simultaneously. If this is not possible, the most recent payment from the participant lacking liquidity is de-activated. This process is iterated until all payments have been de-activated or until a solution has been found. The 'FIFO bilateral optimisation algorithm' is described in detail in Appendix 1.

The fact that PNS and TARGET2 rely on the same bilateral optimisation algorithm comes as no surprise. It is indeed quite easy to show that the bilateral optimisation algorithm used in PNS and TARGET2 is the best algorithm that abides by the bilateral-FIFO rule, in the sense that it will always provide the unique solution maximising both the volume and value settled (Bech and Soramäki, 2001).

3.2.2 The bilateral Greedy algorithm

The bilateral Greedy algorithm was proposed by Güntzer et al in 1998. Payments are not retained according to their arrival order but according to their value. As in the FIFO bilateral optimisation algorithm, the Greedy algorithm first tries to settle all payments queued between the two banks simultaneously. If this is not possible, all payments from the participant lacking liquidity are de-activated and are then re-activated whenever possible given the liquidity constraint in the decreasing order of their value. This process is iterated until all payments have been de-activated or until a solution has been found. The details of the algorithm can be found in Appendix 1. Compared to the FIFO algorithm used in PNS and TARGET2, bigger payments are favoured at the expense of payments that entered the system early. One of the advantages of the Greedy algorithm over the FIFO algorithm is that queues will not be blocked due to a single very large payment that would prevent all subsequent payments from settling.

A very interesting property of the Greedy algorithm is that it yields a solution that maximises the value of payments settled when the sequences of values of the queued payments are superincreasing, that is to say when every queued payment from A to B is larger than the sum of all the smaller queued payments from A to B and every queued payment from B to A is larger than the sum of all the smaller queued payments from B to A. A proof of this claim is presented in Appendix 3. In the case of the PNS system, it can be shown that any average set of 3 payments has a 95% chance of forming a superincreasing sequence. This probability drops to 65% if we consider a set of 5 payments and to only 2% if we consider a set of 10 payments. The

ideal case of superincreasing sequences is therefore not unrealistic when there are only few queued payments between two given participants (as is often the case in PNS in normal working conditions). It is also important to keep in mind that the Greedy algorithm can very well provide the best solution even if the payment sequences are not superincreasing, although this is not guaranteed in this case.

Another interesting feature of the Greedy algorithm lies in its simplicity and speed. Indeed, once queued payments have been ordered according to their value, the number of operations to perform is only proportional to the number of queued payments, that is to say the Greedy algorithm is not slower than the simple FIFO algorithm used in PNS and TARGET2. The time needed to order a set of N payments is typically proportional to N.log(N) but such a task only needs to be performed once. Furthermore, the tests showed that compared to the FIFO algorithm, fewer iterations were needed for Greedy to produce a solution.

3.2.3 New ideas regarding bilateral optimisation

We present some new ideas regarding bilateral optimisation. The Greedy algorithm is already very efficient but is not guaranteed to give the best solution when payment values are not superincreasing. Is it possible to improve on Greedy?

Two distinct ideas were investigated. The first idea is to introduce some flexibility to Greedy, which always re-activates payments in the decreasing order of their value. We consider the problem of bilateral optimisation between bank A and bank B and denote payments from A to B as the $(a_i)_{i=1...N}$ where a_1 is the biggest payment and a_N is the smallest payment. It is clearly optimal to re-activate a payment a_i satisfying $a_i \geq \sum_{k=i+1}^{N} a_k$ as we know that the Greedy algorithm will yield the best answer for a superincreasing payment sequence (see Appendix 3). On the other hand, if the sequence is not locally superincreasing, ie if $a_i < \sum_{k=i+1}^{N} a_k$, it is unclear whether the payment a_i should be re-activated or not. The idea behind the Las Vegas Greedy bilateral optimisation algorithm is to try both solutions, stochastically. The algorithm is presented in more detail in Appendix 1. It is important to note that for superincreasing payment value sequences, the Las Vegas Greedy algorithm degenerates into Greedy.

The other idea investigated was to try and benefit from the ever-increasing computational power available to try more settlement possibilities than Greedy does. While this computational power might not be sufficient to try each of the 2^{N+M} possibilities involved (N is the number of queued payments from A to B and M is the number of queued payments from B to A), it is reasonable to consider that some of them might be tested to the limit of, say, a thousand cases. The key question is now how to select those cases to be tested. The idea behind Greedy++ is to run the Greedy algorithm, and after each iteration that does not yield a solution because one of the participants does not have the needed liquidity, to test all possibilities involving the 10 payments closest to the error (ie the $2^{10} = 1024$ cases obtained when considering the re-activation/de-activation of the 10 payments closest to the error, all other payments staying in the same state). If a solution is found, then the solution maximising the settled value will be chosen. When this treatment yields no solution, a Greedy iteration is applied, hence the name of Greedy++ for this algorithm. The algorithm is presented in more detail in Appendix 1.

3.2.4 A few basic examples

The aim of this section is to help understand concretely how the algorithms work on practical examples.

Example 1

		Bank A	Bank B
Cash balance		10	10
Queued payments, in order of arrival	1st	500	20
	2nd	20	20
	3rd	20	20
	4th	20	20

In this example a large payment from bank A to bank B (of value 500) is preventing subsequent payments from settling. Clearly, nothing can be settled with a FIFO algorithm as any solution would involve the by-passing of bank A's earliest-sent payment. The Greedy algorithm, as well as Greedy++ and the bilateral Las Vegas Greedy, will however find the value maximising solution (settle bank A's second, third and fourth payment together with three of bank B's payments).

Example 2

		Bank A	Bank B
Cash balance		0	10
Queued payments, in order of arrival	1	140	20
	2		20
	3		30
	4		100

This example is typical of non-superincreasing payment sequences (here 30 < 20+20 so the sequence of Bank B's payments is not superincreasing). The Greedy algorithm will start by activating all payments, and as bank B has a negative virtual position (-30), will de-activate all payments from B to A and re-activate them in the decreasing order of their value. By re-activating the payment of value 30, Greedy will miss the trivial solution (140 = 100 + 20 + 20) and terminate without settling any payment.

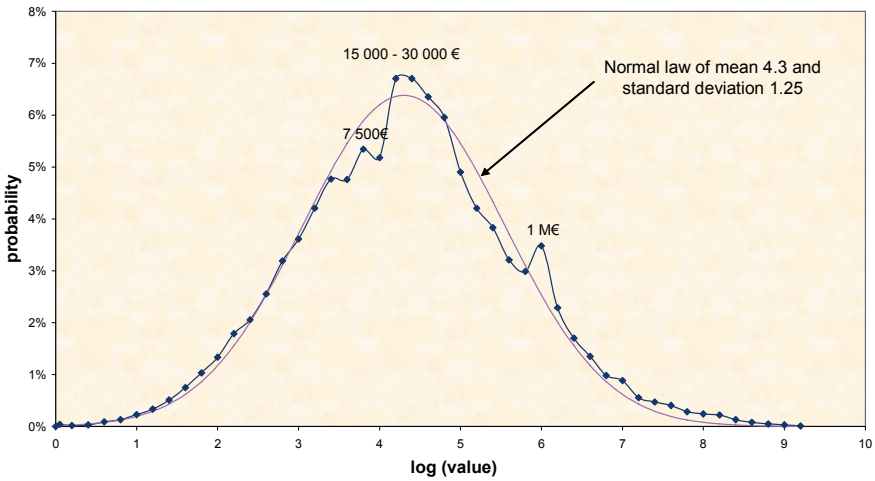
The Greedy++ algorithm will start by activating all payments, and as settlement is impossible, will examine all possibilities involving the 10 payments closest to the error (here the error is equal to 30, and as there are only 5 payments in the queue, the $2^5 = 32$ possibilities will be tried, and the value maximising solution will be retained). Greedy++ will therefore find the correct solution – as always when the number of queued payments is fewer than 10.

The bilateral Las Vegas Greedy algorithm will also start by activating all payments, and after noticing that B has a negative virtual position, will de-activate all payments from B to A. Payments from B to A will then be considered for re-activation in the decreasing order of their value, up to a total cumulated value of 140 (the sum of the activated payments from A to B + B's position). Bank B's biggest payment, of value 100, will be re-activated with a probability of 100% since the cumulated value of the lower payments, 70, is strictly lower than B's virtual position of 140. The payment of value 30 is then considered for re-activation. It will be re-activated with a probability equal to $\frac{30}{20+20} = 75\%$. If the algorithm is launched 10 times, the probability for the value-maximising solution to be found is then close to 95%.

3.2.5 Relative efficiency of bilateral optimisation algorithms

In order to compare the different bilateral optimisation algorithms presented in the previous pages, the following test was developed:

Figure 3.1 **Payment value distribution in the PNS system**



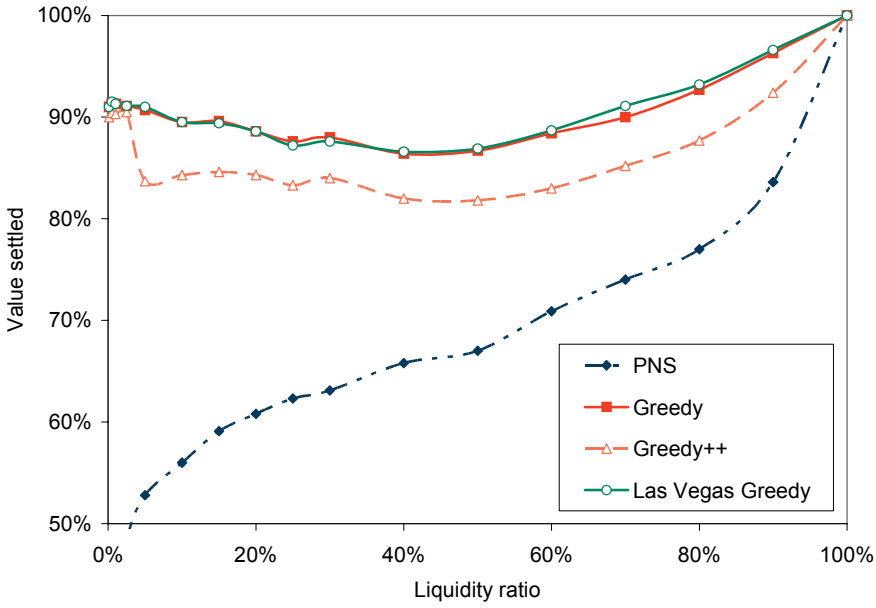
- We considered two banks, A and B. We assumed that there are N queued payments from bank A to bank B, the $(a_i)_{i=1...N}$, and N queued payments from bank B to bank A, the $(b_i)_{i=1...N}$. The value of each of these 2N payments was generated randomly according to the observed payment value distribution in the PNS system: as shown in the above graph, the payment distribution in PNS can be approximated by a log normal law of mean 4.3 and of standard deviation 1.25 with great accuracy.

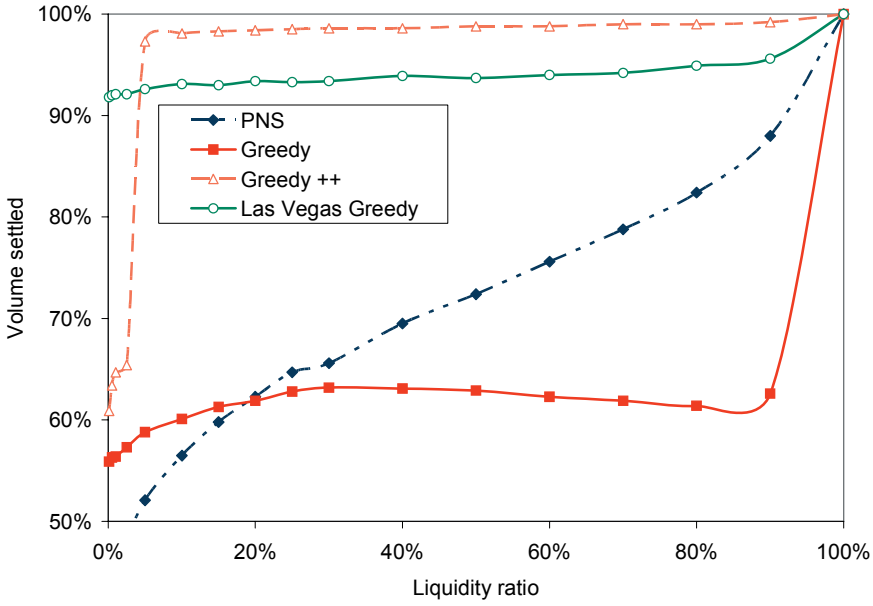
- We can assume without any loss of generality that the sum of the values of the payments emitted by A, designated by $G = \sum_i a_i$ exceeds the sum of the values of the payments emitted by B, noted as $H = \sum_i b_i$. The starting balance of bank B, S_B is then set to zero, while the starting balance of bank A, the net emitter, is set to $S_A = \alpha(G - H)$, where α is a parameter ranging from 0 (no liquidity is present at all), to 1 (all queued payments can be settled simultaneously).

- The presented problem of bilateral optimisation was run with the PNS/T2 FIFO algorithm, Greedy, Greedy++ and Las Vegas Greedy bilateral algorithm. Regarding the Las Vegas Greedy bilateral algorithm, it was applied 5 times in a row (ie it was applied a first time to the initial problem, then it was applied a second time to what had not settled the first time, and so on…). The results were averaged over 5,000 different payment distributions generated randomly, according to the presented log normal law.

- The results obtained in terms of value and volume settled are shown in Figure 3.2. While the volume efficiency is defined simply as the ratio between the number of settled payments and the total number of queued payments 2N, it was thought more significant to define the value efficiency as the ratio between the cumulated value of settled payments and the maximum amount that can be settled if payments can be split: $2H + S_A$.

Figure 3.2 **Bilateral optimisation, value (top) and volume (bottom) settled versus liquidity**





3.2.6 Conclusion regarding bilateral optimisation

In terms of settled value, while the three presented non-FIFO algorithms perform significantly better than the standard FIFO algorithm, especially at low liquidity levels, the use of the most complicated algorithms (Las Vegas Greedy and Greedy++) does not yield better results than the use of the simple Greedy algorithm.

In terms of settled volume, however, the Greedy algorithm performs significantly worse than the standard FIFO algorithm, with only 60% of the total number of payments settled when 90% of liquidity is available. On the other hand, the Greedy++ algorithm is basically able to settle 99% of all queued payments whenever more than 5% of the liquidity needed to settle all payments is present.

The best overall performance is arguably realised by the Las Vegas Greedy bilateral algorithm, which performs extremely well both in terms of volume and value. If only the settled value matters, the simple Greedy algorithm – simpler and faster than Las Vegas Greedy and Greedy++ – is the natural choice. Finally, the bilateral optimisation algorithm implemented in TARGET2 and PNS provides satisfactory results given the strong constraint represented by the FIFO rule.

3.3 Multilateral optimisation

This section focuses on multilateral optimisation. This time, all participants and all queued payments are considered simultaneously. The aim of multilateral optimisation is to find a set of payments – as far as possible with the largest cumulated value – that can be settled simultaneously.

3.3.1 Multilateral optimisation in PNS and TARGET2

The multilateral optimisation algorithm of both PNS and TARGET2 starts by activating all queued payments. Of course, if all participants have a positive virtual cash balance,[4] all the payments are settled

[4] Throughout this paper, the virtual cash balance of a participant designates its cash balance if all the activated payments of all participants in the system are settled simultaneously. Clearly a necessary condition for all activated payments to be settled is that all participants should have a positive virtual cash balance.

simultaneously. In the opposite situation, the participant with the largest net debit position is considered by the multilateral optimisation algorithm of both TARGET2 and PNS. The approach followed is then slightly different in the two systems.

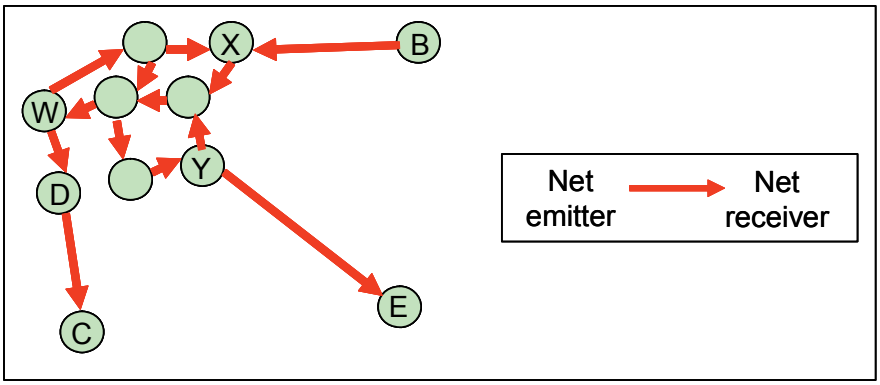
In TARGET2, the algorithm will simply de-activate the most recent payment of the participant with the largest net debit position.

In PNS, the algorithm will consider the participant with the largest net debit position but this time de-activate the smallest payment whose value exceeds the value of its net debit position. (In case no payment exceeds the value of the net debit position, the biggest payment of the participant having the largest net debit position is de-activated, then the second biggest, and so on until one payment exceeds the value of the participant's net debit position.)

3.3.2 A new concept: pre-conditioning

The concept of pre-conditioning is a new idea in the field of multilateral optimisation. The basic idea is to make the most of the existing liquidity by simply letting it flow towards the central core of the payment network. In order to do so, we de-activate as many as possible of the queued payments towards the peripheral participants, who only exchange payments with a single other bank. (In the sketch below, the peripheral participants are B, E and C. By recursion, once the payment from D to C has been de-activated, D will also become a peripheral participant and the payment from W to D will be de-activated.)

Figure 3.3 **Pre-conditioning algorithm**



There are two kinds of peripheral participants:

- The net emitters such as B (B is a net emitter because the cumulated value of queued payments from B to X is larger than the cumulated value of queued payments from X to B), which are a source of liquidity for the network. However, the reason for some payments between X and B being held in the queue is that B does not have the necessary liquidity to settle its net position. As B cannot receive liquidity from any other participant, the set of queued payments between X and B will never be settled as a whole.

- The net receivers (such as C and E) are liquidity traps for the network (C is a net receiver because the cumulated value of queued payments from C to D is smaller than the cumulated value of queued payments from D to C). Indeed the liquidity transmitted from Y to E will not be used again for further settlement.

We can then conclude that whatever their net position (net emitters or net receivers), peripheral participants always have a negative impact on the network. We can therefore try and improve the efficiency of a multilateral optimisation algorithm by removing them before the algorithm is launched.

In the example presented in Figure 3.3, the pre-conditioning algorithm will therefore de-activate all payments from or towards participants B, E, C and then D. Once the multilateral optimisation algorithm has been applied to the network, payments involving peripheral participants will be dealt with separately with the help of bilateral optimisation algorithms.

This pre-conditioning algorithm was implemented in the following algorithms presented in this paper: the Multilateral Greedy Las Vegas, the Multilateral PNS Las Vegas and the OPM1010 algorithm.

3.3.3 The multilateral Las-Vegas algorithms

As in bilateral optimisation, some algorithms trying to use randomly generated numbers to improve on the efficiency of standard algorithms were developed, such as the Multilateral Greedy Las Vegas and the Multilateral PNS Las Vegas algorithms.

The Multilateral PNS Las Vegas algorithm is based on the algorithm used in PNS. However, instead of de-activating the smallest payment that is larger than the deficit of the bank with the largest

debit position, the algorithm randomly chooses which payment to de-activate. In order to so, each payment is affected by a certain 'de-activation probability' based on three different criteria: payments whose value is close to the net debit position of the emitter, payments whose de-activation allows the emitter to reach a net credit position and, finally, payments whose de-activation neither creates nor aggravates the deficit of another participant are de-activated with a higher probability. Appendix 2 provides more insight on the details of the algorithm.

The Multilateral Greedy Las Vegas algorithm is somewhat similar, with the exception that instead of de-activating payments of participants with a net debit position payment by payment, a 'Greedy approach' is followed. All payments originating from the considered participant with a net debit position are de-activated and are considered for re-activation in decreasing order of their value, as in Greedy, but also taking into account the position of the receiver of the payment (payments towards participants with a net debit position are re-activated with a higher probability).

As in bilateral optimisation, the use of random numbers is a way to create algorithms which can be run several times. In the following tests, the Las Vegas algorithms were applied five times in a row (ie the algorithm was applied a first time to the initial problem, then it was applied a second time to what had not settled the first time, and so on…).

The last algorithm tested is the OPM1010 algorithm. It is quite close to the Multilateral PNS Las Vegas algorithm, with the difference that the payments are not de-activated randomly but in a deterministic way. For each payment, a 'de-activation score' is calculated by considering the net positions of the emitter and of the receiver, and the payment with the higher score is de-activated.

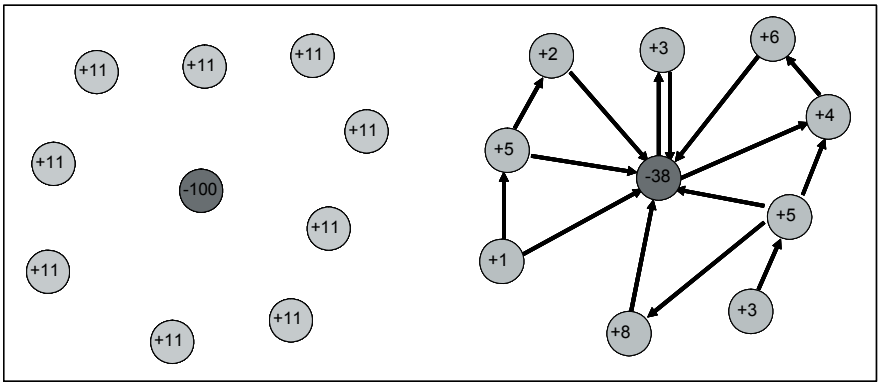
3.3.4 Relative efficiencies of multilateral optimisation algorithms

A test case was derived to assess and compare the settlement efficiency of the presented multilateral optimisation algorithms. We considered ten banks participating in a large value payment system and assumed that a severe operational problem affecting the payment system IT infrastructure had resulted in the unavailability of the

banks' cash balances.[5] As a consequence, the cash position of every participant was considered to be zero until some fresh collateral was provided by the banks.

This liquidity shortage prevented a highly urgent 'all or nothing' ancillary system from settling. We assumed the net position of the banks within the ancillary system was as shown on the left part of Figure 3.4, with nine participants being equally long in the system with a net credit position of EUR 11 million, and only one short participant with a net debit position of EUR 100 million.

Figure 3.4 **Multilateral optimisation test case: settlement of an urgent Ancillary System in a LVPS**



The Central Bank operating the large value payment system wished to speed up as much as possible the settlement of the highly urgent ancillary system and to do so asked the participant with a net debit position in the ancillary system to provide some additional collateral. As fresh collateral might have been scarce in a period of crisis, the Central Bank was interested in trying to reduce the liquidity burden affecting the participant with a net debit position in the AS. To achieve this goal, the system operator could have made use of normal priority payments that were held in the queue due to the lack of available liquidity in the system. It is clear that simultaneously settling the pending AS with some normal priority payments from the long participants to the short participant could lower the amount of collateral the short participant has to find in order to be able to settle

[5] In the context of Target2, such situation could occur for example in case of a regional disaster.

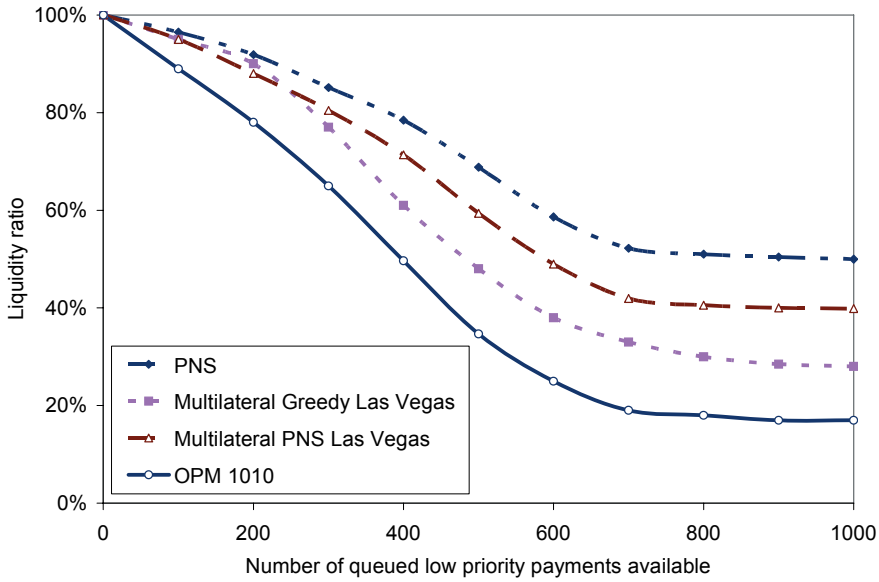
the ancillary system. To do so, we can use a multilateral optimisation algorithm that 'locks the AS settlement', meaning that the algorithm can not settle any payment unless the highly urgent AS is settled simultaneously with it. Such an approach is in particular used in TARGET2, with algorithm 4[6] ('Partial optimisation with ancillary system').

In our test case, we assumed a given number of low priority payments were queued between participants. The low priority payments were generated randomly according to the log normal law that describes the payments value distribution in the PNS system (see Section 3.2.5), and choosing the emitter and the receiver of the payments from the list of the participants with an equal probability.

This test case was run with the presented multilateral optimisation algorithms. The obtained results, averaged over 100 randomly generated low priority payment distributions, are shown in Figure 3.5. The liquidity ratio, defined as the ratio between the remaining collateral value that the short participant has to find, and its net debit position in the ancillary system (EUR 100 million), is plotted on the y-axis against the total number of available low priority payments at the beginning. As an example, in the graph provided in Figure 3.4, the obtained liquidity ratio is 38%. Clearly, when no low priority payments are present to offset the AS, the short participant has to provide the entire 100 millions and the liquidity ratio is one, whatever the algorithm used. When more low priority payments are available, the collateral needs of the short participant are reduced, to an extent that depends on the chosen algorithm.

[6] TARGET2 User Detailed Functional Specifications, first book version 2.0, page 161.

Figure 3.5 **Multilateral optimisation test case: Results**



The results clearly show the interest of multilateral optimisation for the settlement of ancillary systems. When many payments are available, the best algorithm is able to divide by five the value of fresh collateral the short participant has to provide. The best algorithm is OPM 1010, followed by the Multilateral Las Vegas Greedy Algorithm, the Multilateral Las Vegas PNS algorithm and the algorithm implemented in the PNS system.

3.4 Optimisation in PNS in case of an operational failure

3.4.1 The PNS system

PNS (Paris Net Settlement) is a French LVPS which operates alongside TBF, the French RTGS component of the TARGET system. It provides real-time settlement of transactions on central bank money accounts that must always remain in credit. In 2006, 17 banks and credit institutions were participating in the PNS system and exchanging an average of 27,000 payments on a daily basis, with a total value between EUR 45 and EUR 90 billion per day. A cash link

established between PNS and TBF allows the participants to transfer liquidity between their TBF account and their PNS account at any time of the day, depending on their cash needs.

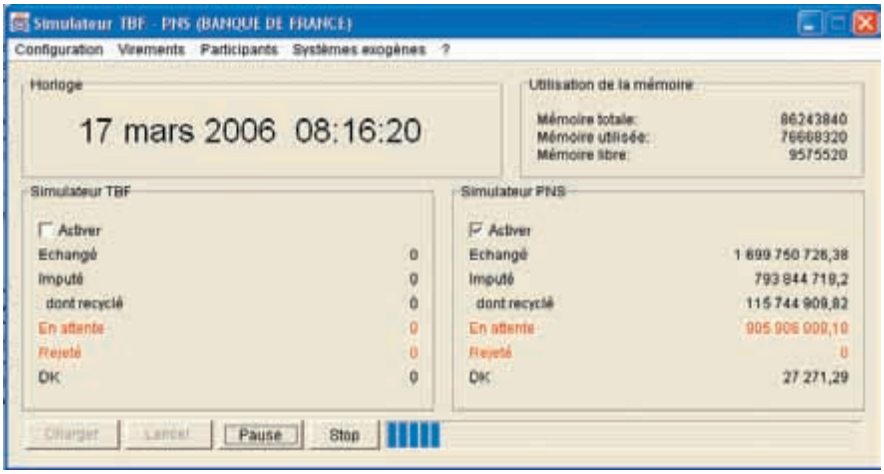
PNS is often presented as a hybrid system because it is equipped with efficient optimisation algorithms that are able to settle simultaneously several queued payments, thus allowing the system to operate at lower liquidity levels. The study of the PNS system is of special interest to the central banks of the Eurosystem because the algorithms implemented in the PNS system are extremely similar to the ones that will be used in the future pan-European TARGET2 system. Moreover as in TARGET2, bilateral sender limits (which can be defined as the maximum net amount a participant is willing to pay to another participant before being paid in return) can be set and modified freely by each participant of the PNS system vis-à-vis its counterparties.

3.4.2 Simulating the technical default of the largest participant in the system

Following Banque de France's previous paper on the PNS system,[7] we investigated the role of optimisation mechanisms under special crisis circumstances. A previous study showed that an operational problem preventing a major participant from issuing payments could lead to a liquidity shortage within the PNS system and finally to the rejection of several queued payments at the closure of the system. Indeed, as the biggest participant is still able to receive payments, but can no longer issue payments, it turns into a 'liquidity trap', depriving the system of the liquidity needed to settle the pending payments. The settlement delay thus increases and eventually some payments can even be rejected at the end of the day. Being able to use advanced non-FIFO algorithms at this point could allow a significant reduction in the number of rejected payments. The case of the technical default of the biggest participant in PNS was therefore revisited after the algorithms presented in this paper had been implemented in Banque de France's PNS/TBF simulator.

[7] Analysis by simulation, of the impact of a technical default of a payment system participant (Liquidity, risks and speed in payment and settlement systems – a simulation approach, Bank of Finland Studies, 2005), Mazars, E and Woefel, G.

Figure 3.6 **Screenshot of Banque de France's PNS simulator**



The simulator used for those simulations is able to reproduce the exact functioning of PNS (bilateral limits, optimisation algorithms, liquidity transfers between PNS and the TBF) and processes the operations one by one as in the real system. In contrast to the test cases presented in Section 3.2.5 (bilateral optimisation) and Section 3.3.4 (multilateral optimisation), real transactions data was used this time.

3.4.3 Results

The month of March 2006 was selected and for each day of the month the consequences of the technical default of the largest participant were investigated with Banque de France's PNS simulator. We assumed that the other participants would not retain their payments in reaction to the technical default of the biggest participant and would not change their behaviour in any way. The most severe consequences were observed for 17 March. Indeed, on this day, provided the default had no influence at all on the behaviour of the other participants, the technical default of the biggest participant would have resulted in 32 payments, representing a total value of EUR 14 billions or 28% of the total value of the submitted payments being rejected at the end of the day. The consequences of the technical default of the biggest participant appear, therefore, to be extremely strong. In reality, however, it is likely that the non-defaulting participants would have tried to mitigate the consequences of the crisis by injecting more

liquidity into the system, thus reducing the number and value of the rejected payments.

The potential impact of the implementation of the presented advanced algorithms into the PNS system as replacements for the original algorithms was investigated with the PNS simulator for 17 March 2006.

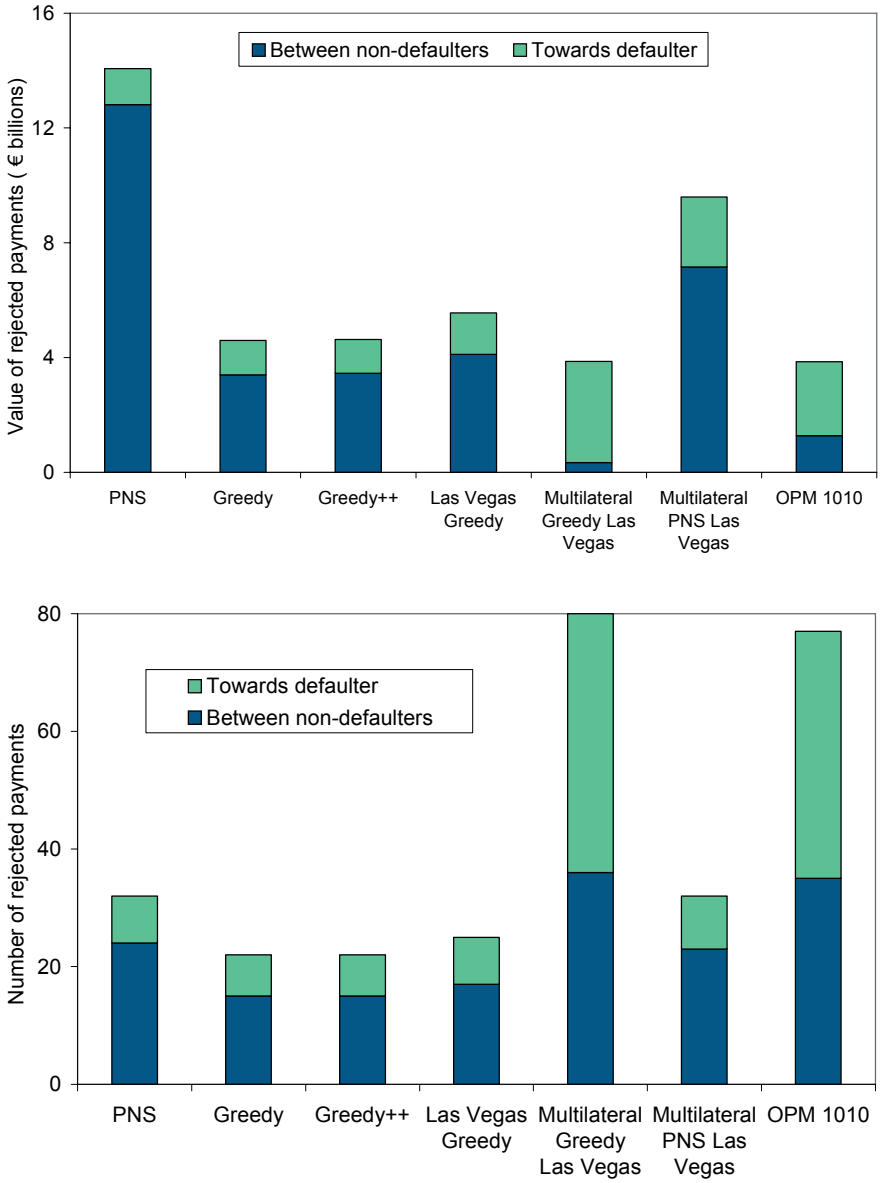
Simulations were made using:

- PNS bilateral optimisation algorithm and PNS multilateral optimisation algorithm (pure PNS);
- Greedy bilateral optimisation algorithm and PNS multilateral optimisation algorithm;
- Greedy++ bilateral optimisation algorithm and PNS multilateral optimisation algorithm;
- Las Vegas Greedy bilateral optimisation algorithm and PNS multilateral optimisation algorithm;
- PNS bilateral optimisation algorithm and the multilateral Las Vegas Greedy algorithm;
- PNS bilateral optimisation algorithm and the multilateral Las Vegas PNS algorithm;
- PNS bilateral optimisation algorithm and the OPM 10–10 algorithm.

The algorithms making use of random numbers for optimisation (bilateral Las Vegas Greedy, Multilateral Las Vegas Greedy, Multilateral PNS Las Vegas) were run five times. No significantly better results were found by increasing the number of iterations.

Figure 3.7 shows the impact of the various optimisation algorithms on the number and total value of the payments rejected at the end of the day. In this given case, it appears that non-FIFO algorithms presented in this paper perform significantly better than the algorithms used in the PNS system. We can also note that the chosen algorithm can also significantly shift the outcome of the settlement, either towards an outcome less favourable to the defaulter (with the multilateral Greedy Las Vegas algorithm) or characterised by a decreased average value of rejected payments (with OPM 1010 or the Multilateral Greedy Las Vegas algorithm).

Figure 3.7 **Effect of the technical default of the biggest participant in the PNS system. Rejected payments towards the defaulter (light) and between non-defaulters (dark) at the end of the day according to the algorithms implemented, in terms of value (top) and volume (bottom).**



The influence of optimisation algorithms on the settlement delay is shown Figure 3.8. In normal conditions (ie without any technical default), the use of non-FIFO optimisation algorithms lowered the settlement delay by about 50% in terms of value, while the settlement delay in terms of volume remained constant. Multilateral optimisation algorithms have a much smaller influence on the settlement delay, as in PNS the multilateral optimisation algorithm is called only three times a day, at 10:30, 12:30 and 16:00. When the biggest participant defaults, the edge given by the non-FIFO algorithms in terms of value becomes significantly bigger.

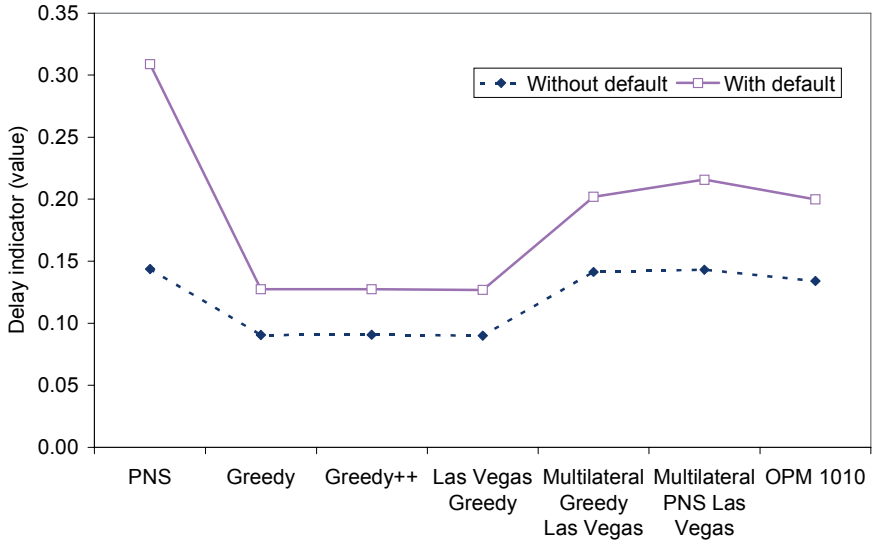
3.4.4 Payments rejected at the end of closure

In order to provide the reader with a clearer insight of the effect of the optimisation algorithms on settlement efficiency, Table 3.1 presents the list of payments rejected between two participants in the PNS system, designated here as participant A and participant B. It appears that the cumulated value of rejected payments between those two participants is extremely high, and represents the main part of the total value of rejected payments.

The PNS bilateral optimisation algorithm is unable to settle any of those payments, given the cash balances of participants A and B. However, it is easy to see that the Greedy bilateral algorithm will simultaneously settle the payments with a value of EUR 1,500 and EUR 2,000 million from A to B and EUR 3,500 million from B to A. In this situation, the Greedy++ algorithm will simultaneously settle the payment with a value of EUR 3,500 million from B to A and the payments with a value of EUR 313, EUR 956, EUR 2,000, EUR 51 and EUR 180 million. As none of those solutions complies with the FIFO rule, the PNS FIFO bilateral optimisation algorithm will not consider them.

In this situation, the use of an advanced optimisation algorithm results in a reduction of EUR 7 billion reduction in the total value of the payments rejected at the end of the day.

Figure 3.8 **Effect of the technical default of the biggest participant in the PNS system. Settlement delay in terms of value (top) and volume (bottom), in normal conditions (dotted line) and in case of the technical default of the biggest participant (solid line).**



DELTA (volume)

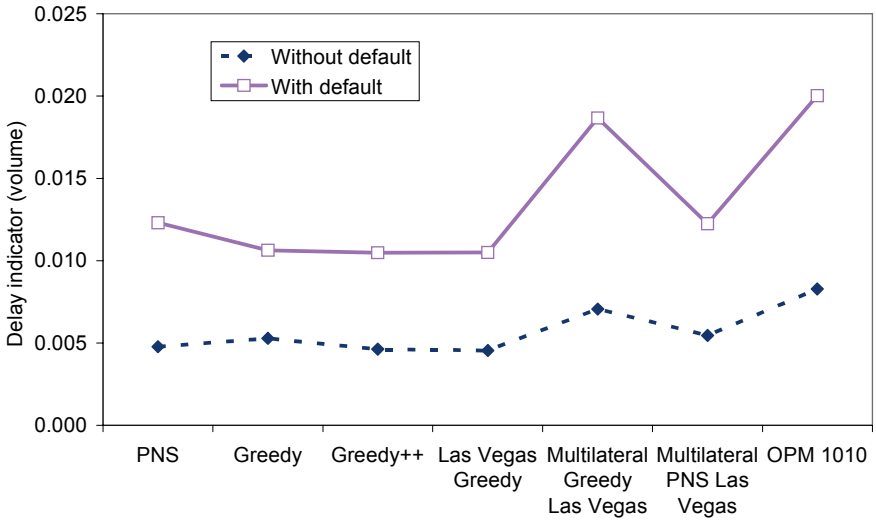


Table 3.1 **Simulation of the technical default of the biggest participant in the PNS system (17 March 2006, standard PNS algorithms), rejected payments between two selected participants.**

		Participant A	Participant B
Cash balance in EUR at closure		3.5 million	22.5 million
Queued payments between participants A and B rejected at closure, in order of arrival	1	160 million	1,000 million
	2	313 million	3,500 million
	3	956 million	87 million
	4	1,500 million	
	5	2,000 million	
	6	51 million	
	7	180 million	

3.5 Conclusion

In this paper several optimisation algorithms that do not follow the FIFO constraint, ie algorithms that are allowed to settle queued payments irrespective of their order of arrival, were presented and their efficiency was investigated in several tests. The results of these tests suggest that the simple Greedy algorithm of Güntzer et al and the suggested OPM1010 algorithm are able to improve respectively on their bilateral and multilateral FIFO counterparts.

Of course, the choice of an optimisation algorithm in a RTGS involves many other considerations than the mere settlement efficiency of the algorithms. In particular, the rules of the system have to be legally sound and have to match the needs of the users as much as possible. No definitive conclusion regarding the use of non-FIFO algorithms in RTGS can therefore be drawn from this paper.

The standard case of the technical default of the biggest participant in a RTGS was also revisited in the context of the PNS system and with several different optimisation algorithms. On the business day chosen for this exercise, (chosen as the 'worst day' of the month of March 2006 in terms of rejected payments resulting from the technical default of the biggest participant), the use of non-FIFO algorithms was shown to greatly reduce the value of rejected payments at the end of the day while shortening the settlement delay. However, when the same exercise was carried out for certain other days of the same month, the use of non-FIFO algorithms did not bring any improvement. It was even the case that the use of non-FIFO

algorithms led to a slight deterioration of the situation at the end of the day. This is due to the fact that efficient algorithms tend to settle payments earlier, as shown by Figure 3.8. Sometimes a slightly less efficient algorithm stockpiles many payments in the queue during the day and is then able to profit from the optimisation opportunities created by the large number of queued payments, resulting in better end-of-the-day results.

This observation having been made, it could make sense to imagine an RTGS in which FIFO algorithms, which combine the advantages of being fast, reasonably efficient, predictable and perfectly transparent to the users, would be used throughout the day, while some more advanced, non-FIFO algorithms could be used in case of a liquidity shortage. In the case of TARGET2, for example, those algorithms would be launched at the closure of the system in case some payments remain in the queue. If the advanced algorithms are then able to settle some additional payments, the number and cumulated value of the rejected payments would be lowered. In the opposite case, the use of those algorithms would not have affected the functioning of the system.

Advanced non-FIFO algorithms could also be useful to accelerate the settlement of a highly urgent ancillary system in the context of a liquidity shortage, as presented in Section 3.3.4. Such specially designed non-FIFO algorithms would only be run in case the standard AS settlement procedure has failed and the settlement delay is creating concerns.

References

Bech, M – Soramäki, K (2001) **Gridlock resolution in interbank payment systems.** Discussion Paper 9/2001. Bank of Finland.

Bech, M – Garratt, R (2003) **The intraday liquidity management game.** Journal of Economic Theory 109, 198–219.

Bech, M – Soramäki, K (2005) **Gridlock resolution and bank failures in interbank payment systems, in Liquidity, risks and speed in payment and settlement systems – a simulation approach.** Bank of Finland Studies E:31.

Güntzer, M – Jungnickel, D – Leclerc, M (1998) **Efficient algorithms for the clearing of interbank payments.** European Journal of Operational Research 106, 212–219.

Koponen, R – Soramäki, K (2005) **Intraday liquidity needs in a modern interbank payment system – a simulation approach.** In Liquidity, risks and speed in payment and settlement systems – a simulation approach. Bank of Finland Studies E:31.

Leinonen, H – Soramäki, K (2005) **Optimising liquidity usage and settlement speed in payment systems.** In Liquidity, risks and speed in payment and settlement systems – a simulation approach. Bank of Finland Studies E:31.

Shafransky, Y – Doudkin, A (2006) **An optimisation algorithm for the clearing of interbank payments.** European Journal of Operational Research 171, 743–749.

Appendix 1

Bilateral optimisation algorithms

Bilateral optimisation: notations

We denote the payments from bank A to bank B (respectively from bank B to bank A) as the (a_i) and the (b_i). The (x_i) and the (y_i) are two vectors of indicators. For each k, $x_k = 0$ (resp. $x_k = 1$) means that the payment a_k is not activated (resp. activated); similarly for each k, $y_k = 0$ (resp. $y_k = 1$) means that the payment b_k is not activated (resp. activated).

S_A is the initial cash balance of bank A and S_B is the initial cash balance of bank B. For given (x_i) and (y_i), bank A's virtual cash balance is equal to $B_A = S_A - \sum_i a_i x_i + \sum_i b_i y_i$ and bank B's virtual cash balance is $B_B = S_B + \sum_i a_i x_i - \sum_i b_i y_i$.

FIFO bilateral optimisation algorithm (PNS, TARGET2 …)

Activate all payments between the two considered banks.
WHILE the simultaneous settlement of all activated payments is impossible
- De-activate the most recent activated payment from the deficient bank.

END WHILE
Settle all activated payments

LAS VEGAS GREEDY bilateral optimisation algorithm

Activate all payments between the two considered banks.

WHILE one of the two banks has a negative Virtual Cash Balance
- De-activate all payments from the deficient bank (let us suppose it is bank A).
- Go through the payments of bank A, the $(a_i)_{i=1...N}$ from A's biggest payment a_1 to A's smallest payment a_N. When considering payment a_k for re-activation:
  - IF a_k > Bank A's Virtual Cash Balance THEN a_k can not be re-activated.

- ELSE:

  - Calculate $R_k = \sum_{i=k+1}^{N} a_i$ sum of all the payments smaller than a_k
  - IF R_k < Bank A's Virtual Cash Balance THEN a_k is activated
  - ELSE:
    - Let $p_k = \min\left(\frac{a_k}{R_k};1\right)$
    - Re-activate payment a_k with a probability of p_k

- Next a_k

END WHILE

Settle all activated payments.

The point of this algorithm is to launch it several times. In the tests presented in this paper, it was applied 5 times in a row (ie it was applied a first time to the initial problem, then it was applied a second time to what had not settled the first time, and so on). Another possible use is to run it a certain number of times on the initial problem and to retain the best solution.

GREEDY bilateral optimisation algorithm (Güntzer et al, 1998)

Activate all payments between the two considered banks.

WHILE one of the two banks has a negative Virtual Cash Balance

- De-activate all payments from the deficient bank (let us suppose it is bank A).
- Go through the payments of bank A, the $(a_i)_{i=1...N}$ from A's biggest payment a_1 to A's smallest payment a_N. When considering payment a_k for re-activation:
  - IF a_k > Bank A's Virtual Cash Balance THEN a_k can not be re-activated.
  - ELSE re-activate a_k
  - Next a_k

END WHILE

Settle all activated payments

GREEDY++ bilateral optimisation algorithm

Activate all payments between the two considered banks, bank A and bank B.

WHILE one of the two banks has a negative virtual cash balance

- Let $G = \sum_i a_i x_i$ and $H = \sum_i b_i y_i$
- The error is defined as: $\Delta = \left| G - H - \frac{1}{2}(S_A - S_B) \right|$
- Pick up the 10 payments (from either bank, selected or not) closest to the error Δ (we pick up the a_i and b_i that minimise $\left| \log\left(\frac{a_i}{\Delta}\right) \right|$)
- Try all possibilities involving the 10 picked up payments (1024 possibilities)
- IF at least one of the possibilities allows settlement
  - o THEN choose the possibility that maximises the value settled.
  - o ELSE:

    - De-activate all payments from the deficient bank (let us suppose it is bank A).
    - Go through the payments of bank A, the $(a_i)_{i=1...N}$ from A's biggest payment a_1 to A's smallest payment a_N. When considering payment a_k for re-activation:
      - IF a_k > Bank A's Virtual Cash Balance THEN a_k can not be re-activated.
      - ELSE re-activate a_k
      - Next a_k

END WHILE
Settle all activated payments

Appendix 2

Multilateral optimisation algorithms

Multilateral PNS Las Vegas

Activate all payments

1. Attempt to settle all queued payments simultaneously (‘*all or nothing*’)

2. As long as there is a peripheral participant:
   De-activate all payments to or from a peripheral participant.

3. WHILE there is a participant with a negative Virtual Cash Balance

   3.1 Randomly choose a participant with a negative Virtual Cash Balance (Uniform law).

   3.2 The chosen participant, bank i, has a negative Virtual Cash Balance B_i. We are then going to de-activate one of bank i’s outgoing payments. In order to do so, for each activated payment k emitted by bank i, we calculate the coefficient $b_{i,k} = \gamma_{i,k}^{suf}\gamma_{i,k}^{def}\gamma_{i,k}^{cre}$ where:

   - $\gamma_{i,k}^{suf} = 2$ if the inactivation of payment k makes bank i’s Virtual Cash Balance positive, else $\gamma_{i,k}^{suf} = 1$.
   - $\gamma_{i,k}^{def} = \max\left[\min\left(\frac{|B_i|}{p_i^k};\frac{p_i^k}{|B_i|}\right);0.1\right]$ so that payments whose value are close to the deficit are de-activated with a higher probability.
   - $\gamma_{i,k}^{cre} = 4$ if the inactivation of payment k does not create nor aggravate the deficit of another participant, else $\gamma_{i,k}^{cre} = 1$.
   - We then randomly select one of bank i’s outgoing payments so that payment k has a probability $\frac{b_{i,k}}{\sum_k b_{i,k}}$ to be de-activated.

   3.3 If bank i now has a positive Virtual Cash Balance, attempt to re-activate some of bank i’s outgoing payments in the order of their decreasing amount.

END WHILE

4. When all participants have a positive Virtual Cash Balance, all activated payments are settled.

Re-activate all de-activated payments (including those involving peripheral participants) and go through all bilateral relations, from the most balanced to the most unbalanced and run the Las Vegas Greedy bilateral optimisation algorithm.

The point of this algorithm is to launch it several times. In the tests presented in this paper, it was applied 5 times in a row (ie it was applied a first time to the initial problem, then it was applied a second time to what had not settled the first time, and so on). Another possible use is to run it a certain number of times on the initial problem and to retain the best solution.

Multilateral Greedy Las Vegas

Activate all payments

1. Attempt to settle all queued payments simultaneously (‘*all or nothing*’)

2. As long as there is a peripheral participant:
   De-activate all payments to or from a peripheral participant.

3. WHILE there is a participant with a negative Virtual Cash Balance

   3.1. Consider all the banks with a negative Virtual Cash Balance in the increasing order of the number of participants they send payments to, then in the decreasing order of their deficit (we then start by the banks which emit payments towards a single counterparty).

   3.2 The considered bank i, has a negative Virtual Cash Balance B_i. De-activate all its outgoing payments, then consider them for re-activation in the decreasing order of their value, under the constraint that the virtual position of bank i remains positive. We then have, for payment number l of bank i, p_i^l:

   a. IF $p_i^l > B_i$ then payment number l can not be re-activated.

   b. ELSE IF $R_i^l < B_i$ then payment number l is re-activated (where $R_i^l = \sum_{k \geq l+1} p_i^k$ is the cumulated value of Bank i’s payments smaller than p_i^l)

c. ELSE the payment p_i^l is re-activated with a probability equal to $\min\left(\frac{p_i^l}{R_i^l} b^{\pm};1\right)$, where:

- $b^+ = \frac{m^+ + m^-}{2m^+ + m^-}$ $\quad b^- = \frac{m^+ + m^-}{m^-} b^+$
- m^+ (resp. m^-) is the number of participants receiving payments from bank i whose Virtual Cash Balance is positive (respectively negative).
- $b^{\pm} = b^+$ if the receiver of the payment has a positive Virtual Cash Balance, else $b^{\pm} = b^-$.

END WHILE

4. When all participants have a positive Virtual Cash Balance, all activated payments are settled.

Re-activate all de-activated payments (including those involving peripheral participants) and go through all bilateral relations, from the most balanced to the most unbalanced and run the Las Vegas Greedy bilateral optimisation algorithm.

The point of this algorithm is to launch it several times. In the tests presented in this paper, it was applied 5 times in a row (ie it was applied a first time to the initial problem, then it was applied a second time to what had not settled the first time, and so on). Another possible use is to run it a certain number of times on the initial problem and to retain the best solution.

Algorithm OPM 1010

Activate all payments

1. Attempt to settle all queued payments simultaneously (‘*all or nothing*’)

2. As long as there is a peripheral participant:
   De-activate all payments to or from a peripheral participant.

3. WHILE there is a participant with a negative Virtual Cash Balance

   3.1 Go through all banks with a positive Virtual Cash Balance and re-activate the payments that can be re-activated.
   3.2 Randomly choose a bank with a negative Virtual Cash Balance.
   WHILE the chosen bank i has a negative Virtual Cash Balance B_i:
   Calculate for each activated payment k sent by bank i, the coefficient $b_{i,k} = \gamma_{i,k}^{suf}\gamma_{i,k}^{def}\gamma_{i,k}^{cre}$ where:

   - $\gamma_{i,k}^{suf} = A$ if the inactivation of payment k makes bank i’s virtual position positive, else $\gamma_{i,k}^{suf} = 1$.
   - $\gamma_{i,k}^{def} = \max\left[\min\left(\frac{|B_i|}{p_i^k};\frac{p_i^k}{|B_i|}\right);0.1\right]$ in order to favour the payments whose value is close to $|B_i|$ the net debit position of bank i.
   - $\gamma_{i,k}^{cre} = C$ if the inactivation of payment k does not create nor aggravate the deficit of another participant, else $\gamma_{i,k}^{cre} = 1$.
   - A sensitivity study performed at several levels of liquidity concluded that the OPM alrorithm have better results with A = C = 10, hence the name of OPM1010 for this given variation of the algorithm.

   De activate the payment with the highest coefficient $b_{i,k}$.

   END WHILE

   If there are some of bank i’s de-activated payments can be re-activated, re-activate them in the decreasing order of their value.

END WHILE

4. When all participants have a positive Virtual Cash Balance,
   4.1 Go through all banks and re-activate the payments that can be re-activated in the decreasing order of their value.

5. Settle all activated payments.

Re-activate all de-activated payments (including those involving special participants) and go through all bilateral relations, from the most balanced to the most unbalanced and run the Las Vegas Greedy bilateral optimisation algorithm.

The point of this algorithm is to launch it several times. In the tests presented in this paper, it was applied 5 times in a row (ie it was applied a first time to the initial problem, then it was applied a second time to what had not settled the first time, and so on). Another possible use is to run it a certain number of times on the initial problem and to retain the best solution.

Appendix 3

The Greedy algorithm of Güntzer et al[8] and superincreasing payment values distributions

Superincreasing sequences

Let p be a strictly positive integer. A sequence of positive reals $(u_i)_{i=1...p} \in \Re^p$ is said to be superincreasing when: $\forall k \in \{1...p-1\}$, $u_{k+1} > \sum_{i=1}^{k} u_i$. For a central banker, a good example of a superincreasing sequence is the sequence of the values of the euro banknotes (5 euros, 10 euros, 20 euros, 50 euros, 100 euros, 200 euros and 500 euros). Indeed, any banknote is worth more than the sum of the smaller banknotes. This highly desirable property ensures that a cashier can minimise the number of banknotes to be given back to a customer by simply following a Greedy type of algorithm, that is to say by always using the biggest banknote whose value is lower than the remaining amount of money to be handed back. Should a 400-euro banknote be introduced, the Greedy solution (500+200+100, 3 banknotes) would be beaten by a non-Greedy solution (400+400, 2 banknotes) if 800 euros had to be handed back by the cashier. This property is actually closely related to the aim of this demonstration.

Notations

- Let there be two banks A and B, characterised by their respective liquidity S_A and S_B. There are N queued payments from A to B and M queued payments from bank B to bank A.

- We assume that the sequences of the queued payments from A to B and from B to A, respectively the $(a_i)_{i=1...N}$ and $(b_i)_{i=1...M}$ are superincreasing sequences, that is to say that we have $\forall i \in \{1...N-1\}, a_i > \sum_{k=i+1}^{N} a_k$ and $\forall i \in \{2...M-1\}, b_i > \sum_{k=i+1}^{M} b_k$ (a_1 is

[8] Güntzer, M – Jungnickel, D – Leclerc M (1998) Efficient algorithms fort he clearing of interbank payments. European Journal of Operational Research 106, 212–219.

therefore the biggest payment from bank A to bank B, and a_N is the smallest).

- The $(x)_{i=1...N} \in \{0,1\}^N$ and the $(y)_{i=1...M} \in \{0,1\}^M$ are two vectors of indicators. For each k, $x_k = 0$ (resp. $x_k = 1$) means that the payment a_k is not activated (resp. activated); similarly for each k, $y_k = 0$ (resp. $y_k = 1$) means that the payment b_k is not activated (resp. activated).

- The Greedy algorithm is as defined in Appendix 1.

Lemma

Provided the sequence of the payment values is superincreasing, the Greedy algorithm re-activates at each iteration the payments whose cumulated value is maximal.

Proof

Without any loss of generality, we can assume that the bank in deficit is bank A. At the beginning of an iteration, all payments emitted by bank A are de-activated and are then considered for re-activation in the decreasing order of their value. It is clear that the total cumulated value of the re-activated payments can not exceed a ceiling of $S_A + \sum_{i=1}^{M} b_i y_i$ where y_i indicates whether the ith payment of bank B is activated.

The Greedy algorithm first considers bank A's biggest payment a_1 for re-activation. If $a_1 > S_A + \sum_{i=1}^{M} b_i y_i$ then a_1 can clearly not be re-activated, whatever the algorithm used. Let us now suppose that $a_1 \leq S_A + \sum_{i=1}^{M} b_i y_i$, the Greedy algorithm will therefore re-activate payment a_1. Any algorithm which would choose not to re-activate this payment would yield a poorer solution than Greedy's since as $a_1 > \sum_{k=2}^{N} a_k$ (because the sequence is superincreasing), any solution not retaining a_1 would be worse than any solution retaining a_1.

By induction, the same result applies to all of bank A's payments, hence we can conclude that for a given iteration, the value of the payments re-activated by the Greedy algorithm is maximal.

Proposition

The Greedy algorithm is the most efficient in terms of settled payment value provided the sequence of the payment values is superincreasing.

Proof

Let $G_k = \sum_{i=1}^{N} a_i x_i$ be the cumulated value of the activated payments from A to B after the kth iteration of the *Greedy* algorithm where bank A is in deficit. Similarly let $H_k = \sum_{i=1}^{M} b_i y_i$ be the cumulated value of the activated payments from B to A after the kth iteration of the *Greedy* algorithm where bank B is in deficit.

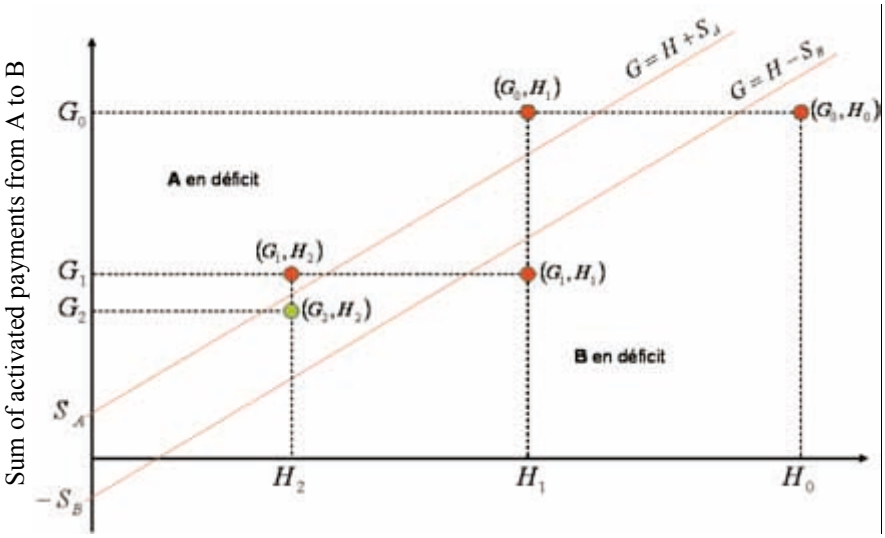
The settlement condition can be written as the dual inequality: $-S_B \leq G - H \leq S_A$.

At the start of the algorithm, all payments are activated, hence $G_0 = \sum_{i=1}^{N} a_i$ and $H_0 = \sum_{i=1}^{M} b_i$. Without any loss of generality, we can assume that bank B will be the first bank to be in deficit. The pairs of payment flows that will be considered will then be: (G_0, H_0), (G_0, H_1), (G_1, H_1) ...

It is easy to demonstrate that the Greedy algorithm will terminate, by noticing that the G_k and H_k are two strictly decreasing sequences taking only a finite number of positive values. We denote as t the subscript of the last iteration of the Greedy algorithm. The final state will therefore be either (G_t, H_t) or (G_{t-1}, H_t).

Given the characteristics of the Greedy algorithm, we already know that we will have: $G_0 > G_1 > .... > G_t$ and $H_0 > H_1 > .... > H_t$. The underneath sketch shows how the pairs (G_k, H_k) converge towards a solution satisfying the settlement condition (the pairs satisfying the settlement condition are located between the two parallel red lines).

Figure 3.9 **Illustration of the convergence of the Greedy algorithm**



Let G* and H* be the values of the payment flows characterising the solution maximising the settled value. This value maximising solution trivially exists (at worse we have G* = H* = 0).

Let us show by induction that $\forall k \in \{0...t\}, \begin{cases} G_k \geq G^* \\ H_k \geq H^* \end{cases}$

Basis: trivially, we have $\begin{cases} G_0 \geq G^* \\ H_0 \geq H^* \end{cases}$

Inductive step: Let be $k \in \{0...t\}$. Suppose that $\begin{cases} G_k \geq G^* \\ H_k \geq H^* \end{cases}$

As we assumed that B was initially in deficit, after k iterations on G and k iterations on H, B is still the bank in deficit. Greedy then has to evaluate the new cumulated payment flows of bank B, H_{k+1}. Let us show that $H_{k+1} \geq H^*$.

According to the lemma, H_{k+1} is the highest possible value that can take the cumulated sum of the activated payments of bank B under the constraint: $H_{k+1} \leq S_B + G_k$.

Now we also have $G_k \geq G^*$ according to our inductive hypothesis and we know in addition that H^* verifies the settlement condition $H^* \leq S_B + G^*$, since the pair (G^*, H^*) is a solution to the problem. That gives us the inequality $H^* \leq S_B + G_k$ and as H_{k+1} is the highest possible value lower than $S_B + G_k$ we can then conclude that $H_{k+1} \geq H^*$.

Now A is in deficit and the same demonstration applies to prove that $G_{k+1} \geq G^*$. We can then conclude.

We have then shown that $\forall k \in \{0...t\}, \begin{cases} G_k \geq G^* \\ H_k \geq H^* \end{cases}$

In particular we have $\begin{cases} G_t \geq G^* \\ H_t \geq H^* \end{cases}$

G^* et H^* being by construction the best solution, we have $\begin{cases} G_t = G^* \\ H_t = H^* \end{cases}$

When the payment value sequences are superincreasing, the Greedy algorithm thus yields the solution that maximises the settled value.

Chapter 4

Examining the tradeoff between settlement delay and intraday liquidity in Canada's LVTS: a simulation approach

Neville M Arjani

4 Examining the tradeoff between settlement delay and intraday liquidity in Canada's LVTS: a simulation approach 88

Abstract 88

4.1 Introduction 88
4.2 Settlement delay and intraday liquidity in an LVPS.....91
4.3 A simple graphical framework 95
4.3.1 Description of the framework.....95
4.3.2 Innovation: a complex queue-release algorithm.....98
4.4 Empirical study: estimating the tradeoff in Canada's LVTS 99
4.4.1 Background on the LVTS.....100
4.4.2 Settlement delay and intraday liquidity in T2: tradeoff and improvement 102
4.5 Data description and simulation methodology 104
4.5.1 Description of data 104
4.5.2 Simulation description and methodology 105
4.6 Simulation results and discussion.....108
4.6.1 The delay-liquidity tradeoff in the T2 payment stream 108
4.6.2 Discussion 111
4.7 Conclusions and caveats.....113

References 117

Appendix Is the T2 payment stream technically efficient? 119

4 Examining the tradeoff between settlement delay and intraday liquidity in Canada's LVTS: a simulation approach

Abstract

The paper explores a fundamental tradeoff occurring in the daily operation of large-value payment systems (LVPS) – between settlement delay and intraday liquidity – with specific application to Canada's Large-Value Transfer System (LVTS). To reduce settlement delay, participants generally must maintain greater intraday liquidity in the system. Intraday liquidity and settlement delay can be costly for LVPS participants, and improvements in the tradeoff are desirable. The replacement of standard queuing arrangements with a complex queue-release algorithm represents one such improvement. These algorithms are expected to lower intraday liquidity needs and speed up payment processing in an LVPS. Simulation analysis is used to empirically test this proposition for the case of Canada's LVTS. The analysis is conducted using a payment system simulator developed by the Bank of Finland, called the BoF-PSS2. It is shown that increased use of the LVTS central queue (which contains a complex queue-release algorithm) reduces settlement delay associated with each level of intraday liquidity considered, relative to a standard queuing arrangement. Some important issues for discussion emerge from these results.

4.1 Introduction

A well-functioning large-value payment system (LVPS) is an integral component of any advanced financial system. In a market economy such as Canada's, virtually all economic transactions ultimately involve a transfer of funds between a buyer and a seller. An LVPS provides the electronic infrastructure necessary to facilitate such an exchange of funds between financial institutions in order to discharge large-value payment obligations on behalf of their own business and that of their customers. There are different designs of LVPS currently

operating around the world, with each achieving a different balance between the minimisation of systemic risk, the speed of payment settlement, and the liquidity and operational costs of settlement.

This paper examines a fundamental tradeoff occurring in the daily operation of an LVPS – between settlement delay and intraday liquidity – with particular application to Canada's LVTS.[1] Settlement delay refers to a potential time lag occurring between a participant's intended submission of a payment to the system and when it is processed by the LVPS with finality.[2] Intraday liquidity refers to a participant's ability to meet its outgoing payment obligations immediately when intended. Generally speaking, to achieve shorter settlement delay participants must maintain greater intraday liquidity in the system. When sufficient intraday liquidity is not maintained, payments will be queued and will be released only when the participant's liquidity position improves. Settlement delay, then, reflects the amount of time that a payment is queued before being processed by the system.

Intraday credit is an important source of liquidity. To control credit risk, grantors of intraday credit (typically central banks) usually require eligible collateral, which is likely to entail a cost for participants. At the same time, settlement delay may also be expensive for participants. The cost of settlement delay may be borne both internally by the participant that delays sending the payment and externally by the receiving participant. Participants generally must tradeoff the cost of settlement delay and the cost of intraday liquidity in conducting their daily payment operations. It follows that a reduction in the amount of intraday credit provision to participants will entail both a benefit and cost. The benefit is that participants' liquidity (ie collateral) cost can be reduced, but possibly only at the expense of a higher settlement delay cost.

A simple graphical framework of the general risk-efficiency tradeoff in payment systems, inspired by Berger, Hancock and Marquardt (1996), is useful when thinking about the nature of the tradeoff between settlement delay and intraday liquidity in an LVPS. Given the cost to participants of both settlement delay and intraday

[1] The LVTS is owned and operated by the Canadian Payments Association (CPA). For a more thorough description of the LVTS, including an overview of the Bank of Canada's multiple roles within the system, see Dingle (1998) and Arjani and McVanel (2006).

[2] Use of the term 'intended' is made so that this definition of settlement delay could apply to LVPS designs with and without a central queue. Under the latter design, a participant may intend to submit a payment to the LVPS at a certain time but, due to lack of intraday liquidity and the absence of a central queue, must hold the payment internally until it can be successfully processed by the system.

liquidity, improvements in the tradeoff are desirable. An improvement in the tradeoff is characterised by this paper as reduced settlement delay associated with each level of intraday liquidity, for the same value of payment activity. Innovations in LVPS design may make this possible. The replacement of standard queuing arrangements with a complex queue-release algorithm represents such an innovation. The potential benefit of such algorithms includes both lower liquidity needs for the release of queued payments and thus faster processing of these payments by the LVPS.

A simulation approach is used to empirically test the proposition that a complex queue-release algorithm can lower liquidity costs and speed payments processing relative to a standard queuing arrangement – that is, improve the tradeoff between settlement delay and intraday liquidity. Using actual intraday transaction and credit limit data, simulation analysis is employed to quantify the current tradeoff between settlement delay and intraday liquidity in the Canadian LVTS. Then, improvements in this tradeoff are sought by simulating an alternative LVTS environment in which current restrictions on use of the LVTS central queue are relaxed. The LVTS queue employs a complex queue-release algorithm that seeks to partially offset batches of queued payments on a multilateral basis throughout the day. However, under current system rules, participants' excessive use of the central queue is not encouraged.[3] Instead, standard internal queuing arrangements are typically employed by participants.

The analysis reveals that a tradeoff does indeed exist between settlement delay and intraday liquidity in Canada's LVTS. Moreover, the results indicate that increased use of the central queue will reduce settlement delay in the LVTS for each level of intraday liquidity considered according to three different settlement delay measures. Some important discussion points also emerge from these results.

The remainder of this paper is as follows. Section 4.2 discusses the nature of the tradeoff between settlement delay and intraday liquidity in greater detail. The graphical framework is presented in Section 4.3, and potential improvements in the tradeoff are also discussed in that section. Section 4.4 contains relevant background information on the LVTS. Section 4.5 provides an overview of the simulation

[3] See LVTS Rule No. 7. There are several hypothesised reasons for this. Perhaps the foremost reason pertains to the issue of whether queue transparency may cause participants to take on credit risk by crediting clients' accounts with expected incoming funds prior to these payments actually being received. This was a major concern of central banks at the time the LVTS was being developed. See RTGS (1997) and discussion in Section 4.6.2.

methodology as well as a description of the data. Section 4.6 presents results from the simulations and related discussion. Section 4.7 offers concluding remarks and some caveats to the analysis.

4.2 Settlement delay and intraday liquidity in an LVPS

Participants in an LVPS typically maintain a daily schedule of payments which they must send through the system on behalf of their own business and that of their customers. Included in this schedule is the time that each payment is due to be sent. For example, certain payments are considered 'time-sensitive' and thus have to be sent by a specific time during the day. The remaining majority of payments is considered 'non-time-sensitive' and simply must be sent by the end of the day. In practice, however, participants generally do not wait until the end of the day to submit all of their non-time-sensitive payments for reasons that will be outlined below.

In Real-Time Gross Settlement (RTGS) and RTGS-equivalent LVPS (such as Canada's LVTS), participants must maintain intraday funds in the system to send a payment to another bank. Hence, the concept of intraday liquidity in an LVPS specifically refers to a participant's ability to access sufficient intraday funds to meet its outgoing payment obligations in a timely manner. There are two main sources of intraday funds available to an LVPS participant: 1) funds acquired from other participants due to either regular transaction activity or through an interbank loan arrangement and 2) funds acquired through an intraday credit extension. Incoming funds from regular transaction activity are the cheapest source of liquidity for participants, and it is expected that participant banks will try to use these funds as much as possible to finance their own payment activity.[4] For various reasons (eg the differing nature of individual participants' business), however, it may not always be possible for participants to coordinate their daily payment activity so that incoming payments largely finance their outgoing payment needs.

The inability of participants to perfectly coordinate their incoming and outgoing payment activity creates a role for the provision of intraday credit. Martin (2005) emphasises the importance of intraday

[4] See McAndrews and Rajan (2000) and McAndrews and Potter (2002) for discussion and identification of this type of coordination behaviour among participants in the US Fedwire system.

credit as a source of intraday funding for participants. The author argues that the coordination of incoming payments to meet outgoing obligations is often difficult (especially for time-sensitive payments), and therefore a well-designed LVPS should allow participants to acquire funds when necessary through intraday credit. Where intraday credit is available to participants on a free and unlimited basis, participants can borrow funds any time that a payment is due, thus eliminating potential settlement delay in the LVPS. However, although settlement delay would cease to exist in this case, lenders of intraday credit (typically central banks) could face large risk exposures vis-à-vis borrowers, which is not desirable from a public policy perspective. Consequently, intraday credit in RTGS and equivalent systems is not free and unlimited, but rather is often subject to net debit caps, (eligible) collateral requirements which typically entail an opportunity cost, and in certain cases an explicit interest charge, eg the US Fedwire system. Maintaining intraday liquidity in the system can therefore be costly for participants.

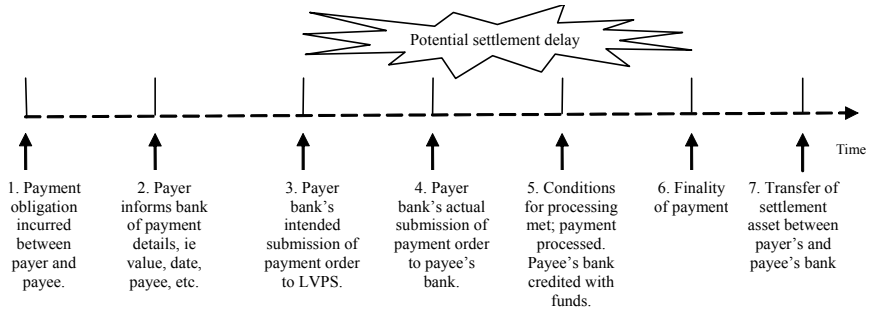
Where a participant does not have sufficient funds available to meet a payment obligation upon intended submission, processing of the payment by the LVPS will be delayed. Settlement delay can be defined as a time lag occurring between a participant's intended submission of a payment to the LVPS, and when the payment is processed by the LVPS with finality, ie when intraday funds are exchanged between participants on an unconditional and irrevocable basis in order to discharge the payment obligation.[5] Payments that cannot be processed because of a participant's lack of intraday liquidity may be held in that participant's internal queue. Alternatively, these payments could be submitted to the LVPS and held in the system's central queue if one is available. Under standard queuing procedures, internally and centrally queued payments are released and processed by the LVPS on an *individual* basis when a sending participant's intraday liquidity improves to the extent that these payments can be passed.[6] The settlement delay associated with an individual payment essentially reflects the amount of time that the payment must wait in the queue before being processed by the LVPS.

[5] A key feature of RTGS and equivalent LVPS is that these systems offer immediate intraday finality. Payments in these systems are considered final upon being processed.

[6] This liquidity improvement could occur as a result of the participant receiving a payment, or gaining access to more intraday credit.

Figure 4.1 provides a graphical characterisation of settlement delay within the context of the life-cycle of a large value payment.[7]

Figure 4.1 **The life-cycle of a large-value payment**



Just as there is a cost associated with maintaining intraday liquidity in the system, given the high speed and high value of daily payments processed by an LVPS, settlement delay may also entail a significant cost for participants. Further, the nature of this delay cost is likely to depend on whether a payment is time-sensitive or not. Time-sensitive payments may include those related to the final funds settlement of other important national and international clearing and settlement systems, large government receipts and disbursements, and also payments related to the daily implementation of monetary policy. A participant that is unable to meet a time-sensitive payment obligation when due may therefore face large internally borne costs because of the delay, such as reputation damage with its peers and, possibly, a loss of its clients' business. Explicit penalty charges may even be imposed by the system operator since the delay of these payments could cause a disruption elsewhere in the financial system.

For the remaining majority of (non-time-sensitive) payments, there is no formal intraday deadline to submit these payments. It is not expected that a participant will incur an (immediate) reputation loss or penalty charge, nor a loss of its clients' business, if processing of these

[7] The paper recognises that achieving payment finality need not encompass the transfer of the settlement asset. Therefore, the notion of settlement delay applies equally to RTGS and RTGS-equivalent LVPS, where this transfer occurs on a multilateral net basis at the end of the day in the latter.

payments is delayed until the end of the day.[8] However, there may be other external costs imposed on the system in this case. Despite being non-time-sensitive, intended receiving banks may be expecting these payments by a certain time of day, and such a delay will result in a shortfall in their intraday funds position. If these participants are planning on using these funds to send their own payments, then they may have to incur additional costs in order to replace these funds on short notice. Where they cannot find other funds in time to meet their obligations, additional settlement delay is created in the system. Settlement delay created by one participant in an LVPS could quickly spread to others in the system. Moreover, a comparable disruption to the liquidity position of a receiving bank's client may also occur (where a delayed payment is ultimately intended for this customer), resulting in potentially broader consequences for economic activity.

Prolonged delay of these payments may also intensify the potential losses associated with other risks in the system, such as operational risk. An operational event (such as a computer outage that prevents one or more participants from sending payments) will likely have a larger impact in a case where a number of payments remain unprocessed at the time that the incident occurs.[9] At the same time, a large backlog of payments being submitted all at once to the LVPS late in the day could increase the potential likelihood that an operational event occurs in the first place. Lastly, where the potential for settlement delay could discourage use of an LVPS in favor of systems that are not as well risk-proofed, the existence of settlement delay may translate to higher systemic risk in the broader financial system.

It follows that, to eliminate the potential costs associated with settlement delay, participants will likely have to borrow a large amount of intraday credit and thus incur high liquidity costs. Conversely, participants need not incur any intraday liquidity cost, but will then have to bear (possibly along with other participants in the

---

[8] Prolonged delay of non-time-sensitive payments is unlikely to cause reputation loss immediately, but such a loss could occur if repeated over time. In a relatively concentrated payment system like Canada's LVTS, participants maintain frequent communication with each other throughout the day and are able to develop fairly accurate forecasts of certain incoming payment flows based on historical payment patterns with other participants. Thus, a participant that often delays its non-time-sensitive payments in favor of lower liquidity costs is unlikely to go unnoticed among its peers in the system.

[9] Conversely, an operational disruption could also lead to settlement delay in an LVPS since it may result in a participant's inability to send payments through the system. For this reason, contingency measures are usually available in an LVPS for the release of time sensitive payments in the event of a disruption.

system) the costs of the accompanying settlement delay. It is unlikely that participants will not maintain sufficient liquidity to meet their time-sensitive payment obligations since the cost of delaying these payments is very high. Consequently, the discussion of a tradeoff between settlement delay and intraday liquidity may not apply to time-sensitive payments in practice. However, for non-time-sensitive payments, the tradeoff is likely to exist. Since settlement delay may entail costs and repercussions for the system as a whole, any innovation in LVPS design that can increase settlement speed for a given level of intraday liquidity is desirable.

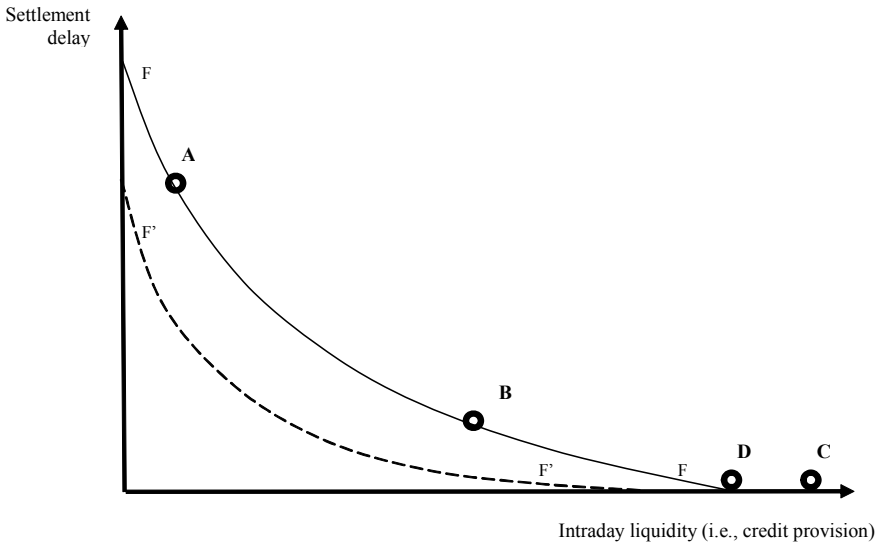
4.3 A simple graphical framework

4.3.1 Description of the framework

The expected relationship between settlement delay and intraday liquidity in an LVPS is illustrated in Figure 4.2 below. Figure 4.2 is inspired by the concept of an 'efficient frontier' presented by Berger, Hancock and Marquardt (1996).[10] This framework will help in interpreting the empirical results later in the paper.

[10] In describing this framework, the terms 'intraday liquidity' and 'intraday credit' are used synonymously.

Figure 4.2 **The LVPS delay-liquidity efficient frontier**



The framework is presented in delay-liquidity space. All points in the space represent possible settlement delay-intraday liquidity combinations necessary to produce a given level of payment activity. The vertical axis measures the magnitude of overall settlement delay in the LVPS while the horizontal axis measures the provision of intraday credit. It is useful to think of the magnitude of settlement delay in an LVPS as reflecting both the number of payments entering the queue upon intended submission and also each payment's duration in the queue until being processed. The tradeoff is captured by the curve denoted FF, and this curve is generated based on the existing technology for processing payments (ie the existing LVPS design). Specifically, the curve shows how settlement delay and intraday credit provision can be traded off against each other for a given level of payment activity under current LVPS arrangements. The slope of FF captures the reduction in settlement delay that can be achieved by participants following a unit increase in the provision of intraday credit.

The decreasing convex shape of the tradeoff curve reflects the assumption of diminishing marginal returns to liquidity. An increase in intraday credit provision is anticipated to have a lesser impact in terms of reduced settlement delay when moving further along the frontier from left to right. This assumption is attributed to the positively skewed nature of the distribution of individual payment

values in an LVPS.[11] At a very low level of liquidity (point A), a small increase in intraday credit provision will lead to a higher reduction in settlement delay since many smaller payments that would otherwise have been delayed can now be immediately processed upon intended submission. As intraday credit provision is continuously increased, it is expected that more payments will be processed upon intended submission and the delayed finality of these payments will be averted. However, even at higher levels of intraday credit provision (such as point B), it is expected that a few very large payments will still be delayed. Only a substantial injection of intraday credit would allow these payments to be processed immediately.

All combinations along the curve, and also above and to the right of the curve, represent feasible combinations of settlement delay and intraday liquidity for a given level of payment activity under the existing LVPS design. The tradeoff curve is the most technologically efficient of these feasible combinations and, therefore, an LVPS is considered to be technically efficient if it is processing payments anywhere along the curve. This notion of efficiency captures the idea that, when operating along the curve, reductions in settlement delay can only be achieved by an increase in intraday credit provision, and vice versa, for a given level of payment activity. Processing the same level of payment activity at a point above, or to the right, of the tradeoff curve represents inefficiency. For instance, producing at a point like C in Figure 4.2 means that intraday credit provision could be reduced and participants' liquidity costs lowered without causing any increase in settlement delay. In fact, intraday credit provision could be lowered from point C all the way to point D before any further reductions lead to increased settlement delay in the LVPS. Point D represents the familiar upper bound of liquidity as described in Leinonen and Soramäki (1999, 2003). Points below the efficient frontier are currently unattainable given the existing LVPS technology and can only be achieved through some form of innovation.

[11] For instance, in Canada's LVTS, the average payment value is around CAD 7.5 million while the median value is around CAD 50,000. Moreover, the value of some payments in the LVTS is well over CAD 100 million.

4.3.2 Innovation: a complex queue-release algorithm

As mentioned above, points below the tradeoff curve are not attainable given the existing LVPS technology. An improvement that allows lower settlement delay for any given level of intraday liquidity, or vice versa, is required to attain such an outcome. The impact of this improvement appears in Figure 4.2 as a shift of the tradeoff curve FF to its new position closer towards the origin at F'F'. Along the new curve, the same amount of payment activity can be produced with lower settlement delay for each level of intraday liquidity, and therefore at a lower overall cost to participants.

Such an improvement can be achieved through a technological innovation in LVPS design. Reductions in settlement delay can be achieved through either faster processing of queued payments or fewer payments entering the queue upon submission, where the latter may occur as a result of the former. Faster processing of queued payments means that intended receivers will obtain incoming funds more quickly, reducing the likelihood that their own subsequent outgoing payments will become queued upon submission. It is argued that the replacement of standard queuing arrangements with the introduction of central queuing with a complex queue-release algorithm represents such an innovation. The benefit of these types of algorithms, in terms of both reduced settlement delay and intraday liquidity needs in an LVPS, are frequently highlighted throughout the payments literature. For example, see McAndrews and Trundle (2001), BIS (2005), Leinonen and Soramäki (1999), Bech and Soramäki (2001), Güntzer, Jungnickel, and Leclerc (1998) and Koponen and Soramäki (1998).

These algorithms are designed to simultaneously search for and offset batches of queued payments, thus serving as an effective coordination device for participants' incoming and outgoing payments. Recall, under standard queuing procedures, payments are released from the queue *individually* when a participant's intraday liquidity is sufficient for them to be processed. In contrast, under central queuing with a complex queue-release algorithm, the simultaneous processing and release of a batch of queued payments is attempted at regular intraday intervals. In this latter case, LVPS participants no longer must wait to obtain sufficient intraday funds for their queued payments to be released individually, but rather they only need to hold the amount of intraday funds necessary to settle any net debit position resulting from the payment offset. The anticipated benefits to LVPS participants from this innovation include lower intraday liquidity needs and related costs for the release of queued

payments, faster processing times for these queued payments, and a reduction in average intraday queue length, when compared to a standard queuing arrangement.

The addition of a complex queue-release algorithm will not necessarily represent a new development in all LVPS, since these algorithms have been used in some systems in the past as a gridlock resolution mechanism. However, over the last decade increases in computing power have led to the improved design and more frequent use of these algorithms within an LVPS central queue. The complexity of these algorithms has also risen considerably; the choice of full or partial optimisation is available and offsetting may take place on a bilateral and/or multilateral basis; BIS (2005).

To sum up, it is expected that the addition of a central queue with a complex queue-release algorithm will lead to an improvement in the tradeoff between settlement delay and intraday liquidity in an LVPS and will allow participants to complete the same level of payment activity at a lower overall cost, relative to a standard queuing arrangement.

4.4 Empirical study: estimating the tradeoff in Canada's LVTS

This empirical exercise considers the tradeoff between settlement delay and intraday liquidity in Canada's LVTS. Some questions that may arise are: What does the tradeoff curve look like for the LVTS? Does it have the same shape as outlined above? Are there possible LVTS design changes, relating to queuing arrangements or otherwise, that could potentially improve this tradeoff, where the same level of payment activity can be processed with either reduced settlement delay or lower intraday liquidity needs or both? The remainder of this paper is devoted to answering these questions using simulation analysis. Simulation analysis is a recent development in payment systems research. Simulation models are a valuable tool since they often can be calibrated to replicate a specific LVPS environment. These models can then be used to assess the impact of changes in the structural arrangements and decision parameters of an LVPS without causing any costly disruption to the operation of the actual system.

4.4.1 Background on the LVTS[12]

The LVTS is an RTGS-equivalent system, where individual payment messages are processed on a gross basis in real-time and settlement of the system occurs on a multilateral net basis at the end of the day. The LVTS's risk controls and collateral arrangements, coupled with a settlement guarantee provided by the Bank of Canada, provide certainty of settlement for the system.[13] Certainty of settlement facilitates intraday finality for all individual payments sent through the LVTS. Recipients of LVTS payments can make use of these funds immediately upon receipt without any possibility that a payment will become unwound. The LVTS consists of two payment streams – Tranche 1 (T1) and Tranche 2 (T2) – and participants may use either stream when sending payments through the system. Each stream has its own real-time risk controls and collateral arrangements. The focus of this analysis is on the T2 payment stream since, due to its more economical collateral requirements relative to T1, it is the dominant stream for LVTS activity.[14]

Intraday liquidity in T2 is facilitated by T2 payments previously received and also by drawing on a T2 intraday line of credit. This intraday line of credit is subject to both a (indirect) collateral requirement and a net debit cap. Specifically, LVTS participants grant bilateral credit limits (BCLs) to each other, where the value of a BCL represents the maximum bilateral T2 net debit position that a grantee (credit line recipient) may incur vis-à-vis the grantor (credit line provider) at any time during the payment cycle. A participant's T2 intraday credit limit, known as its T2 Net Debit Cap (T2NDC), is calculated as the sum of all BCLs granted to it by others in the system multiplied by a system-wide parameter (SWP), which is currently equal to 0.24.[15] The T2NDC represents the maximum multilateral T2

[12] Only LVTS background information relevant to the analysis is provided here. For more information on the LVTS, see Dingle (1998) and Arjani and McVanel (2006).

[13] In the extremely remote event of multiple participant defaults in the LVTS, and if collateral value pledged by participants to the Bank of Canada is not sufficient to cover the final net debit positions of all defaulters, the Bank stands ready to exercise its settlement guarantee by realising on available collateral and absorbing any residual loss.

[14] Approximately 87% of daily LVTS value and 98% of daily LVTS volume are sent through the T2 payment stream, on average. T1 consists of mostly time-sensitive payments between LVTS participants and the Bank of Canada.

[15] The SWP is an exogenous parameter established by the CPA. When the LVTS began operations in February 1999, the SWP was equal to 0.30. Since then, it has been gradually reduced and has been equal to 0.24 since March 2000. The choice of SWP value (SWP < 1) reflects the effect of multilateral netting; Engert (1993). See LVTS Rule No. 2 for information on the SWP.

net debit position that a participant can incur during the LVTS payment cycle. The T2NDC of hypothetical bank n (where n = 1,..,N) is calculated as follows

$$T2NDC^n = \sum_{j \neq n}^{N-1} BCL_{jn} \cdot SWP$$

It follows that two real-time risk controls are applied to payments submitted to the T2 payment stream. A payment will only be processed if it does not result in the sending participant exceeding either its BCL vis-à-vis the receiver or its T2NDC.

A survivors-pay collateral pool is used in T2 to facilitate LVTS settlement in the event of participant default. Eligible collateral consists mainly of government securities and also high-quality corporate debt. Participants are required to pledge T2 collateral equal to the value of the largest BCL that they grant to any other participant, multiplied by the SWP. The value of this T2 collateral obligation is referred to as a participant's Maximum Additional Settlement Obligation, or MaxASO. Essentially, a participant's MaxASO represents its maximum financial loss allocation as a result of another participant's default in the LVTS. Hypothetical bank n's MaxASO is calculated as follows

$$MaxASO^n = \max(BCL_{n,j \neq n}) \cdot SWP$$

The LVTS employs a central queue. Submitted payments to the LVTS failing the real-time risk controls are stored in this queue.[16] The queue is equipped with an offsetting algorithm that runs at frequent intervals (every 15 minutes) throughout the payment cycle. This complex queue-release algorithm, called the Jumbo algorithm, searches for and offsets full or partial batches of queued payments on a multilateral and/or bilateral basis.[17] Payments successfully released by this mechanism are processed by the LVTS as normal. However, current LVTS rules state that excessive use of the central queue is not

[16] Payments are stored on a First-In First-Out (FIFO) basis within each tranche type. Currently, only 'Jumbo' payments (> CAD 100 million) failing the real-time risk controls become centrally queued in the LVTS.

[17] For queued T2 payments, the Jumbo algorithm applies partial offsetting on both a bilateral and multilateral basis over two stages. See Arjani and McVanel (2006) for more information on this algorithm.

encouraged.[18] Instead, participants utilise internal queues to store payments that are unable to pass the real-time risk controls upon intended submission. Internally queued payments are typically re-submitted against the LVTS's risk controls (within a participant's internal LVTS workstation) *individually* on a by-pass FIFO basis each time that its intraday liquidity position is increased.[19] If this process reveals that an internally queued payment can pass the risk controls, it is automatically released to the LVTS for processing.

4.4.2 Settlement delay and intraday liquidity in T2: tradeoff and improvement

Deciding on how to hypothetically impose a reduction in participants' intraday liquidity represents a key aspect of the analysis. For the LVTS T2 payment stream, one way to accomplish this is to constrain the intraday credit available to participants by lowering the value of the SWP.[20] As in the earlier discussion, a reduction of the SWP will entail both a benefit and cost for LVTS participants, holding BCL values constant. The benefit is that a reduction in the value of the SWP will lower participants' T2 collateral requirement and related liquidity cost. However, assuming that no migration of payments from T2 to T1 occurs, reducing the SWP will likely also increase the level of settlement delay in the T2 payment stream. This is because participants' T2NDCs will decline, lowering T2 intraday liquidity in the system, and causing more payments to become queued upon their intended submission. Under current queuing arrangements, delayed payments will accumulate in participants' internal queues until the sending participants' T2 liquidity is sufficient for these payments to be processed by the LVTS.

The tradeoff curve between settlement delay and intraday liquidity in the LVTS is expected to have a decreasing convex shape as outlined in the earlier graphical framework. As the SWP is reduced further, overall settlement delay in the system is expected to rise at an

[18] LVTS Rule No. 7 states that participants are able to track their bilateral and multilateral positions in real-time through their internal LVTS workstations and are expected not to submit payments that will fail the risk controls.

[19] Under bypass-FIFO, a participant's first (earliest) queued payment will be re-tried against the risk-controls. If it does not pass, this payment will be by-passed and the participant's second queued payment will be re-tried, and so on.

[20] Alternatively, such reductions in intraday credit availability can also be achieved through reductions in the value of BCLs that participants grant to each other, while maintaining the current SWP value of 0.24.

increasing rate. Participants will become constrained by their T2NDC more quickly and frequently throughout the day when trying to send payments. In the extreme case, an SWP equal to zero will result in a state of payments deadlock where settlement delay reaches a maximum. No participant will have access to T2 intraday credit and therefore will not be able to incur a T2 net debit position. Consequently, no payments will be sent and all will remain unsettled in participants' internal queues until the end of the day.

It has been argued that an improvement in the tradeoff between settlement delay and intraday liquidity can be achieved with the introduction of a complex queue-release algorithm in the central queue. The LVTS already contains a central queue with a partial offsetting algorithm, but use of this queue is currently discouraged. It is anticipated that, by allowing increased use of the LVTS central queue (and this algorithm), overall settlement delay could be reduced for each hypothetical level of T2 intraday credit provision. Under this alternative scenario, participants would no longer need to manage an internal payments queue and instead would submit all payments to the LVTS at the time they are intended regardless of whether these payments could be immediately processed by the system. Release of these queued payments could then be attempted on a multilateral net basis rather than individually.[21] This proposed change in queuing regime is expected to increase the efficiency of the system since, even where the amount of T2 intraday credit available to participants (and related cost) is lowered, the processing time for queued payments can be faster, and average queue length could decrease, compared with current internal queuing arrangements.

In the next sections, a simulation approach will be utilised to shed light on the following questions:

- Under current internal queuing arrangements, what does the tradeoff between settlement delay and intraday liquidity in the LVTS look like? Is it consistent with the assumptions of the graphical framework presented above?
- Could increased use of the LVTS central queue improve this tradeoff? In other words, can the level of settlement delay associated with each amount of intraday credit be reduced for a given level of payment activity?

[21] The key benefit of central queuing compared to internal queuing is that multilateral offsetting of payments is only possible in the former case.

4.5 Data description and simulation methodology

4.5.1 Description of data

Three months of LVTS T2 transaction and credit limit data have been extracted over the period July-September 2004. Transaction data include the date and time that each transaction was submitted to the LVTS as well as the value of each payment and the counterparties involved in the transaction. It is assumed that the time stamp attached to each payment represents the intended submission time of the payment. Transactions data include only those payments processed by the LVTS and do not include rejected or unsettled payments. Data on credit limits include the value of the T2NDC available to each participant as well as the date and time that the value of the T2NDC is effective. These data represent 64 business days and approximately 1.05 million transactions and are believed to be representative of normal LVTS activity. Table 4.1 provides a summary of the transaction data.[22]

Table 4.1 **Summary of LVTS T2 transaction data**

	Jul 2004	Aug 2004	Sep 2004
Total value of T2 payments (CAD billion)	2,283.0	2,203.5	2,446.5
(% of LVTS total)	(87.8)	(87.9)	(86.3)
Total volume of T2 payments	349,948	344,357	356,676
(% of LVTS total)	(98.0)	(98.0)	(98.1)
Daily average value (CAD billion)	108.7	100.2[23]	116.5
Daily average volume	16,664	15,653	16,985
Average payment value (CAD million)	6.52	6.40	6.86
Median payment value (CAD)	42,436	40,377	45,719

[22] In addition, the Hirschman-Herfindahl Index (HHI) suggests that payment activity over the sample period is somewhat concentrated. The HHI will vary between 0.50 (concentration among only two banks) and 1/N (equal distribution of payment activity among all participants), where N represents the number of banks in the sample. In this case, 1/N = 0.08. The average HHI value for the sample is 0.1944 and 0.1813 for T2 payments value and volume, respectively. A value in this range is consistent with payment activity being distributed evenly across approximately 5–6 banks. Indeed, the largest five Canadian banks account for between 85–90% of daily LVTS value and volume.

[23] A lower average daily T2 payments value in August is expected given that the Canadian civic holiday occurs during this month. Total value reached only CAD 6.9 billion on this holiday in 2004.

4.5.2 Simulation description and methodology

The simulation analysis is conducted using a payment and settlement simulator developed by the Bank of Finland (the BoF-PSS2). This software application is currently being used by over thirty central banks. It should be noted that the version of the BoF-PSS2 used for this analysis does not contain BCL functionality, which is an important component of the LVTS.[24] As a result, the methodology in this paper includes the assumption that BCL values remain constant in light of proposed changes to LVTS rules on queue usage. Further, participants' payment-sending behaviour is also treated as exogenous and therefore the same transactions data are used throughout the analysis. Potential implications associated with these assumptions are addressed later in the paper.

Two batches of simulations will be run where each batch is intended to replicate a different LVPS design. In particular, batch one replicates the current internal queuing arrangement in the LVTS, while batch two replicates the alternative central queuing arrangement. Each batch consists of eight individual simulations (s = 1,2,..,8), where each simulation is distinguished by tighter constraints on participants' intraday liquidity. Changes in intraday liquidity are introduced by altering the value of each participant's T2NDC. Since it is assumed that BCLs remain constant, a reduction in each participant's T2NDC is achieved by hypothetically lowering the value of the SWP. Specifically, each individual participant n's T2NDC in simulation s is calculated as follows

$$T2NDC_s^n = SWP_s \cdot \sum_{j \neq n}^{N-1} BCL_{jn}$$

where $SWP_{1,...,8}$ = 0.24, 0.21, 0.18, 0.15, 0.12, 0.09, 0.06, 0.03.[25]

In specifying the first batch of simulations, the objective is to mimic participants' decision to either submit a payment to the LVTS for processing or hold the payment internally when sufficient intraday funds are unavailable. Settlement delay occurring in this batch

[24] A version of the BoF-PSS2 was released in 2006 that includes both multilateral and bilateral credit limits functionality. Bank of Canada staff were involved in the development and testing of this new version.

[25] Transactions data include only processed payments under the current SWP value of 0.24. Thus, it is not possible to observe potential reductions in settlement delay from an SWP value greater than 0.24, due to a lack of readily available data on delayed or unsettled transactions for this SWP value.

represents payments being held internally by participants, ie the simulator's queue is replicating participants' internal queues. A bypass-FIFO queue-release algorithm is specified to imitate current internal queuing practices of LVTS participants. When this algorithm is applied, a participant's queued payments are re-submitted from the queue and re-tried against the risk controls on an individual bypass-FIFO basis whenever its intraday liquidity position improves. In the real LVTS, this occurs within the participant's internal workstation. Internally queued payments that can successfully pass the risk controls are assumed to be released from the participant's queue and submitted to the LVTS for processing. In interpreting the simulation results for this first batch, settled transactions are assumed to be those that participants were able to submit to the LVTS for processing, while unsettled transactions represent those remaining in participants' internal queues due to lack of intraday liquidity.

Specification of the second batch is intended to replicate a central queuing regime similar to that available in the LVTS. In these simulations, two queue-release algorithms are specified that closely match the LVTS's actual release mechanisms. The first of these algorithms is a FIFO (no by-pass) queue-release algorithm which re-submits a participant's centrally queued payments against the risk controls on an individual FIFO basis each time its intraday liquidity position improves. The second is a complex queue-release algorithm which employs partial offsetting on a multilateral basis and is scheduled to run every twenty minutes, similar to the LVTS's Jumbo algorithm.[26] Settlement delay captured in this second batch of simulations is meant to represent payments being held in the system's central queue, ie the simulator's queue is replicating the LVTS central queue. In the simulation results for this batch, all payments in the sample are assumed to have been submitted to the LVTS at their intended time of submission, and unsettled transactions are those remaining in the central queue which cannot be processed due to a sender's lack of intraday liquidity.

[26] At the time that the analysis was conducted, the frequency of the Jumbo algorithm was every 20 minutes. The frequency of this algorithm increased to every 15 minutes in December 2005. Since bilateral credit limit functionality is currently not incorporated in the simulation application, the partial offsetting algorithm used in the simulations does not exactly replicate the LVTS Jumbo algorithm for T2 payments. Despite this limitation, the results generated by the simulations are still expected to be useful and relevant. Further, in specifying this second batch of simulations, it is also assumed that the LVTS's queue expiry algorithm is no longer utilised and *all* payments failing the risk control check become centrally queued (not just 'Jumbo' payments).

Three alternative measures of settlement delay are calculated for each simulation within each batch. These measures are intended to capture the daily level of settlement delay associated with each amount of intraday credit provision under both the current and alternative queuing environments described above for the same level of payment activity. They are described as follows

1. Daily proportion of unsettled transaction value (PU):

$$PU_t^N = \left(\frac{\text{Value of unsettled transactions}_t^N}{\text{Value of submitted transactions}_t^N} \right)$$

This indicator is calculated on an aggregate level (ie across all participants) for each day t in the sample, where t = (1,…,64). This measure represents the occurrence of the maximum settlement delay possible for a payment in this analysis. Unsettled transactions represent those that enter the queue upon intended submission and remain there until the end of the day.

2. Daily system-wide delay indicator (DI):

$$DI_t^N = \left(\sum_{n=1}^{N} \omega^n \rho^n \right)$$

$$\text{where } \rho^n = \left(\frac{\sum_{i=1}^{T} Q_i^n}{\sum_{i=1}^{T} V_i^n} \right) \text{ and } 0 \le \omega^n, \rho^n, DI^N \le 1$$

Adapted from Leinonen and Soramäki (1999) and commonly used in payment simulation analyses, this indicator is calculated on an aggregate level and is based on a weighted average of each individual (n) participant's daily delay indicator (ρ). This indicator (and the ratio ρ) can take on any value between 0 and 1, where a value of 0 is achieved when all payments are successfully processed by the LVPS upon intended submission and no settlement delay occurs. A value of 1 is calculated where all payments become queued upon intended submission and remain unsettled at the end of the day. Weights (ω) are based on participants' average share of

total transaction value over the 64-day sample period. Calculation of this measure requires dividing each LVTS business day into T=108 ten-minute intervals (i = 1,..,T). The numerator of ρ represents the sum of a participant's queued payment value (Q) over all T ten-minute intervals throughout the day. The denominator represents the sum of the cumulative value of a participant's submitted payments (V) over all T ten-minute intervals throughout the day. It follows that this indicator is influenced by both the value and delay duration of each payment in the queue calculated for each intraday interval.

3. Average intraday (interval) queue value (AQV):

$$AQV_t^N = \left(\frac{\sum_{i=1}^{T} Q_i^N}{T} \right)$$

This is an aggregate measure which calculates the average value of queued payments in an interval over day t. It is found by dividing the sum of total queued payment value (Q) over all T ten-minute intervals on each day by the number of intervals per day (T=108).

4.6 Simulation results and discussion

4.6.1 The delay-liquidity tradeoff in the T2 payment stream

Simulation results for each of the three delay measures are presented in Figures 4.3 through 4.5. Two curves are presented in each graph corresponding to each batch of simulations. The curve denoted 'internal queuing' portrays the simulation results estimated under current LVTS (internal) queuing arrangements. The curve denoted 'central queuing' depicts results estimated under the alternative LVTS (central) queuing environment.

Figure 4.3 **Average daily proportion of unsettled transaction value**

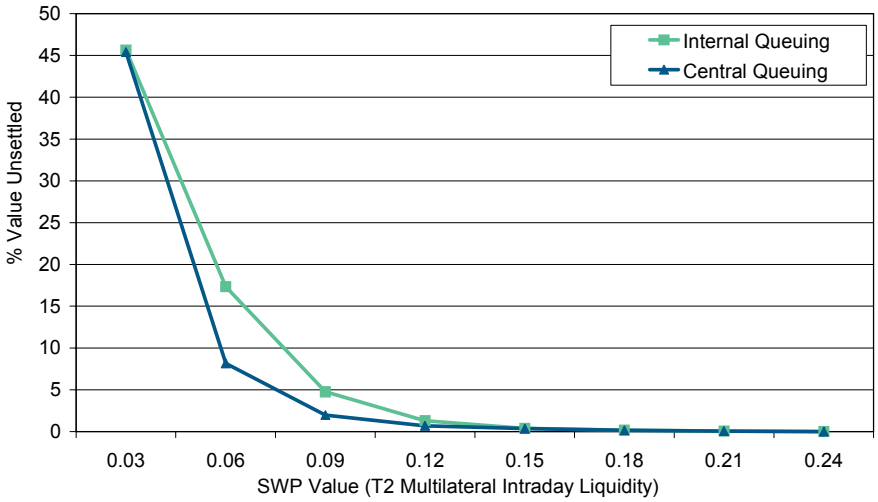


Figure 4.4 **Average daily system-wide payments delay**

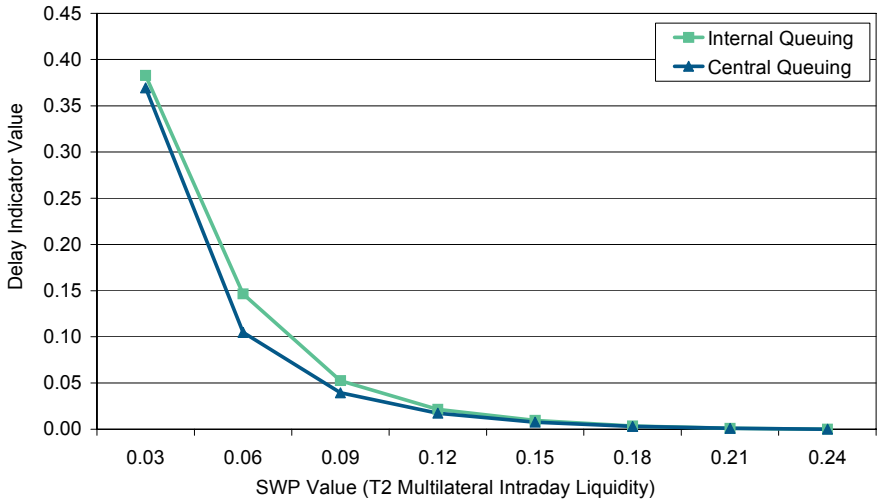
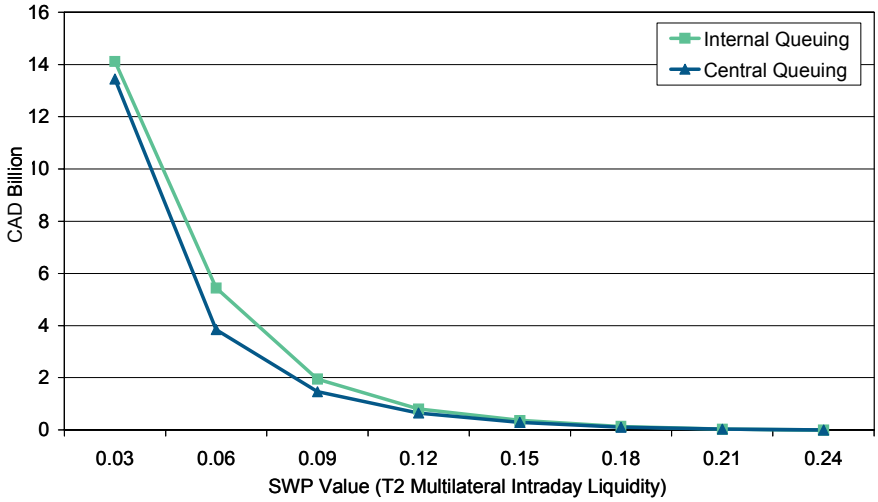


Figure 4.5 **Average intraday (interval) queue value**



Earlier hypotheses regarding the tradeoff between settlement delay and intraday liquidity are confirmed by the simulation results. Under current LVTS queuing arrangements, a tradeoff exists in the LVTS's T2 payment stream according to all three delay measures. Like the earlier graphical framework, the curve is convex; as intraday credit constraints are further tightened (by lowering the value of the SWP), participants' intraday liquidity becomes more scarce and settlement delay in the system rises at an increasing rate. The slope of this curve increases substantially at low amounts of intraday credit provision.

The introduction of a design innovation – allowing increased use of the LVTS central queue – results in an improvement to this tradeoff and the curve shifts closer towards the origin according to all three measures. Settlement delay associated with each level of intraday credit provision is reduced following the introduction of the partial offsetting algorithm. The relative benefit of partial offsetting (in terms of reduced delay) increases gradually as intraday liquidity is further constrained. At the SWP value of 0.06, the difference in settlement delay between the two queuing regimes is greatest. In this case, the average proportion of unsettled transactions value is reduced by 9 percentage points or about CAD 10 billion (Figure 4.3), the system-wide delay indicator is reduced by 28% (Figure 4.4) and average intraday queue value is reduced by 29% or about CAD 1.6 billion (Figure 4.5), relative to the first batch of simulations.

Gains from the alternative central queuing design begin to decline when the SWP is reduced beyond 0.06, as the system begins to

approach a state of deadlock. When the SWP value is 0.03, settlement delay is only slightly reduced following the introduction of a partial offsetting algorithm, which could mean that participants' intraday liquidity levels are so low that only very small batches of queued payments can be processed each time this algorithm runs. At this level of SWP, close to half of all daily payment value remains unsettled on average under both queuing regimes (Figure 4.3).

The simulation results also reveal another finding that is closely related to the notion of technical efficiency described earlier. The above results suggest that, under current queuing arrangements, settlement delay in T2 increases when the SWP value is lowered from 0.24 to 0.21. However, it remains to be seen whether reductions in the SWP below 0.24 but still greater than 0.21 can be achieved without inducing any further settlement delay in the LVTS. In other words, can a lower amount of T2 intraday credit (and an associated reduction in T2 collateral requirements) be accommodated without increasing the level of settlement delay for payment activity during the three-month sample period, holding all other factors constant? If this were the case, it would be similar to operating at point C in the graphical framework. Indeed, the simulation results suggest that the current value of SWP (= 0.24) is needed to process payments in this sample and cannot be reduced further without increasing the level of settlement delay. This is not necessarily a surprising result since one might expect participants to conform to this value of SWP when sending payments through the system. A complete discussion of this analysis, including full details of the simulation methodology used, is provided in Appendix 1.

4.6.2 Discussion

Some other interesting discussion points emerge from these results, offering areas for future research. First, the simulation results suggest that, under both existing LVTS queuing arrangements and also under the alternative central queuing arrangement, settlement delay in T2 will increase only marginally as the SWP is initially reduced from its current value of 0.24, holding all other factors constant. For example, a reduction in the SWP from 0.24 to 0.18 is estimated to increase the average proportion of unsettled daily transaction value by only 0.15 per cent under the current queuing regime and 0.14 per cent under a central queuing arrangement (Figure 4.3). Similar results are also observed according to the other two delay measures. Reducing the SWP entails a benefit for LVTS participants in the form of lower T2

collateral requirements and related liquidity cost, as has already been mentioned. Specifically, a reduction in the SWP to 0.18 reduces the aggregate value of T2 collateral required by about CAD 750 million per day on average over the sample period, holding BCL values constant. On one particular day in the sample, the value of T2 collateral required is about CAD 1 billion less when the SWP is equal to 0.18.

This raises the question as to whether or not a lower-cost combination of intraday credit provision and settlement delay currently exists for LVTS participants in the T2 payment stream.[27] Put differently, is it the case that the marginal settlement delay cost incurred by moving to an SWP value of 0.18 equals the marginal cost of additional intraday credit provision (and collateral) associated with the current value of 0.24? If the former cost is less than the latter, then lowering the SWP to 0.18 could lead to overall cost-savings for participants. Of course, answering this question entails, among other things, the difficult task of quantifying the cost of the additional settlement delay associated with moving to a SWP value of 0.18.

Secondly, the analysis highlights the possible benefit of central queuing with a complex queue-release algorithm with respect to settlement delay and intraday credit provision. Nonetheless, participants face other types of risk and cost in the LVPS environment, and such a change in LVTS queuing arrangements could increase participants' other costs. For example, as outlined in BIS (1997), a possible implication of permitting unrestricted use of the central queue pertains to the issue of queue transparency and specifically whether the reduction in settlement delay could be replaced by an increase in credit risk taken on by participants. A participant, upon observing an incoming payment in the central queue, may choose to provisionally credit its client's account with these expected funds before the payment actually arrives, thus exposing itself to credit risk until the payment is successfully received. If these funds do not eventually arrive for some reason, the participant would seek to unwind this payment, which would be costly for both the participant and its client. This issue is pertinent to the LVTS because participants have the ability to track expected incoming and outgoing payments in the queue in real-time through their internal participant workstations. Although details regarding client recipients of incoming queued payments are not included in these workstation reports,

[27] Alternatively, the question could instead be posed as whether current values of BCLs granted by participants to each other are cost-minimising holding the current SWP value constant.

participants could informally access this information. However, it is not clear that LVTS participants would be willing to incur this credit risk in any case.[28]

4.7 Conclusions and caveats

The objective of this paper has been to gain a better understanding of the tradeoff between settlement delay and intraday liquidity in an LVPS, with a specific focus on the Canadian LVTS. Simulation analysis shows that a tradeoff exists in the LVTS between settlement delay and intraday liquidity, and that this tradeoff exhibits a decreasing convex shape. Further, allowing increased use of the LVTS central queue (and the Jumbo algorithm) is expected to improve this tradeoff, ie, reduce settlement delay in the system for all levels of intraday liquidity considered. Such an innovation improves the efficiency of the system, leading to overall cost-savings for participants.

At the same time, it was found that under both the current and proposed queuing regimes, a modest reduction in the SWP below its current value results in only a marginal increase in the level of settlement delay in the LVTS, while providing substantial T2 collateral cost-savings for system participants. Further research is necessary to quantify whether this collateral cost-saving benefit is worth the associated increase in settlement delay cost. It was also argued that, although increased use of the central queue is expected to reduce total settlement delay and liquidity costs for participants, this may result in a potential increase in credit risk taken on by participants. However, LVTS participants may not necessarily react to a change in LVTS queuing arrangements in this manner.

These results are preliminary, and certain caveats exist. These caveats are raised here with the intention of motivating further research. The first caveat relates to behavioural assumptions made throughout the analysis. Significant changes to LVTS queuing arrangements were proposed in the analysis. However, despite these changes, the current simulation methodology assumes that LVTS

[28] This credit risk issue may also be avoided in the LVTS since a client beneficiary of funds can always request a Payment Confirmation Reference Number (PCRN) from its participant bank. All payments processed by the LVTS are assigned a PCRN indicating that the payment has successfully passed all LVTS risk control tests and is thus considered final and irrevocable. Upon obtaining the PCRN, the beneficiary does not have to worry about the funds being revoked at a later time.

participants' payment sending and bilateral credit granting behaviour remains unchanged. One must question whether this is a realistic assumption. For example, following discussion in McAndrews and Trundle (2001), the availability of netting is likely to increase the incentive for participants to submit payments to the system earlier in the day, relative to these payments' current intended submission times, essentially increasing the scope for multilateral netting of payment messages. The benefit of netting is expected to increase with the number and value of payments in the queue at the time that it occurs. Anecdotal evidence suggests that LVTS participants typically receive information regarding outgoing payment requests well in advance of their intended submission time. Participants' collective submission of as many payments as early as possible to the system under a central queuing regime is anticipated to result in a greater turnover of intraday funds, a lesser need for costly intraday credit, and faster processing of these payments. This may result in a further downward shift of the tradeoff curve closer to the origin thus leading to further cost-savings for participants.

At the same time, it is argued that participants, in granting BCLs to each other, strive to minimise the value of their T2 collateral requirement subject to achieving an established level of throughput efficiency, ie an acceptable level of settlement delay. It is likely that payment activity under current internal queuing arrangements may already reflect participants' acceptable levels of settlement delay. Thus, participants may not perceive the benefit of central queuing to be a further reduction in settlement delay, but instead may treat this as an opportunity to realise lower T2 collateral requirements (and costs) while maintaining the same level of settlement delay in the system. This suggests that, under the central queuing arrangement, participants may collectively choose to reduce the BCLs they grant to each other in order to achieve these cost-savings. This reduction in BCLs is expected to continue to the extent that any decline in settlement delay resulting from increased use of the central queue is fully offset.[29]

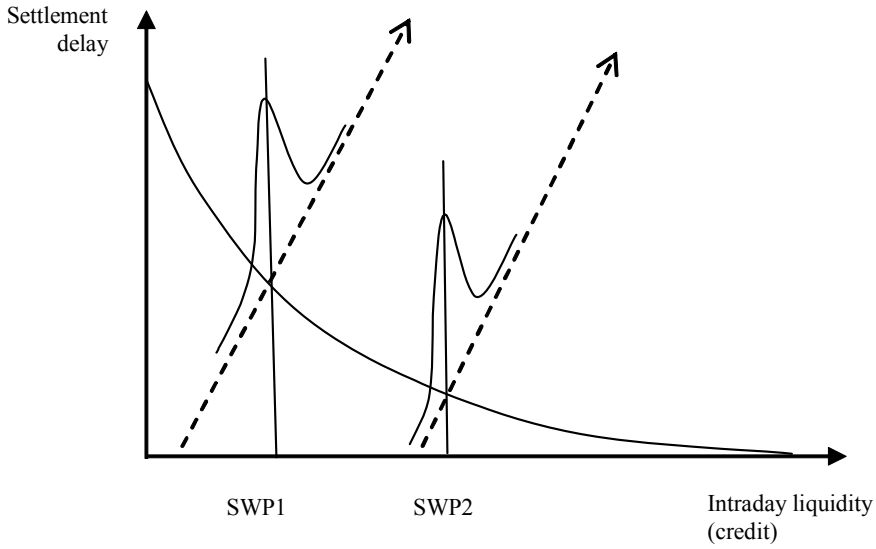
A second caveat follows closely with a discussion found in Bedford, Millard and Yang (2005) and relates to the statistical robustness of the simulation findings. The simulation analysis is

[29] Initially, participants are not likely to know exactly how much BCLs must be reduced to achieve the same level of settlement delay under the alternative central queuing regime. Instead, this will be an iterative process that eventually converges to the equilibrium of a perfect offset. In the interim, it may be the case that participants 'overshoot' this target level of BCL reduction, temporarily resulting in a higher level of settlement delay in the system relative to the existing level.

intended to estimate the increase in settlement delay brought on by a reduction in LVTS participants' intraday liquidity over a three-month sample period. Point-estimates of this impact for each amount of intraday liquidity are used to generate the tradeoff curves presented in Figures 4.3 through 4.5. Previous internal research conducted by the Bank of Canada shows that annual LVTS payment activity is affected by specific calendar events and also monthly trends. Consequently, the estimated impact on settlement delay following reductions in intraday liquidity is expected to take on different values based on the specific dataset used in the analysis. Although using a three-month sample helps to capture the effect of certain monthly and quarterly calendar effects occurring during this period, there is a desire to reduce the risk of small-sample bias and to obtain more statistically robust results. For example, it has been observed that the same calendar event may yield a different effect on LVTS payment activity depending on when it occurs throughout the year. Similarly, use of a single three-month sample may not capture the effect that semi-annual and/or annual calendar events may have on the simulation results. Nor will it capture the potential impact of monthly trends in LVTS T2 payment activity.

In order to achieve more statistically robust results, it is suggested that the same simulation methodology be repeated as many times as is feasible using real and/or artificially generated LVTS payment flow data over some fixed sample duration. Grouping the point-estimates of the impact on settlement delay for each amount of intraday liquidity from all of the samples will facilitate generation of an empirical distribution of this potential impact (Figure 4.6). It follows that the shape of the empirical distribution may be different for each amount of intraday liquidity. For example, the impact on settlement delay may be more volatile and will thus deviate from its mean value more often at lower amounts of intraday credit provision. The shape of the empirical distribution may also change over time.

Figure 4.6 **Plotting distribution of settlement delay outcomes**



A third and final caveat pertains to the absence of BCL functionality in the version of the BoF-PSS2 used in this analysis. This absence creates the possibility that the estimated tradeoff curves provided in Figures 4.3 through 4.5 represent a 'lower bound' of the impact on settlement delay resulting from reduced intraday liquidity. As the value of the SWP is reduced and payments become delayed upon failing the T2 multilateral risk-control test, intended receivers of these payments may consequently be prohibited from sending their own payments when due. All of this will result in added volatility in bilateral net positions, possibly to a point where some participants' bilateral net debit positions are greater than the BCLs granted to them. In the LVTS, this cannot occur due to a bilateral risk control test being applied to every payment which guarantees that participants do not exceed their BCL vis-à-vis a receiving participant. Payments failing the bilateral risk control test become queued until the sending participants' bilateral liquidity position improves. This added delay is not captured in the results generated by the current version of the simulator. This forces the assumption that all LVTS payments, when processed by the simulator, have passed not only the multilateral risk control test, but also the bilateral risk control test. Thus, it would be useful to repeat the analysis again with Version 2.0 of BoF-PSS2 to compare how much greater is potential settlement delay in the system when bilateral risk controls are also taken into account.

References

Arjani, N – McVanel, D (2006) **A Primer on Canada's Large-Value Transfer System.** Bank of Canada Web Document, www.bankofcanada.ca/en/financial/lvts_neville.pdf.

Bank for International Settlements (2005) **New developments in large-value payment systems.** May.

Bank of International Settlements (1997) **Real-Time Gross Settlement Systems.** March.

Bech, M – Soramäki, K (2001) **Gridlock Resolution in Interbank Payment Systems.** Bank of Finland Discussion Papers 9/2001.

Bedford, P – Millard, S – Yang, J (2005) **Analysing the impact of operational incidents in large-value payment systems: A simulation approach.** In Leinonen, H (ed.) Liquidity, risks and speed in payment and settlement systems – a simulation approach. Bank of Finland Studies, 247–274.

Berger, A N – Hancock, D – Marquardt, J (1996) **A Framework for Analyzing Efficiency, Risks, Costs, and Innovations in the Payment System.** Journal of Money, Credit and Banking 28(4), 696–732.

Dingle, J (1998) **The LVTS – Canada's large-value transfer system.** Bank of Canada Review (Autumn), 39–55.

Engert, W (1993) **Certainty of Settlement and Loss Allocation with a Minimum of Collateral.** Bank of Canada Working Papers No. 93–14.

Güntzer, M M – Jungnickel, D – Leclerc, M (1998) **Efficient algorithms for the clearing of interbank payments.** European Journal of Operational Research 106, 212–219.

Koponen, R – Soramäki, K (1998) **Intraday Liquidity Needs in a Modern Interbank Payment System: A Simulation Approach.** Bank of Finland Studies E:14.

Leinonen, H – Soramäki, K (1999) **Optimising Liquidity Usage and Settlement Speed in Payment Systems.** Bank of Finland Discussion Papers 16/99.

Martin, A (2005) **Recent Evolution of Large-Value Payment Systems: Balancing Liquidity and Risk.** Economic Review (First Quarter). Federal Reserve Bank of Kansas City, 33–57.

McAndrews, J J – Trundle, J (2001) **New payment system designs: causes and consequences.** Financial Stability Review (December). Bank of England, 127–136.

Appendix 1

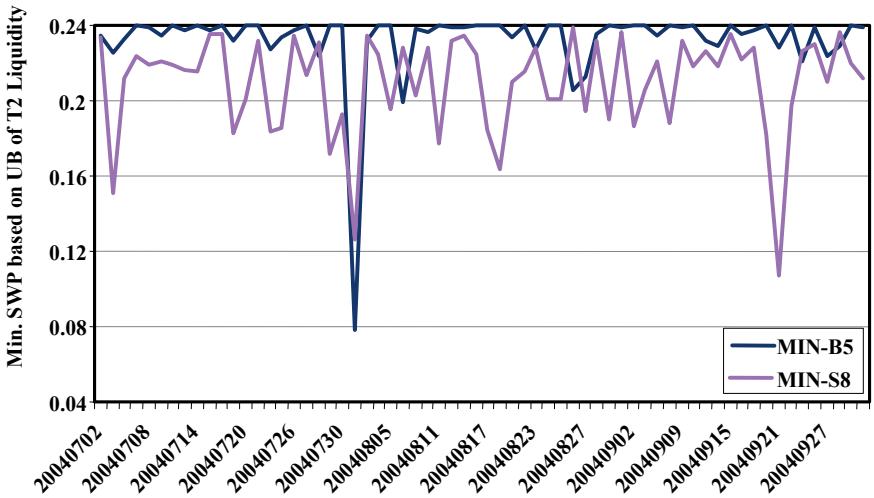
Is the T2 payment stream technically efficient?

The objective of this supplemental analysis is to find the minimum SWP (call this SWP*) necessary to process all payments in the sample without delay, holding all other factors constant. It may be the case that SWP* < 0.24, which means that existing levels of T2 intraday credit, and perhaps more importantly for participants, T2 collateral requirements could be lowered without inducing additional settlement delay during the three-month sample period (recall point C in Figure 4.2).

Simulation results produced by the BoF-PSS2 can provide insight into this issue. Treating participants' payment-sending behaviour as exogenous, a simulation is run using the same sample data but this time specifying unlimited intraday credit. Under this simulation scenario, all payments will pass the risk controls immediately upon submission and therefore no queuing algorithms need to be specified. The daily T2NDC each participant actually needs in order for its payments to be passed without delay can be derived from these simulation results, and is equal to the largest multilateral net debit (negative) position incurred by each participant during the day. This value is defined as a participant's upper bound (UB) of T2 liquidity. The daily UB of T2 liquidity for each participant can then be used to calculate a value of SWP* that, when multiplied by the sum of the actual BCLs granted to each participant, will produce this UB value. It follows that the highest value of SWP* calculated for any participant on any day is considered the minimum SWP* value necessary to send all payments in the sample through the system without delay. This SWP* can then be compared with the current value of 0.24.

The results from this simulation analysis reveal that on 45 of the 64 days, SWP* reached 0.24 for at least one LVTS participant. This means that the current value of SWP was necessary for the immediate processing of T2 payment activity during this three-month sample period. Hence, further T2 collateral cost-savings could not be realised without an increase in the level of settlement delay, holding payment activity constant. The results also indicate that the T2NDC constraint (when SWP=0.24) is binding more often for large LVTS participants (denoted 'B5' in Figure 4.7). Figure 4.7 below shows that on 42 days in the sample at least one of the major Canadian banks reached their T2NDC at some point in the day.

Figure 4.7 **Minimum SWP required – B5 vs. S8**



Focusing on the large LVTS participants, the simulation results show that, on these 42 days, four different institutions bumped up against their T2NDC at least once intraday. One of these participants reached its T2NDC at least once on 37 different days, while the three others reached this limit on 10, 2 and 1 day(s), respectively. The results also indicate that participants did not reach their T2NDC constraint at the same time each day. For example, regarding the first two large participants mentioned above, the LVTS day has been divided into four periods and the time that each of these participants reached its T2NDC has been located in the simulation results and tabulated. A summary of these findings is provided in Table 4.2.

Table 4.2 **Percentage of instances where T2NDC is binding by time of day**

Time of day	Bank 1 (37 instances)	Bank 2 (10 instances)
00:30–06:00	0	0
06:00–12:00	19	0
12:00–17:00	73	40
17:00–18:30	8	60

It also deserves mention that, where a high number of instances occur within a certain period (eg 27 instances for Bank 1 during the interval between 12:00 and 17:00 hours), these occurrences typically do not

take place at the same time within the interval, but rather were scattered throughout the period.

It is not necessarily surprising that SWP* reaches 0.24 on most days in the sample period. The gradual reduction of the SWP from 0.30 to 0.24 between February 1999 and March 2000 was influenced by participants' preferences, and this value has held steady at 0.24 since that time. Given participants' perceived contentment with this SWP value, one might expect participants' to conform to it, meaning that they choose to structure their payment submission behaviour in a certain way so as to make full use of their available T2 intraday credit when sending payments through the system.

Some discussion is also warranted regarding results for the eight smaller LVTS participants (denoted 'S8' in Figure 4.7). On only 4 of the 45 days, SWP* reached 0.24 for one of these participants. Further, this occurred for a different participant in each of these four instances. There exist a variety of possible explanations for these results. It may be the case that larger LVTS participants, in sending a higher volume of payments earlier in the day, are 'subsidising' smaller participants' intraday liquidity in the system, to the extent that smaller participants need to rely less on intraday credit as a source of funding for their outgoing payments. Indeed, SWP* was equal to zero (ie no T2 intraday credit was drawn upon) for at least one small participant on 18 of 45 days in the sample. In contrast, this did not occur on any day for large LVTS participants. A second possible explanation could be that, for various reasons, small LVTS participants may tend to bump up against their BCLs far more frequently relative to their T2NDC. Of course, further research is necessary before either of these explanations can be confirmed.

Chapter 5

Funding levels for the new accounts in the BOJ-NET

Kei Imakubo – James J McAndrews

5 Funding levels for the new accounts in the BOJ-NET 124

Abstract 124

5.1 Introduction 124
5.2 Large-value payments in Japan 126
5.2.1 Current structure of large-value payment systems .126
5.2.2 Future structure of large-value payment systems ...127
5.3 Liquidity effects of combining FXYCS, Zengin and BOJ-NET payments 129
5.4 Optimising funding levels 132
5.5 Simulations and results.....134
5.5.1 Four baseline simulations.....135
5.5.2 Distributional funding simulations 139
5.5.3 Progress-payment simulations.....142
5.6 Concluding remarks 144

References 145

Appendix 1 146

5 Funding levels for the new accounts in the BOJ-NET

Abstract

The Bank of Japan decided to implement the next-generation RTGS project of the BOJ-NET Funds Transfer System. Under the project, the new system will have liquidity-saving features and will incorporate large-value payments that are currently handled by two private-sector designated-time net settlement systems, the Foreign Exchange Yen Clearing System and the Zengin System. We analyse characteristics of the optimal funding levels under the new features using simulation analysis and find that the optimal funding levels can be described with the total balances in the system, the distribution of the total balances across participants, and the timing of funding.

5.1 Introduction

In February 2006, the Bank of Japan decided to implement the next-generation RTGS (RTGS-XG) project of the BOJ-NET Funds Transfer System (BOJ-NET), its primary large-value payment system.[1] Under the RTGS-XG project, BOJ-NET will introduce liquidity-saving features in a current real-time gross settlement (RTGS) mode. The new system will also incorporate payments from three different streams of the current payment activities, two of which now settle toward the end of the processing day in private-sector designated-time net settlement (DNS) systems. The project will be implemented in two phases, with the first phase scheduled for fiscal 2008 (April 2008 to March 2009) and the second for 2011. One of the primary motivations for the development of the new system is to quicken settlement of large-value payments relative to the current pattern and to reduce intraday settlement exposure of those payments by allowing for intraday settlement finality and liquidity-saving at the same time.

Much of the design work for the new system is already completed, while some decisions related to the implementation still remain. In the

[1] See Bank of Japan (2006b) for an overview of the RTGS-XG project.

paper, we focus on one aspect of the new system – the levels of funding for newly-developed accounts that will be drawn on to effect settlement throughout the day in a liquidity-saving mode.

The first issue that we explore is whether the plan to incorporate the payments that are currently settled on the two private-sector DNS systems and most payments on the current BOJ-NET into the new system will yield liquidity-saving under a certain level of funding. It is plausible to think that maintaining separate systems might require less liquidity or might result in speedier settlement for a given level of liquidity. If incorporating the payments in the three systems turns to be liquidity-saving, then it can be said that there are liquidity complementarities among the three systems to be combined. As demonstrated in the paper, strong complementarities do exist among the three systems.

Second, we simulate the performance of the new system using several levels of initial balances for the new accounts. In general, there is a clear trade-off between the rate of settlement of a group of payments and the level of funding devoted to those settlements. With a large level of funding, settlement can be made more quickly. Firstly, the total level of funding of initial balances is important in establishing how much value is settled prior to the end of the settlement period. Once the total level of funding is determined, participants can seek to optimise the distribution of initial balances across participants. The optimum distribution of balances across participants leads to the greatest value of settlement within the settlement period for that total level of funding used. A characteristic of the optimum distribution of balances across participants is that additional balances placed in any participant's account yield equal increases in amounts settled. This 'equalisation of marginal benefits' is a characteristic common to many allocation problems in economics.

We examine how changes in a level of initial balances affect the value of payments settled, the amounts left unsettled after a particular time, and the average time of settlement. This information can be useful to participants and planners in seeking the right balance between the value settled during the day and the liquidity-saving potential of the new system. In the context of Japan's payment activities, this is the first examination studying effects of liquidity on intraday settlement.

The paper is organised as follows. We begin in Section 5.2 by briefly describing the current large-value payment landscape in Japan, and how the design of the new system is expected to alter that landscape. We also provide a rough description of the planned new system and explain the purpose of the new account and its funding. In

Section 5.3 we examine changes in liquidity efficiency of combining the two new payment streams with the payments on the current BOJ-NET. In Section 5.4 we describe the problem of finding optimum funding levels, and in Section 5.5 we present the results of simulation analysis. In Section 5.6 we provide a short summary and conclusion.

5.2 Large-value payments in Japan

5.2.1 Current structure of large-value payment systems

BOJ-NET plans to incorporate payments currently made on BOJ-NET, the Foreign Exchange Yen Clearing System (FXYCS) and the large-value payments on the Zengin Data Telecommunication System (Zengin). We briefly describe some aspects of these three systems.[2]

BOJ-NET is a pure RTGS system for the Japanese yen, owned and operated by the Bank of Japan. The system is one of the core financial infrastructures supporting economic and financial activities in Japan. It settles almost JPY 100 trillion daily with annual turnover ranging 40 times as high as Japan's nominal GDP.

BOJ-NET handles both Japanese government Securities (JGSs) and funds transfers. The latter mainly consist of money-market transactions, but also include the settlement payments for various payment and securities settlement systems that use BOJ-NET to transfer the final settlement payments and the cash legs. In addition, money-market operations of the Bank of Japan are carried out using BOJ-NET. There are a limited number of third-party, or customer, payments settled on BOJ-NET, and those are very high-value payments, indicating that these are also money-market transactions conducted by market participants that do not have accounts with the Bank of Japan. Settlement amounts in 2005 indicated that on a daily average basis BOJ-NET settled 21,641 transfers with a total value of JPY 88.3 trillion. The average value per settlement was JPY 4.1 billion.

FXYCS is basically a DNS system that handles yen legs of foreign exchange trades. It conducts the final settlement at 14:30 using BOJ-NET. The volume and value of its daily average activities in 2005 indicated that it settled 28,022 transactions per day with a total value of JPY 16.4 trillion. The average value per transaction was JPY 586

[2] For an overview of payment systems in Japan, see the Japan section of BIS (2003).

million. The net amount transferred on BOJ-NET in 2005 averaged JPY 4.1 trillion. FXYCS has not only a DNS mode but also an RTGS mode, although its use is rather limited.

Finally, Zengin is a simple DNS system, whose final payment takes place at 16:15. In 2005, Zengin averaged 5.4 million transactions per day with a total daily average value of JPY 9.5 trillion. The average size of payments was JPY 1.8 million. It is mainly used for commercial payments. On average, the daily settlement amounts made through BOJ-NET were JPY 1.8 trillion per day in 2005. It is estimated that roughly two-thirds of the value transferred on Zengin, approximately JPY 6 trillion per day, is made up of payments that are larger than JPY 100 million.

5.2.2 Future structure of large-value payment systems

The new system plans to operate as a queue-augmented RTGS system.[3] The new liquidity-saving features will be provided on a new type of accounts as shown in Table 1. Participants will be able to designate payment instructions to be settled either via the new accounts, that will not offer intraday overdrafts capability, or via the standard accounts, on which collateralised overdrafts will remain available. The intent of both participants and the Bank of Japan is that most of the three payment streams described above will be settled via the new accounts. The standard accounts and the dedicated accounts for simultaneous processing of delivery-versus-payment and collateralisation, known as SPDC, will still operate and are intended to be used for the rest of settlements.[4]

The new system will operate the new accounts as follows. The new accounts will be funded by participants each morning at the start of the processing day (9:00) with an infusion of funding from the standard accounts. That establishes the participants' initial balances in the new accounts, because the new accounts will have a zero balance overnight. Participants will then submit payment instructions to the

[3] See BIS (1997), McAndrews and Trundle (2001), and BIS (2005) for basic ideas on a queue-augmented RTGS.

[4] The SPDC facility is another type of liquidity-saving facility used only for settlement of cash legs of JGSs transactions. It allows the receiver of JGSs to pledge the incoming securities as collateral for intraday overdrafts while using the overdrafts to pay for the incoming securities. Similarly, the deliverer of JGSs is able to withdraw the securities pledged with the Bank of Japan for delivery to the receiver while using the funds received to repay the overdrafts.

new accounts, and a bilateral offsetting algorithm will initiate a search for bilaterally offsetting payments on a FIFO basis. If a pair of bilaterally offsetting payments is found, and if funds are sufficient to settle the payments, settlement of the selected payments takes place simultaneously. At designated times, a multilateral offsetting algorithm will attempt to find the largest set of payments that can be settled using available balances.[5] See Appendix 1 for the details of bilateral and multilateral offsetting algorithms in the new system.

Table 5.1 **Account structure in the new system**

	Standard account	SPDC account	New account
Types of transactions settled	– interbank transfers (eg money market, foreign exchange) – third-party transfers – the cash legs of securities transactions – settlement obligations arising from clearing systems – transactions with BOJ/ government	– the cash legs of JGSs transactions using the SPDC facility	– interbank transfers (eg money market, foreign exchange) – third-party transfers (including large-value Zengin payments)
Liquidity supply	Intraday overdrafts	Intraday overdrafts, liquidity transfers from standard account	Liquidity transfers from standard account
Liquidity saving	Not applicable (pure RTGS)	SPDC facility	Queuing and offsetting mechanisms
Account management	Overnight	Intraday (zero balance at the end of the processing day)	Intraday (zero balance at the end of the processing day)
Opening and closing times	9:00–17:00*	9:00–16:30	9:00–16:30

* Closing time is 19:00 for participants that have applied for access to extended hours.

Participants will be able to transfer funds between their new accounts and their standard accounts freely throughout the day. Payment instructions remaining in the queue will be rejected if insufficient

[5] The algorithm will include all queued payments in the initial offsetting and successively drop the largest payment from the participant with the largest funding shortfall until a set of payments that have no funding shortfalls is found. Bech and Soramäki (2001) show that this algorithm finds the largest set of payments that can be settled using a multilateral offsetting given that one breaks a FIFO ordering rule.

funds are submitted to the new accounts by 16:30. The standard accounts will remain open until 17:00.

5.3 Liquidity effects of combining FXYCS, Zengin and BOJ-NET payments

As described above, the new system plans to incorporate payments currently made on BOJ-NET and FXYCS, and the large-value payments on Zengin. The question is whether the combination of these payment streams increases liquidity efficiency by aggregating the currently fragmented payment systems or reduce it by eliminating the DNS systems but with the obvious benefit of permitting intraday settlement of payments. We examine this question by first simulating operations of the new system with payments that are currently settled in BOJ-NET. Then we conduct simulations of the performance of FXYCS and the large-value Zengin, using the settlement method of the new system, while assuming (contrary to the planned design) that they were separately operated from BOJ-NET. Adding liquidity required in each of these two simulations provides an indication of liquidity that would be used if BOJ-NET, FXYCS, and Zengin remained separate systems, but all adopt an intraday finality capability. Finally, we simulate the performance of the new system when payment streams from all these systems are combined and settled in the same system. If liquidity required to settle the combined payment streams is lower than that required to settle the payments when the systems are operated separately (for a fixed level of delay), then it can be expected that there are liquidity complementarities, or scale economies in liquidity use, in combining the payment streams. If, on the other hand, liquidity use is lower with the systems operated separately, then there are diseconomies in liquidity use in combining the systems.

For each system, we conduct three treatments on each day's data (the ten days of historical data in September 2003 are used in the simulations that we report on here).[6] The first treatment is to endow participants with sufficient liquidity to settle the day's payments without delay. The second is to endow them with sufficient liquidity only to settle their multilateral net debit, with which the payments will be settled as quickly as possible (using the new settlement method).

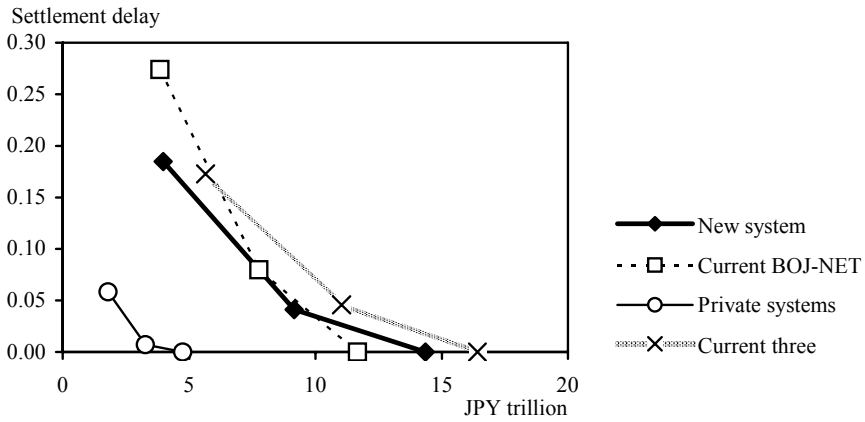
[6] See Appendix 1 for the summary statistics of simulation data.

Finally, in the third treatment, participants are endowed with the average of the two other levels of liquidity – in other words, they are endowed with liquidity that is halfway between the level sufficient to settle payments without delay and the level of multilateral net debits.

We examine a trade-off between liquidity necessary to settle the payments and delay with which the payments are settled. If the locus of points that describes this trade-off shifts inward or outward as the different payment streams are added, it can be said that there are liquidity efficiencies or costs respectively in combining the different payment streams.

The results of these simulations, using the ten days of historical data and the settlement method of the new system, are shown in Figure 5.1. On average it is found that there are significant liquidity complementarities in combining the payment streams. This can be seen clearly in the inward shift of the black line (new system), which illustrates the performance of the new system, relative to the grey line (current three), which illustrates the total liquidity requirements of the three systems when operated separately. The inward shifts show that at all the three levels of delay simulated the new system requires less liquidity to settle the payments.

Figure 5.1 **Delay indicator and liquidity for the separate systems, the sum of the separate systems operating in isolation, and for the new system**



Source: Authors' calculation.

Table 5.2 provides more details on each of the ten days of simulated data and presents both the delay indicator measure and the value-weighted average time of settlement.[7] In every simulation, and for any average time of settlement or any indicator of delay of settlement, the new system requires less liquidity to settle the payments. The results therefore suggest that there are significant liquidity complementarities, or economies of scale in liquidity use associated with the combination of the payment streams from the three systems. On average, across the treatments and the days, combining the payment streams results in 20% reduction in liquidity use.

Table 5.2 **Liquidity use, delay indicator and value-weighted average time of settlement for the separate systems and for the new system**

JPY billion; hh:mm

	Level (1)	Level (2)	Level (3)
New system			
Liquidity	3,975	9,159	14,344
Delay	0.185	0.041	0.000
Average time	12:22	11:38	11:26
Current three systems			
Liquidity	5,649	11,032	16,415
Delay	0.173	0.042	0.000
Average time	12:17	11:39	11:26
Current BOJ-NET			
Liquidity	3,850	7,760	11,670
Delay	0.274	0.042	0.000
Average time	12:56	11:39	11:34
Two private systems			
Liquidity	1,799	3,272	4,745
Delay	0.058	0.007	0.000
Average time	11:34	11:18	11:16

Source: Authors' calculation.

Note: Level (1) endows participants with sufficient liquidity only to settle their multilateral net debit, Level (2) with liquidity that is halfway between the level sufficient to settle payments without delay and the level of the multilateral net debits, and Level (3) with sufficient liquidity to settle payments without delay.

[7] Specific definitions of these indicators are described in Appendix 1.

It is an interesting feature of the system that the current BOJ-NET requires less liquidity than the new system to process its payments without delay, but requires almost the same level of liquidity as the new system to settle its payments on a multilateral net basis. This suggests that as some of FXYCS and large-value Zengin payments arrive later in the day, they offset with some current BOJ-NET payments that arrive earlier in the day but still remain in the queue. As the current BOJ-NET payments are settled with a slight delay, they settle with less liquidity when combined with payment streams from the other two systems. Again, this indicates particularly strong liquidity complementarities among the systems. It should also be noted that while the combined payments settle without delay using more liquidity, a close examination of Table 5.2 shows that the new system settles at an earlier hour of the day than the current BOJ-NET where participants are endowed with sufficient liquidity to settle payments without delay.

5.4 Optimising funding levels

The funding levels in the new accounts will be determined by a choice of participants. In general, the higher the funding levels, the greater a proportion of those payments that are submitted to the new accounts can be settled. In addition, the higher the funding levels, the more quickly settlements will occur.

A feature of the new system is that funding for the new accounts can be supplied from the standard accounts at any time of the day. To some degree, this option simplifies the problem for participants regarding the amount of funding to transfer to the new accounts at the start of the processing day as any shortfalls or overages in funding can be corrected during the day.

When designing a payment system that uses a liquidity-saving mode of operations as well as a pure RTGS mode of operations, one question designers face is whether to create another account, as in the BOJ-NET's new accounts. One choice is simply to rely on a single account and have participants decide on the priority of the payment, in other words, decide whether to send the payment instruction in a pure RTGS or in a liquidity-saving mode. The liquidity-saving mode then relies on incoming funds over a period of time as well as offsetting. Such a choice is described by Johnson, McAndrews, and Soramäki (2004). In the case of the new system, the computational requirements

of BOJ-NET are reduced considerably with the introduction of the new accounts.

The efficiency of the new system could potentially be negatively affected if participants were to transfer funds into and out of their new accounts often during the day. The multilateral offsetting algorithm, for example, might not find many payments that can be settled if some participants had withdrawn funds immediately prior to operations of the algorithm. Because of this potential negative effect of rapid changes in funding levels, it may be useful to conduct the following thought experiment. Suppose, contrary to the design of the new system, that participants could only fund their new accounts twice during the day, at the opening of the processing day and for settlement of their unsettled queued payment instructions at 16:00. Under that counterfactual assumption, what would be efficient levels of initial funding?

Higher levels of initial funding will be associated with a faster rate of intraday settlement and a higher proportion of payments settled prior to 16:00. There is, however, no clear answer to the question of how to value an increased rate of intraday settlement as there is no easily observable intraday rate of interest that would provide a benchmark level of benefits from a faster rate of intraday settlement and a benchmark level of costs of intraday funds. Similarly, there is no clear measure of increases in credit and liquidity risks caused by leaving more payments unsettled until 16:00.

In the following exercises we investigate levels of initial funding that are sufficiently high so as to quicken the overall settlement of large-value payments in Japan. In addition, we investigate funding levels high enough to assure that a level of unsettled payments at 16:00 is no greater than it is in today's large-value payment systems.

Consider the following problem.

$$\min \sum_i b_i \text{, subject to } \{P_{ij}\},\ \forall i, j;\ i \neq j$$

$$b_i \geq 0$$

$$\sum_{t=t_k}^{t_k+h} \sum_i \sum_j s_{ij}^t \geq S,\ \forall 0 \leq k \leq \bar{k},\ \bar{h} > h > 0.$$

It seeks to minimise the sum of initial balances of each participant i in the new account (b_i), under the constraints that a set of payments that day is fixed and given by P_{ij}, that the balances are non-negative and that settlement (in a value term) under the new system procedures over a given time interval during processing is at least as high as a rate

of settlement S, where S is some yen-rate of settlement per h minutes of the day.

By examining the structure of the problem, we can infer that the optimal levels of initial balances satisfy the following 'equalisation of marginal benefit condition'. An extra yen added to any participant's initial balance has the same incremental effect on the total settlement as an extra yen added to any other participant's initial balance. We can infer that because the variables of initial balances enter the objective function in an additively separable way, there cannot be any way, at the optimal level of balances, to shift balances among accounts (holding fixed the sum of balances) and increase a rate of settlement. Otherwise we could reduce the sum of balances from the minimum level, which contradicts that the level is at a minimum. From that, it must then be the case that an extra yen of initial balances increases a rate of settlement by the same amount regardless of into whose account that yen is added.

The problem outlined above is not fully specified as it does not contain full richness and complexity of the settlement algorithms used by the new system. Nonetheless, an examination of the problem clarifies the heuristic strategy we employ in seeking the efficient levels of initial funding for the new accounts. First, notice that a rate of settlement is specified as the sum of all payments settled. The goal is therefore not to increase a particular participant's rate of settlement but to increase a rate of settlement for the whole system. Second, the problem seeks to minimise the sum of initial balances, not any participant's initial balance. Thus the efficient levels of funding we discuss are characterised by the following three factors: the total level of funding, the distribution of balances across participants, and the timing of funding.

5.5 Simulations and results

To find a locally optimum distribution of balances using simulations on historical data would require a large number of simulations. It is rational that we rely on that feature of the optimum levels of initial balances to guide the following heuristic strategy to characterise the efficient levels of balances. We first simulate the working of the new system starting with various levels of initial balances. After each simulation we examine the performance of the system in terms of the value of payments settled prior to 16:00, the value of the remaining unsettled payments at that time, the value of additional amounts that

need to be paid in to settle all the remaining unsettled payments, and the value-weighted average time of settlement. We also examine the effects of alternative levels of balances on the system as a whole, and on a separate basis, for the five largest banks and all the other participants. We then investigate the intertemporal distribution of balances as we seek a local optimum distribution of balances.

The results of these simulations give participants and planners a sense of how the alternative levels of balances would affect the system's performance.

5.5.1 Four baseline simulations

We perform simulations using the ten days of historical data in September 2003. We conduct four sets of baseline simulations. The first scenario is to simulate the performance of the current situation in which BOJ-NET, FXYCS and the large-value Zengin independently operate as they operate now. The scenario endows participants with sufficient liquidity to settle their payments without delay (although it treats FXYCS and Zengin as simple DNS systems) and uses the time of entry of payments. As a result, these baseline simulations provide a measure of current liquidity usage in the systems. These simulations are referred to as *current baseline* simulations.

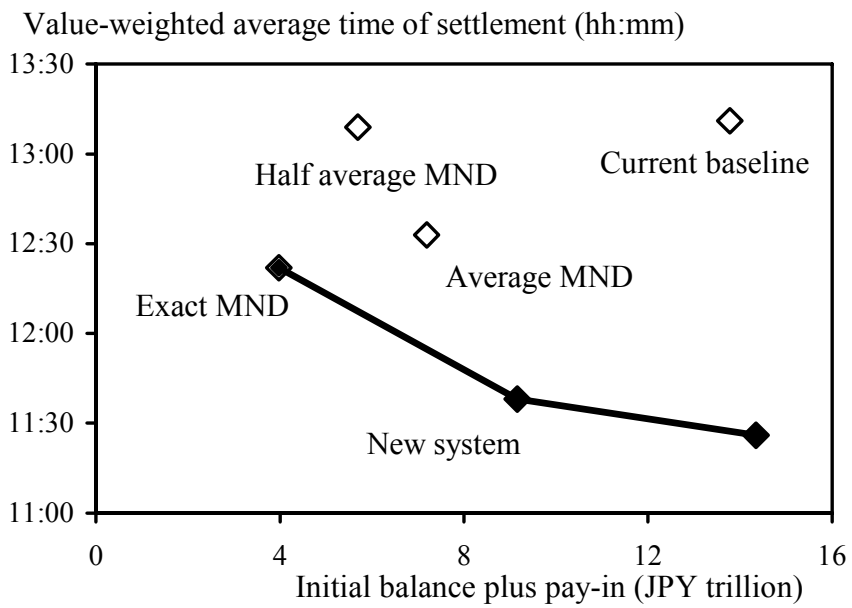
Another baseline simulation is to endow participants with the exact amount of funds (in the new accounts) equal to that day's multilateral net debit of each participant, given that day's payments history. A participant's multilateral net debit is the amount it would owe to settle its payments if the system were a DNS system. In general, participants do not necessarily know their own multilateral net debits in advance. This scenario can be thought of approximating the case in which participants make pay-ins throughout the day as they gradually learn the exact size of their multilateral net debit. The multilateral offsetting operations may be one way participants do learn the amount of their multilateral net debits, and this scenario approximates the learning process by assuming that they know the amounts with certainty in advance. These simulations are referred to as *exact multilateral net debit (MND) funding* simulations or *progress-payment approximation* simulations.

The third baseline simulation endows participants with their average multilateral net debit funding, where the average is taken over the ten days of the sample period. This scenario is first to assume that participants fund their new accounts in the morning and then make another pay-ins to the new accounts after 16:00 to settle the payments

that remain unsettled at that time. The average multilateral net debit is, of course, quite close in size to the exact multilateral net debit amount used in the *exact MND funding* simulations. However, because it is an average, some payments on some days will remain unsettled at 16:00. These simulations are referred as *average multilateral net debit (MND) funding* simulations.

The fourth baseline simulation endows participants with half the amount of funding as in the *average MND funding* simulations. These simulations are referred as *half average multilateral net debit (MND) funding* simulations.

Figure 5.2 **Overview of the performance of the new system**



Source: Authors' calculation.

Figure 5.2 summarises the performance of the new system described in Section 5.3 and of these four baseline simulations. Points in the lower-left corner of the figure are more desirable combinations of total balances and settlement time. It can be found that conducting these baseline simulations attempts to search the local optimum level around the point at which participants are endowed with sufficient liquidity only to settle their multilateral net debits.

Table 5.3 shows the performance of these four baseline simulations on average across the ten days of the sample period with regard to the amounts of initial balances used in the simulations, the additional amounts of pay-ins to the new accounts that would be required after 16:00 to settle those payments that still remain unsettled at that time, the cumulative amounts settled by 16:00, the gross amounts unsettled at 16:00, and the value-weighted average time of settlement. Because the analysis of only ten days yields a small sample, we simply examine averages without considering the statistical significance.

Table 5.3 **Averages from the baseline simulations**

JPY billion; hh:mm

	Initial balances	Five LBs' balances	End-of-day pay-ins	Cumulative value settled at 16:00	Gross value unsettled at 16:00	Average time of settlement
Current baseline	13,780	3,460	0	56,673	12,625	13:11
	(-)	(-)	(-)	(-)	(-)	
Exact MND	3,975	492	0	61,106	8,192	12:22
	(0.288)	(0.142)	(-)	(1,078)	(0.649)	
Average MND	3,964	492	3,224	55,954	13,344	12:33
	(0.288)	(0.142)	(-)	(0.987)	(1.057)	
Half average MND	1,982	246	3,712	48,119	21,180	13:09
	(0.144)	(0.071)	(-)	(0.849)	(1.678)	

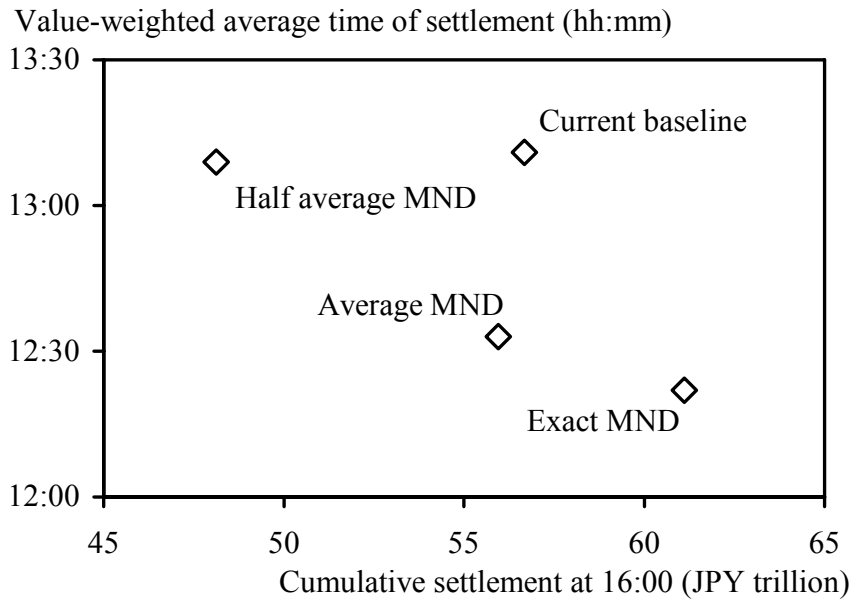
Source: Authors' calculation.
Note: Figures in brackets are ratios of each item to that of the *current baseline* simulations. 'Five LBs' stands for five largest banks.

The exact *MND funding* simulation clearly settles more payments by 16:00 with the initial balances as small as one-third of those the *current baseline* simulation requires. The *average MND funding* simulation also has the same qualitative results relative to the *current baseline* simulation, using fewer initial balances than the *current baseline* simulation. The *average MND funding* simulation results that payments unsettled at 16:00 reach up about 20% of that day's total payments. These payments would be settled with an additional pay-in of JPY 3.2 trillion, so that the total liquidity used in these simulations is about twice as high as in the *exact MND funding* simulation. The amounts settled by 16:00 in the *half average MND funding* simulation are far below those in the three other scenarios, though economising too much of initial balances. The *half average MND funding* simulation settles on average only slightly more quickly than the *current baseline* simulation, using much less liquidity than the *current baseline* simulation. Because of its larger pay-in after 16:00, the *half*

average MND funding simulation uses almost as much liquidity in total as the *average MND funding* simulation.

Figure 5.3 shows the value-weighted average time of settlement and the cumulative settlement by 16:00 for the various cases. The settlement performance improves as the outcome plotted on the figure moves toward the bottom right, meaning a larger value settled in a quicker manner and vice versa. The four scenarios can be roughly arranged in the desirable order as the *exact MND funding* simulation, the *average MND funding* simulation, the *current baseline* simulation and the *half average MND funding* simulation.[8]

Figure 5.3 **Value-weighted average time of settlement and total value settled by 16:00**



Source: Authors' calculation.

Overall, the *exact MND funding* simulation settles payments most quickly and extensively and uses less liquidity than the *average MND funding* simulation. This suggests that if participants were to make pay-ins during the day in line with their multilateral net debit

[8] The *current baseline* simulation may be better than the *average MND funding* simulation, depending on the shape of indifference curves assumed. For example, the former improves if a high preference is given to settlement completion by 16:00.

positions, they might be able to have fewer payments unsettled after 16:00. In comparing the performance of the *average MND funding* simulation and the *half average MND funding* simulation, the latter settles fewer payments by 16:00 and has a later average time of settlement (although it also settles payments more quickly than the *current baseline* simulation on average). It has approximately 25% of the payments unsettled at 16:00. Settlement of these payments requires an additional pay-in of JPY 3.7 trillion. The *half average MND funding* simulation, after all, uses about 80% of liquidity used in the *average MND funding* simulation, after taking into account the large pay-ins at the end of the day. This result reminds one that as one limits the initial amount of liquidity available to the system, larger pay-ins will be required later in the day.

The results of these four baseline simulations suggest that the new system may perform quite satisfactorily with levels of liquidity that are significantly lower than those currently used in settlement of the three systems. In addition, the behaviour of a rough approximation to the progress payments suggests that participants may be better able to conserve funding by making pay-ins to the system during the day as they learn the multilateral net debit resulting from that day's payments.

5.5.2 Distributional funding simulations

As the results of the *exact MND funding* and *average MND funding* simulations have suggested, the different distribution of initial balances across participants leads to the different performance of intraday settlement even when the total balances in the system are the same.

It is well known that there are a few hub-like participants in Japan's interbank payment network.[9] They play a significant role in the redistribution of liquidity in the system by making outgoing payments and receiving incoming payments continuously during the day. Therefore the malfunctioning of these hub-like participants potentially has negative effects on the performance of the system as a whole.

In this section, in addition to the baseline simulations, we perform some additional simulations that show the effects of small changes in

[9] For the structure of Japan's payment network, see Inaoka et al (2004) and Bank of Japan (2006a).

the funding provided by the five largest banks, which are known to work as hub-like participants in BOJ-NET. These simulations are conducted with the other participants in the system being endowed first with the exact multilateral net debit funding and, for the second set of these simulations, with half that level of funding. Because those participants are endowed with the exact amount of their multilateral net debit, these simulations are probably best compared with the *exact MND funding* simulation. The amounts that the five largest banks are endowed with are quite small amounts equal to the 90th percentile of the size of the payments they each send and receive on the current BOJ-NET alone. So these simulations are indicative of a situation in which all but the five largest banks make regular progress payments in the amounts of their multilateral net debits, and the five largest banks supply very little in the initial funding amounts. These simulations are not meant to model the actual behaviour of participants but rather to investigate the possible behaviour of the new system as we vary the funding of some particular participants in different ways.

These simulations are quite illustrative of the effects of small changes in particular participants' funding levels. To investigate these effects for individual participants would be quite time-consuming and require many simulations. Because of those resource requirements, we forego such an investigation in the paper.

The first set of simulations shows that reducing the five largest banks' total funding from JPY 492 billion, as in the *exact MND funding* simulation, to JPY 18 billion does not substantially reduce the speed of settlement in the system (see Table 5.4). The value-weighted average time of settlement changes from 12:22 to 12:34. Nor is the total amount settled by 16:00 reduced appreciably, even though the largest five banks had multilateral net debits of approximately JPY 500 billion on the sample days. These results show that individual participants, or even groups of participants, may significantly reduce their initial level of funding without necessarily causing proportional changes in the amounts settled. Note that these results come at the cost of large amount of end-of-day pay-ins. Further research could determine the local optimum in the initial funding amounts.

Table 5.4 **Averages from the simulations with the 90th percentile funding**

JPY billion; hh:mm

	Initial balances	Five LBs' balances	End-of-day pay-ins	Cumulative value settled at 16:00	Gross value at 16:00	Average time of settlement
(1) Exact MND	3,975	492	0	61,106	8,192	12:22
	(-)	(-)	(-)	(-)	(-)	(-)
+90 percentile	3,500	18	1,527	58,170	11,129	12:34
	(-475)	(-474)	(+1,527)	(-2,936)	(+2,937)	(+0:12)
+90 percentile*2	3,518	35	1,452	58,495	10,803	12:34
	(-457)	(-457)	(+1,452)	(-2,611)	(+2,611)	(+0:12)
+90 percentile*3	3,535	53	1,405	59,025	10,274	12:33
	(-440)	(-439)	(+1,405)	(-2,081)	(+2,082)	(+0:11)
(2) Average MND	3,964	492	3,224	55,954	13,344	12:33
	(-)	(-)	(-)	(-)	(-)	(-)
+90 percentile	3,490	18	3,398	54,172	15,128	12:43
	(-474)	(-474)	(+174)	(-1,782)	(+1,784)	(+0:10)
+90 percentile*2	3,507	35	3,371	54,056	15,243	12:42
	(-457)	(-457)	(+147)	(-1,898)	(+1,899)	(+0:09)
+90 percentile*3	3,525	53	3,366	54,621	14,678	12:41
	(-439)	(-439)	(+142)	(-1,333)	(+1,334)	(+0:08)
(3) Half average MND	1,982	246	3,712	48,119	21,180	13:09
	(-)	(-)	(-)	(-)	(-)	(-)
+90 percentile	1,754	18	3,756	46,017	23,282	13:19
	(-228)	(-228)	(+44)	(-2,102)	(+2,102)	(+0:10)
+90 percentile*2	1,772	35	3,724	46,350	22,948	13:18
	(-210)	(-211)	(+12)	(-1,769)	(+1,768)	(+0:09)
+90 percentile*3	1,789	53	3,720	46,494	22,804	13:17
	(-193)	(-193)	(+8)	(-1,625)	(+1,624)	(+0:08)

Source: Authors' calculation.
Note: 'Five LBs' stands for five largest banks. Figures in brackets are differences from the benchmark level of each sub-scenario.

The second set of simulations endows all but the largest five banks with their average multilateral net debit amounts, as in the *average MND funding* simulations (see Table 5.4). The largest five banks are again endowed with an amount that is equal to the size of the payment that is at the 90th percentile of their payment size distribution on the current BOJ-NET alone. In this simulation, which is best compared with the *average MND funding* simulations, we see that the performance of the system remains quite good even though the largest five banks' funding levels are reduced substantially. The amounts settled by 16:00 falls by only 3%, and the value-weighted average time of settlement occurs 10 minutes later.

A final set of these simulations, in which participants other than the largest five banks have their initial funding levels set at half of the day's multilateral net debit, confirms the result that dramatically

reducing the funding levels of the largest five banks does not reduce settlement by that proportion (see Table 5.4).

In each set of the simulations just discussed, we vary the funding levels of the five largest banks by endowing them with multiples of JPY 18 billion, namely 35 (doubled) and 53 (tripled) for their initial balances. These increases in the levels of initial balances do not appreciably change the outcome. One reason is that liquidity-saving features effectively reduce some distortions from optimal balances by running offsetting mechanisms continuously during the course of the day. Offsetting mechanisms can relax conditions for gross settlement in comparison with a pure RTGS mode and then achieve relatively smoother flow of payments despite the distortions of initial distribution of balances.

In general, there tends to be a greater amount settled as the initial funding levels of the largest five banks increases, but this is not always true. For example, raising the largest five banks' initial funding from JPY 18 billion to 35 slightly reduces the amounts settled by 16:00 in the second set of simulations. This result implies that the amount settled by 16:00 is not a monotone increasing function of some particular participants' initial balances.

5.5.3 Progress-payment simulations

The *exact MND funding* simulation has endowed participants with the exact amounts of the multilateral net debit at the beginning of the processing day. This simulation can also approximate the case in which participants make pay-ins continuously during the day as they learn the size of their multilateral net debit in that day. The question is how the performance in the system can be affected if the timing of intraday pay-ins is changed.

It has been already described that the *half average MND funding* simulation substantially underperforms the *exact MND funding* simulation because of the severe liquidity constraints in the system. In the progress-payment simulations, starting with the half average multilateral net debits and then making intraday pay-ins at 10:00 or 12:00, both the value settled by 16:00 and average time of settlement can approach those of the *exact MND funding* simulation (see Table 5.5). The high performance of the progress-payment simulations with intraday pay-ins comes at the cost of twice as large amount of the total liquidity in the *exact MND funding* simulation.