Designing large value payment systems: An agent-based approach

Amadeo Alentorn[*] *Sheri Markose*[*], *Stephen Millard*[#] *and Jing Yang*[#],

[*] University of Essex , Economics Department and Centre For Computational Finance and Economic Agents (CCFEA)
[#] Bank of England
E-mail: stephen.millard@bankofengland.co.uk
jing.yang@bankofengland.co.uk

1st Draft February 2004
This draft December 2004

ISSN 1368-5562

Contents

Abstract

In this paper, we report on the main building blocks of an ongoing project to develop a computational agent-based simulator for a generic real-time large-value interbank payment system with a central processor that can implement different rules for payment settlement. The main types of payment system in their polar forms are Real Time Gross Settlement (RTGS) and Deferred Net Settlement (DNS). DNS generates large quantities of settlement risk; in contrast, the elimination of settlement risk in RTGS comes with excessive demands for liquidity on banks. This could lead them to adopt various delaying tactics to minimise liquidity needs with free-riding and other 'bad' equilibria as potential outcomes. The introduction of hybrid systems with real-time netting is viewed as a means by which liquidity costs can be reduced while settlement risk is unchanged. Proposed reforms for settlement rules make it imperative to have a methodology to assess the efficiency of the different variants along three dimensions: the cost of liquidity to the individual banks and the system as a whole, settlement risk at both bank and system levels, and how early in the day payments are processed, since this proxies the impact of an operational incident. In this paper, we build a simulator for interbank payments capable of handling real time payment records along with autonomous bank behaviour and show that it can be used to evaluate different payment system designs against these three criteria.

Keywords: Real Time Gross Settlement; Deferred Net Settlement; Agent-based simulation; Payment Concentration; Liquidity; Systemic Risk

Summary

The smooth functioning of payment systems is clearly important for financial stability and, given this, the Central Bank should have an understanding of the risks associated with different systems and seek to minimise these. In this paper we report on the main building blocks of an ongoing project whose main objective is to develop a computational agent-based simulator of a generic real-time large-value interbank payment system with a central processor that can implement different rules for payment settlement. In order to assess the risks associated with different systems, we require such a simulator to be able to carry out experiments such that we can assess the speed of processing, – a proxy for operational risk since the effect of an operational incident will be larger the more payments remaining to be processed when the incident occurs – the liquidity required for the system to operate smoothly and the credit risk that arises from a settlement failure by one or more of the banks.

In our simulator, we model banks as agents, capable of a degree of autonomy with which to respond to system rules and adopt a strategy that determines when to send payment orders to the central processor and what priority to attach to each payment. An interbank payment system with costly liquidity requires banks to solve an intraday cash management problem, minimising their liquidity and delay costs. In this paper, we assume that banks use simple rules of thumb to do this.

As an example of how our simulator can be used, we report on a set of experiments that throw some light on the relative merits of two variants of an RTGS system: one in which banks generate all the liquidity they need to use the system by posting collateral with the Central Bank at the beginning of the day – opening liquidity (OL) – and one in which they generate liquidity, by borrowing from the Central Bank, as and when they need it – just in time (JIT). In the JIT system banks weigh up the costs of delaying a payment against the interest they would need to pay on a loan from the central bank in order to determine when liquidity is used. The performance of these systems in terms of the liquidity-delay trade off is evaluated when banks do not reorder payments requests and follow a first in first out (FIFO) rule and when they follow a delay strategy where small payments are settled first unless overridden by a priority cost.

At the level of individual banks, an attempt to handle the trade-off between liquidity and delay costs may result in behaviour where banks delay settlement in anticipation that other banks will make payments to them first. Further, it appears to be intuitive that at an individual level larger payments should be delayed if their delay costs are not too high. We found that though this appears to be an individually rational response at the level of banks, it leads to a deterioration in the collective performance of the RTGS system whether in the OL case or the JIT variant. However, we found that the JIT system is more prone to rapid deterioration of its liquidity recycling capabilities than the OL system. In fact, a key message of our experiments is that, at any given level of liquidity, the JIT system would generate more delayed payments than an otherwise identical system in which banks posted their liquidity at the beginning of the day, and this would be bad from an operational risk point of view.

However, there are some important caveats that need to be born in mind with our results. (The purpose of the experiment was more to demonstrate what can be done with the IPSS.) In particular, our experiments relied on one particular stochastic simulation of data based on one day's worth of actual payments. To get a clearer picture one would need to run multiple simulations based on data from a large number of days. In addition, our experiments on the OL system imposed a given level of opening liquidity. In order to get a real understanding of how such systems work, one would need to postulate behavioural rules to explain the decision of how much liquidity to post at opening. Such a rule would relate liquidity posted to such variables as the opportunity cost of collateral, the average delay costs expected by the bank, the total expected value and volume of payments coming in or going out and the uncertainty surrounding these. For the JIT system, a possible extension might be to consider behavioural rules that allow banks to take into account the possibility of using liquidity from incoming payments to make their own future payments when they choose how long to delay payments.

1 Introduction

The interbank flow of large-value payments increased substantially in the 1980's and 1990's as a result of financial innovation and the deregulation and globalisation of financial markets. On a daily basis it has been estimated that close to 20% of a country's GDP typically comes up for settlement in the interbank payment networks of each of the G10 countries.[1] Given that the smooth functioning of payment systems is clearly important for financial stability, the Central Bank should have an understanding of the risks associated with different systems and seek to minimise these. Bank of England (2000) discusses four types of risk in payments systems:

- Credit risk: the risk that a bank will not actually meet a payment obligation incurred by it either when the obligation is due or at a later stage
- Liquidity risk: the risk that a bank won't meet an obligation at the time it is due, although it will at some point thereafter (as a result of being 'short of liquidity')
- Operational risk: the risk that the system breaks down or fails to function and this results in possible financial losses
- Legal risk: the risk that unexpected legal decisions or legal uncertainty more generally will leave the system or its members with unforeseen obligations and possible losses

The primary objective of this project is to develop a computational agent-based simulator of a generic real-time large-value interbank payment system with a central processor that can implement different rules for payment settlement. For short, the simulator will be referred to as the Interbank Payment System Simulator (IPSS). We would expect such a simulator to be able to carry out experiments such that we can assess the risks associated with different systems along three dimensions:

(i) Operational risk, which will be inversely related to the speed of processing – measured from the time of initiation by the customer of payments and their final settlement – since a given operational event will have a larger impact the more payments remain to be settled that day
(ii) Liquidity risk, which will be greater the greater is the liquidity needed by the individual banks and the system as a whole
(iii) The credit risk that arises from a settlement failure by one or more of the banks.

Banks, modelled as agents, are capable of a degree of autonomy with which to respond to system rules and adopt a strategy that determines when to send payment orders to the central processor and what priority to attach to each payment. An interbank payment system with costly liquidity requires banks to solve an intraday cash management problem, minimising their liquidity and delay costs. In this paper, we assume that banks use simple rules of thumb and leave adaptive learning for future work.

Attempts at comparative analyses of settlement rules in the literature so far have been hampered in a number of ways. A number of theoretical papers (Angelini 1998, Bech and Garratt, 2003,

[1] In the United Kingdom, CHAPS processes some £200 billion per day with a transactions volume of about 100,000; the resulting average value per transaction is about £2 million.

Willison, 2003) have used a game-theoretic perspective to understand the differences in incentives for the banks created by different credit and settlement arrangements in interbank payments. These papers are insightful and give qualitative suggestions on design issues. However, they cannot address the above three design objectives in the quantitative fashion that is needed for a realistic comparison of different 'real-life' interbank payment systems. Further, these authors typically make a number of simplifying assumptions that are not innocuous. For instance, most assume that banks are of equal size and know in advance what payments are coming in to them over the course of the day.

Leinonen and Soramäki (1999) use a simulator developed by the Bank of Finland (BoF PSS for short) to examine different hybrid systems that combine the advantages of netting in multilateral or bilateral form with real time settlement processing. Koponen and Soramäki (1998) and Bech and Soramäki (2002) take the experiments with the BoF PSS a step further. They allow the banks to post varying amounts of liquidity at opening and take the recorded time that payments are submitted to the central processor as being identical to the time of arrival of requests at the bank. The simulator first determines at each minute the settlement balances of each bank and then operates an automatic system of settlement which follows a first in first out rule. The delays in payment settlement are evaluated at different levels of opening liquidity. The BoF simulator can also simultaneously operate a bilateral or multilateral gridlock resolution algorithm based on a real time netting framework for non-settled payments with the condition that settlement reserves for each bank at no time become negative. The delay in settlement is measured by the difference between the time a payment request arrives at the central processor and the time of settlement. The trade-offs between liquidity and delay are compared with and without the hybrid gridlock resolution algorithms.

In many ways, this approach is similar to ours. However, because the data put into the BoF PSS only contains the times that banks actually submitted payments to the central processor, it cannot deal with the strategic decision of the banks to delay payments or to reorder the time of settlement that departs from a first in first out basis. In addition, it assumes that bank behaviour and network interconnections in the system remain unchanged across experiments. In our work, we allow payments to randomly arrive at the bank prior to the time they were submitted to the central processor and allow the banks to use a simple rule of thumb to decide when to actually submit payments. Furthermore, it is possible to run 'stochastic simulations' with our simulator while the BoF PSS is only able to run deterministic simulations based on actual data fed into it. Stochastic simulations enable the experimenter to vary the statistical properties of the interbank system in terms of the size, arrival times of payment requests and distribution of the payment flows in the interbank system. For instance, in the IPSS simulator, a menu-driven command converts the interbank system to one of perfect symmetry with identical banks making the same number of equal-sized payments to one other. The implications of this for liquidity requirements and systemic risk can then be contrasted with more realistic asymmetric structures of the interbank system.

As an example of how our simulator can be used, we report on a set of experiments that throw some light on the relative merits of two variants of an RTGS system: one in which banks generate all the liquidity they need to use the system by posting collateral with the Central Bank

at the beginning of the day – opening liquidity (OL) – and one in which they generate liquidity, by borrowing from the Central Bank, as and when they need it – just in time (JIT). In the JIT system banks weigh up the costs of delaying a payment against the interest they would need to pay on a loan from the central bank in order to determine when liquidity is used. The performance of these systems in terms of the liquidity-delay trade off is evaluated when banks do not reorder payments requests and follow a first in first out (FIFO) rule and when they follow a delay strategy where small payments are settled first unless overridden by a priority cost.

The rest of the paper is organized as follows. In Section 2 we discuss the main issues in payment system design. In Section 3 we set out the computational modelling framework and discuss the features of IPSS Simulator. In addition, we discuss the scope for behavioural rules for banks to respond to liquidity and delay costs. In Sections 4 and 5 the results of our example experiment are reported. Section 6 concludes.

2 Payment Systems in Practice

Historically, interbank payments following the clearing house tradition for paper based IOUs such as cheques have involved central processing with multilateral net settlement at the end of the day. Such end-of-day netting systems were the norm when the process of transmitting payments was expensive and the physicality of the IOUs militated against real-time settlement. But these Deferred Net Settlement (DNS) systems can generate large intraday credit exposures. Notification by payer banks of payment requests to customers of payee banks, result in the latter processing payments, granting *de facto* credit extensions to the initiating/payer banks until final settlement occurs at the end of day. Further, in a DNS system, as banks treat the promised inflows with a substantial degree of finality they make no explicit arrangements for any liquidity in excess of the end of day multilateral netted amount. Thus, as the size and volume of payments grew larger, the corresponding increase of risk of non-settlement by payer banks in DNS led to the introduction of Real Time Gross Settlement (RTGS) systems in the 1990's by all EU and G10 countries (with the exception of Canada). In a RTGS system all payment requests arriving at the central processor are processed individually, with immediacy and finality using the balance in a bank's settlement account. Those payments that do not satisfy the criteria set out by the rules are returned to the sender.[(2)] While settlement risk can in principle be eliminated completely by RTGS such systems require large quantities of intraday liquidity. As an alternative, banks can use payment inflows to finance subsequent outflows, strategically delaying the settlement of payments requested in anticipation of offset. Such delaying tactics, which result in hidden queues of unsettled payments, though individually rational, can result in free riding and 'bad' equilibrium outcomes which can compromise the efficiency and the capacity of RTGS to be free of settlement risk (see, for example, Angelini, 1998, or Bech and Garratt, 2003).

The intraday liquidity needed in a payment system is a non-trivial function of the random arrival times of payments as well as the size and distribution of payments among banks. For a DNS system, if every bank is assumed to owe every other bank the same value of payments, with multilateral or bilateral netting done at end of day, the liquidity needed will be zero. Given the

[(2)] For a fuller discussion of the different variants of the RTGS in practice, see McAndrews and Trundle (2001). McAndrews and Rajan (2000) discuss the settlement process in Fedwire.

lack of symmetry in the value (and volume) of real world interbank payment flows, the end of day multilateral netted amount is positive and typically about 2.5% - 3 % of the total value of payments. This amount, calculated by a multilateral netting algorithm on the end of day liability matrix of the interbank settlement system, is generally referred to as the 'lower bound' level of liquidity needed by a payment system. The 'upper bound' level of liquidity needed by a payment system, on the other hand, is equal to the amount of liquidity that banks have to post on a just in time basis so that all payment requests are settled immediately at the time they are requested with no payments being queued. This value is typically far less than the total value of payments processed in the system and reflects the speed with which liquidity is recycled. It has been found (see, for example, James, 2003, or Bech and Soramäki, 2002) that because of a combination of regulatory requirements and the low opportunity cost of collateral for banks, the UK and the European systems are liquidity rich, with the upper bounds being far less than the liquidity that banks actually post.

However, one can imagine situations in which banks post less than their upper bound level of liquidity. In this case, the possibility arises of payment gridlock. Bech and Soramäki (2002) define a gridlock as one where the (possibly hidden) queues of payment requests of banks can be eliminated if they can be simultaneously netted with no *additional* posting of liquidity.[3]

The relative advantages of DNS and RTGS systems have led to the recent development of hybrid systems that combine the bilateral or multilateral netting features of DNS with RTGS to reduce liquidity requirements and to rid RTGS of the potential for free riding and gridlocks. The maximum benefits from a hybrid system arise when the system operates at the lower bound levels of liquidity, all payments are cleared in full at the time they are made, and all payments are made early in the day. (See Willison (2004) for a discussion of this.)

It is worth commenting at this point on how liquidity is generated within a payment system. In many RTGS systems liquidity is obtained by banks posting collateral with the Central Bank and receiving cash on their settlement account at the beginning of the day. At the end of the day, the Central Bank returns this collateral to the settlement banks. This process is equivalent to the Central Bank giving the settlement banks fully collateralised but otherwise free loans intraday. One could think of alternatives to this approach. In particular, the Central Bank could charge an interest rate on these loans and/or could provide these loans on an uncollateralised basis. In the case of uncollateralised loans, this could work through the provision of overdraft facilities at the Central Bank.

Other features of a payment system that will be of interest to us include the types of payments that are made through it. For instance, we can distinguish between systems that process wholesale financial market transactions, such as CHAPS Sterling, and those that process retail payments, such as BACS or credit card schemes. Related to this will be the issue of size of payments. The wholesale systems are likely to process fewer payments but with much higher values than the retail systems. Finally, we could also consider the extent to which banks access

[3] Situations in which additional liquidity is needed to assist in the elimination of payment queues with simultaneous or multilateral netting are referred to in Bech and Soramäki (2002) as *deadlock*.

the system. In the United Kingdom, the payment systems are highly tiered. That is, there are a small number of settlement banks that process payments not only on their own account but on behalf of a large number of other banks that do not have settlement accounts.

In the model that we construct, we focus on large-value payment systems. But we make the model general enough that it can handle different rules as to how liquidity is obtained and on what terms. In particular, our example experiment compares two large-value payment systems:

1) Banks obtain liquidity by posting collateral at the start of the day; this liquidity is provided free of charge by the Central Bank – OL
2) Banks obtain liquidity as and when they need it by borrowing uncollateralised from the Central Bank at a cost – JIT

3 The Interbank Payment System Simulator (IPSS)

The structure of the IPSS is as follows. We have a central bank that operates the central processor of the payment system. Then there are a set of settlement banks that have direct access to the payment system and have settlement accounts at the central bank. All other banks and non-banks have indirect access to the payment system via correspondent or other relationships with the settlement banks; these all submit payment requests at random times. Settlement banks can also initiate payments on their own account. For the rest of this paper we use the term 'banks' to refer only to IPSS settlement banks.

In principle, there are three arrival times that need to be stipulated for payments: let t_R be the time when the customer of bank i has made the request for a payment, t_C be the time the payment request arrives at the central processor where it is either settled immediately or put in the central queue (if this facility exists), and t_E be the time when the system settles the payment with finality.

Let $X^{ij}_{t_R}$ denote a payment request made by a customer from bank *i* to bank *j* at time t_R and is known only to bank *i*. Let $X^{ij}_{t_C}$ denote a payment from bank *i* that has been submitted to the central processor for settlement and hence is known to the central bank and to bank *j*. In the absence of the facility of central queues, the time between t_C and t_E would be zero as only payments capable of being executed are submitted to the central processor. On the other hand for those payments that were requested at time t_R but are forwarded for final execution at $t_C = t_E$, $t_C > t_R$, banks have effectively maintained 'hidden queues' denoted by $\boldsymbol{X}_i^{HQ}(0, t)$, which is a vector of time stamped non-settled payment requests being held at each bank *i*.

If the opening time is t=0, then a bank's settlement account balance at the central bank at t-1 is denoted by B_{it-1}

$$B_{i,t-1} = LP_{i,0} + \sum_{s=1}^{t}\sum_{j} X_{t_C - s}^{\ ji} - \sum_{s=1}^{t}\sum_{j} X_{t_C - s}^{\ ij} \qquad \textbf{(1)}$$

$LP_{i,0}$ denotes liquidity posted at opening by bank *i*; the second term is the sum of all payments made to bank *i* by all other banks *j*; and the third term is bank *i*'s payments to all other banks.

3.1 Process for arrival of payments

The IPSS is capable of handling the full record of intraday payments. Using a menu driven process, the payments data arrival process can take the following forms. Form (i) is deterministic whereas forms (ii) and (iii) are stochastic.

(i) The data can come in the form of records of real time payments data at the central processor, say based on a day's worth of CHAPS Sterling transactions data. In the simulator, experiments in this category are denoted 'Real'. Here it is assumed that payments arrive at the level of banks and are immediately forwarded to the central processor with zero delay, i.e. $t_C = t_R$.

(ii) An alternative that still relies on using real payments data involves allowing the payment requests to arrive at banks prior to the time they were submitted to the central processor. The arrival times are drawn from an identical and independent distribution (iid) random arrival process, $t_R \sim iid$, subject to $t_R < t_C$. In the simulator, experiments in this category are denoted 'IID Real'. Clearly, these payment requests will have the same empirical distribution in terms of volume and value of payments as in (i). However, the arrival time of payment requests at the bank is stochastic and constrained to be before the recorded time of arrival at the central processor. A large number of experiments with the same intraday payments data regarding value and size of banks' inflows and outflows can be useful in determining how banks process their payment requests for settlement.

(iii) A final alternative for the data arrival process is denoted 'Proxied Data'. Proxied data can take a number of forms. The option on the menu denoted 'Equal Banks' produces data involving an equal number and value of payments for each bank with iid random arrival times at the bank. Under the rubric 'Large/Small Banks' we proxy for the asymmetry of the real world payments data. Here, the payments are generated by an iid arrival process as discussed in (ii) above with the volumes and values of payment requests being generated to match the statistical distribution of CHAPS Sterling payments in 2003 set out in Table A. This includes the payment concentration statistics of the CHAPS Sterling interbank system discussed in the next subsection.

Table A: Summary statistics for CHAPS Sterling data

	2002	2003
Average total value of daily payments	£176 billion	£211 billion
Mean payment size	£1.9 million	£2.5 million
Median payment size	£17,000	£16,000
Average total volume of daily payments	95,000	84,725

Source: CHAPSCo

3.2 The concentration of payment activity

Our approach to modelling the concentration of payment activity in the interbank system relies on the Herfindahl Index. In the case of payments this index is given by

$$\mathrm{HI}_{\mathrm{Payments}} = \sum_i \left(\frac{Bank_i\ Payments}{Total\,Value\,of\ Payments} \right)^2 \qquad \textbf{(2)}$$

and measures the concentration of payment activity. James (2003) reports that the Herfindahl Index for CHAPS Sterling is about 0.2 which is what one would observe if payment activity were divided evenly between five to six banks. In general, the Herfindahl Index will lie between 0.5 and 1/*n*, where *n* is the number of banks. It would equal 0.5 when activity is equally divided between only 2 banks while it would equal 1/*n* where payment activity is equally divided between the *n* banks.

In the cases marked as 'Large/Small Banks' in the IPSS, we assume that the simulated payment activity is distributed between the banks in a manner that is also consistent with the payments Herfindahl index reported for CHAPS Sterling by James (2003).

3.3 Strategies for submitting payments

In this sub-section, we define strategies for the banks. In particular, we need to propose rules governing the amount of liquidity they post at the beginning of the day and the order and timing of payment submission to the central queue. In what follows, we assume that the amount of liquidity posted at opening is exogenously given though we will want to change this in future work. Since delaying payments is costly, banks will always settle payments when they have the liquidity to do it. That is, if $B_{i,t-1} > 0$ and greater than any of its payment requests in its hidden queue, $\boldsymbol{X}_{\mathrm{i}}^{\mathrm{HQ}}(0, t)$, the bank will select such payments for settlement without delay on a first-in first-out (FIFO) basis. This is called the 'Automatic FIFO Settlement Rule' and the IPSS implements this as a matter of course. However, it must be noted that, in general, banks have discretion to override the FIFO rule and reorder payments in their queues for submission to the central processor.

Now, a benchmark case is the one where banks promptly despatch payments at time of arrival to the central processor. This requires that banks post additional liquidity if settlement balances defined in equation **(1)** at the central bank are insufficient to make the payment. In IPSS this is denoted 'No strategy' and the total amount of liquidity used by banks under these conditions gives the upper bound of liquidity needed by the interbank system.

From the point of view of minimising costs, banks would operate the zero delay rule in a system in which they faced a zero cost of liquidity. As posting liquidity to process payments forwarded to the central processor typically has an opportunity cost, we also consider an alternative queuing rule for banks to determine when to forward payment requests to the central processor.

We start by noting that the total costs incurred from making a given payment, X, will be the sum of two components: a delay cost and a liquidity cost. We assume that delay costs can be represented by an exponential function of the length of time the payment is delayed, $t_E - t_R$. In addition, we assume that different payments have different priorities with high priority payments carrying a greater delay cost than low priority payments for a given delay. The priority function associated with every incoming payment request $X^{ij}_{t_R}$ from bank i to bank j, denoted by $\beta(X^{ij}_{t_R})$, is assumed to be drawn from a uniform distribution (0,10). That is, the incoming payment requests to banks are randomly assigned to the ten priority bands with 10% probability. Liquidity costs depend solely on the interest rate charged by the Central Bank. For ease, we assume that when banks calculate the expected liquidity cost of making a payment, they assume that any liquidity borrowed from the central bank will only be repaid at the end of the day. The implicit assumption is that any liquidity they obtain from incoming payments carries an opportunity cost in the market equal to the rate charged by the central bank.[4] We should note that this assumption, in effect, rules out free-riding.

Putting all this together suggests that the total cost of making a payment of value X will be given by:

$$Cost = aXe^{\frac{b\beta(t_E - t_R)}{T}} + \frac{iX(T - t_E)}{T} \qquad \textbf{(3)}$$

where T is the number of minutes in a trading day and i is the intraday interest rate. Note that this is expressed as the rate charged for a loan taken out at the beginning of the day and paid back at the end of the day, the rate a bank would need to pay if it borrowed from the central bank in order to post liquidity at the beginning of the day.

Minimising the total cost implies an optimal time to execute the payment that will be given by:

$$t_E^* = t_R + \frac{T}{b\beta}\ln\left(\frac{i}{ab\beta}\right) \qquad \textbf{(4)}$$

Our rule – denoted 'Rule of Thumb' in the IPSS – is simply to execute payments immediately if the liquidity is available and otherwise at the time suggested by equation **(4)**. We can note that this time will be independent of the size of the payment, depending only on its priority and the cost of liquidity (interest rate). Furthermore, banks operating this rule will make queued payments early, highest-priority first, if the necessary liquidity to do it comes into them. The parameters a and b can be set by the users of the IPSS as can the intraday interest rate. In the experiment we report below, we set the intraday interest rate to 0.1 basis points, a to 10^{-6} and varied b in order to achieve different levels for the liquidity-delay trade-off.

[4] To illustrate, suppose a payment of £100 came in at time t. The bank could either lend this out in the market and obtain $100i(T - t)/T$ or use it to reduce its overdraft balance with the central bank, saving $100i(T - t)/T$. So the net gain is zero. This enables the bank to ignore possible incoming payments when calculating the optimal delay time.

The analysis of the liquidity-delay trade off in the different variants of RTGS relies not only on the absolute values of liquidity posted and the number and value of delayed payments but also on the time-weighted values for these. These are important in order to see for what proportion of the day payments were delayed or settlement balances were positive. The time delay on payments uses the time stamps converted to the closest number of minutes from opening for payment requests, t_{isR}, and payment execution, t_{isE}, for each payment indexed by s for bank i such that when $(t_{isE} - t_{isR}) > 0$. The time weighted delay is given by dividing each of these numbers by T where T is total time in minutes from opening to closing. The aggregate time weighted value of payments for N banks is given by

$$X^{TWD}(£) = \sum_{i=1}^{N} \sum_{s} X_{is} \left[\frac{t_{isE} - t_{isR}}{T} \right] \quad \textbf{(5)}$$

It is clear that if all payments are requested at opening and delayed till end of day, $X^{TWD}(£)$ will equal the total value of payments requested in the day and, in percentage terms, the time-weighted proportion of delayed payments will be 100%.

The aggregate time weighted value of liquidity (L^{TW}) used is a useful measure to be contrasted with the total absolute value of payments made and the liquidity used in so doing. This is defined as :

$$L^{TW} = \sum_{i=1}^{N} \sum_{s} L_{is} \left[\frac{T - t_{isL}}{T} \right] \quad \textbf{(6)}$$

4 An example experiment

As an example of how one might go about using the simulator we consider an experiment comparing the liquidity-delay trade offs for two variants of the RTGS system. In the first payment system, banks can post liquidity only at the start of the day as there is an implicit assumption there is a large penalty if opening liquidity plus payment inflows cannot cover all payment requests received by closing time and that the cost of posting additional liquidity intraday is large relative to the potential cost of not using liquidity posted at the beginning of the day. In the second system banks obtain liquidity as and when they need it by borrowing uncollateralised funds from the central bank at a cost. To make the experiments being run for the two variants of RTGS comparable, identical sets of payments data (with payments arriving at the same time at the level of banks) need to be used so that both systems have the same upper and lower bound liquidity requirements. In the experiments reported below, we use the 'IID Real' arrival process for payments. In other words, we used data on payments settled within CHAPS Sterling on a particular day but allowed payment requests to arrive at banks earlier than the time they were submitted to the central processor, assuming an iid random arrival process.

4.1 Liquidity posted at opening (OL)

As we do not specify a decision rule by which banks post opening liquidity, we follow Bech and Soramäkii (2002) and specify exogenous amounts of opening liquidity. Six liquidity levels are operated for simulation purposes. These lie between the upper bound and lower bound levels of liquidity for each bank, which were defined in Section 2.

The six levels of liquidity are calculated as follows

$$L(\alpha) = UB - \alpha(UB - LB) \qquad \textbf{(6)}$$

where $\alpha = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$

As the opening liquidity operates like an externally verifiable budget constraint with the automatic FIFO settlement rule applied to equation **(1)** and the experimenter in control of the payments data arrival process, what payments are delayed and for how long can be tracked by IPSS. Payments will be delayed only if the exogenously posted opening liquidity is less than the upper bound. Due to the asynchronous nature of the arrival and size of payments, if banks post opening liquidity far short of what is needed to settle all payments as and when they arrive (the upper bound), payments may need to be settled as a group at the end of the day. We assume that this happens on a multilateral net basis. We note that opening liquidity close to the lower bound levels can be inadequate for full execution of all payments on an RTGS basis. Such payments are referred to as 'failed payments' with their gross value being given. Finally, based on our earlier discussion, banks may alter the order of payments and delay settlement in ways that differ from the automatic settlement rule of FIFO. To make the experimental results for the two RTGS variants comparable, we assume that banks reorder payments for settlement in the OL system by priority. We also looked at what would happen if the banks re-ordered their queues by size – lowest-value first – as well as priority.

4.2 Liquidity is raised just in time (JIT)

For a payment system in which liquidity is raised just-in-time – through borrowing from the central bank – we assume that banks post no liquidity at opening. But we do need to specify a behavioural rule for banks to decide when to settle payments. We suppose that banks operate the rule given by equation **(4)**. We also consider the alternative where banks re-order any delayed payments by size (rather than FIFO), again making payments as the liquidity becomes available. The contrast in the build up of queues of payments within each bank when banks reorder payments and when they do not is interesting to note. Such queues can be viewed in the IPSS using the menu button market 'Queues'.

5 Results

In this section, we report the results of our example experiment. Within the IPSS, all results can be obtained from the menu marked 'Intraday Statistics' which gives the full breakdown, at the level of banks, of liquidity posted/used and controlled, the number and value of delayed payments, length of time delayed, value of failed payments etc. At an aggregate level of the system the results of the experiment are found in tabular form under the function 'Statistics Collator'.

5.1 Liquidity-delay trade off results in RTGS

Here we first implement the Bech-Soramäki methodology for determining the liquidity-delay trade off in RTGS where all liquidity is posted up front at opening and delayed payments are settled on a FIFO basis. Table B gives the results of this experiment. Table C, in contrast, reports the results for the OL system when banks submit the payments for settlement smallest in value first. What is interesting is that as the OL system is squeezed for liquidity, i.e., when liquidity posted is at the lower bound value, there are some 10 unsettled payments of about £7.8 billion when banks 'strategically' reorder payments for settlement. This compares with zero unsettled payments in the FIFO case for the OL system in Table B. It was found that when all banks followed the strategy of postponing large payments till end of day at the lower bound value of liquidity, this left two banks needing to make six and four payments to each other, respectively. As neither had an adequate settlement balance to make the smallest payment they owed the other, we were left with a gridlock situation in which £7.8 billion of payments were left unsettled. Further, at the lower end of liquidity posted, on delaying larger payments for longer, the time weighted value of delayed payments in Table C is over twice that of the case in Table B when banks follow the FIFO rule.

Table B: Liquidity-delay statistics for Opening Liquidity, FIFO

Alpha	Liquidity (£ billion)	TW Liquidity (£ billion)	Number of delays	Number of delays (%)	Value of delayed payments (£ billion)	TW value of delayed payments (£ million)	TW value of delayed payments (%)
0.0	17.6	17.6	0	0.00	0.0	0	0.00
0.2	15.2	15.2	60	0.07	4.7	68	0.03
0.4	12.8	12.8	303	0.36	12.3	200	0.11
0.6	10.4	10.4	796	0.94	23.2	500	0.24
0.8	8.0	8.0	3370	3.98	49.2	1,300	0.61
1.0	5.6	5.6	12319	14.54	87.6	3,800	1.80

Table C: Liquidity-delay statistics for Opening Liquidity, Order by size, smallest first

Alpha	Liquidity (£ billion)	TW Liquidity (£ billion)	Number of delays	Number of delays (%)	Value of delayed payments (£ billion)	TW value of delayed payments (£ million)	TW value of delayed payments (%)
0.0	17.6	17.6	0	0.00	0.0	0	0.00
0.2	15.2	15.2	33	0.04	4.5	72	0.03
0.4	12.8	12.8	79	0.09	10.9	300	0.12
0.6	10.4	10.4	188	0.22	20.5	500	0.25
0.8	8.0	8.0	314	0.37	38.3	1,800	0.86
1.0*	5.6	5.6	848	1.00	65.2*	8,500*	4.04*

* Includes value of the 10 failed payments totalling £7.8 bn.

These results are in line with some unpublished work that two of the authors carried out using the BoF simulator. They suggest that if the system operators are worried about controlling operational risk by minimising the total value of payments in the queue, they would always prefer banks to use the standard FIFO by priority method of sorting their payments. If, alternatively, they were most concerned about the volume of payments in the queue, they would always prefer the banks to use the 'order by size, smallest first' method of sorting their payments. In practice, banks choose to post far more liquidity than the lower bound value (indeed they post far more liquidity than the upper bound value), probably to counter the possibility of being unable to make time-sensitive payments.[7] In particular, this is more likely to be the economical option when liquidity costs are relatively low and posting additional liquidity would not make large inroads into bank profitability. In addition, it is also likely that in a gridlock situation the banks concerned would negotiate an interbank loan so as to enable one bank to make the first payment needed for the gridlock situation to unwind; this possibility is not modelled in the IPSS currently.

Tables D and E report the liquidity-delay trade offs for the JIT system conditional on different threshold values for *b*. With $b = 10$, banks use a total of £17.6 billion, the 'upper bound' value of liquidity needed by the system. With $b = 0.85$ (at which point all payments are delayed), banks use a total of £11.2 billion if they order their queues by FIFO and £11.6 billion if they order their queues by value. At all levels of liquidity, fewer payments are delayed when queues are ordered by value than when they are ordered by FIFO but the value of delayed payments is higher when queues are ordered by value than when they are ordered by FIFO. This is the same as we found for OL systems and again suggests that if the system operators are worried about controlling operational risk by minimising the total value of payments in the queue, they would always prefer banks to use the standard FIFO by priority method of sorting their payments.

Table D: Liquidity-delay statistics for Just in time, FIFO

b	Liquidity (£ billion)	TW Liquidity (£ billion)	Number of delays	Number of delays (%)	Value of delayed payments (£ billion)	TW value of delayed payments (£ billion)	TW value of delayed payments (%)
0.85	11.2	8.8	3509	4.14	46.7	7.0	3.30
1	12.8	10.4	1218	1.44	33.7	4.8	2.29
2	15.4	13.3	462	0.55	8.8	1.1	0.54
3	16.7	14.4	272	0.32	3.7	0.3	0.14
4	17.5	15.1	172	0.20	1.8	0.1	0.05

[7] See James and Willison (2004) for much more on this.

10	17.6	15.3	0	0.00	0.0	0.0	0.00

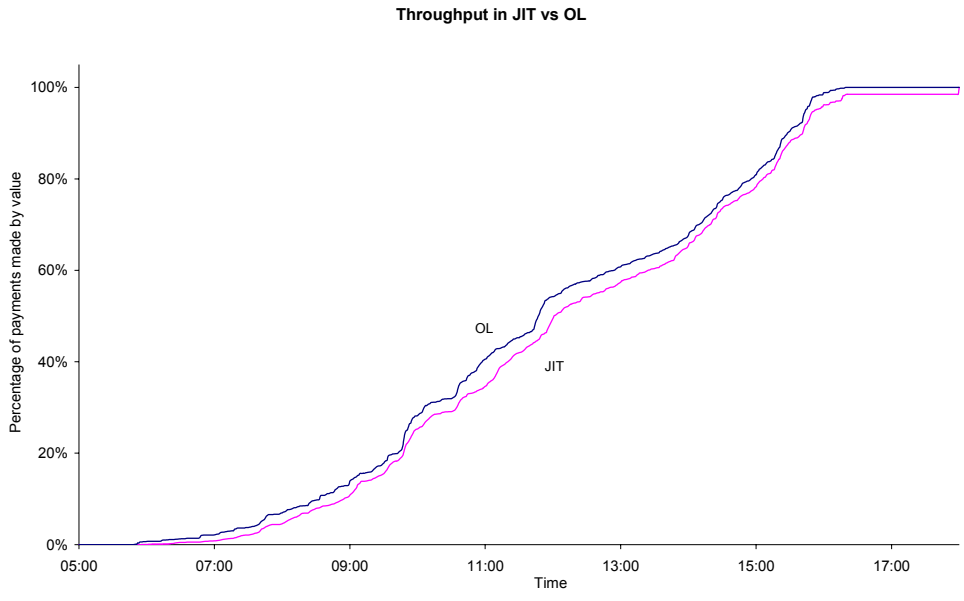
Table E: Liquidity-delay statistics for Just in time, Order by size, smallest first

b	Liquidity (£ billion)	TW Liquidity (£ billion)	Number of delays	Number of delays (%)	Value of delayed payments (£ billion)	TW value of delayed payments (£ billion)	TW value of delayed payments (%)
0.85	11.6	8.9	3310	3.91	47.4	7.2	3.39
1	13.3	10.8	1019	1.20	36.4	5.5	2.60
2	15.3	13.3	409	0.48	9.5	1.2	0.58
3	16.6	14.4	249	0.29	3.9	0.3	0.16
4	17.4	15.1	157	0.19	2.1	0.1	0.06
10	17.6	15.3	0	0.00	0.0	0.0	0.00

5.2 The impact of operational events in the two systems

As we said earlier, one way of gauging the effect of an operational event is to examine how quickly payments are submitted to the central processor and settled within the system. The logic is that should an operational event occur, it will affect the ability of the system to sort out any remaining payments for the day. The more payments have already been processed before the event happens, the less this will be a problem. Figure 1 shows the throughput for the two systems, viz. the percentage of payments by value that are settled prior to any given time for each of our two systems. To make the comparison fair, we assumed that £12.8 billion of liquidity was used in each of the two systems. From Tables B and D, respectively, we see that in the FIFO case, at this level of liquidity, the JIT system delays £33.7 billion worth of payments with time weighted value of £4.8 billion. In contrast, the OL system delays payments valued at about £12.3 billion with time weighted value of £0.2 billion. Figure 1 shows that, at each point in time, the OL system has settled more payments than the JIT one.

Figure 1: Throughput in the two RTGS systems



6 Concluding remarks and future work

In this paper, we have built a simulator for interbank payments capable of handling real time payment records along with autonomous bank behaviour and shown that it can be used to evaluate different payment system designs against these three criteria. In RTGS systems, payment requests to banks are not fully and individually financed as then they would need an amount of liquidity equal to the total value of payments made; rather, the bulk of the liquidity used for settling comes in the form of incoming payments. The efficiency in recycling the liquidity posted by banks is the key to the design of RTGS system. We showed that our simulator could be used, in principle, to evaluate different designs of RTGS systems by carrying out an example experiment using two systems: one in which liquidity was posted at the beginning of the day and one where it could be borrowed 'just in time'.

At the level of individual banks, an attempt to handle the trade-off between liquidity and delay costs may result in behaviour where banks delay settlement in anticipation that other banks will make payments to them first. Further, it appears to be intuitive that at an individual level larger payments should be delayed if their delay costs are not too high. Thus, in principle banks always have the discretion to reorder payments for submission at the central processor influencing the liquidity that they need to post and the ability of the system to settle payments. What appears to be an individually rational response at the level of banks, viz. to delay large payments (with low priority), leads to a deterioration in the collective performance of the RTGS system whether in the OL case or the JIT variant. However, we found that the JIT system is more prone to rapid deterioration of its liquidity recycling capabilities than the OL system. In fact, a key message of our experiments is that, at any given level of liquidity, the JIT system would generate more delayed payments than an otherwise identical system in which banks posted their liquidity at the beginning of the day, and this would be bad from an operational risk point of view.

However, the results of our experiment should not be taken too seriously since there are some important caveats that need to be born in mind. (The purpose of the experiment was more to demonstrate what can be done with the IPSS.) In particular, our experiments relied on one particular stochastic simulation of data based on one day's worth of actual payments. To get a clearer picture one would need to run multiple simulations based on data from a large number of days. In addition, our experiments on the OL system imposed a given level of opening liquidity. In order to get a real understanding of how such systems work, one would need to postulate behavioural rules to explain the decision of how much liquidity to post at opening. Such a rule would relate liquidity posted to such variables as the opportunity cost of collateral, the average delay costs expected by the bank, the total expected value and volume of payments coming in or going out and the uncertainty surrounding these. For the JIT system, a possible extension might be to consider behavioural rules that allow banks to take into account the possibility of using liquidity from incoming payments to make their own future payments when they choose how long to delay payments.

The main issue relating to mechanism design in real time interbank settlement systems is 'How can the socially efficient outcome be achieved by design, with banks having to behave

autonomously and faced by asymmetric information?' In the philosophy of agent based modelling, however, the prior question is 'Can banks behaving autonomously adaptively learn to achieve the liquidity savings associated with cooperative outcomes?' An obvious extension of the work presented in this paper would be to focus on the question of whether or not banks, as autonomous and adaptively intelligent agents playing a repeated game, can move to the efficient and stable point of the 'good' equilibrium; that is, could they co-operate in such a way as to enable them to make payments with little delay in an OL system operating at liquidity levels close to the lower bound. Computational experiments of this kind may yield invaluable normative insights into the complex intraday liquidity management game by banks within the context of bank profitability and solvency.

References

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

Bank of England (2000), *Oversight of payment systems*

Bech, M L, and Soramäki, K (2002), 'Liquidity, gridlocks and bank failures in large value payment systems', in R. Pringle and M. Robinson (eds.) *E-Money and Payment System Review*, London: Central Banking Publications.

Bech, M L, and Garratt, R (2003), 'The intraday liquidity management game' *Journal of Economic Theory,* Vol. 109, pages 198-219.

James, K, (2003), 'A statistical overview of Chaps Sterling', Bank of England *Financial Stability Review*, June.

James, K, and Willison, M (2004), 'Collateral posting decisions in CHAPS Sterling', Bank of England *Financial Stability Review*, December.

Leinonen, H, and Soramäki, K (1999), 'Optimizing Liquidity Usage and Settlement Speed in Payment System', Bank of Finland *Discussion Paper* 16/1999.

McAndrews, J, and Trundle, J (2001), 'New payment system design: Causes and consequences', Bank of England Financial Stability Review, December, pages 127-36

McAndrews, J, and Rajan, S (2000), 'The timing and funding of Fedwire funds transfers', Federal Reserve Bank of New York *Economic Policy Review*, July, pages 17-32

Willison, M (2004), 'Real-time gross settlement and hybrid payment systems: A comparison', forthcoming Bank of England *Working Paper*.