Moving Towards Immediate Payments

By Tom Hay, Head of Payments

March 2015

An Icon Solutions Whitepaper

Moving Towards Immediate Payments

Contents

1. What Is Immediate Payments?

Immediate Payments – Defining Characteristics

Immediate payments is also called Faster Payments and Real-Time Payments. Although there is no standardised definition of immediate payments, it is generally accepted that systems have the following characteristics:

- **Immediate credit.**
The funds become available in the payee's account immediately (within a few seconds) of the payment being initiated by the payer.
- **Irrevocability.**
Once the payer has initiated the payment, they cannot cancel it.
- **Certainty of fate.**
When the payer initiates the payment, they are informed immediately (within a few seconds) whether the payment has successfully reached the payee's account or not.
- **Straight Through Processing (STP).**
An immediate payment completes end-to-end. Problems result in immediate rejection, allowing the payer to correct and retry – there are no manual repair queues.

The world is moving towards immediate payments. A recent **survey** identified around 30 systems globally that might be classed as immediate payments, with many more countries in the planning stage. There is still no consensus around the standards, features or technology that such a system should use. Nevertheless, there are some common issues that will face immediate payments systems irrespective of the specific implementation details.

As Fast As ATM Withdrawals

The aspiration is for immediate payments systems to operate at a similar speed and in a similar way to using a debit card in an ATM – when you initiate the transaction you get a response typically within a few seconds. You cannot cancel the transaction once it has been initiated, and if the transaction is successful, you receive the cash in your hand, and those funds are no longer available in your account.

Does It Achieve These Goals?

The **UK Faster Payments system** generally meets these expectations in practice, even though the scheme rules allow for end-to-end transaction times of up to 15 seconds, and allow institutions up to two hours to carry out fraud and sanctions checking before crediting funds to payee accounts. This does weaken the "certainty of fate", as does asynchronous processing discussed later, both of which may result in a payment being returned to the payee even after being confirmed as accepted.

What Is A Switch?

Immediate payments systems generally use a central "switch" system to which all institutions submit payments. As well as validating, logging and routing the payment instructions, the central system also handles settlement of payments across the central bank.

2. Is Immediate Payments Just Fast Batches?

A Certainty Of Fate

Immediate payments is very different from fast batches. One of the key characteristics of immediate payments is “certainty of fate”. This means that the payer gets an immediate confirmation from the payee’s bank that the payment has been received and credited to the account. This cannot be achieved by batches, no matter how fast.

To achieve immediate confirmation to the payer, instead of sending files containing batches of multiple payments, an immediate payments system sends messages that carry a single payment.

Anyone who is familiar with card processing will recognise this approach, and indeed the UK Faster Payments system adopted and adapted the ISO8583 message format that is used in card processing, and the request-response message flow and the end-to-end integrity model are also borrowed from the world of POS and ATMs.

Understanding The Impacts

This has a number of impacts on the system.

Immediate payments must use a different communications protocol from batch systems: Instead of a file transfer protocol such as FTP or C:D, an immediate payments system will use a message-oriented protocol such as https, or even a low-level protocol across TCP/IP socket connections.

Payments have to be processed as received, within seconds: Instead of handling a scheduled batch load, the system must process an unpredictable and “spikey” load. An immediate payments system does not have a quiet period where online processing stops and batches may be processed – many users make immediate payments outside of regular office hours. This will impact the sizing and design of the system.

The system has to be available to process payments 24x7: If payments are to be credited to the account immediately, the core banking system cannot run an offline “end of day”; or if it does, the immediate payments must be processed against a “stand-in” system. This allows the funds to be received and made available, even though not yet updated on the system of record. This approach has been used for many years by debit card systems.

24x7 availability also means that the system must be highly resilient (to protect against unscheduled outages) and cannot be taken out of service for scheduled maintenance.

These considerations make the architecture and design of immediate payments quite different from traditional batch payment systems, and may require new skills or even a new IT platform to deliver the service.

3. Is Immediate Payments The Same As RTGS (Real Time Gross Settlement?)

The key feature of immediate payments is immediate **clearing** (i.e. as soon as the payment is complete, the funds are available in the payee's account). It does not have to be an immediate **settlement** system (i.e. the settlement between banks does not have to happen synchronously with the clearing). This is what distinguishes immediate payments from RTGS, which settles every transaction across central bank accounts in real time.

Settlement Risk…

UK Faster Payments settles three times per banking day, while Mexico's SPEI settles every few seconds. In cases where settlement is not synchronous with clearing, there is the possibility of a settlement risk. If a bank has sent payment instructions to other banks, who have credited their customers, the payee banks are exposed to the risk of the payer bank failing to settle. A number of mechanisms are available to limit this risk.

… And Mitigations

One mechanism is the use of collateralised liability limits, set on a per-bank basis and enforced by the central switch. A bank starts at a zero position, and every payment it makes increases its liability, while every payment it receives from another bank reduces its liability.

During the course of a day a bank's position will fluctuate as payments flow in and out; if the bank is a net sender of payments its position will drift towards the liability limit, if it is a net receiver it will drift away from the liability limit.

If the central switch detects that a bank has reached its liability limit, it will block any further payments from that bank until in-payments or settlement bring its position back below the limit. When settlement takes place each bank's liability position is adjusted by the settled amount.

Another possibility is to use pre-funded liability limits. This ties up working capital, but does allow funds to be added (hence limits to be increased) intra-day if necessary. It means that the central bank can immediately draw on the pre-funding to settle the liabilities of a failed bank.

Settling In Real Time

Of course it is possible for an immediate payments service to settle in real time, and in Australia's 'New Payments Platform' the RBA's **Fast Settlement Service (FSS)** is designed to do just that.

While this eliminates settlement risk, it puts much higher technical demands on the RTGS system.

The Central Bank's RTGS system has to be able to process payments at the same rate as the immediate payments system, and it introduces another source of latency, an additional node that complicates "undo" processing, and another potential point of failure into the system.

> "Immediate payments systems are not necessarily RTGS systems…"

The flow illustrated in the RBA paper indicates that the payment cannot be considered complete (debit and credit to payer and payee finalised) until the settlement messages are received (settlement may be rejected); but the paper also says that even if "the FSS is unavailable clearing may still continue".

This sounds as if it will compromise the "certainty of fate" criterion of immediate payments, as the payer will not know whether that payment has succeeded until the settlement message is received (which might never happen).

In summary, immediate payments systems are not necessarily RTGS systems, and mechanisms are available to limit settlement risk to an acceptable level without incurring the inevitable penalties of trying to eliminate it entirely.

4. How Similar Is Immediate Payments To Card Systems?

Similarities And Differences

There are many similarities between card systems and immediate payments systems. The key concept of an end-to-end message flow that provides certainty of fate originated in card systems with the ISO8583 message specification, which is also used in the UK Faster Payments system.

The main difference is the direction of message flow. In card systems the payee's system sends a request message to the payer's system which validates the request, checks availability of funds in the payer's account, and returns a response to the payee. Immediate payments reverses the flow, implementing a "push" payment initiated by the payer, rather than a "pull" payment initiated by the payee.

Aside from the direction of message flow, there are a number of important differences between the payment processes:

Message routing: Card systems use the card number prefix to determine the issuer (payer's system) to which to route the request message. Immediate payments systems must use the destination sortcode (in the UK) or BIC (in mainland Europe and elsewhere) to determine the payee institution. This works well as long as each BIC/sortcode maps to a single payee institution, but more elaborate logic is required when this simple mapping does not exist.

Authentication: In card systems the payer must present their credentials to the payee's system. The strength of authentication varies widely according to the card acceptance environment, from mag stripe and signature, through eCommerce card number and CVV2 (possibly with 3D Secure), to EMV chip and PIN.

The inherent insecurity of passing confidential information through payee's systems has been amply demonstrated by the repeated massive compromises of merchant systems, despite the introduction and repeated strengthening of PCI-DSS security standards.

In an immediate payments "push" system, the payment is initiated by the payer authenticating themself with their own chosen financial institution. Security credentials are not shared with a third party, increasing the inherent security of the system. (This situation may change with the introduction of PSD2 and TPPs, the legislation is still unclear on this point).

Reversal handling in the system: Another difference is around the handling of transaction co-ordination and "undo" processing. In a card system, technical problems resulting in message loss or timeout can be resolved by the payee sending a reversal message.

In a "pull" system, it is simple for the payer's institution to re-credit their account when a reversal happens. In the immediate payments "push" system, the payee's institution may not be able to reverse the credit to the payee's account, as the funds may already have been withdrawn.

This presents a problem to the payee institution, as there is no clearly identifiable moment at which the payee institution may consider the credit definitive. Such occurrences do not happen frequently, so banks may choose to deal with those few occasions manually. If this is not considered acceptable, it is possible to eliminate the risk entirely by modifications to the card-style message flows, though this will increase the number of messages needed to complete a payment.

5. Is ISO20022 Suitable For Immediate Payments?

The Difference In Message Flows

The ISO20022 message standards arose from an asynchronous messaging environment (SWIFT) where messages are sent via a "guaranteed delivery" network transport. Messages are generally treated as "fire and forget"; there is no need to wait for an application level response to a message.

This type of operation is very well suited to file-based processing, where a file containing multiple (possibly hundreds of thousands) payment instructions is sent, and some time (possibly hours) later response file(s) are received confirming the fate of each of the payment instructions.

The message flows in immediate payments are quite different. A payer initiates a single payment, and expects an immediate response confirming the success or failure of the payment.

This type of request-response messaging is very similar to those used by card processing systems (albeit in the opposite direction – card messages are "pull" payments, initiated by the payee's agent and responded to by the payer's agent, while immediate payments are generally "push" payments, initiated by the payer's agent and responded to by the payee's agent).

ISO8583 – The Card Message Standard

The messaging standard used in the cards world is called ISO8583, and it consists of two main components: message formats and message flows.

The message formats are based on "bitmaps" and were designed to minimise message sizes in the days when bandwidth was severely constrained (the first version of ISO8583 was released in 1987 when communication bandwidth was around 1,000 times lower than today).

The message flows are designed to provide certainty of fate of an end-to-end transaction, even when the transport mechanism does not guarantee delivery. This is achieved by having the nodes in the message path wait for application-level responses, and pass error messages upstream and downstream in case of a timeout (allowing other nodes to achieve a common view of the outcome of the transaction).

> "While the ISO8583 flows provide a good model for Immediate Payments, the message formats sacrifice richness and flexibility to achieve a size reduction that is not necessary in today's world."

ISO20022 – The Best Of Both Worlds

It has **been suggested** that ISO20022 cannot be used for immediate payments, due to its asynchronous heritage. This is definitely not the case – it is possible to use ISO20022 message formats conforming to ISO8583-like flows – this is how Singapore FAST has been implemented. Certainly XML-based ISO20022 is a more verbose message format than bitmapped ISO8583, but this is compensated for by the ability to include rich data such as remittance information and the wealth of tools available for processing XML, with hardware "appliances" to boost performance if necessary.

ISO20022 is the emerging standard for payment processing across the world, and there is no reason why immediate payments systems should not fully participate in this trend.

6. Why Is Settlement Difficult In Immediate payments Systems?

A Matter Of Timings

In a batch payments system, there is a well defined cut-off time for submission of files of payments. Payment instructions received before the cut-off time are processed on that processing day, while payments received after the cut-off time are processed and settled the following processing day. (Most systems operate a calendar, so the "following" day refers to the next processing day, which is not necessarily the next calendar day).

All payments within a processing day can be settled together by a net settlement process; there is no problem to determine which processing day a particular payment belongs to.

Immediate payments systems that run 24x7 cannot implement a simple cut-off time. Even if all participant system clocks were closely synchronised (which would be possible in principle by using Network Time Protocol), network and processing delays could easily result in a message being sent in one processing day, but received the next processing day.

This would result in mismatches between the settlement positions calculated by participants, and the actual settlement carried out by the payment system operator.

Overcoming The Challenge

To avoid this problem a settlement indicator can be carried in the message, specifying which processing day the message belongs to. The indicator could be inserted by the message sender or by the central switch.

If set by the sender, it would be unwise to rely on a fixed schedule, as that would limit the flexibility needed to cope with unexpected operational issues. A better solution would be for the central switch to broadcast an "end of day" message to ensure that all participants start using the new processing day; but this in turn leads to problems if a participant "misses" the broadcast for any reason.

The alternative approach, where the central switch inserts the settlement indicator in the message, is simpler, but has implications for the type of message security. For example, a simple digital signature across the whole message, passed through from sender to receiver, cannot be used.

Mind The Gap

Whichever approach is used, there will be a short period of time following cut-off when in-flight payments from the previous settlement period are still being processed while new payments are being initiated in the new settlement period.

Care must be taken to accumulate the payment values into the appropriate settlement total. Particular care must be taken with payments that are initiated in one period, and a reversal is generated (perhaps only a few seconds later) in the next period.

Note that there is no reason why settlement should be carried out on a daily basis. More frequent settlements are perfectly possible, down to a granularity of a few minutes. The frequency of such settlements, and the hours and days on which settlements can be executed, may well be constrained by limitations on the central bank's settlement system and/or processes, which must be taken into account.

7. Should all payments be immediate payments?

What's The End Goal?

If some payments can be made immediately, it is reasonable to ask why all payments shouldn't be immediate. Of course, it would take a while to convert all batch payment systems – not only in banks, but also in companies, government and other organisations – but in principle, is the end goal to have all payments made immediately?

Certainly not all payments **need** to be immediate. Some payments are scheduled well in advance, so there is no benefit to users in making them immediate (it was a quirk of the UK Faster Payments implementation that scheduled Standing Orders were the first category of payments to be migrated to Faster Payments – they are arguably a completely different class of payments that happen to run on the same "rails" as Faster Payments). For other unattended payments, such as supplier payments, payroll, benefits etc. – arguably the majority – users would benefit from having the payments executed same day i.e. credited to the payee's account on the same day that the payer issues the instruction, but strictly immediate execution is not a requirement.

Planning For The Future

It is when one looks to the future that truly immediate payments become really important. As digitisation sweeps the world of banking, immediate payments stand out as the perfect mechanism for executing P2P and C2B payments. These are attended payments where the payer is waiting for completion of the transaction.

Until now such payments have relied on card systems, which have been the only systems to offer immediate confirmation (albeit with deferred clearing and settlement). Of course in the digital world the idea of a "card" is an anachronism, and immediate payments eliminate cards from the equation.

Irrespective of whether users **need** payments to be immediate, if the cost of processing immediate payments were the same as the cost of processing batch payments, it would be reasonable to make all payments immediate – it would simplify technical infrastructures and operational processes if all payments flowed across just one set of "rails". Unfortunately, for the time being at least, it is significantly more expensive to process single payments than it is to process bulk payments.

Why The Challenge?

The reasons for this are not hard to find. There are many technical processing steps associated with making a payment – digital signing, logging, transmitting and so on – that can be executed just once for a batch of payments, but which must still be executed for even a single payment.

Even if the overhead of these operations is very small per transaction, the cumulative saving of processing, say, 100,000 payments in a single batch is very considerable. Processing payments individually also increases probability of communications-related problems, simply by the law of large numbers.

A Changing Landscape

Although batch processing is still more cost-effective than single message processing, the gap is closing. Since the advent of the internet, IT systems have been increasingly optimised to handle single transactions and HTTP message exchanges rather than batches of transactions.

Technologies such as NoSQL databases are much better suited to single message processing than SQL databases that were explicitly designed to handle data as sets rather than individual rows. Techniques such as reactive programming and use of functional languages promise to further enhance single message processing.

So while it is not possible to eliminate the additional overhead of single message processing, the cost of the overhead will soon become negligible. At that point it will make sense for all payments to be immediate.

8. Should immediate payments support asynchronous messaging?

A Contradiction?

The idea of immediate payments being handled asynchronously seems like a contradiction in terms. If a payment is immediate, it cannot be asynchronous because that implies an indeterminate time between payment initiation by the payer and payment receipt by the payee. While this is true strictly speaking, it is one of the enhancements that makes immediate payments more usable in the real world.

Synchronous payments (indeed any synchronous messaging) require sender and receiver both to be available when the message is sent. The requirement of immediate payments systems to be available 24x7 is meant to guarantee this simultaneous availability by ensuring that all parties are available at all times. In practice all participants are likely to suffer periods of unavailability from time to time, which would result in degraded service. Worse, it would be the payer that suffered inconvenience, even though the problem is with the payee's institution.

The Solution

A solution to this problem has been available for many years in card payment systems. The idea of "stand-in" processing allows an intermediate system to handle the payment and generate a response in case the primary processing system is unavailable.

Typically this service is offered by the central switch, either on explicit request by the destination system (typically used to cover scheduled outages) or dynamically, in case the destination system fails to respond within a specified time.

There are limitations on stand-in processing. The stand-in system does not have access to core banking systems, so must make an approve/reject decision based on limited information. This is a bigger risk in a card system using debit requests (where an incorrect decision may result in payment being approved when the payer's account does not have sufficient funds) than in an immediate payments system using push credits, where the worst case outcome is that the payment cannot be applied to the destination account, so must be returned to the payer.

Other Opportunities

Another possible use of asynchronous processing is to allow an immediate payments system to handle less-urgent payments that require same-day processing but not immediate credit to the payee. In this case the central system can provide an immediate response back to the payer system, and store-and-forward the payment to the payee system.

This allows delivery of less urgent payments to be paced or throttled, and treated with a lower priority than true immediate payments, preventing the non-urgent payments from negatively impacting processing of urgent payments.

This latter use of asynchronous single message processing is not the only approach to processing urgent and less urgent payments across the same payments system, and may not be the best. Sending multiple payments in a single larger message, similar to a SWIFT MT-102, but using a common settlement approach, may be a better use of system resources for less urgent payments.

9. What About Immediate Debits?

The UK Faster Payments scheme only supports credit transfers, while Singapore's FAST system also supports debits. What are the pros and cons of including debits in an immediate payments system?

Credits And Debits

- A credit transfer is initiated by the payer's institution (push), and the classic use case for immediate payments is where the payer initiates the transfer via a digital channel (internet or mobile banking). In this case the payer has authenticated themselves to their financial institution, so there is no question that the payer has authorised the transaction.
- A debit is initiated by the payee's institution (pull) on behalf of the payer. When a card is used at POS, the payer provides some proof of identity (PIN or signature), and the merchant's terminal sends the debit request to the card scheme. Debits are also used for recurring payments such as subscriptions or utility bills, which the payer does not authorise on a per-transaction basis.

For recurring payments such as subscriptions or utility bills, there would be no particular benefit in making them immediate payments. Billers collect these payments under a framework agreement with the payer authorising them to collect funds according to a set of agreed rules. The UK Direct Debit scheme provides such a framework and is extremely popular. The rules may limit payment amount and/or frequency, and potentially require a pre-advice period allowing the payer to block the debit if they disagree with it. The framework and rules are reflected in a mandate held and checked by the payer's financial institution (or a central system). Since these are scheduled payments there is no need to process them via immediate payments, though like scheduled credits they could be processed this way if it were cost-effective to do so.

The situation is very different where the payer is making a non-scheduled or one-off payment, which covers the majority of bricks-and-mortar, eCommerce and mCommerce transitions. Historically this scenario has been handled by credit cards or debit cards, and it seems an ideal use case for immediate payments. However, the fraudulent use of cards and theft of credentials as they pass through merchant systems has become a huge problem for the industry, and simply copying the card model in an immediate payments system would invite the same problems.

Problems With Card Security

Over the past 20 years an incalculable amount of time and effort has gone into trying to solve the problem of ensuring that the payer has really authorised a card transaction. The story of CVV2, 3DSecure, chip and PIN, PCI-DSS etc. is much too long to explore here, but despite the billions spent on prevention, the global card fraud rate continues to grow, with over $11bn worth of annual losses. The recent Target data breach, in which 100 million card details were stolen, highlights the inherent risk of passing customer authentication details through payee systems.

However, there is another way of handling debits in an immediate payments system that does not require a debit "pull". A credit card is a dumb piece of plastic, but the consumer's banking app on their smartphone can receive a "request for payment" from the merchant and alert the consumer. The consumer then provides an appropriate level of authentication to the banking app to make the push payment to the merchant. This keeps consumer details confidential between the consumer and their own chosen financial institution, completely bypassing the problems experienced by card systems. Some additional work is needed to route messages between merchant and consumer, but these can be solved in a way that provides a very slick consumer experience as **Zapp** has done.

Planning For Tomorrow, Today

The conclusion is that immediate payments systems should support debits in anticipation of the cost equation favouring migration of scheduled debts to immediate payments. They should also consider how the system can introduce new messages types, such as the "request for payment" message (which is neither a debit nor a credit), to support future innovation.

10. How Can Fraud Checking Be Handled In Immediate Payments?

When Faster Payments was launched in the UK, some people condemned it as inviting ‘Faster Fraud’. Its immediate and irrevocable character certainly raises the bar on fraud checking, but is by no means an insuperable challenge.

Tackling Fraud Checking Issues Before They Arise

It is important to address the problem with the correct mindset. Immediate payments is a platform for the future, and any fraud prevention measures that rely on limitations of the initial use cases are likely to cause problems in the future. For example, some UK banks viewed Faster Payments as “immediate bill payments”. This led them to introduce a beneficiary whitelist, and each time a payer makes a payment to a new beneficiary the bank makes an out-of-band confirmation such as an automated phone call before releasing the payment and adding that beneficiary to the whitelist. This is a good approach until ‘spontaneous’ payments such as mobile person-to-person, eCommerce or mCommerce payments are introduced, which makes the confirmation step intrusive and irritating for the consumer.

Another Way?

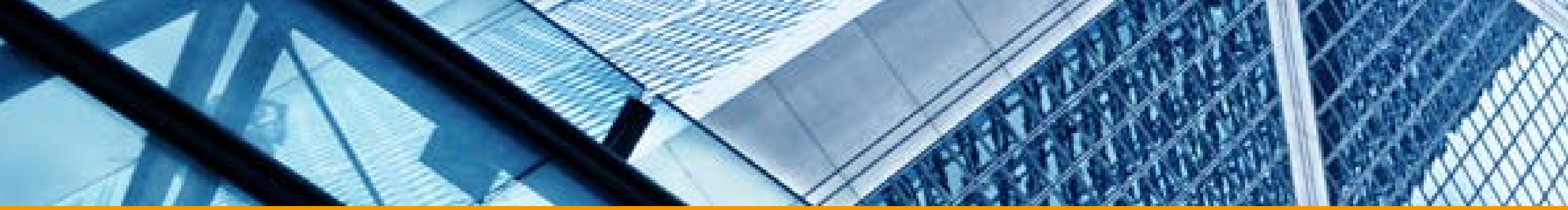
The alternative to explicit payer confirmation of new beneficiaries is monitoring of payments by a real-time fraud scoring engine, with the option to automatically block payments whose score exceeds a particular threshold. This approach is commonly used in card systems, but there are some significant differences between the two environments. Card transactions typically carry significant amounts of contextual information, such as the card acceptor location, merchant type, terminal type, data input method and so on. These are not present in an immediate payment, which may contain little data beyond the source and destination account numbers. Card transactions also have some easily detected suspicious patterns. Sophisticated fraud systems use rule-based detection that can be calibrated against pooled community data supplemented with neural networks that can be trained using historical data. Neither community data nor historical data are available in a new immediate payments system.

Luckily some next-generation fraud detection systems are becoming available that support rapid self-optimising detection algorithms. These systems can generate risk scores in real time, allowing suspect transactions to be halted prior to execution, and can achieve a high level of accurate fraud detection with a low rate of time-wasting ‘false positives’. The adaptive algorithms allow new patterns of fraud to be recognised and defeated much more quickly than older systems. Implementing this kind of fraud detection engine can provide a high level of protection to immediate payments systems, and if implemented as part of the central switch, costs per participant can be minimised.

Compete Or Co-operate?

Banks continue to debate whether fraud detection should be a competitive or a co-operative space. From a functional point of view, in principle there are certain patterns of fraudulent payments that could only be detected by a central system, but so far no overwhelming advantage of centralised versus per-bank systems has been demonstrated.

However, from a cost point of view the case for implementing a shared central fraud detection function is overwhelming, compared with each participant implementing their own solution. In a mature market some banks may argue that they have already invested in sophisticated fraud detection systems that they see as a competitive advantage, so have no incentive to invest in a co-operative solution. This is likely to change as the next-generation systems discussed above increasingly demonstrate their superiority over existing solutions, and banks see the need to upgrade.



11. When Can You Be Sure That A Payment Has Been Made?

In an ideal world the question of when a payment has been made would never arise – the aspiration of immediate payments systems is to credit the payee within the transaction flow, so that when the response is returned to the payer, they know that the payee's account has already been credited. Unfortunately in the real world this is not always possible. Uncertainty can arise at a technical level or at a business level.

The Technical Level

In the 'happy path' scenario the request message (payment instruction) flows from the payer's institution, via the central switch, to the payee's institution. The payee's account is credited and the payee's institution sends back the response message via the central switch to the payer's institution.

Now consider a case where the response message is lost between the payee's institution and the central switch due to a network glitch. The switch will time out waiting for the response, and must therefore assume that the payment has not been made.

It sends a rejection to the payer's institution, and a cancellation message to the payee's institution (to ensure that if the payment was actually applied, it is reversed).

If the payee's institution applies the credit as soon as the request message is received, there is a very small chance that the credit will have to be reversed if a timeout occurs.

On the other hand, it significantly complicates the system if the payee's institution has to wait until there is no possibility of a timeout. Adding extra steps to the message flow would improve the situation, but the additional complexity is not justified given how rare message failures are in practice.

Uncertainty Arising At The Business Level

Moving to the business scenarios, the uncertainty occurs on the payer's side. Suppose that the payee institution's system flags the payment for fraud or AML concerns. It cannot reject the payment – that would run the risk of tipping off – so it must accept the payment, but not apply it until the investigation is complete.

Therefore an accepted response from the payee cannot be taken as 100% confirmation that the credit has been applied. This is the main reason why the UK Faster Payments scheme allows payee banks up to 2 hours to apply the payment, and payer banks advise their customers that faster payment may take up to 2 hours to be credited, even though the vast majority of payments are in fact credited immediately.

12. What Sort Of Security Should Immediate Payments Use?

Card systems typically sign messages with a message authentication code (MAC) generated using symmetrical encryption keys. This requires point-to-point encryption between banks and the central switch using pre-shared keys. The keys are changed regularly.

This arrangement ensures that the sender of a message is valid, and that the message has not been tampered with. It does not provide non-repudiability – in other words, if there is a later disagreement between sender and receiver regarding message content, neither side can prove that it has not changed the message because both use the same key and could therefore change their record of the message and re-sign it.

Public Key Encryption

Non-repudiation can be provided by a different type of encryption called asymmetric or public key encryption. This uses a pair of keys, one of which is secret and is retained by the message sender, the other is made public in the form of a digital certificate. Any receiver can use the public key to verify that a message was signed by the private key, but cannot use the public key to re-sign a changed version of the message.

Public key encryption has another advantage. Digital signatures can be passed end-to-end and validated by the final recipient rather than having to be validated and re-encrypted by the central switch.

If the design of the system does not require the central switch to make changes to the message as it passes through, this is simple. If the central switch **does** have to make changes – inserting a settlement indicator for example – then the digital signature has to encompass specific fields rather than the entire message. This is catered for by standard signing protocols, but requires a little more design work.

The Validation Problem

One drawback of public key encryption is that the recipient needs to be sure that the sender's private key is still valid. If the key has been compromised the sender will revoke it.

There are two ways for a recipient to check whether a key is still valid: check a Certificate Revocation List (CRL), or make an online call using the Online Certificate Status Protocol (OCSP). The CRL approach can only detect compromised keys when the revocation list is circulated, which may be 24 hours or more after the compromise – this is clearly unacceptable.

The OCSP approach gives immediate confirmation of certificate revocation, but the OCSP call must be made in real time. The time taken to execute the call is added into the overall latency of the immediate payment, so the speed of response of the OCSP provider is critical.

The decision whether to use symmetrical encryption or public key encryption is one of the many important considerations in the design of an immediate payments system.

13. What's Next For Immediate Payments?

No-one has a crystal ball, and given the nature of immediate payments as a platform for innovation in payments, it is impossible to foresee all of the ways in which it might evolve. Nevertheless, there are a number of developments that can be predicted with a high level of confidence.

International Expansion

As more countries introduce domestic immediate payments systems, pressure will grow to introduce immediate cross-border payments. The European Retail Payments Board is driving a **pan-European instant payments initiative**, though arguably this would be better viewed as an extended domestic system since all participating countries share the Euro as a common currency.

Other regions with heterogeneous currencies have also expressed an interest in regional immediate payments schemes.

> "As more countries introduce domestic immediate payments systems, pressure will grow to introduce immediate cross-border payments."

Overcoming International Issues

A number of issues need to be addressed in extending immediate payments internationally. The first and most obvious is the need for agreed standards in technical areas such as message formats, networks, security standards etc, and operating rules covering service levels, liabilities, dispute resolution procedures etc.

Another area is the need for currency conversion. This could be carried out by the sender, the receiver, or the intermediate switch(es) – there is likely to be considerable discussion over this point given the margin that can be derived from FX.

A third issue concerns settlement – which institution can act as a settlement agent with which all member institutions could hold accounts? And what settlement schedule would work across time zones? There is also the fundamental question of fees and business model – what financial arrangements would encourage participation by all countries, despite local variations in charging arrangements for payments?

International card schemes operate with a single central authority to resolve such issues. Given the development of national schemes, it is unlikely that a similar international organisation will emerge to govern immediate payments. It will be interesting to see whether a federated organisation can act rapidly enough, and get sufficient co-operation from its members, to govern effectively and drive the system forwards.

Improving Flexibility

Another area where immediate payments systems need to evolve is in terms of flexibility. We have only seen the beginning of the kind of digital services that can be built on an immediate payments platform, such as the Pingit and Paym mobile payment services, and the forthcoming Zapp eCommerce and mCommerce service in the UK.

The Australian New Payments Platform requirements explicitly refer to 'overlay services' as a separate layer from the base message handling.

> "We have only seen the beginning…"

Elsewhere in this paper, the need for a 'request for payment' message is discussed. This is just one example of a message that is needed to enable a value-added service – it is not possible to predict in advance what those messages might be.

Of course it is possible to create a parallel network to carry such messages, but doing so is costly and complex. It would be unnecessary if the network and switch were designed upfront to support such messages, but design for flexibility is much more demanding than simply meeting today's requirements.

It's clear that immediate payments is the right platform for the digital age, but it will require good governance and top-flight design to realise its full potential. Icon Solutions is excited by the opportunity it brings and we're looking forward to working with customers to help them leverage the benefits of immediate payments.

About Us

Icon Solutions is a specialised IT consultancy inspired by the vision of delivering simplified solutions for complex technology challenges.

Our experience working on multiple implementations of immediate payments (e.g. UK Faster Payments and Singapore G3 FAST) means that we can identify the key issues and possible solutions.

We can help you draw the map, avoid the obstacles, and guide you on your way.

For more information contact:

Tom Hay
Head of Payments
T: +44(0) 207 147 9955
E: tom.hay@iconsolutions.com

Icon Solutions (UK) Limited,
Tuition House, 27-37 St George’s Road,
Wimbledon, London, SW19 4EU

www.iconsolutions.com