the Availability Digest

www.availabilitydigest.com

Real-Time Fraud Detection

December 2009

Much of today's commerce depends on plastic. People use their credit cards and debit cards to purchase products, get cash, and pay bills. If card services come to a halt, much of today's commerce comes to a halt as well.

Credit and debit cards are issued by banks. Each card transaction must be approved by the issuing bank before it can be accepted by a merchant or an ATM.

The heart of credit card and debit card services is the authorization networks that relay card transactions to the banks that issued the cards. Transaction authorization must be returned in real-time because the customer is waiting for it. Without these networks, a bank cannot authorize card transactions; and card transactions cease (unless a merchant is willing to take a card transaction on faith and enter it when the network is restored).

It is for this reason that many of these authorization networks use continuously available active/active switches to route card transactions to the issuing banks. When a card is used at a merchant's point-of-sale (POS) device, the transaction is sent via an authorization network to the bank that issued the card. Likewise, when a bank servicing its own ATMs receives an ATM request for a card issued by another bank, it uses an authorization network to send the transaction to the issuing bank.

Debit-Card and Credit-Card Authorization

The issuing bank determines whether the transaction should be allowed and returns an authorization or rejection response to the POS device or ATM, which then takes appropriate action. A transaction may be denied for many reasons, including exceeding the account balance for a debit card, exceeding the credit limit or daily limit for a credit card, a card that has been reported to be lost or stolen, or suspected fraudulent activity.

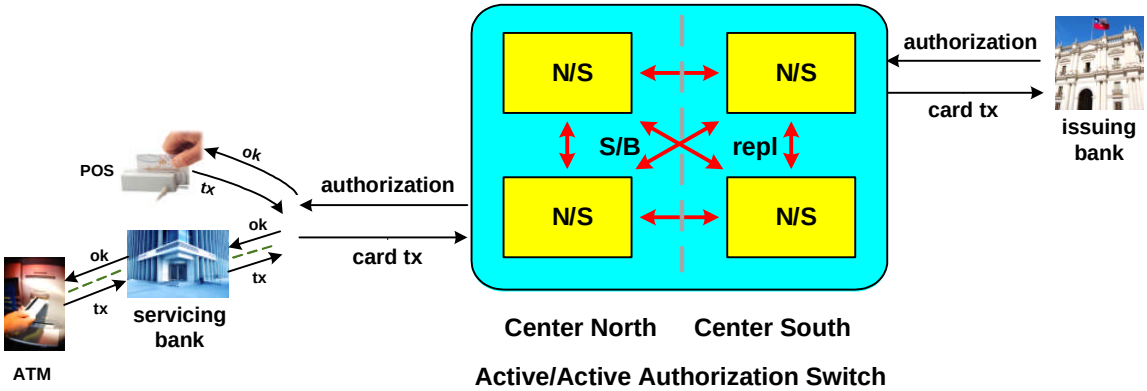
Determining fraudulent activity is a massively complex problem that depends upon a variety of business algorithms. A simple example is a card being used for a purchase at a store in Los Angeles just a few minutes after it was used in New York. Fraud detection typically requires the storage of all recent card transaction activity and a data-mining capability to monitor the transaction history for suspicious activity. As a consequence, it can take hours to determine such

activity and to put a hold on the card, during which time multiple fraudulent transactions may have occurred – all of which may have to be covered by the bank.

A Major Authorization Network

Most fraud detection today is done by the issuing banks, and their systems may require hours or more to detect fraudulent activity. A major provider of card transaction-switching services decided to offer real-time fraud-detection services for its issuing-bank customers. It felt that it could deploy fraud-detection systems that were more powerful, and therefore faster, than could most of its bank customers because it would be sharing the high costs of these systems among many banks. The goal was to detect suspicious activity before a transaction acknowledgement was returned so that suspicious transactions could be denied or could require a call by the customer to the issuing bank before authorization.

However, before we describe the provider's real-time fraud-detection system, let us look at its authorization network.[1] The service provider operates a four-node NonStop active/active[2] switch to route card transactions from merchants and banks servicing ATMs to the banks that issued the cards. For disaster tolerance, the nodes are split between two data centers, one in the Northeast U.S. and one in the Southeast U.S.



Transactions are routed to the switch via the provider's intelligent IP network. POS devices (generally through a common controller operated by the retail merchant) are serviced by a specified node – typically the one that is closest. Online merchants and ATM servicing banks are similarly assigned to specific nodes. By contrast, each issuing bank may have its load distributed across multiple nodes. In this way, the transaction load is spread among the four nodes in the active/active network.

Each node has a configuration database that specifies which merchants, servicing banks, and issuing banks are assigned to it. The nodes are interconnected in a mesh network by Shadowbase bidirectional data replication engines from Gravic, Inc. (www.gravic.com/shadowbase). A database change entered into any one node in the application network is immediately replicated to the other nodes so that all nodes have an up-to-date copy of the configuration database.

The configuration database also performs a very important disaster-recovery function in that it specifies a backup node for each connection should its node fail. Usually, a node fails over to its companion node in the same data center. Should there be a data-center failure in which both nodes are down, failover is to the other data center.

[1] Payment Authorization – A Journey from DR to Active/Active, *Availability Digest*; December 2007.
[2] What is Active/Active? *Availability Digest*; October 2006.

Because of the failover information contained in the configuration database for each node, the data that is replicated is not a one-to-one replication. Rather, the data replicated to each node must be modified to represent the failover sequence for that node. This mapping is done by user exits in the Shadowbase replication engines.

Configuration updates are required when merchants and banks are added or dropped. Furthermore, a merchant or a bank may be moved to another node for load-balancing purposes. Though it is unlikely that two such administrative actions will result in a data collision, it can happen. If this occurs, the later of the two changes is accepted.

Card transactions are not replicated between the nodes since they are transient in nature. Rather, each node batches its transactions and frequently sends transaction batches to the other nodes via the provider's IP network.

Given this architecture, any node can receive a transaction for any card. Large issuing banks have connections to all nodes and receive each transaction from the node that is connected to the ATM or POS device that originated the transaction. If the node receiving the transaction is not the node servicing the issuing bank, the transaction is routed over the IP network to the appropriate switch node, which forwards it to the bank. The issuing bank is determined by the first six digits of the credit or debit card. When the bank returns its authorization response, this response is returned by its node to the requesting retailer or servicing bank.

The architecture of the active/active system is used to advantage to eliminate planned downtime. If a node needs to be upgraded, its connections are assigned to its companion node in its data center. The node is then downed, upgraded, and returned to service. In this way, the upgrade process is within the control of the data center's IT staff.

The provider's switching network provides services for hundreds of thousands of devices. The network will survive any fault in the network, including network components, processing nodes, or an entire data center. Since transactions are automatically failed over to a surviving node following a failure, faults are transparent to the provider's customers; and continuous availability is achieved.

Real-Time Fraud-Detection Service

Fraud Detection – The Old Way

In monitoring fraud detection, most issuing banks provide their own systems. Transactions are written to a log file as they are authorized. Recorded are the card number, the authorization date and time, the location of the POS or ATM activity, and the transaction amount. A separate fraud-detection system periodically inspects this log. Optimized to perform complex analyses on a card's transaction history, the system flags suspicious activity and writes this information to a separate log, which is returned to the authorization system.

The authorization system can then take appropriate action. It may accept the transaction; it may deny the transaction; or, for example, it may require the customer to call a bank representative before the transaction can be authorized.

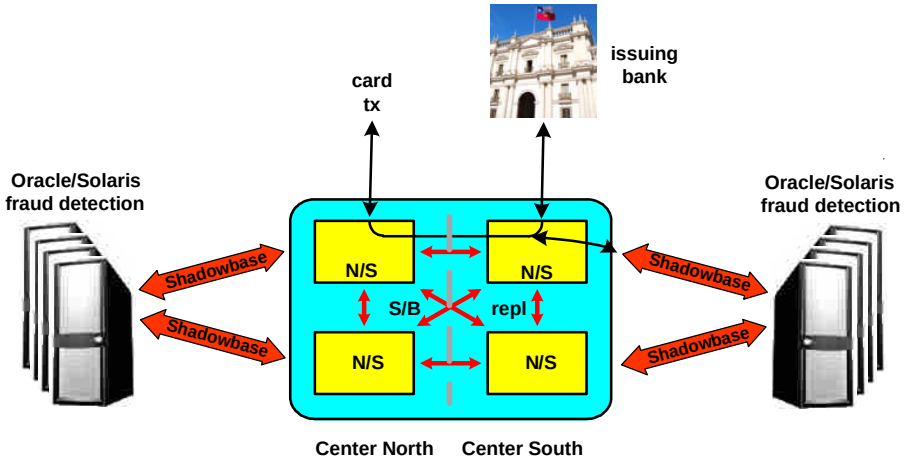
The problem with this method is that it can take hours or even days to detect suspicious activity. In general, the bank is responsible for fraudulent transactions; and this delay can represent a significant cost to the bank.

Catching Fraud On-the-Fly

The transaction switching-service provider realized that catching fraudulent transactions as they were being executed at a retail counter, online, or at an ATM would be a unique and valuable service to offer to its issuing-bank customers, a service that would distinguish its network from other ATM/POS switching networks. Providing the powerful analytical systems to accomplish this complex task in real-time would be expensive, but the costs would be amortized across many issuing banks; and the fraud could be detected sooner.

The provider chose Oracle as the fraud database running on Sun M5000 Solaris servers.[3] The Sun M5000 is a highly scalable server that can contain up to eight SPARC64 quad-core processors. Each data center has its own fraud-detection complex that supports multiple Oracle/Solaris fraud-detection servers. The fraud-detection complex is easily scalable by adding additional fraud-detection servers. Each server within a data center monitors a specified range of card numbers. Should a server fail, another server assumes the card range of the failed server.

As a transaction is received by the switch node that is servicing the issuing bank, the transaction is sent not only to the bank for authorization, but it is also replicated to a local fraud-detection server by a Shadowbase replication engine. Shadowbase determines the fraud server to which to send the transaction according to routing rules based on the card number bound into Shadowbase.



Real-Time Fraud Detection

The fraud server quickly analyzes the transaction for suspicious activity. If it determines that the transaction may be fraudulent, it immediately notifies the switch node of this determination via reverse Shadowbase replication. This notification includes a severity flag that indicates the degree of suspicion determined for the transaction.

The goal is to return the notification before the issuing bank returns its authorization so that real-time action can be taken. If the response from the fraud server is received in time, the switch can take such action as specified by the issuing bank. Based on the severity flag, the switch might allow the transaction, reject it, or require the customer to call a bank representative before the transaction is authorized.

---

[3] Real-Time Credit and Debit Card Fraud Detection: A Shadowbase Real-Time Business Intelligence Solution, Gravic white paper; October 3, 2009.
http://www.gravic.com/shadowbase/pdf/FDC%20-%20Fraud%20Detection%20Case%20Study%202009_10_7.pdf

In any event, the issuing bank is notified of the suspicious activity so that it can notify the card holder and take such other actions on future transactions as it deems appropriate.

Of course, a transaction might pass through any node on its way to the issuing bank depending upon the current failover configuration. Therefore, it is important that both data centers be aware of all recent activity associated with a particular card. To accomplish this, transactions are buffered by each switch node as they are sent to the cards' issuing banks and are transmitted in near-real time to the fraud-detection complex in the other data center. Should one switching node fail, its connections are transferred to the other node in the same data center. However, should both nodes in a data center fail, the fraud systems in the other data center are ready to continue monitoring transactions sent to all issuing banks.

Summary

This real-time fraud-detection system is an excellent example of real-time business intelligence (RTBI).[4] RTBI allows events to control the actions of an enterprise in real time by immediately integrating the independent results of diverse heterogeneous systems into a coherent action.

RTBI will provide the competitive edge to companies in the future. In this example, a NonStop authorization switch, a Solaris/Oracle fraud-detection server, and an issuing bank's mainframe authorization system cooperate to affect the outcome of a transaction request while the transaction is still in progress. This complex interaction is made possible in the example by high-performance, real-time, bidirectional data replication, which is an underlying technology fundamental to real-time business intelligence.

[4] Real-Time Business Intelligence, *Gravic white paper.*
http://www.gravic.com/shadowbase/uses/real-timebusinessintelligenceintroduction.html