CELENT

OPERATIONALIZING FRAUD PREVENTION ON IBM Z16

Reducing Losses in Banking, Cards, and Payments

Neil Katkov

April 5, 2022

This report was commissioned by IBM, which asked Celent to design and execute a Celent study on its behalf. The analysis and conclusions are Celent's alone, and IBM had no editorial control over report contents.

CONTENTS

EXECUTIVE SUMMARY

Advances in artificial intelligence (AI) such as deep learning are enabling significant improvements in fraud detection. However, large banks and payments processors who use AI models often run them on only a fraction of transactions due to throughput and latency constraints with their fraud detection systems. As a result, many fraudulent transactions go unmonitored and undetected.

The IBM Integrated Accelerator for AI, part of IBM's new Telum mainframe processor, is designed to run inferencing for real time workloads at scale and at low latency. The chip is designed to support real time fraud detection even in high-volume bank, card, or payments processing environments.

To help banks and payments processors understand the potential value of this innovation for fraud operations, Celent has developed estimates of the potential reduction in fraud losses if these entities applied AI inferencing to 100% of their transactions.

Quantifiable benefits of AI-based fraud detection on IBM z16 mainframes:

Reduce industry fraud losses by…		Reduce losses per bank by….		Reduce declined card transactions by…
US	Globally	Tier 1 US Bank	Tier 2 US Bank	
5.6¢ per $100	2.0¢ per $100	US$105 million	US$18 million	46%

Celent estimates that applying advanced inferencing models to theoretically all banking, card, and payments transactions running on IBM zSystems mainframes could potentially reduce fraud losses by an estimated US$161 billion globally. In such a case, banks could potentially avoid US$140 billion in losses, and cards and payments could avoid US$21 billion. In the US alone, bank fraud losses could be reduced by a potential US$44 billion and by US$6 billion for cards and payments.

To be sure, there are barriers to adopting AI inferencing on the mainframe for fraud operations, such as model governance issues, rip and replace costs, availability of internal data science resources, and demonstrating the business case.

Still, running advanced AI models directly in the mainframe environment is a powerful innovation in an industry where an estimated 70% of global transaction value runs on IBM mainframes. Fraud detection is an important use case of this new IBM capability, one with demonstrable benefits to both the bottom line and the customer experience.

THE HIGH COST OF FRAUD IN BANKING, CARDS, AND PAYMENTS

Fraud generated an estimated US$385 billion globally in losses to the banking, cards and payments sectors in 2021.

Banking and payments fraud takes many forms across the retail and corporate sectors. Fraud aimed at banks includes account takeover, authorized push payments (APP) fraud, invoice fraud, and a wide range of phishing and social engineering schemes designed to trigger illegitimate money transfers or obtain account credentials. Cards and payments are also vulnerable to account takeover and phishing, as well as specific schemes including synthetic ID, bust-out fraud, and man-in-the-middle fraud.

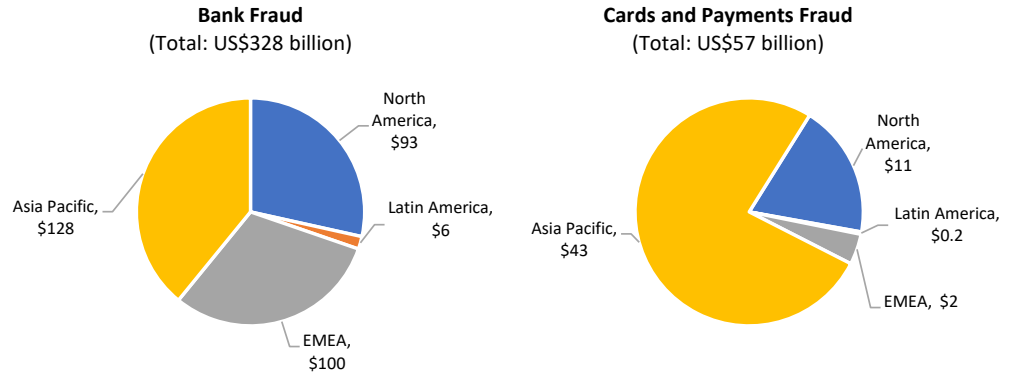
Figure 1: Common Banking and Card Fraud Schemes

Banking Fraud	Card Fraud
Account takeover	Application fraud
APP fraud	Bust-out fraud
Check fraud	Man-in-the-middle
Invoice fraud	Phishing
Social engineering	Synthetic ID

Source: Celent

These and other frauds aimed at bank accounts, cards, and payments are a serious concern to financial institutions. Celent estimates that annual fraud losses average US$209 million for a Tier 1 bank in the US (total assets greater than US$100 billion) and US$35 million for a Tier 2 bank (total assets between US$50 and 100 billion). On an industry scale, banks suffered $328 billion in fraud losses globally in 2021. The cards and payments sectors racked up an additional US$57 billion in losses. All told, fraud generated an estimated US$385 billion in losses to the banking, cards, and payments sectors globally in 2021.

Figure 2: Banking, Cards, and Payments Fraud Losses in 2021



Source: Celent estimates based on BIS transaction data and central bank fraud data.
Note: Bank fraud includes transfers, direct debits, and checks. Cards and payments fraud includes credit and debit cards, e-payments, and other payments.

Although banks and payments processors have been engaged in a decades-long battle to contain fraud with detection systems and chip-based card security, losses have continued to climb as fraudsters stay one step ahead by devising new technology- and social engineering-based schemes.

The COVID-19 pandemic has pushed fraud numbers higher. For banks, a substantial source has been phishing and social engineering schemes exploiting anxieties and medical needs around the pandemic. As for card transactions, the pandemic has led to an increase in digital banking and e-commerce, as consumers have avoided in-branch and in-store transactions. Because card not present (CNP) transactions make up the lion's share of card fraud—around 65%—card fraud losses have increased.

Help on the Way: Deep Learning-Based Fraud Models

Advancements in artificial intelligence, like deep learning, now give banks the tools to fight fraud much more effectively by analyzing data at scale to find patterns that point to fraud, including new, previously unseen typologies.

Deep learning is a type of machine learning model based on a deep neural network (DNN). A DNN consists of computational nodes, or neurons, that use progressive weights to strengthen connections between the nodes. The nodes are arranged in multiple layers—making a "deep" network—that increase the capacity and learning rate of the model. Deep learning models are trained on existing data, such as historical transactions in the case of fraud models. The trained model is then executed on live data, such as a real time transaction, to generate a result, or inference. In the case of fraud models, the inference is typically a score expressing the likelihood that the transaction is fraudulent.

Based on industry conversations and research, Celent estimates that AI inferencing based on deep learning models can increase the accuracy of fraud detection by 60% over existing fraud models.

The potential of inferencing to improve fraud rates is drastically limited, however, by the fact that in high-volume, mainframe environments, these models are often run on only a fraction of transactions—less than 10%—due to latency, cost, and customer friction issues. This means that roughly 90% of potentially preventable fraud is still going undetected. This severely limits the ability of banks to take advantage of AI advances to claw back fraud losses.

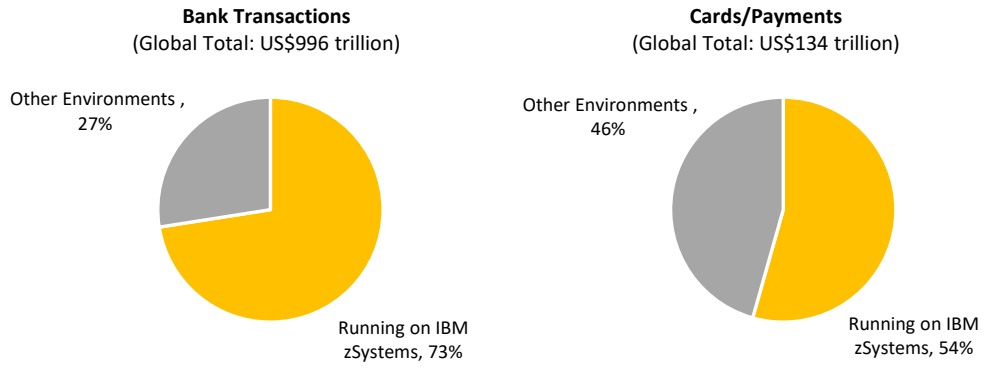
The latency and cost barriers to passing 100% of bank and card transactions through advanced models may now be a thing of the past. The new IBM z16 Telum processor contains an AI accelerator that, in a first for IBM zSystems, can run AI models directly on the chip, in real time. This exponentially improves throughput and response times, making it possible, for the first time, to pass virtually all transactions through deep learning-based fraud detection models.

LIMITATIONS OF STATUS QUO FRAUD DETECTION

Typical fraud detection technology and operational approaches for mainframe environments include running fraud on off-platform systems on selected transactions and/or on a post-transaction basis. This drastically limits the ability of banks and payments processors to run advanced AI models on all transactions.

Many large banks and payments processors run their core systems on mainframe computing environments. IBM estimates that 45 of the top 50 banks globally are running on IBM zSystem mainframes. Most of the major cards and payments processors also run on the platform. Globally, Celent estimates that 70% of bank, cards, and payments transaction value runs on IBM zSystems environments.

Figure 3: Bank, Cards, and Payments Transaction Value on IBM zSystems



Source: Celent

The latency between core systems and off-platform detection systems can be tolerated for some transactions. However, in the case of data-intensive AI inferencing routines applied to real time transactions—such as real time payments, card transactions, and digital banking transactions—latency makes it impractical to pass all transactions through an AI detection platform in high-volume environments. When core system transactions are sent off the mainframe to an off-platform detection system for real time analysis, response times to receive the detection results range from 50 to 80 milliseconds—while the transactions are waiting. This slows down approval times for transactions, which can create customer friction, particularly for card transactions.

More fundamentally, high latency can make it impossible to run all transactions through an off-platform fraud detection system. Latency between the core system and the detection software can delay the core's receipt of detection results to the

extent that real time transactions time out. As a result, some banks only run deep learning models for fraud on a post-transaction basis.

As a result, banks send only a fraction of transactions—less than 10%—through their fraud detection engines in real time. There are serious consequences to this approach. Deep learning models are now enabling significant improvements of about 60% in detection rates. However, banks are not reaping the full benefit because they are running only a sampling of transactions through these models. This means that a higher proportion of fraud will go undetected, increasing fraud losses. As fraud becomes a focus of financial crime compliance, banks may face regulatory risk as well if they are not able to pass all their transactions through anti-fraud detection.

Legacy issues at a Tier 1 US bank

A Tier 1 bank in the US running its core system on an IBM zSystems platform has deployed an off-platform AI-based fraud detection system. Due to cost and latency issues, the bank runs only very high-risk transactions through the AI system. Most transactions are run through rules-based scoring, approved as a convenience to the customer, and then subjected to post-transaction analysis after the fact. The benefits of AI are severely restricted by the inability to run the models on all transactions, meaning that AI is not used to its full potential.

REDUCING FRAUD LOSSES WITH AI INFERENCING ON THE MAINFRAME

IBM has developed a processor for its IBM z16 mainframe computer featuring an accelerator for AI that is designed to run advanced inferencing directly on the chip, at scale. Celent estimates that the new IBM z16 processor can support deep learning-based fraud detection for virtually all transactions, potentially reducing banking, cards, and payments fraud losses by US$161 billion globally.

Deep learning algorithms tend to be more compute intensive than legacy fraud models. As banks implement deep learning-based AI inferencing for fraud, they face challenges in managing these mission critical workloads. When detection is performed on off-platform systems, detection response times can reach upwards of 80 milliseconds, with throughput rates in the 1,000–1,500 transactions per second (tps) range.

Due to these latency and throughput limitations, banks have experienced transactions timing out while they wait for detection results. These and other issues lead banks to send only a fraction of transactions—less than 10%—through their detection engines.

Deep Learning on the Mainframe

Based on a credit card fraud deep learning model, 32 IBM Telum chips running on a single server can deliver up to 3.5 million inferences per second with 1.2 millisecond average response time.

Source: IBM microbenchmark, August 2021

DISCLAIMER: Performance result is extrapolated from IBM internal tests.

IBM has developed an accelerator for its IBM z16 mainframe computer that can run AI inferencing models directly on the chip. According to IBM, the throughput and improvements of running AI models on the mainframe are sufficient to support real-time fraud analysis of virtually all transactions in even high-volume bank, card, or payments processing environments.

Moreover, this can be done with virtually no impact on transaction processing times. IBM claims that its IBM Integrated Accelerator for AI, part of its new Telum processor, can run AI models on the mainframe with a very fast response time of only 1.2 milliseconds for each inference request. In the specific case of card fraud detection, early benchmarks indicated that a configuration of 32 Telum chips can support up to 3.5 million inferences per second.

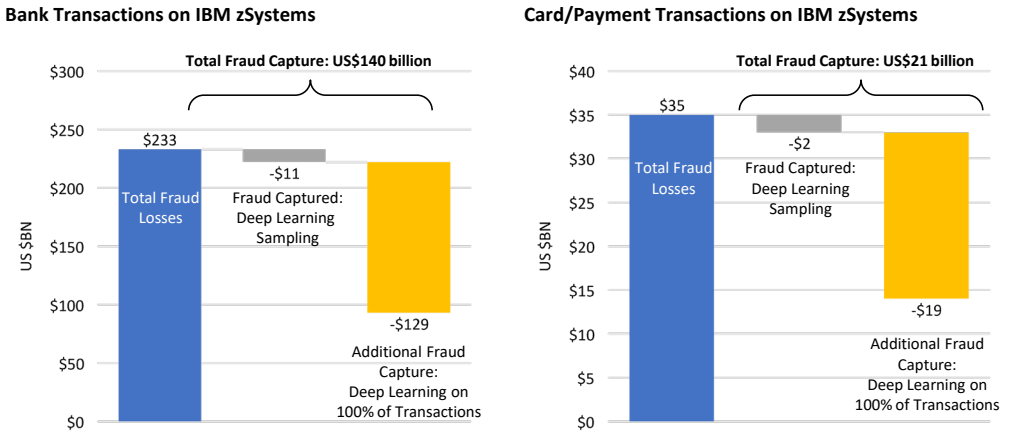
This is scale enough to support even peak transaction flows, making it possible for banks and payments processors to run virtually all transactions through deep learning models.

Banks and card and payment processors can reap the full potential of modern inferencing technology by running advanced models against all transactions. Celent estimates that applying advanced inferencing models to all transactions would potentially reduce fraud losses by 2.0¢ cents for every $100 of transactions globally (2.0 basis points).

In the US, where fraud rates are higher than the global average—9.3¢ for every $100 compared to 3.7¢ globally—fraud losses could be reduced by 5.6¢ for every $100. This is equivalent to saving the bank US$1.33 for an average transaction of US$2,375.

Celent estimates that, theoretically, passing all transactions currently running on IBM zSystems through deep learning models could potentially reduce fraud losses by US$161 billion globally. Banks could avoid US$140 billion in fraud losses; cards and payments could avoid US$21 billion. In the US alone, the potential for fraud loss reduction is US$44 billion for banks and US$6 billion for cards and payments.

Figure 4: Potential Fraud Loss Reduction with Deep Learning Models



Source: Celent

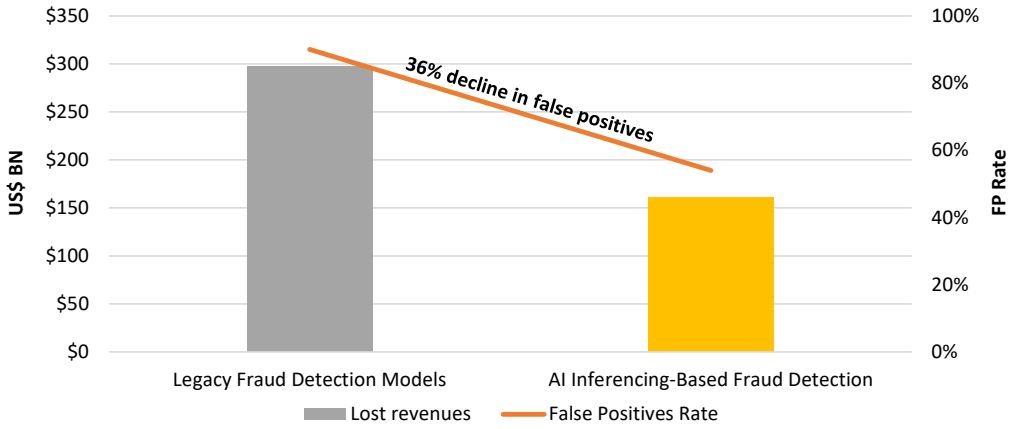
Celent estimates that for a Tier 1 bank on IBM z16, running all transactions through advanced inferencing models—compared to the current best practice of applying AI models to only about 10% of transactions—could reduce fraud losses by an additional US$105 million. A Tier 2 bank could avoid an additional US$18 million in losses. Running all transactions through advanced models would also improve the models themselves. More transactions would produce more data to train the models, resulting in greater accuracy in fraud detection.

REIGNING IN FALSE POSITIVES TO REDUCE CUSTOMER FLIGHT

Legacy anti-fraud models have very high false positive rates—typically 90% of all alerted transactions or higher—that lead banks to reject legitimate transactions. Rampant false positives and denied transactions not only create customer friction but result in hard dollar losses as customers simply pull out their next credit or debit card to make a purchase. Celent estimates that declined credit card transactions cost the industry US$298 billion in lost fee revenue globally.

The need to balance anti-fraud efforts with minimizing customer friction is another reason banks limit fraud detection routines to a sample of all transactions. False positives occur when legitimate transactions are incorrectly flagged by detection software as fraudulent. The increased accuracy of deep learning models can significantly improve the industry's very high false positive rates. This in turn would reduce the number of erroneously rejected transactions. This improves the customer experience and reduces lost revenues due to customer flight. This also means that banks can pass all their transactions through fraud detection with less damage from customer friction.

Figure 5: Deep Learning Models Improve False Positive Rates



Source: Celent

Deep learning models applied to every card transaction could improve false positive rates to around 55%. While still very high, this could potentially result in a reduction in lost card fee revenue by US$137 billion to $161 billion globally.

Fewer false positives would have other benefits too. Fraud analysts would need to work fewer alerts, thereby reducing costs for post-transaction investigation. In terms of reputational benefits, the reduction in customer friction and frustration would enhance goodwill and customer trust.

Advanced models can also lead to improvements in the detection of suspicious behavior that may indicate money laundering. The Bank Secrecy Act in the US, the EU Anti-Money Laundering Directives, and other regulations put banks' anti-money laundering programs (AML) under intense scrutiny by regulators. Regulators in the US are particularly active in citing banks for inadequate AML programs, with fines against some banks exceeding US$1 billion. AML operations also suffer from very high false positive rates, typically more than 95%, which imposes a severe operational burden on banks. Moreover, AML monitoring is typically performed on a post-transaction basis, which subjects banks to increased risk. Leveraging AI-based models for AML operations can help with such issues by improving the accuracy of AML behavior detection and reducing false positives.

PATH FORWARD

Our analysis points to significant, quantifiable benefits from running deep learning models on up to 100% of transactions. IBM claims that its new accelerator can support this for transactions running on IBM z16 mainframes, even in extremely high-volume environments. There remain, however, a number of factors to consider for banks and processors that are taking the leap.

As banks and card and payments processors weigh the advantages of implementing deep learning-based fraud detection on the mainframe, Celent recommends they consider issues such as the following:

- **Model governance.** Regulators and internal auditors require strong governance around fraud models. This means that AI models must be transparent and explainable. While AI platform vendors are generally moving away from "black box" approaches, governance of AI models remains a complex undertaking.
- **Regulatory resistance.** Regulators are comfortable with traditional rules-based detection but are less familiar with advanced deep learning techniques. Banks, data scientists, and their vendors may in some cases need to educate regulators on the efficacy and reliability of advanced AI as they move forward.
- **Cost of replacement.** Many institutions have already implemented AI-based fraud detection systems. Such firms will need to develop the business case for moving detection to the mainframe, including deciding whether to maintain existing systems in some form—for example, to support post-transaction analysis or smaller lines of business—or to scrap them entirely.
- **Data science resources.** IBM's Integrated Accelerator for AI is optimized to execute models, including models built with open source frameworks such as Pytorch and TensorFlow. However, it has not yet been demonstrated to support packaged fraud detection software—although we expect some fraud vendors will eventually step up with packages that can run on the accelerator. Either way, institutions moving AI-based detection to IBM z16 will need the data science capabilities to develop and support advanced deep learning models for fraud, either internally or through specialist model providers.

Financial institutions will want to consider these factors carefully—and do their due diligence on IBM's new AI accelerator. Still, the potential benefits in terms of fewer losses from fraud and declined transactions, as well as reduced friction and improved customer experiences are compelling. Firms running IBM zSystems should take a close look at what might be gained from moving fraud detection to the mainframe.

LEVERAGING CELENT'S EXPERTISE

If you found this report valuable, you might consider engaging with Celent for custom analysis and research. Our collective experience and the knowledge we gained while working on this report can help you streamline the creation, refinement, or execution of your strategies.

Support for Financial Institutions

Typical projects we support include:

Vendor short listing and selection. We perform discovery specific to you and your business to better understand your unique needs. We then create and administer a custom RFI to selected vendors to assist you in making rapid and accurate vendor choices.

Business practice evaluations. We spend time evaluating your business processes and requirements. Based on our knowledge of the market, we identify potential process or technology constraints and provide clear insights that will help you implement industry best practices.

IT and business strategy creation. We collect perspectives from your executive team, your front line business and IT staff, and your customers. We then analyze your current position, institutional capabilities, and technology against your goals. If necessary, we help you reformulate your technology and business plans to address short-term and long-term needs.

Support for Vendors

We provide services that help you refine your product and service offerings. Examples include:

Product and service strategy evaluation. We help you assess your market position in terms of functionality, technology, and services. Our strategy workshops will help you target the right customers and map your offerings to their needs.

Market messaging and collateral review. Based on our extensive experience with your potential clients, we assess your marketing and sales materials—including your website and any collateral.

RELATED CELENT RESEARCH

Remaking Risk: A Taxonomy of Regtech
October 2021

Technology Trends Previsory: Risk, 2022 Edition
October 2021

IT and Operational Spending in AML-KYC: 2021 Edition
December 2021

IT and Operational Spending on Fraud: 2021 Edition
February 2021

Innovation In Risk: A Snapshot Through the Lens of Model Risk Manager 2021
April 2021

Fino Payments Bank: Remote Implementation of Enterprise-Wide Fraud Management During the Pandemic
March 2021

Swedbank: Modernizing Card Fraud Management and Improving Customer Experience
March 2021

COPYRIGHT NOTICE

For more information please contact info@celent.com or:

Neil Katkov nkatkov@celent.com

Americas

USA
99 High Street, 32nd Floor
Boston, MA 02110-2320

+1.617.424.3200

USA
1166 Avenue of the Americas
New York, NY 10036

+1.212.345.8000

USA
Four Embarcadero Center
Suite 1100
San Francisco, CA 94111

+1.415.743.7800

Brazil
Rua Arquiteto Olavo Redig
de Campos, 105
Edifício EZ Tower – Torre B – 26º andar
04711-904 – São Paulo

+55 11 3878 2000

EMEA

Switzerland
Tessinerplatz 5
Zurich 8027

+41.44.5533.333

France
1 Rue Euler
Paris 75008

+33 1 45 02 30 00

Italy
Galleria San Babila 4B
Milan 20122

+39.02.305.771

United Kingdom
55 Baker Street
London W1U 8EW

+44.20.7333.8333

Asia-Pacific

Japan
Midtown Tower 16F
9-7-1, Akasaka
Minato-ku, Tokyo 107-6216

+81.3.6871.7008

Hong Kong
Unit 04, 9th Floor
Central Plaza
18 Harbour Road
Wanchai

+852 2301 7500

Singapore
138 Market Street
#07-01 CapitaGreen
Singapore 048946

+65 6510 9700