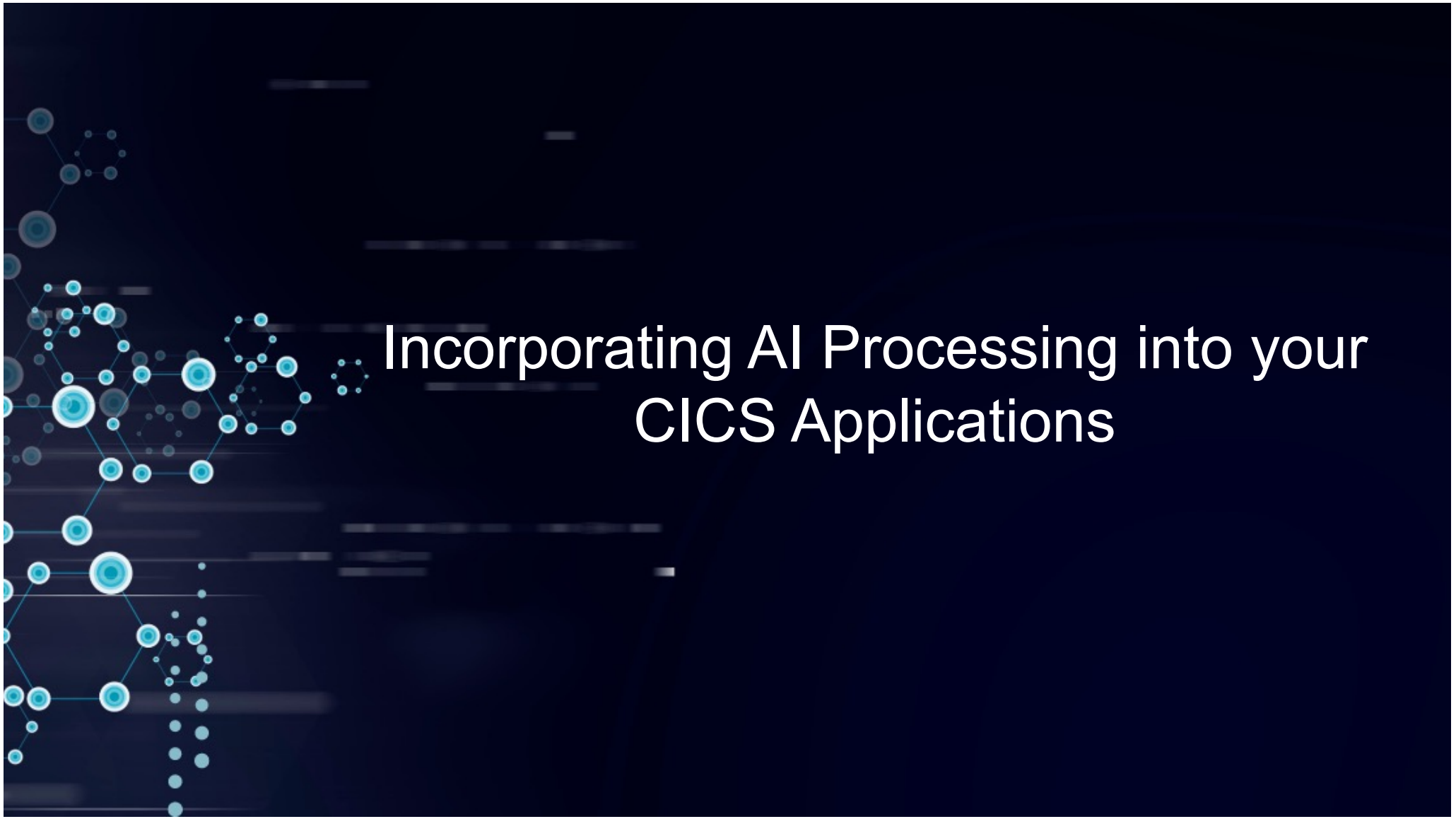

Incorporating AI Processing into your
CICS Applications

Notices and disclaimers

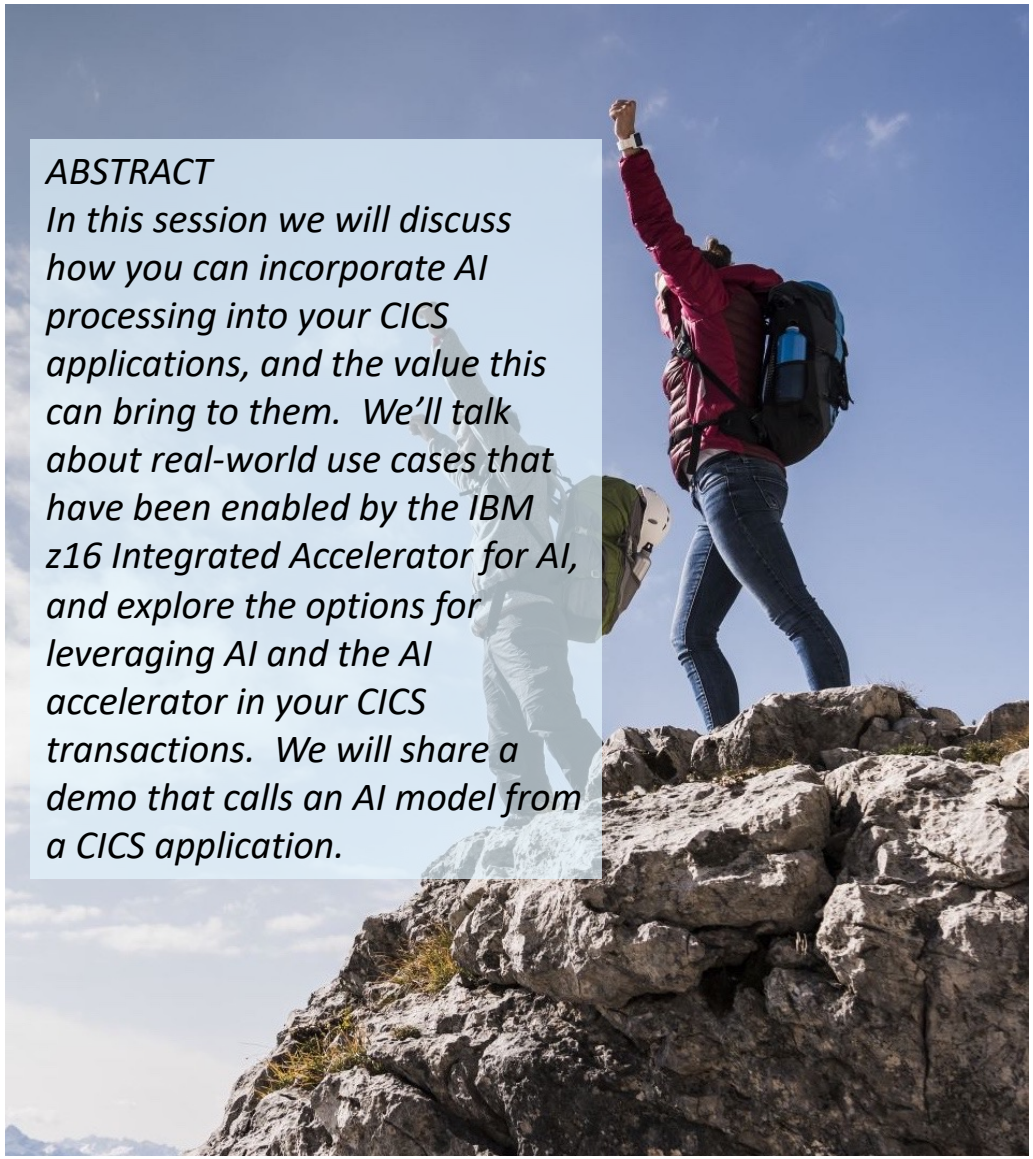
- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.
- IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
- **Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**
- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those
- customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.
- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.
- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.
- It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers

- Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**
- The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.
- IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml

Goals for the session

- Give you understanding of the possibilities for infusing AI in your CICS applications:
  - What we mean by ‘infusing AI’
  - What we mean by ‘AI’
  - Why we are talking about this now
- Discuss use cases and ways you could benefit from infusing AI in your applications
- Discuss what is involved in infusing AI in CICS applications
- Show you a demo of calling AI from CICS
- Answer your questions

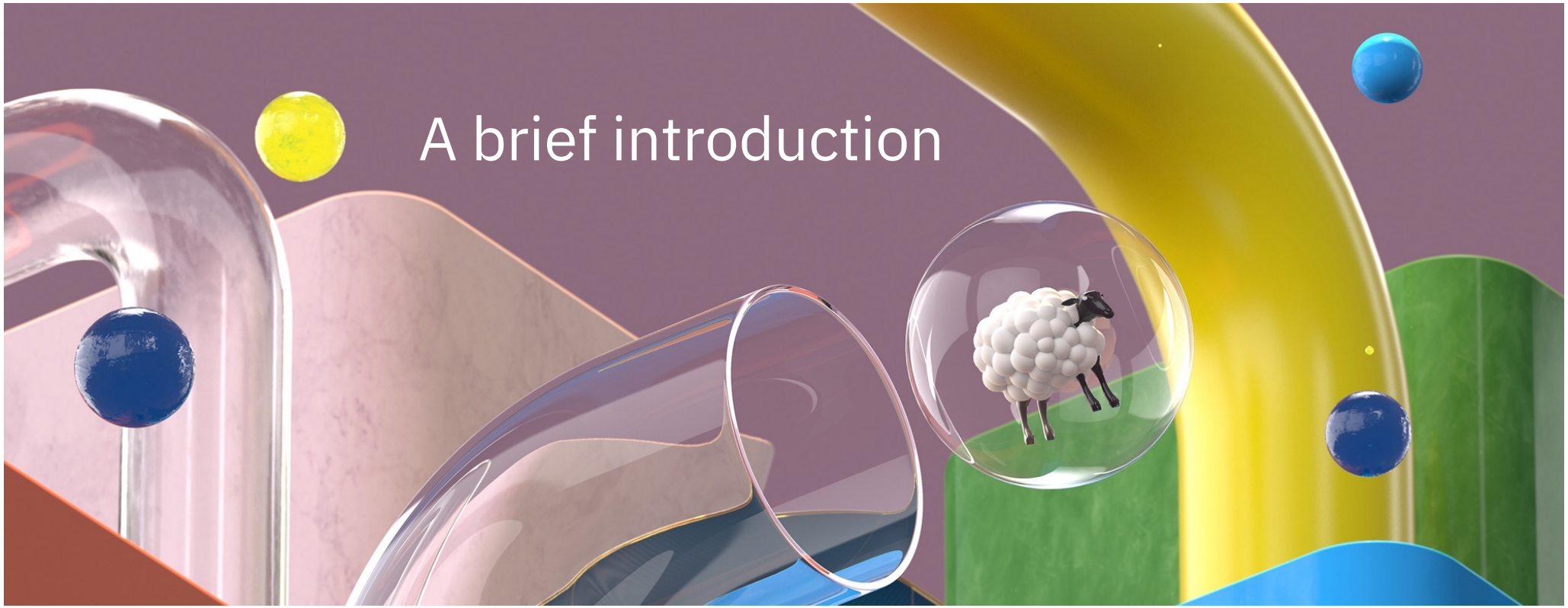


Scope of the session

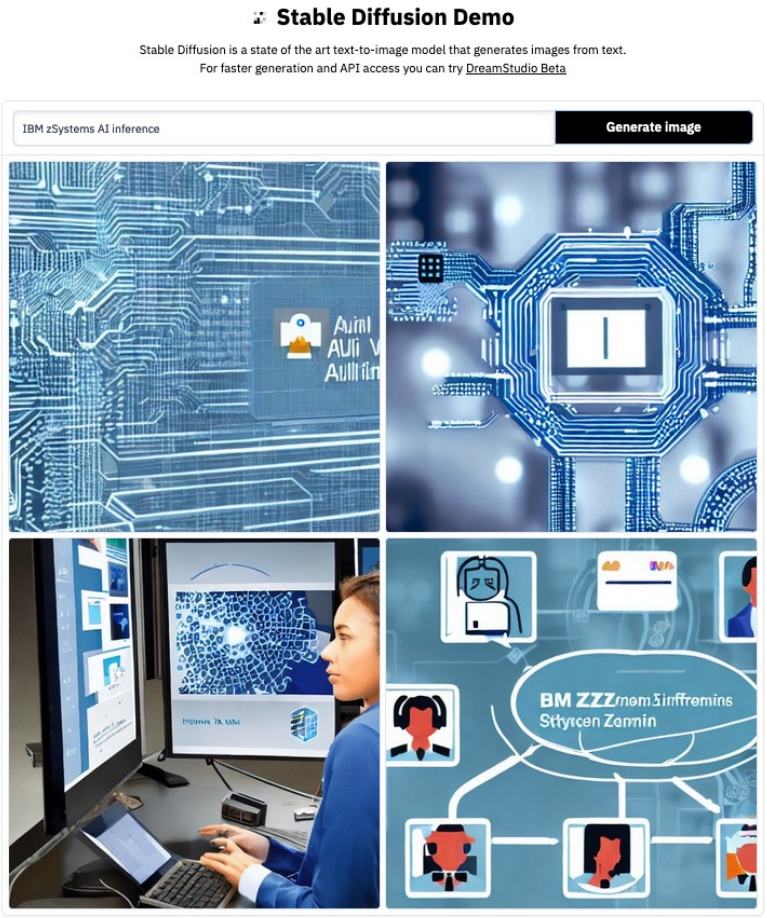
- We are focusing on enabling AI for business
  - Enabling your business to leverage Artificial Intelligence in your applications and transactions
  - Help you to gain better insights and achieve better outcomes by ‘infusing’ AI in your CICS applications

- This is as opposed to other important uses of AI, such as to improve the operations and management of your systems
  - AIOps is an important space with a lot of capability on zSystems
    - Such as IBM Z Operations Analytics, Common Data Provider for z/OS, Watson AIOps & IBM Z Anomaly Analytics, Db2 AI for z/OS, Data Privacy for Diagnostics, z/OS Workload Interaction Correlator and Workload Interaction Navigator and more
  - AI can help improve many aspects of IBM zSystems environments and operations
    - but that’s not the focus for this session

What is artificial intelligence?



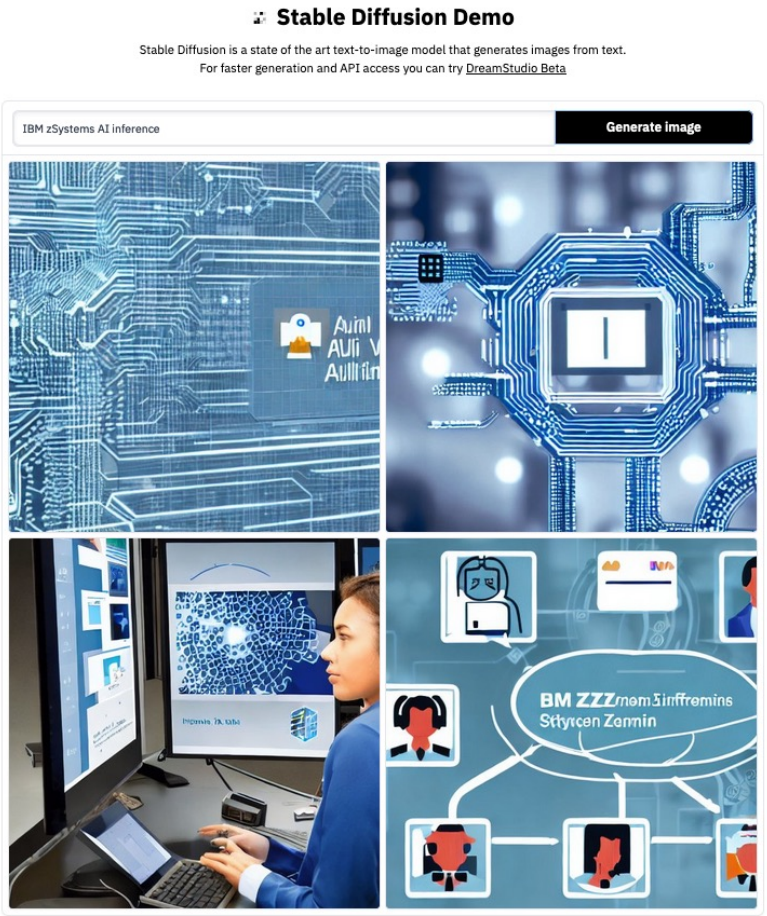
An example of artificial intelligence



Model: Stable Diffusion

“IBM zSystems AI inference”

An example of artificial intelligence



Model: Stable Diffusion

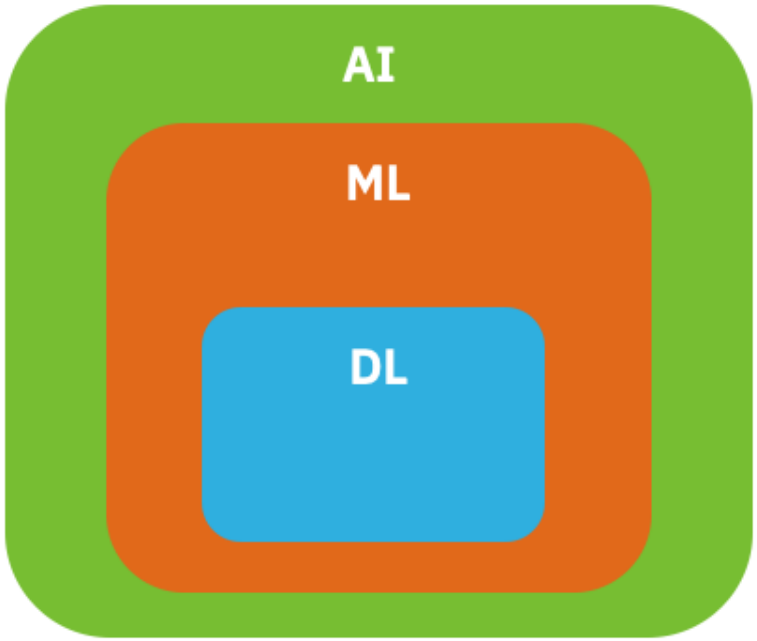
“IBM zSystems AI inference”

Text-to-Image model that generates images from text

- The model was trained on the LAION-5B dataset, which scraped non-curated image-text-pairs from the internet (the exception being the removal of illegal content) and is meant for research purposes.

https://huggingface.co/spaces/stabilityai/stable-diffusion

Categories of artificial intelligence



AI – Artificial Intelligence

- ANI – Artificial Narrow Intelligence
  - Reasoning, Planning, Decision Making
  - Natural Language Processing
- AGI – Artificial General Intelligence
- ASI – Artificial Superintelligence

ML – Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

DL – Deep Learning

- Neural Networks (consisting of more than 3 layers)

- Inferencing: using a trained AI model to make a prediction
- Infusing AI: Applying AI across the enterprise, drawing on predictions, automation, and optimization to improve business decisions and outcomes; operationalizing AI as part of business processing

Why now is the right time to start your AI journey

IBM z16™ is built to build

We built a powerful and secure platform for business. Let's build the future of yours.

Predict and Automate for Increased Decision Velocity

Secure with a Cyber Resilient System

Modernize with Hybrid Cloud

Predict and automate for increased decision velocity

Accelerated AI for insights at scale

80%

of respondents agreed that real-time insights are important[1]

How do you leverage technology to help infuse AI in transactions in real-time and at scale?

49%

getting insights where and when they are needed is a big challenge[2]

How do you apply actionable insights at the right place and at the right time?

Prevent fraud before it happens by scoring up to 100% of transactions in real-time without impacting SLAs

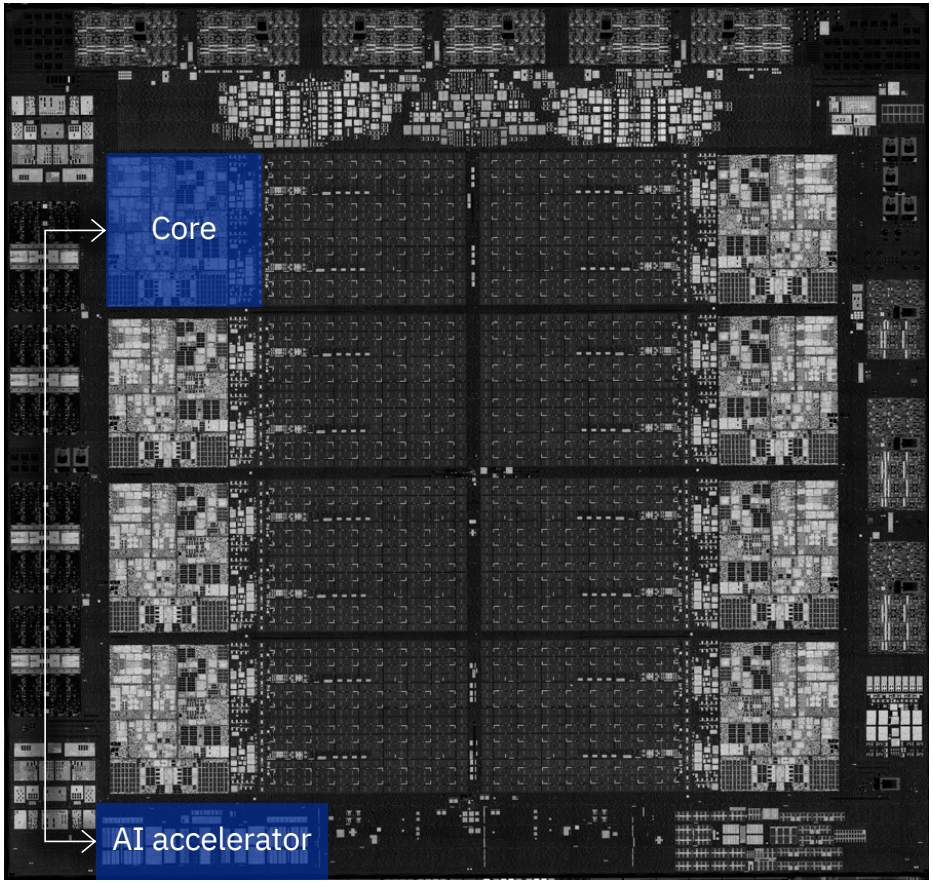
Insights at unprecedented speed and scale mean every customer interaction can now be a personalized experience

1. Forrester: Leverage Data Where It Originates To Drive Substantial Business Benefits Real-time insights are critical to firms' top initiatives, 2020 https://www.ibm.com/downloads/cas/ZEOENRB1
2. https://filecache.mediaroom.com/mr5mr_ibmnews/190846/IBM%27s%20Global%20AI%20Adoption%20Index%202021_Executive-Summary.pdf ; Global AI Adoption Index 2021

IBM z16 & IBM Telum processor

- IBM Telum Processor and Integrated Accelerator for AI
  - Flexible on-chip AI accelerator that works in conjunction with standard cores
  - Fast direct storage access through a gateway for the on-chip L3 cache
- Every CP chip in the IBM z16 system has one Integrated Accelerator for AI built in
- Variety of AI models ranging from traditional Machine Learning to Deep Learning

“IBM ... developed an integrated accelerator, an industry first for data center hardware.”
– Peter Rutten, IDC



Using AI

- What are some of the use cases for infusing AI into CICS and z/OS applications?

Fraud detection

A large bank leverages AI on zSystems to detect fraud patterns and prevent fraudulent transactions. The bank was able to scale on zSystems to examine EVERY transaction in real time. The result was reduced fraud, saved costs, and increased customer satisfaction.

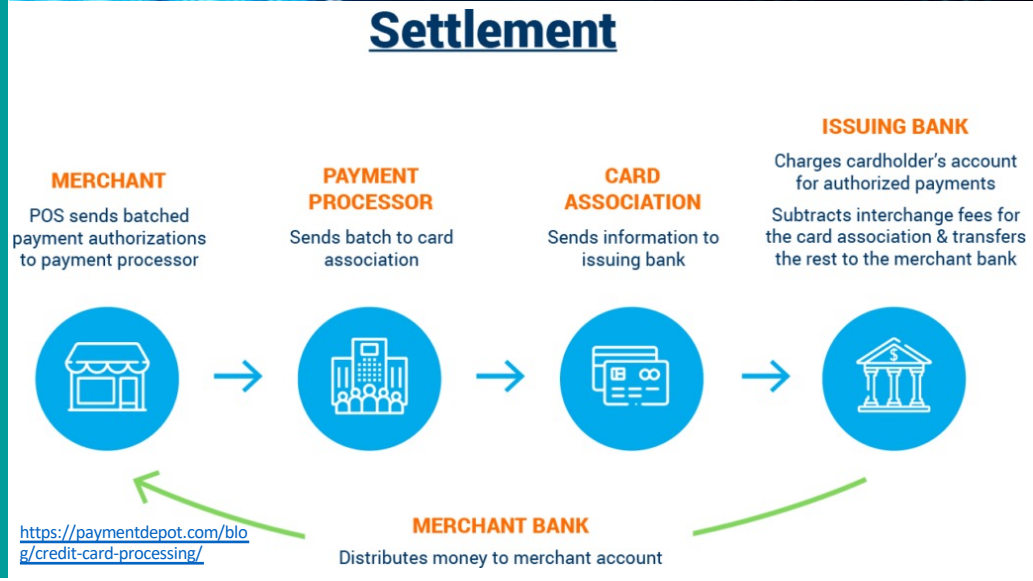
- Achieved consistent response times as well as lower response times – went from >50 ms to 1 ms
- Enabled scoring of 100% of transactions – went from being able to score 1800 TPS to 20,000 TPS

Problem Addressed:
Client discovered that when AI inferencing is performed off platform, they must take unwanted risk and approve transactions without additional fraud checking due to timeout issues.

Clearing and settlement

Leverage AI for Risk Mitigation
Use AI to predict which trades or transactions have high risk exposures and propose solutions for a more efficient settlement process. The expedited remediation of questionable transactions can prevent costly consequences, regulatory violations, and negative business impact.

- No impact to SLAs and batch process window
- Proactively stop losses, lower operational, regulatory, and compliance costs
- Solution is using TensorFlow for high performing low latency scoring



Lots of other use cases, such

MEDICAL IMAGES

Business Challenge
A Health Insurer needs to analyze large volumes of medical records, including training computer vision models, while optimizing for energy efficiency.

Business Impact
Replace x86 server environments with LinuxONE to reduce energy consumption by x%

GEOSPATIAL ANALYSIS & NLP

Business Challenge
A European land registry needs to determine which buildings have been added to, modified, or demolished for land surveys and tax purposes. Also embedding NLP into chat services.

Business Impact
Deploy on IBM z16 to ensure security of data, and optimization of IT and operational operational costs

LOAN APPROVAL

Business Challenge
A large lender needed to speed up their consumer loan approval process, in order to maintain client satisfaction

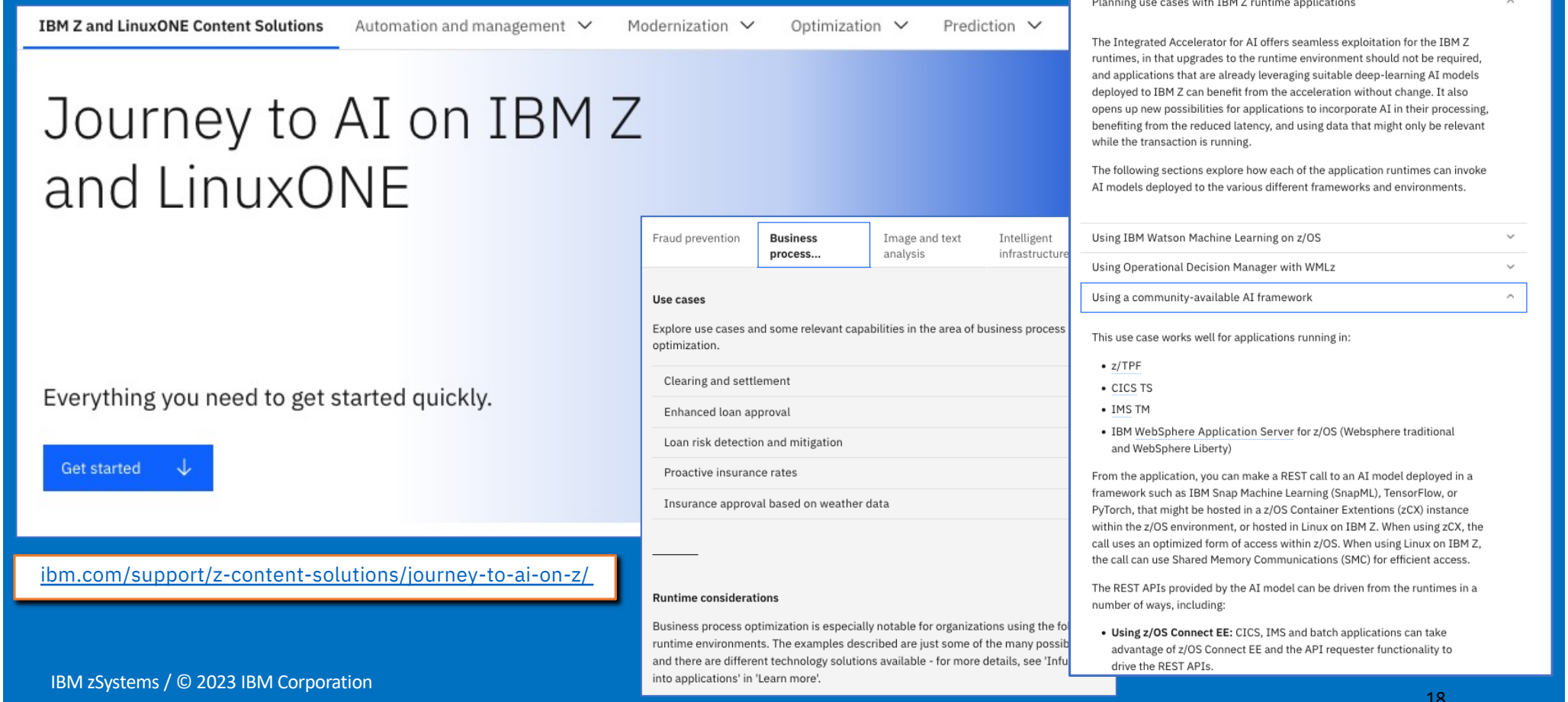
Business Impact
Reduction in latency of over 1000x, with corresponding decrease in lender risk and exposure.

CLAIMS FRAUD

Business Challenge
A state government in the US realized that their process to determine fraudulent claims was manual & intensive and could not scale, taking up to 40 hours per case

Business Impact
Now able to process claims and detect fraud in under 1 minute per case and allocate resources to higher value tasks

More use case ideas in the ‘Journey to AI’ content solution



ibm.com/support/z-content-solutions/journey-to-ai-on-z/

Infusing AI into applications

Infusing AI:

- Applying AI across your enterprise, drawing on predictions, automation, and optimization to improve your business decisions and outcomes
- Operationalizing AI as part of your business processing

What's needed for successful infusion of AI?

Kathryn
Line of Business Manager

Marcus
Data Scientist

Julian
Middleware Systems Programmer

Cara
Chief Data Officer

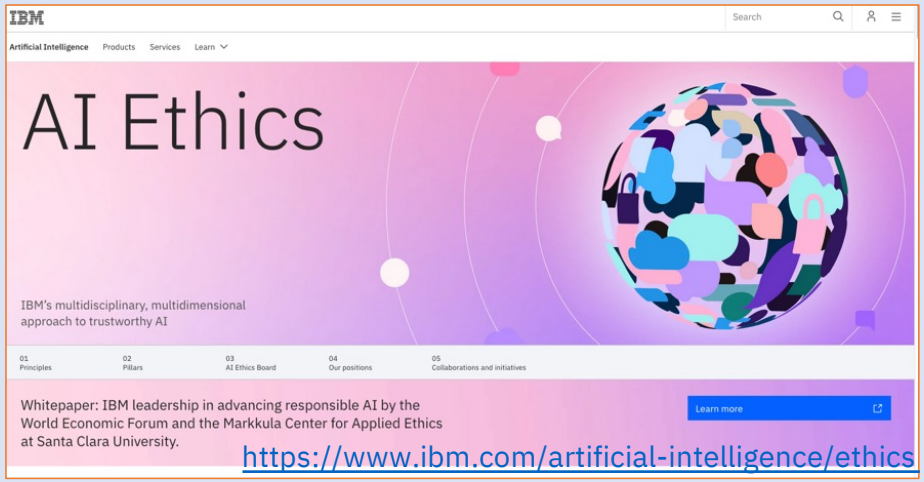
Eric
Application Developer

Toby
Early tenure z/OS Application Tester

- This is a team sport
  - AI/data science team, enterprise architecture, applications team, infrastructure team, DevOps,…
- Identify use cases where in-transaction AI adds value
  - Enhance existing rules/analytics
  - Remove need for off-platform calls
  - Something that wasn't possible before
- Identify relevant data and how it can be accessed
  - From the application
  - From elsewhere
- Build/train AI Model
  - Could be done off-platform, using data scientists' favorite tool
- Determine where the model will be deployed, how it will be invoked
- Consider the lifecycle for the model, ethical considerations, monitoring its effectiveness, retraining, etc

Ethical use of AI

IBM is at the forefront of ensuring that AI is applied in an ethical way, with transparency and without bias

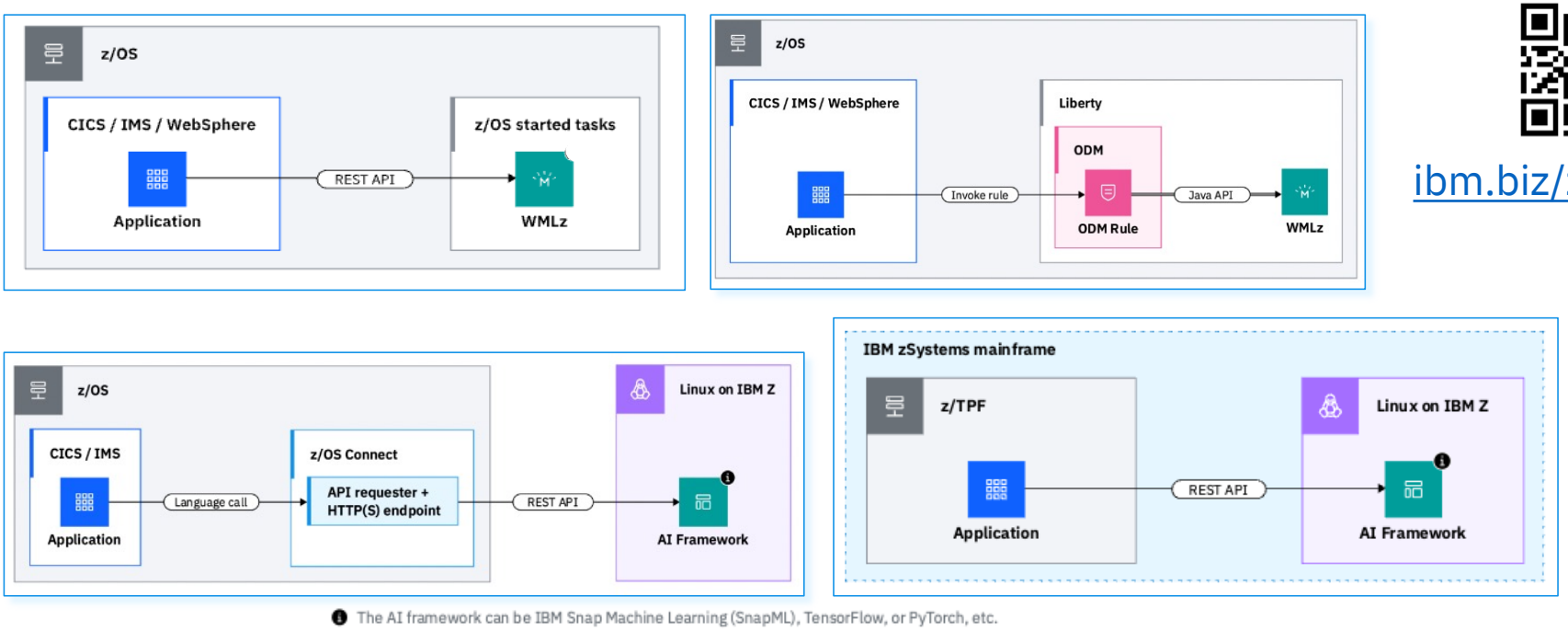


https://www.ibm.com/artificial-intelligence/ethics

Tool kits from IBM Research include AI Explainability 360 and AI Fairness 360

- **Principles for Trust and Transparency**
  - The purpose of AI is to augment human intelligence
  - Data and insights belong to their creator
  - Technology must be transparent and explainable
- As you infuse AI into applications, consider how to ensure *Explainability, Fairness, Robustness, Transparency, Privacy (5 Pillars of Trust)*
  - Model training and the data used as input
  - Inferencing in applications and how the predictions and recommendations are used
- See AI Design Ethics for some pointers

Guidance document:
Planning AI infusion into applications on IBM zSystems



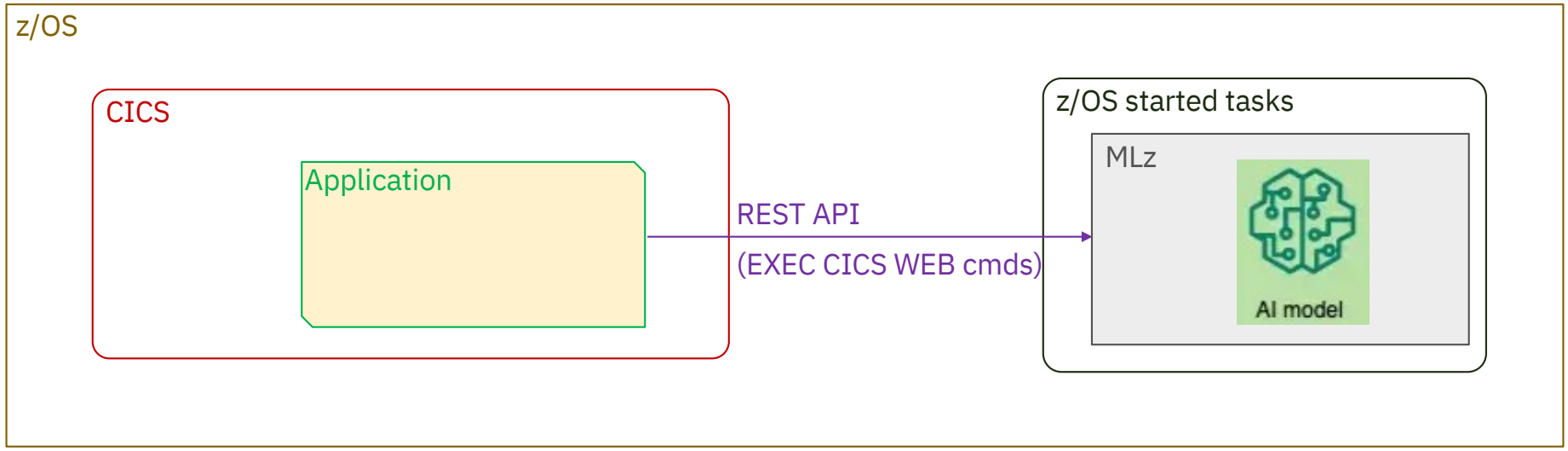
The AI framework can be IBM Snap Machine Learning (SnapML), TensorFlow, or PyTorch, etc.

ibm.biz/zInfuseAI

Infusion Options

Options for CICS applications to invoke AI models and gain insights in their transactions

Using IBM Machine Learning on z/OS with REST



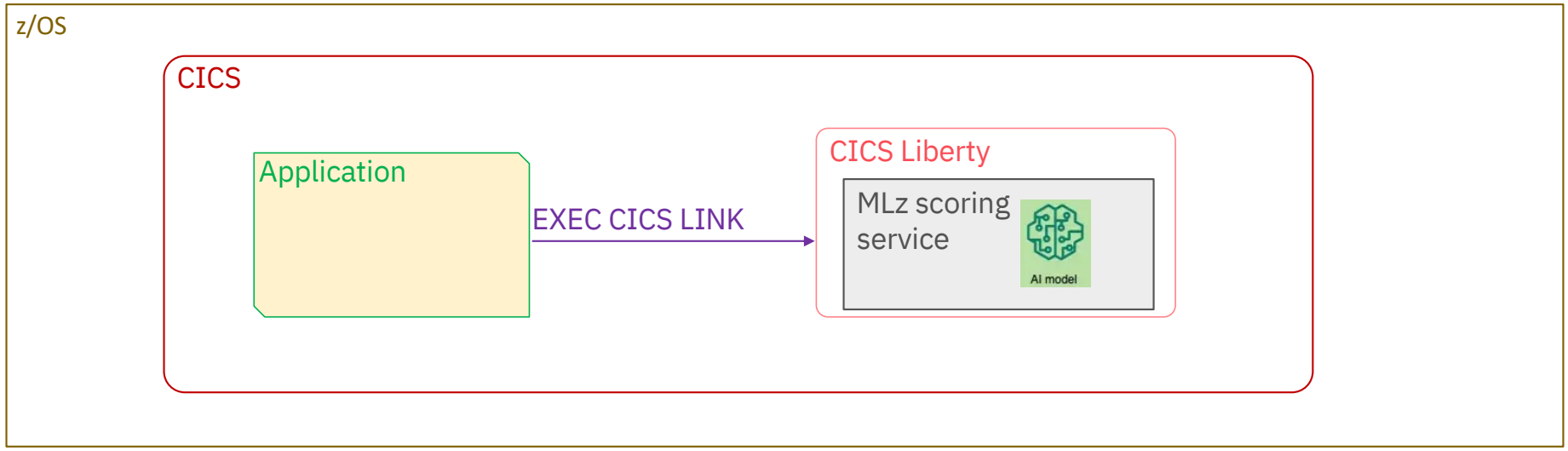
Notes

- MLz base uses IzODA, Spark, and optionally Anaconda and MDS (Mainframe Data Service)
- MLz also provides an Online Scoring Community Edition (OSCE): can be deployed to zCX as a no-charge option to try out the streamlined up-and-running IBM MLz inferencing in-transaction approach

Could be a good choice when

- You want to take advantage of MLz and the full-function solution it offers
- Using models developed using Spark, Scikit-learn, PMML, XGBoost, and ARIMA
- Using ONNX models (which MLz can compile on import using IBM DLC)
  - On IBM z16 with MLz 2.4, can exploit AI Accelerator

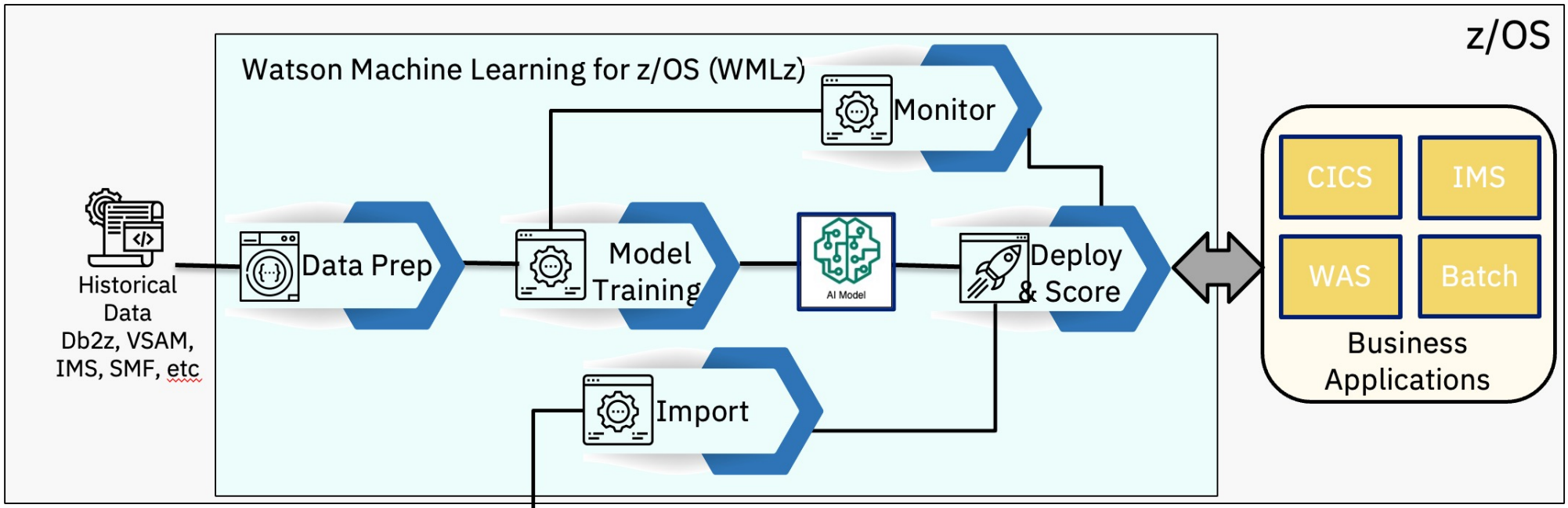
IBM Machine Learning on z/OS with optimized access



Notes
- Program ALNSCORE to interface to the scoring service is provided by MLz

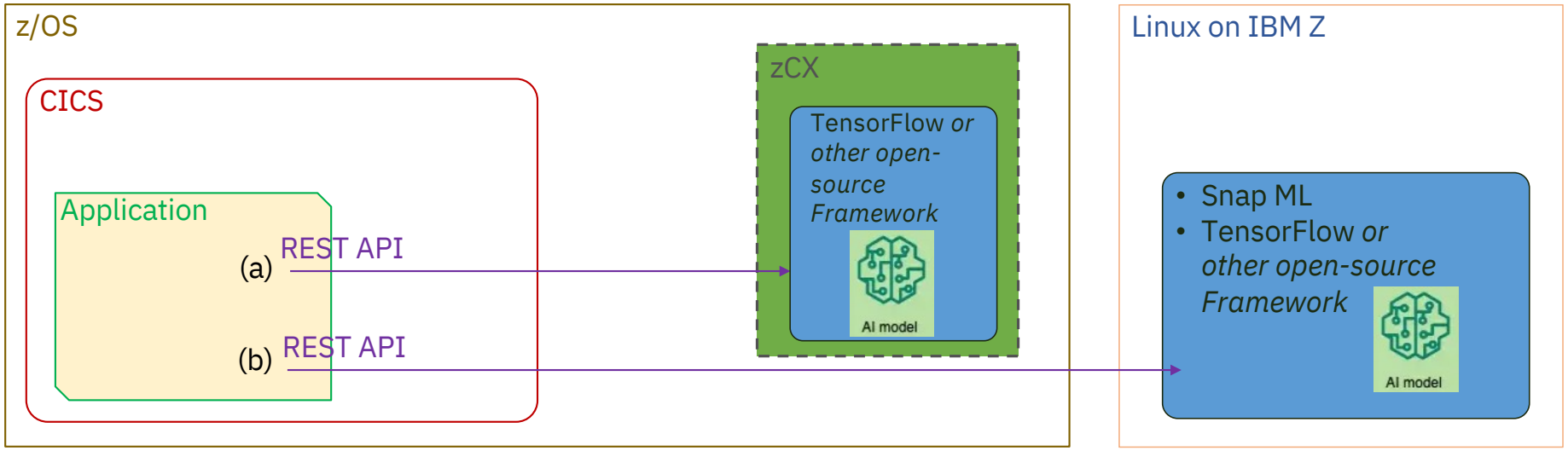
Could be a good choice when
- You want to take advantage of MLz and the full-function solution it offers
- You want to use the EXEC CICS LINK optimized route
- Using models developed using Spark, Scikit-learn, PMML, XGBoost, and ARIMA
- Using ONNX models (which MLz can compile on import using IBM DLC)
  - On IBM z16 with MLz 2.4, can exploit AI Accelerator

IBM Machine Learning for z/OS (MLz)



- Accelerate AI model development – Train anywhere, score on z/OS
- Operationalize AI model lifecycle management
- Infuse real-time AI model inferencing in z/OS applications
- Maintain competitive SLAs and deliver differentiation

Using a community-available AI framework



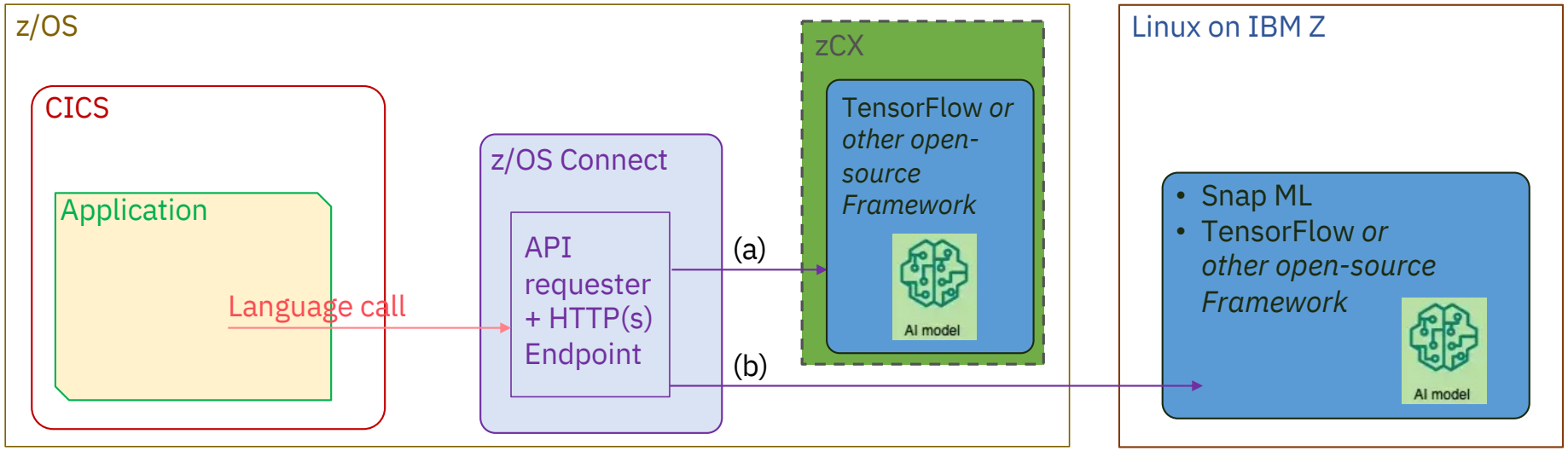
Comments & Notes

- (a) and (b) are alternative options, depending on where model is deployed
- IBM Snap ML is available from the PyPI (Python Package Index) repository
- Open-source AI frameworks include TensorFlow, PyTorch, scikit-learn, Keras, ...
- Calls to zCX use a highly-optimized form of access

Could be a good choice when

- You want to use a framework with which your data science team is familiar, or already have a model that uses the framework
- On IBM z16, Snap ML / TensorFlow can exploit AI Accelerator
- You have a preference for community-available or open-source options

Using a community-available AI framework via z/OS Connect



Comments & Notes

- (a) and (b) are alternative options, depending on where model is deployed
- IBM Snap ML is available from the PyPI (Python Package Index) repository
- Open-source AI frameworks include TensorFlow, PyTorch, scikit-learn, Keras, ...
- Calls to zCX use a highly-optimized form of access

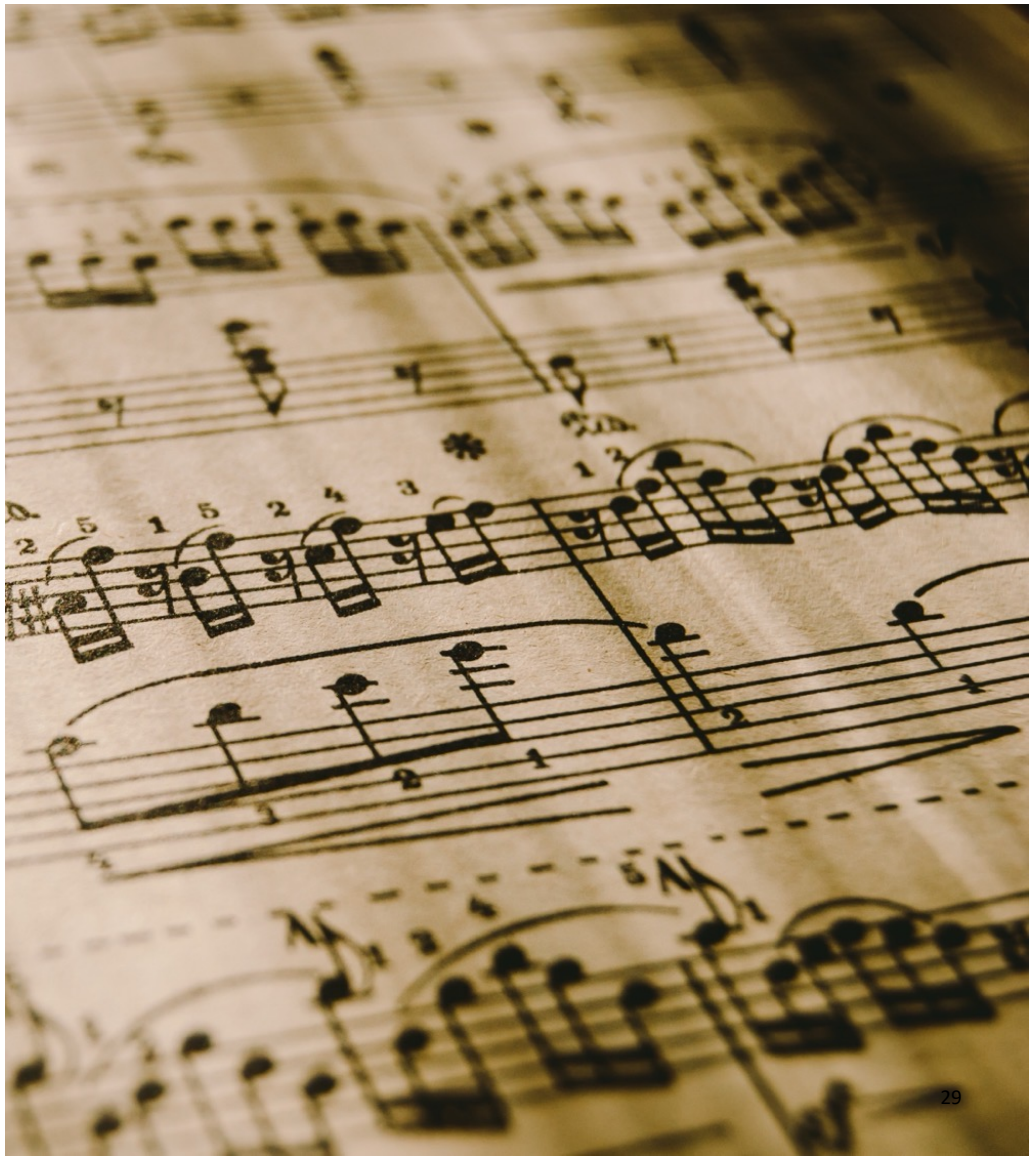
Could be a good choice when

- You already use z/OS Connect, or want to use it to simplify the coding in the application
- You want to use a framework with which your data science team is familiar, or already have a model that uses the framework
- On IBM z16, Snap ML / TensorFlow can exploit AI Accelerator

Infusing AI (an example)

- This example is based on a ‘greener shopping’ scenario

It uses the WMLz scoring service hosted in a CICS Liberty server and invoked via EXEC CICS LINK



Infusing AI (an example) – overall experience

Kathryn
Line of Business Manager

- LoB identifies an opportunity for encouraging greener more ethical shopping
- The transactions that processes items placed in the basket could find the shopper's current cluster, then (where appropriate) suggest a more environmentally or ethically favorable choice
- A data scientist works with the application architect / solutions architect:
- Determines a suitable type of AI model to use

Marcus
Data Scientist

- Data scientists discovers how to get the data needed as input to the model, and gets some training data
  - Might all be data provided from the application, or some could be obtained by the model from other data sources e.g. previous shopping history from Db2
- Uses the provided data to train and test a model

Michael
Senior System Administrator

- In this scenario, the model will be imported into Machine Learning (MLz) and deployed there. The systems administrator works with system programming team:
  - Installs and configures MLz base on z/OS
  - Configures a MLz scoring service in Liberty in CICS

Deb
Early tenure z/OS Application Developer

- Application developer invokes AI model from within the application
- via simple command (details on next slide)

Trish
z/OS Application Tester (lead)

- Tester includes the updated application in automated testing

Stan Cicero
Senior Systems Programmer

Infusing AI (an example) – application details

Deb

Early tenure z/OS Application Developer

- Application developer might be involved in pre-steps
  - to provide data used as input to the model
- When model has been deployed and tested, using the WMLz user interface
  - Obtain the JSON input & output schemas for the model
- Customize and use ALNJS2LS job to generate COBOL copybooks (uses CICS-supplied DFHJS2LS) for input & output to the model
- Also customize and run jobs to generate Java helper classes (used by scoring service to process input & output data structures)
- Update program that processes shopping items to:
  - Pass data features required by the model (current item, basket contents) by populating fields in ALN_xxx containers within a CICS channel
  - LINK to ALNSCORE program, passing the channel
  - Process the results returned in the output container
  - Use the returned *shopper cluster* to select a possible alternative purchase (where this makes sense)
- Test the updates to the application, and include in automated testing

Time for a DEMO

What is this demo?

The demo – a simple scoring model

- Order forecasting
  - requires quick analyses of real-time, or near real-time data
  - predict future quantity demands
- Sequence-dependent prediction model
  - predict the next order quantity based on the past 20 orders

Components

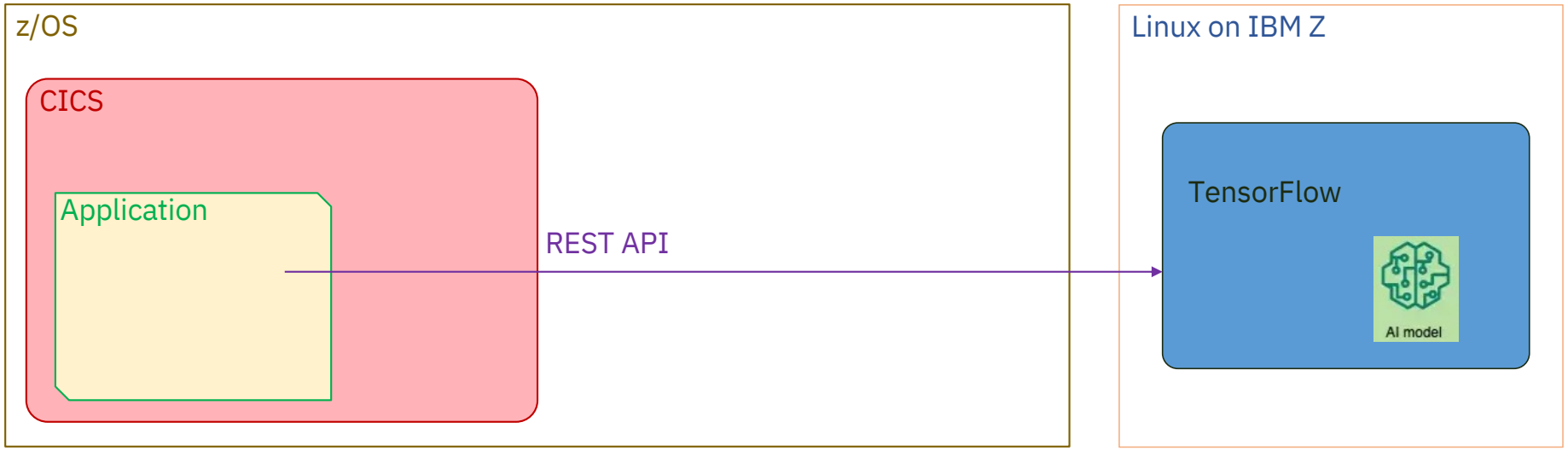
- Demo runs entirely in CICS
- User interface
  - CICS COBOL
  - CICS Web Support (EXEC CICS WEB)
  - Serves web pages (HTML, CSS, JavaScript, images)
- Business logic
  - CICS COBOL
  - Calls AI server
    - REST or LINK
    - REST calls
      - CICS REST client using EXEC CICS WEB
    - LINK calls
      - LINK to ALNSCORE, a Java program

How do we invoke the AI model from CICS application?

Developing the demo

- Development in stages
  - Initial model was built for TensorFlow
    - TensorFlow server running in Docker container in Linux on IBM Z
    - TF server ported to Docker running in zCX
  - Model converted as ONNX and imported into IBM Machine Learning for Z (MLz)
- IBM Machine Learning for z/OS installed on z/OS
  - Two instances
    - Native z/OS Liberty Server
      - Access via REST call
    - Liberty Server running in CICS
      - Access via REST call
      - Access via LINK to ANLSCORE program

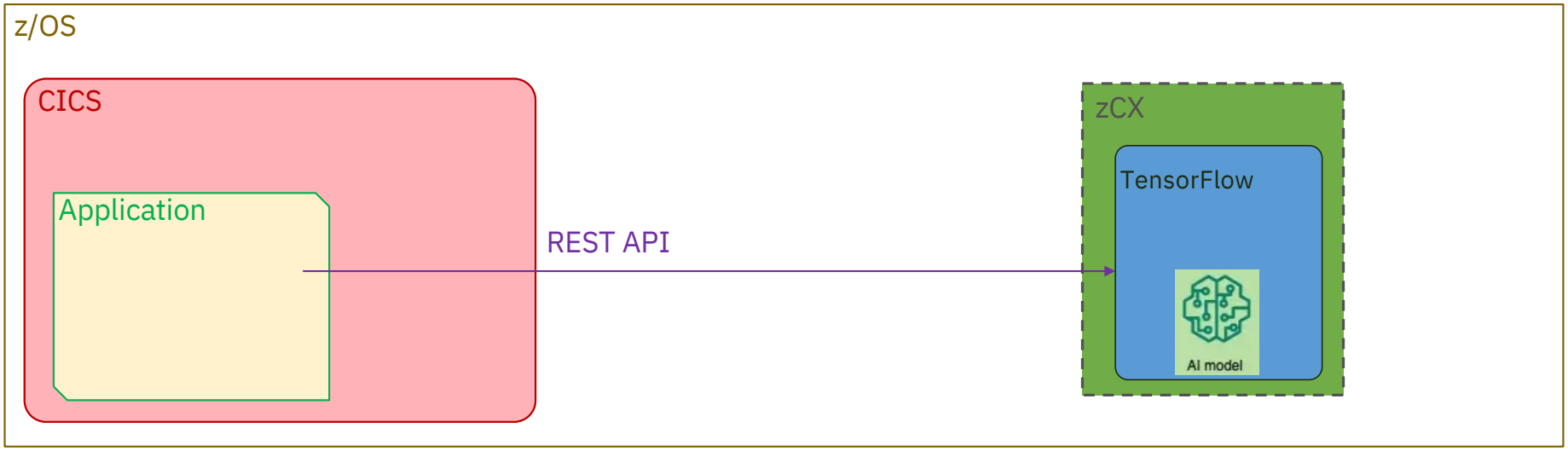
In the beginning …



Comments & Notes

- When the project began, the model had been developed and deployed in a TensorFlow framework running in Linux on IBM Z
- The CICS application used EXEC CICS WEB commands to send the request to the TensorFlow server

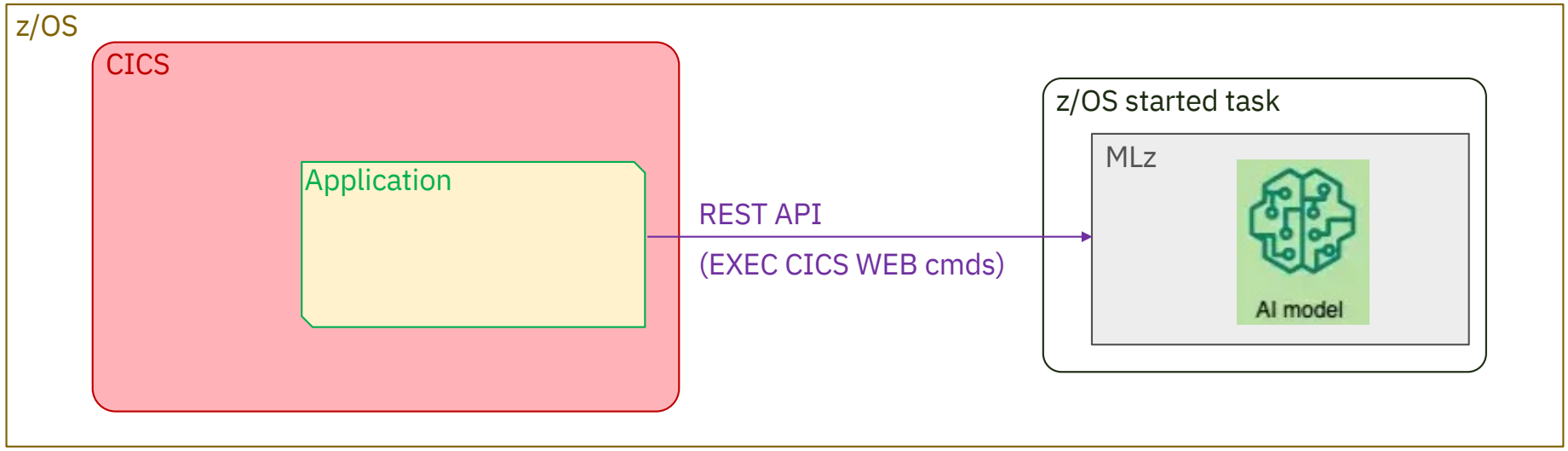
TensorFlow in zCX



Comments & Notes

- IBM® z/OS Container Extensions (zCX) implemented on z/OS and the TensorFlow Docker container deployed into zCX.
- The same CICS application uses EXEC CICS WEB commands to send the request to the TensorFlow server; only changed host/port

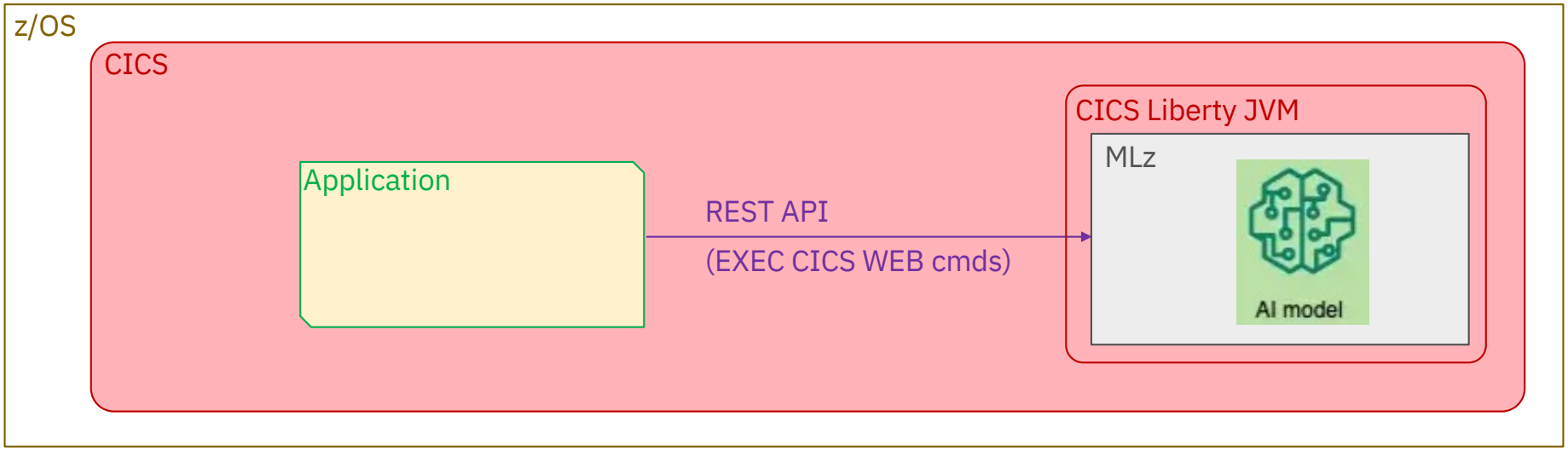
IBM MLz scoring service



Comments & Notes

- IBM® Machine Learning for z/OS® (MLz) installed on z/OS
- Scoring service configured in IBM® WebSphere® Application Server Liberty Profile for z/OS® (WLP)
- The original CICS application using EXEC CICS WEB commands modified due to changes in request/response data formats.

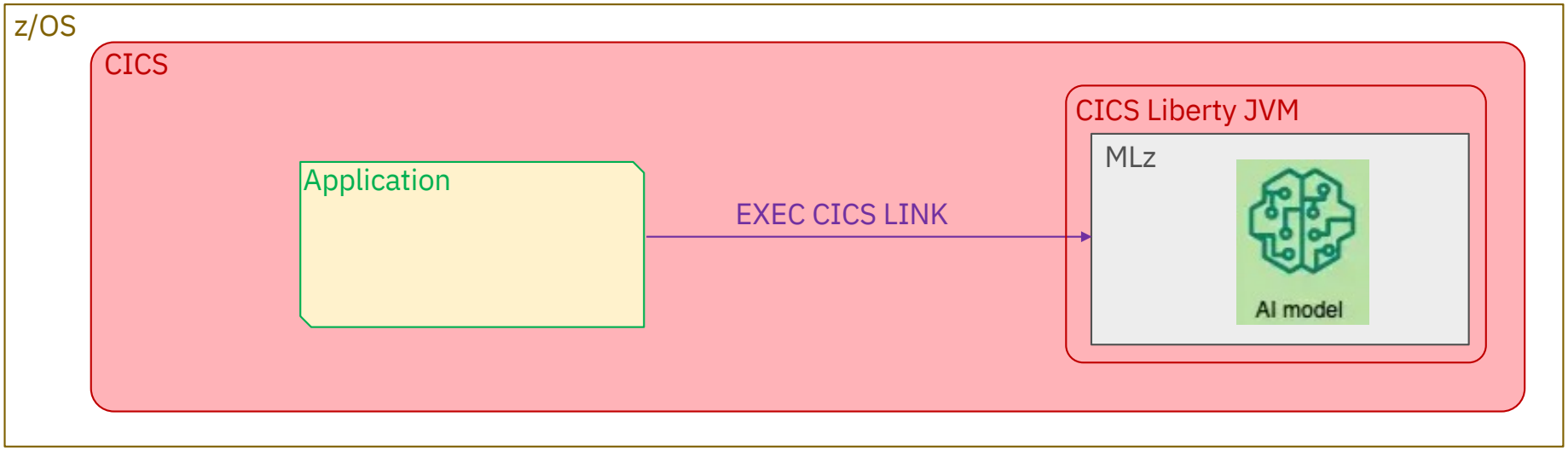
MLz scoring service running in CICS



Comments & Notes

- A second scoring service was configured in CICS Liberty JVM server.
- The now modified CICS application accesses the ONNX model using EXEC CICS WEB commands, but now targeting a different port.

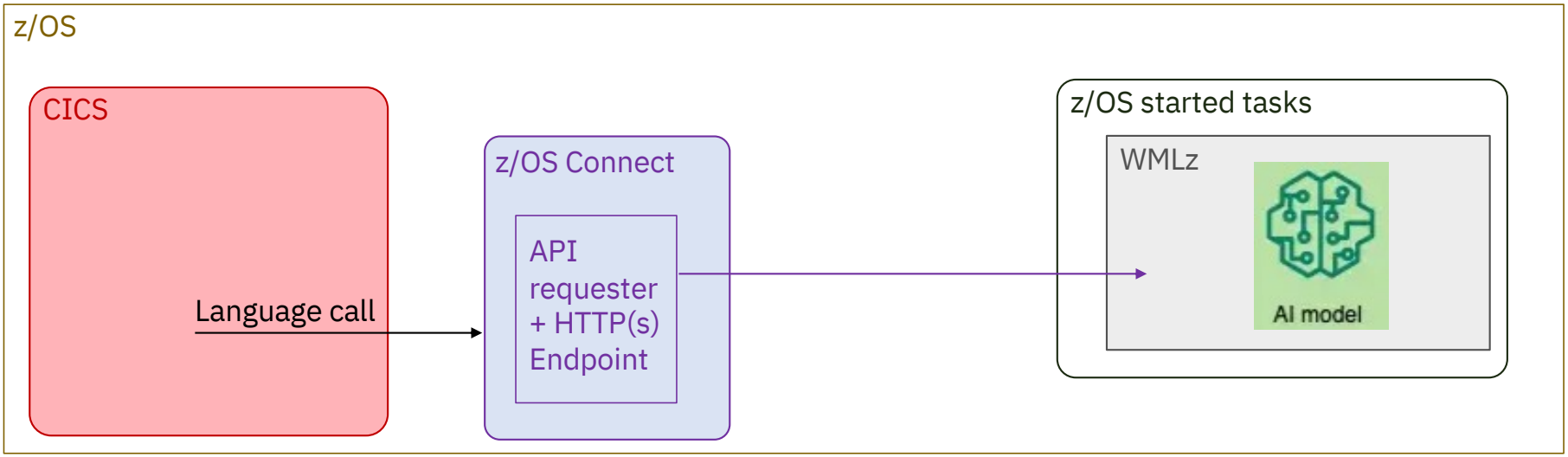
MLz running in CICS



Comments & Notes

- In addition to the REST invocation, when configured within CICS, MLz supplies a LINKable Java program name ALNSCORE to interface to the scoring service
- Utility program is used to generate copybooks to be used by invoking CICS program
- Data is passed in Containers within a CICS Channel

Future development



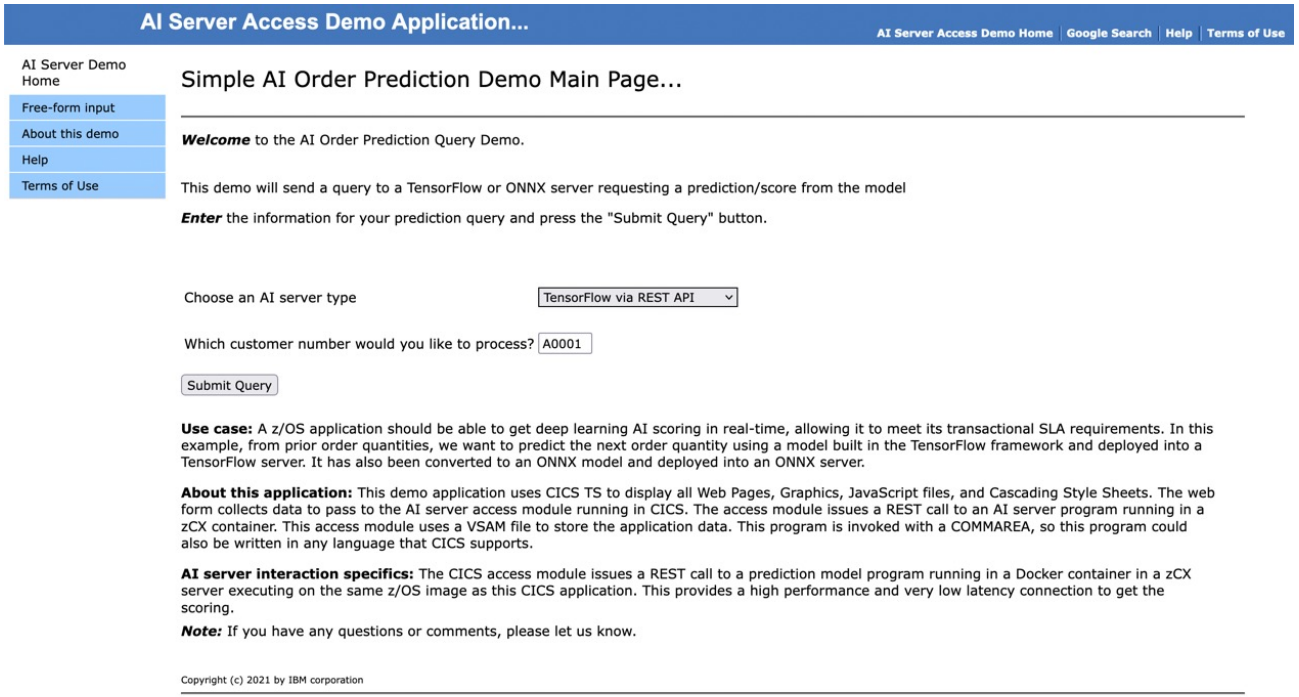
Comments & Notes

- Using API Requester function in z/OS Connect to simplify the coding to call the scoring service
- In development

The demo

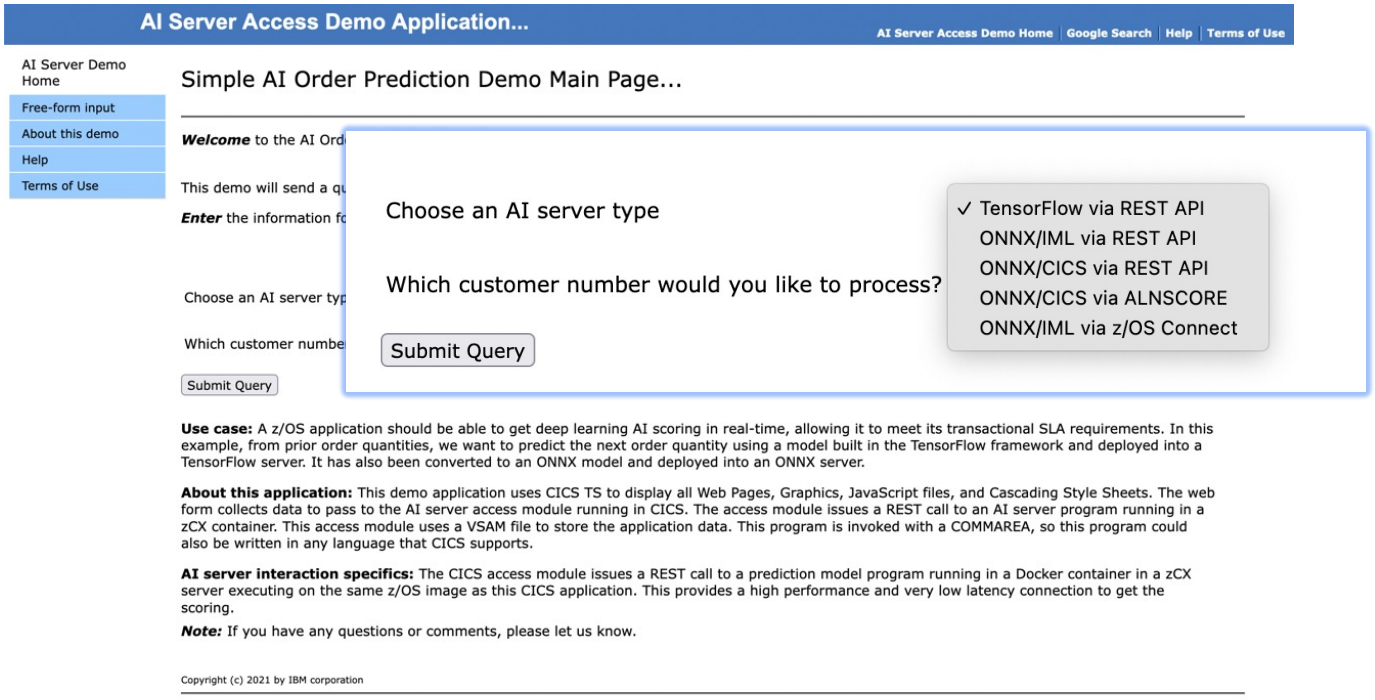
Home page

- The demo implements all the scoring server implementations described. User can select the type of server through pulldown.
- Order history is kept in VSAM file. User selects which customer's data is to be passed to the scoring server.



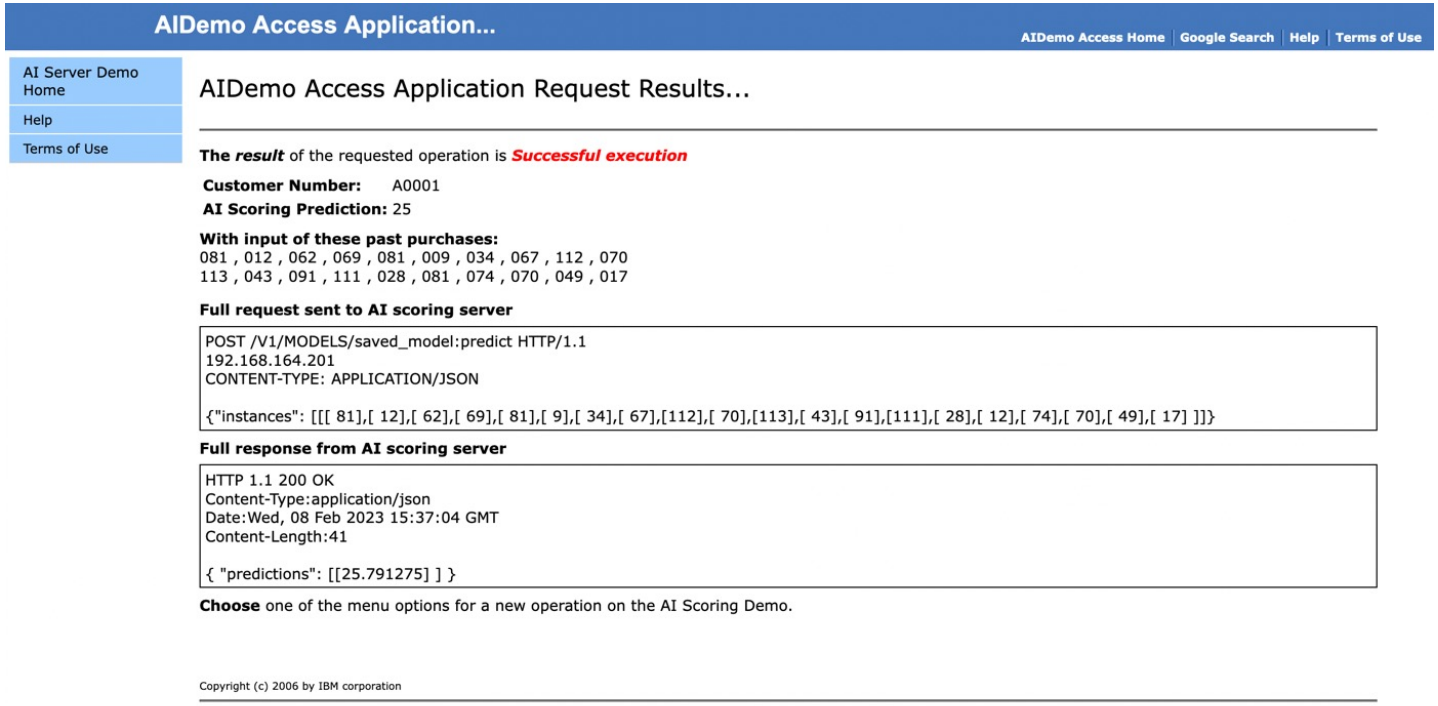
Server access selections

- Each of the scoring service options is offered to the user.

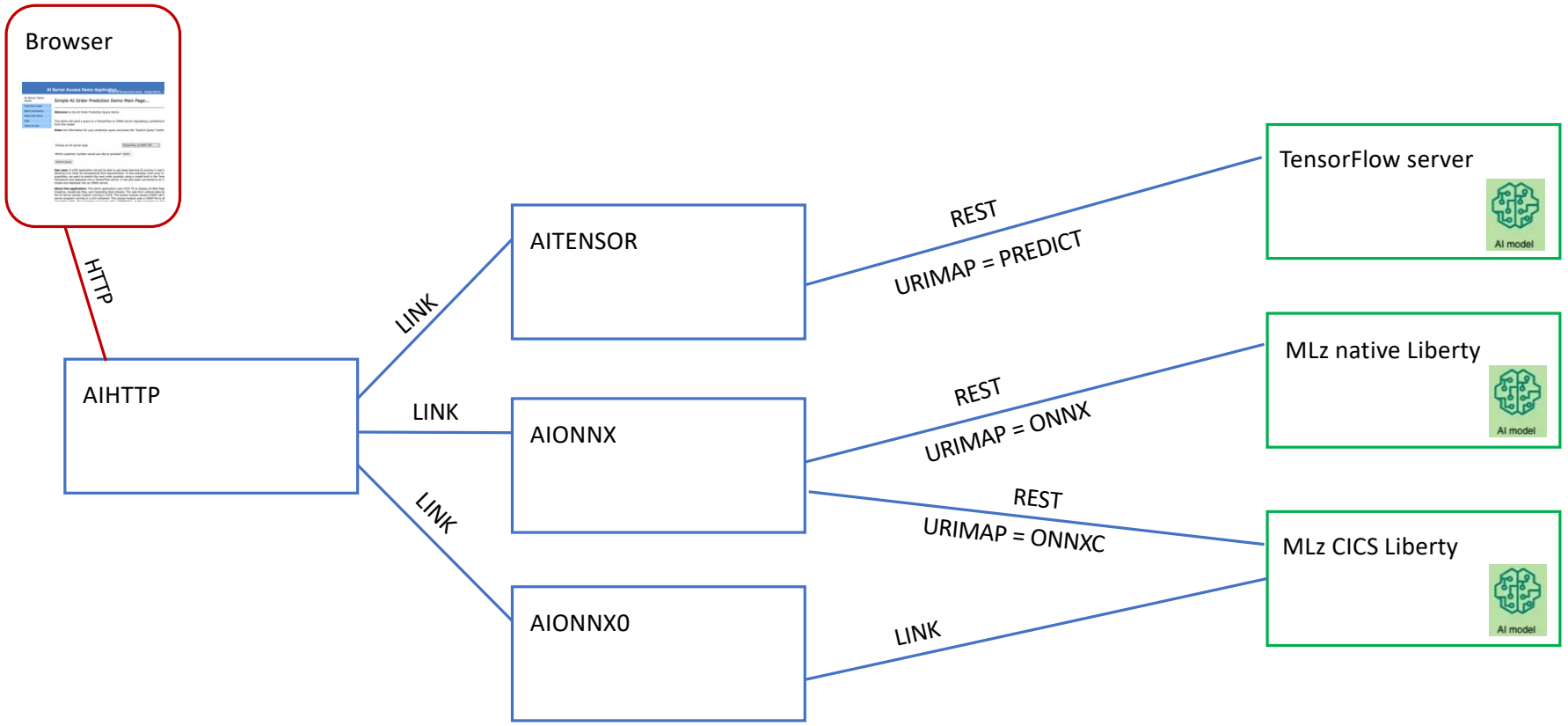


Results

- For all calls via REST, the results page displays the HTTP request that was sent and the response received from the scoring service



CICS Programs & Resources



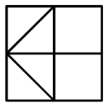
In Summary

My experience

- This has been a fascinating project with an opportunity to experience multiple modern technologies:
- JSON markup; generating and parsing
- Liberty JVM server in CICS
- HTTP/REST client implementation
- … and the opportunity to compare different approaches, particularly when it comes to the programming effort required:
- HTTP via EXEC CICS WEB vs LINK

- Futures:
- Recently developed a bulk/batch client to drive more scoring requests with a large degree of parallelization.
- Will implement API access via z/OS Connect

Summary – AI in CICS TS Applications



Real-time business insights

IBM z16 on-chip AI accelerator for enhanced inferencing offers scope for AI inferencing in every transaction

High throughput, low latency, in-transaction decision making and insights

Train anywhere, run on IBM zSystems, to leverage data and transactional gravity

Seamless exploitation

Applications on the platform, hosted in core middleware such as CICS can exploit the IBM z16 on-chip AI inferencing when using suitable AI models, without requiring changes to the CICS product, and without change to existing AI models

AI infusion options

Several options available to zSystems runtime applications for invoking AI processing as part of the transaction

Options range from using REST API calls to drive open-source AI frameworks in Linux on IBM Z, or zCX, through to calling IBM Machine Learning for z/OS (MLz) hosted in Liberty in CICS

Aids to adoption

Supporting material helps to show the available options, and the path to follow, to successfully incorporate AI inferencing into your CICS applications

Q&A

- Any more Questions on what we discussed?

Want to learn more?

- Some reading material
- AI on IBM zSystems Discovery Workshops
- Get in touch
- Summary
- Q&A
- **DEMO**

Some reading material

- Optimized Inferencing and Integration with AI on IBM Z Introduction, Methodology, and Use Cases (Redpaper, updated November 2022)
- Journey to AI on zSystems (Content Solution)
- Planning AI infusion into applications (Guidance Document – also linked from the 'Journey to AI' content solution)
- Jump-starting your experience with AI on IBM Z (Blog post)
- Announcing IBM z16: Real-time AI for Transaction Processing at Scale and Industry's First Quantum-Safe System (one of many IBM z16 posts and articles)

- Preparing a model for online scoring with CICS program ALNSCORE in MLz documentation
- Deploy AI models for real-time inferencing in your z/OS IMS transactions — IBM Watson Machine Learning for z/OS v2.4 new feature using WOLA
- Integration of WMLz with ODM in ODM documentation
  - Tutorial in ODM documentation that steps through enhancing ODM rules with MLz predictions
  - Make smarter decisions: Apply intelligence to your Z applications with digital decisioning – Webinar on ODM with MLz

Engage with us:

- aionz@us.ibm.com

- AI on IBM Z and LinuxONE Community

- https://ibm.github.io/ai-on-z-101/

- Contact us directly

- Sites
- Journey to AI on IBM Z Content Solution link
- IBM Z and Cloud Mod Center AI Page link
- Real-Time analytics and AI on the IBM mainframe link

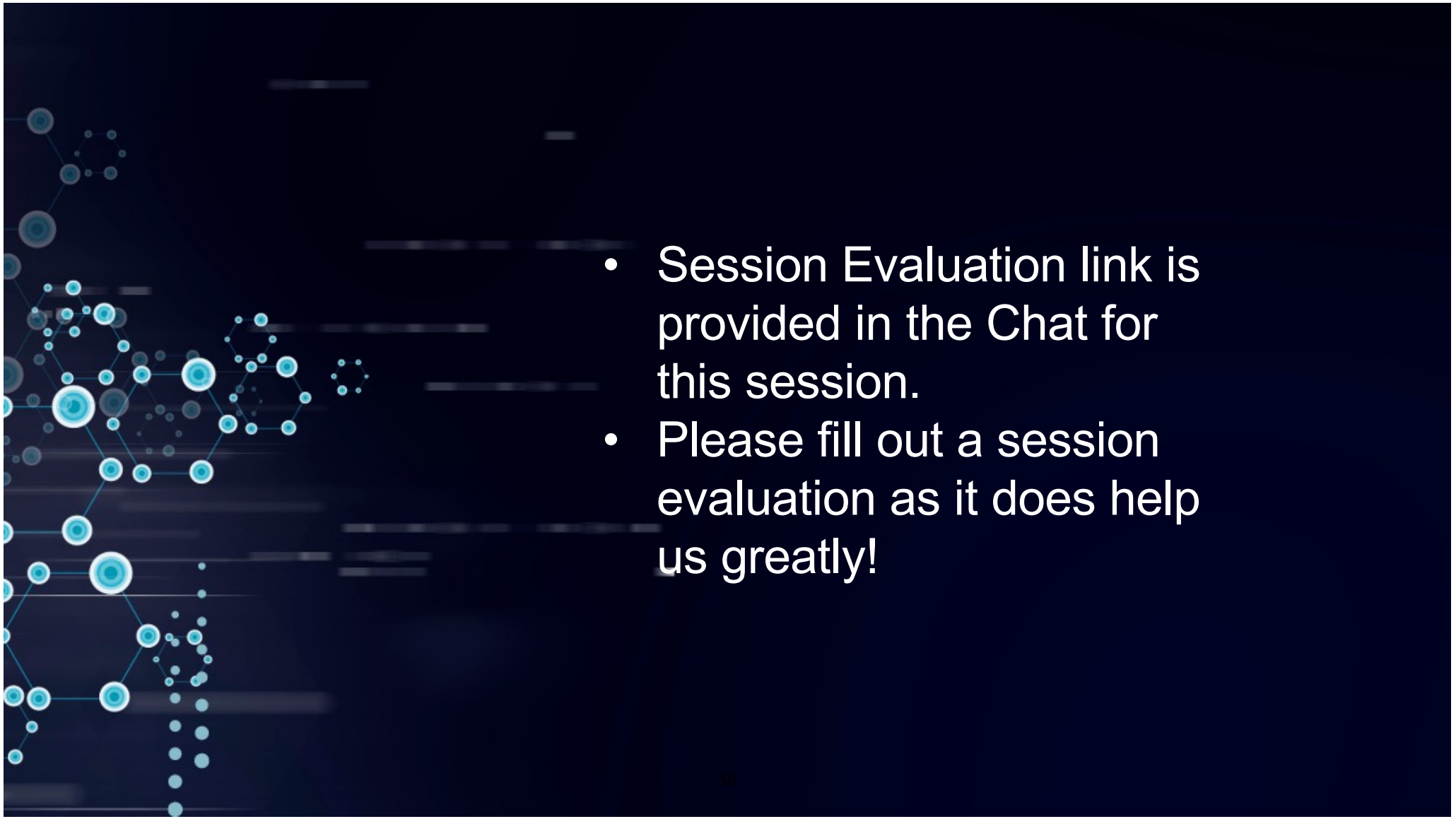
- Blogs
- TensorFlow blog: link
- ONNX blog: link

- Demos
- Watson Machine Learning Demo link
- Anti-Money Laundering with AI on Z link
- Fraud Detection Demo link

- Redbooks
- Optimized Inferencing and Integration with AI on IBM Z Introduction, Methodology, and Use Cases: link
- Demystifying Data with AI on IBM Z –POV: link
- Art of the Possible with AI on IBM zSystems link

- Paper
- IDC: The business value of the transformative mainframe link
- Operationalizing Fraud Prevention on IBM z16: Reducing Losses in Banking, Cards, and Payments link

- Open Source
- IBM Z and LinuxONE container Image Registry: link
- TensorFlow on IBM Z and LinuxONE container Image Registry: link
- Anaconda Partnership link


• Session Evaluation link is provided in the Chat for this session.
• Please fill out a session evaluation as it does help us greatly!