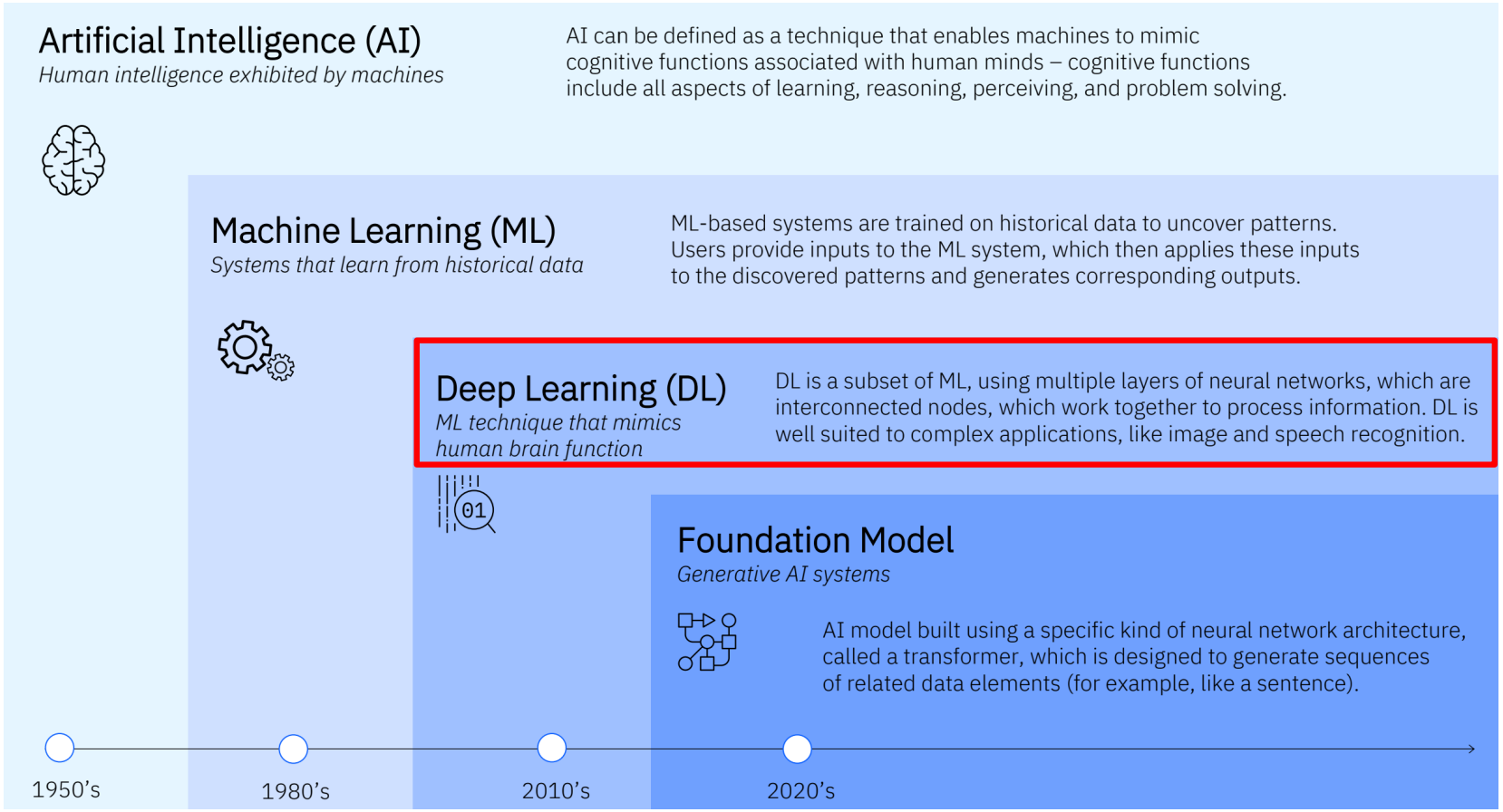
IBM z16
Integrated Accelerator for AI

Agenda

- AI introduction
- Integrated Accelerator for AI overview
- AI model scoring server options to leverage the Integrated Accelerator for AI
- Accelerated credit card fraud model demo

What is AI?

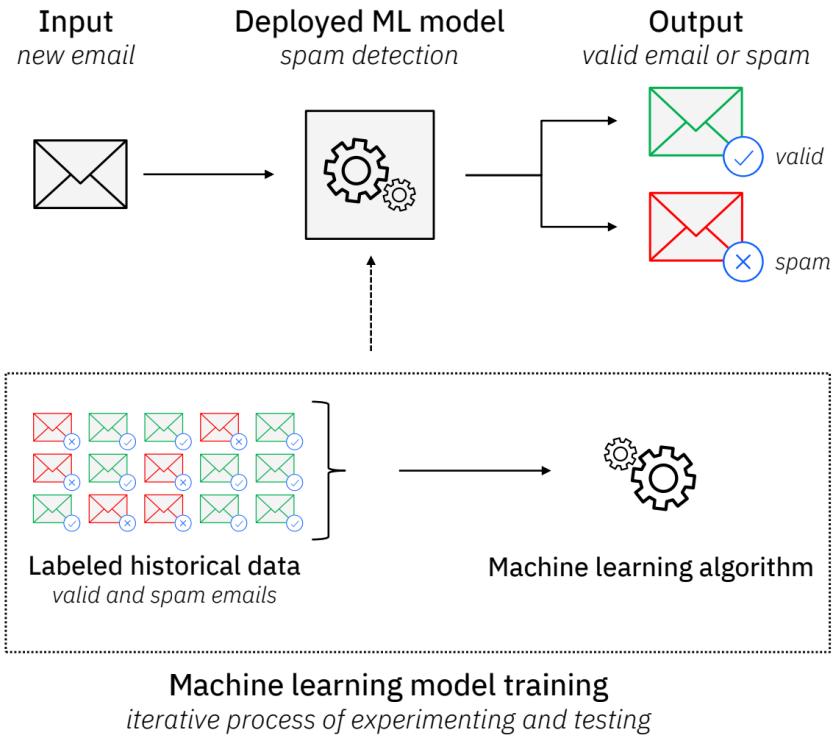


ML/DL Classification Models

Classification models assign labels to model inputs or assign them to specific categories. Common use cases include:

- Fraud detection
- Sentiment analysis
- Medical diagnosis
- Image recognition

Example: Spam detection for email



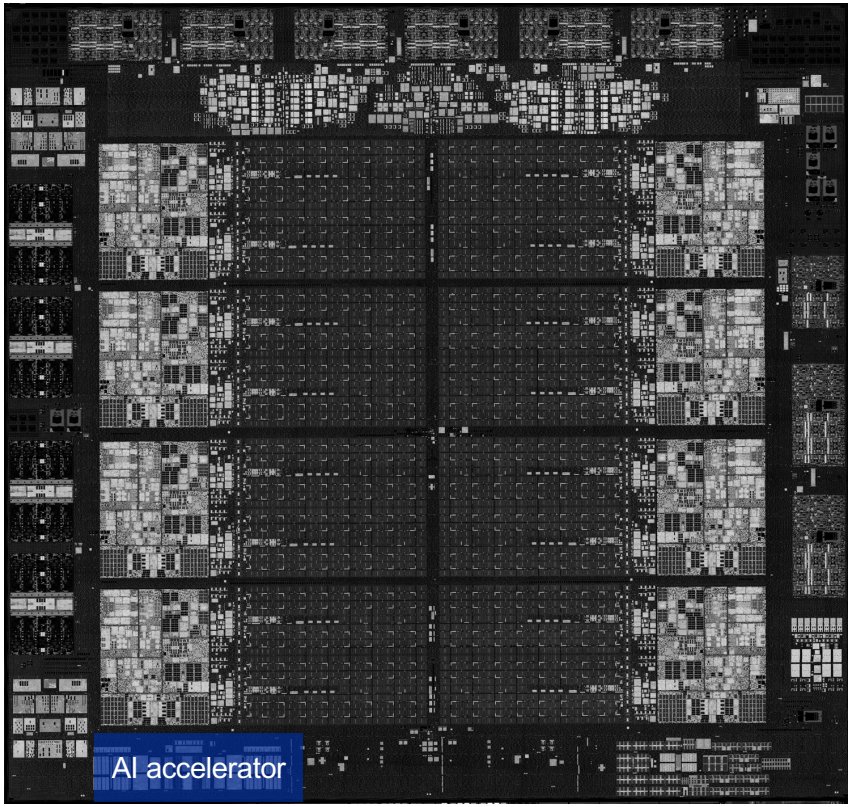
Machine learning model training
iterative process of experimenting and testing

IBM z16 Integrated Accelerator for AI Overview

IBM z16 Integrated Accelerator for AI

Centralized accelerator shared by all cores on-chip

- Very low and consistent inference latency
- Compute capacity for utilization at scale
- Designed for a variety of AI models ranging from traditional ML to RNNs and CNNs
- Security – provide enterprise-grade memory virtualization and protection



Integrated AI Accelerator – combining compute & data movers

On Chip AI Accelerator

Aggregate of >6 TFLOPS / chip

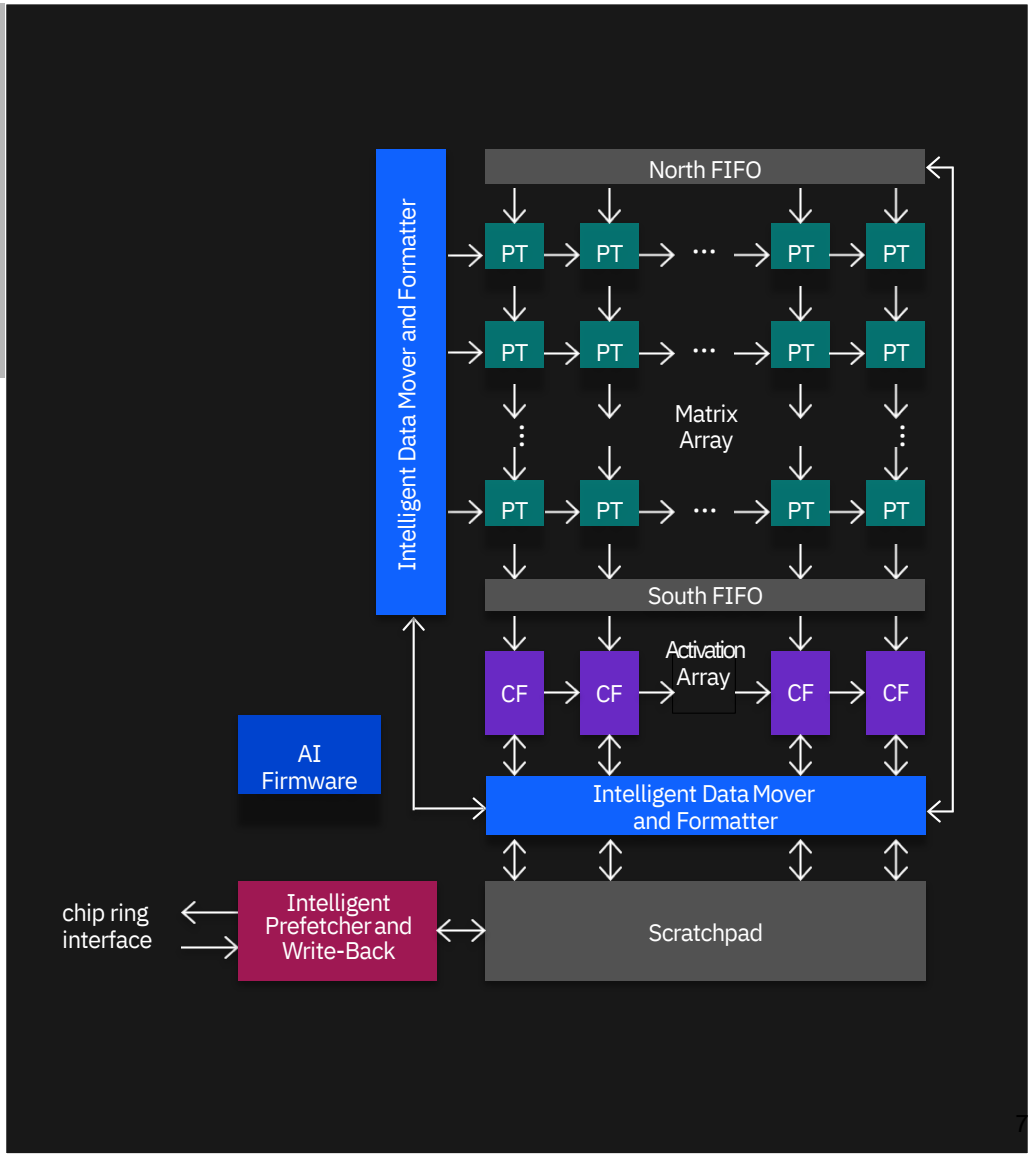
– Over 200 TFLOPS on 32-chip system

Compute Arrays

- 128 processor tiles with 8-way FP-16 FMA SIMD
  - Optimized for matrix multiplication and convolution
- 32 processor tiles with 8-way FP-16/FP-32 SIMD
  - Optimized for activation functions & complex operations

Intelligent Prefetcher and Data Movers

- 200+ GB/s read/store bandwidth from/to cache
- 600+ GB/s bandwidth between engines
- Multi-zone scratchpad for concurrent load, execution and store



Neural Network Processing Assist (NNPA) instruction on IBM z16

- New instruction provided to support the IBM z16 AIU
- AI Functions/Macros are abstracted via NNPA instructions. More than just matrix multiplication!
  - Elementwise, Activation
  - Normalization, Pooling
  - Matrix-multiplication
  - Convolution
  - Conv+Scale+Activate
  - MatMul+Compare/Activate
  - RNN activation
- Functions can be added per firmware update

Function group	#	Function support in GA1
Elementwise ops	0x10	NNPA_EL_ADD
	0x11	NNPA_EL_SUB
	0x12	NNPA_EL_MUL
	0x13	NNPA_EL_DIV
	0x14	NNPA_EL_MIN
	0x15	NNPA_EL_MAX
Activation ops	0x20	NNPA_LOG
	0x21	NNPA_EXP
	0x31	NNPA_RELU
	0x32	NNPA_TANH
	0x33	NNPA_SIGMOID
Norm op.	0x34	NNPA_SOFTMAX
	0x40	NNPA_BATCHNORM
Pooling	0x50	NNPA_AVGPOOL2D
	0x51	NNPA_MAXPOOL2D
Systolic ops	0x70	NNPA_CONVOLUTION
	0x71	NNPA_MATMUL_OP
	0x72	NNPA_MATMUL_OP_BCAST23
RNN	0x60	NNPA_LSTMACT
	0x61	NNPA_GRUACT
	0x00	NNPA_QAF

How can we leverage the Integrated Accelerator for AI?

- Deep Learning Frameworks:
  - ONNX (Open Neural Network Exchange)
  - TensorFlow
- Machine Learning Frameworks:
  - SnapML models
    - Random Forest, Extra Trees, and Gradient Boosting Machines models are accelerated

Model Serving 101

AI Lifecycle: Deploy Phase

Practical use of machine learning (ML) models for inference introduces a different set of requirements and tools than model training.

This is especially true when the insights generated will be consumed by an application.

Model inference servers are a critical component to enable AI in production.

At minimum, a model server provides for:

- Exposing endpoints (i.e., HTTP/REST, gRPC)
- Common API format
- Request management

Other important characteristics include:

Scalability/High Performance

Server-side micro-batching
- With IBM Integrated Accelerator for AI, this capability is critical for optimizing performance.

Support for multiple frameworks

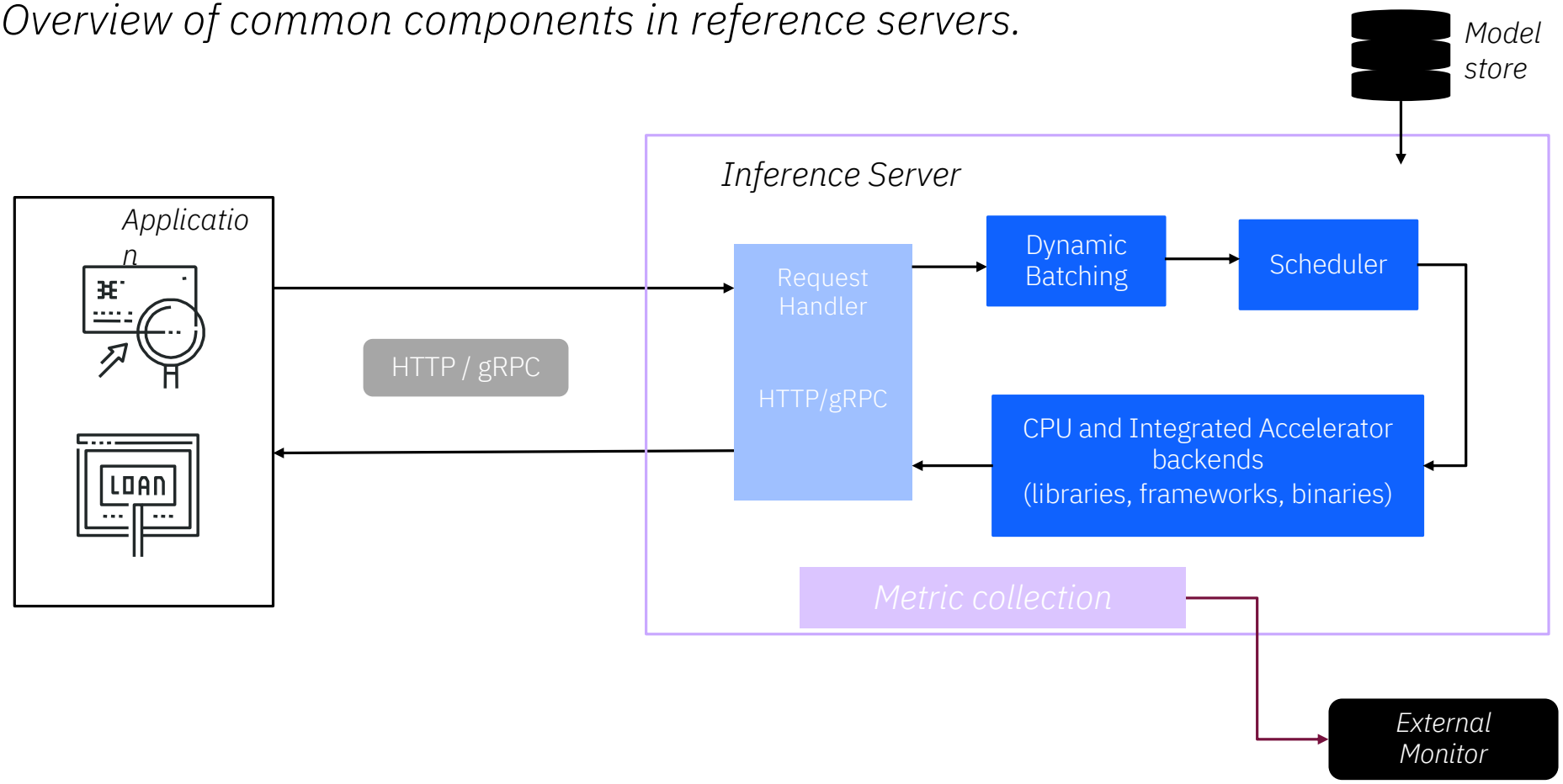
Version control

Metrics/Monitoring Integration

...and more

Inference Server Architecture

Overview of common components in reference servers.



Scoring Server Options*

IBM z/OS

Watson Machine Learning for z/OS

Accelerated Models:
- ONNX

Other Supported Models:
- Spark
- XGBoost
- PMML

Linux on Z

Triton Inferencing

Accelerated Models:
- ONNX
- SnapML

Other Supported Models:
- TensorFlow
- PyTorch
- PMML

TensorFlow Scoring Server

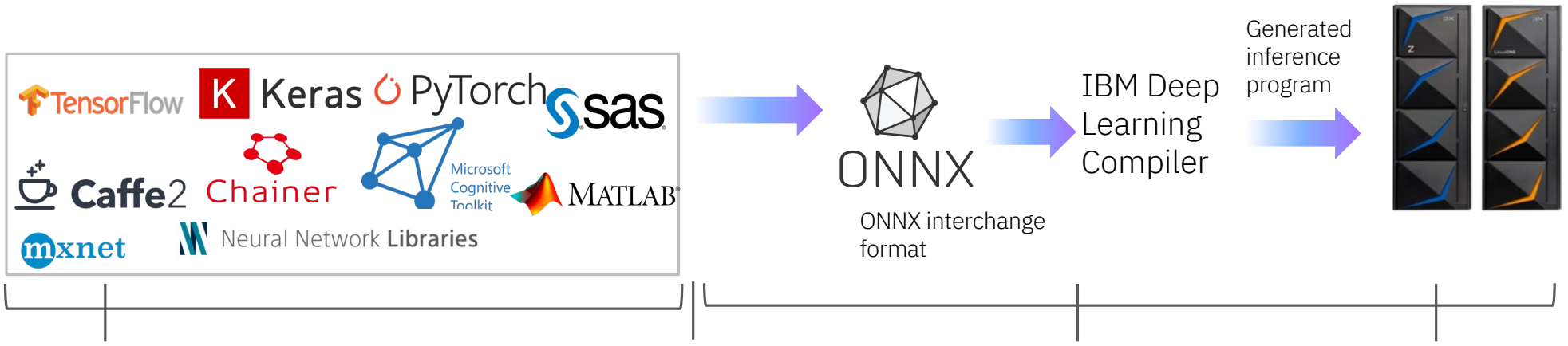
Accelerated Models:
- TensorFlow support under development

Supported Models:
- TensorFlow

* other scoring server options are available

AI Ecosystem:
Seamlessly leverage AI accelerator on IBM z16 with ONNX

- Bring machine learning & deep learning models to IBM z16 with ONNX/DLC
- Exploit IBM Integrated Accelerator for AI for best inference performance.
- Repeatable practice for different vendors to leverage IBM Z Integrated Accelerator for AI



Build and train model in any popular framework on any platform of your choice

Use ONNX, an opensource tool for framework interoperability

Models are converted to the ONNX interchange format

The IBM DLC (Deep Learning Compiler), generates a program from the model for execution on z/OS or Linux® on IBM z16

Deploy on IBM z16 and IBM LinuxONE and infuse model into workload application

Creating a Credit Card Fraud Model

How do we predict credit card fraud with AI?

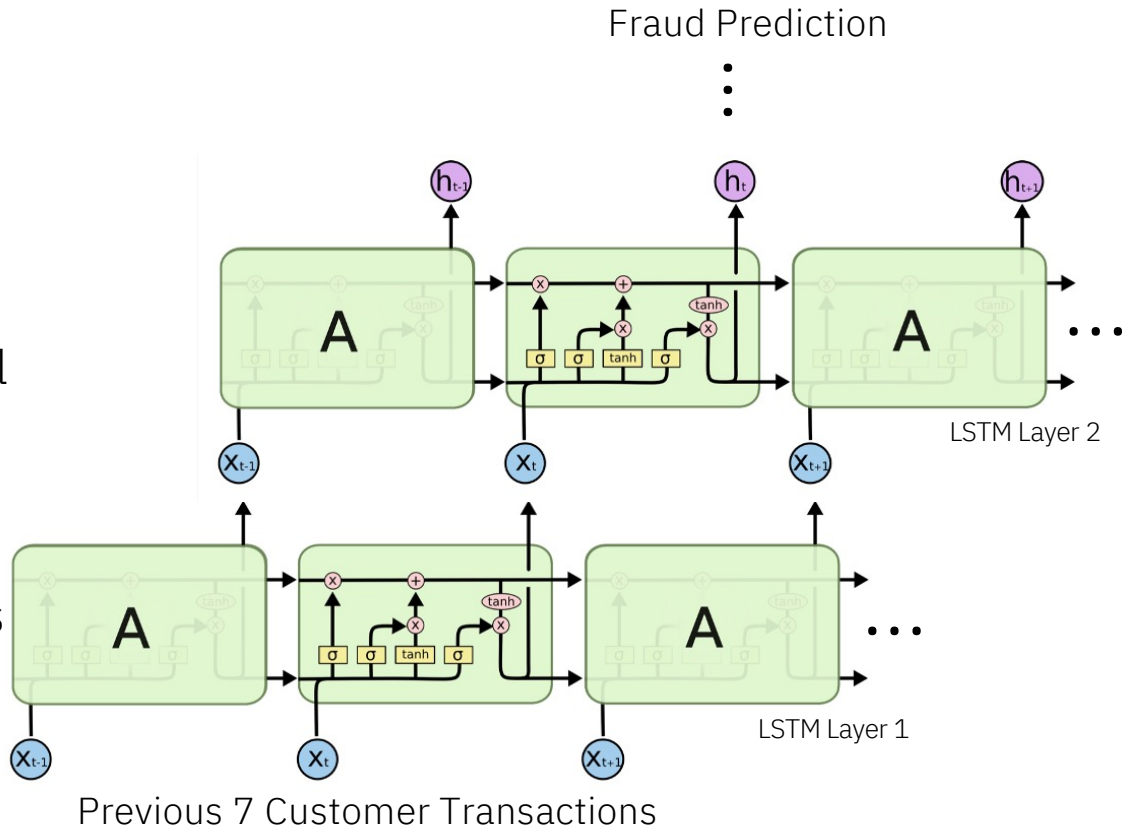
- Create a training dataset from previous customer transactions
  - We used an IBM Research synthetic dataset[1]
  - 2.4M samples with 29,342 labeled as fraudulent (1.22%)
- Label confirmed instances of fraud
- Train an AI model to predict whether a current transaction is likely fraudulent
- Evaluate model performance and deploy

Features Used for AI Training
User
Card
Year
Month
Day
Time
Amount
Use Chip?
Merchant Name
Merchant City
Merchant State
Zip
MCC
Errors?
Fraudster ID
Is Fraud?

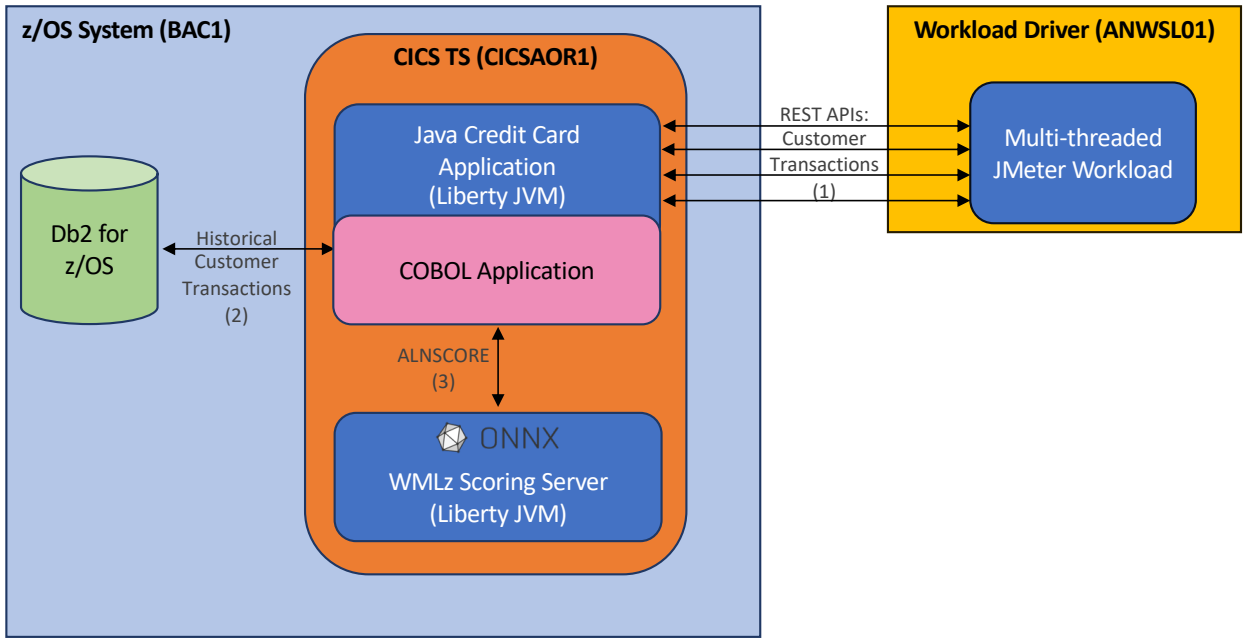
1. https://github.com/IBM/ai-on-z-fraud-detection

LSTM Credit Card Fraud Model

- Long Short-Term Memory model - specialized type of Recurrent Neural Network
- Incorporates feedback to efficiently process sequences of data
- Trained a 2-layer 200 cell LSTM model on ~800,000 customer credit card transaction sequences
- Classifies if the transaction is likely fraudulent or not based on a customer's 7 previous transactions



WSC CICS Transactional Workload Environment

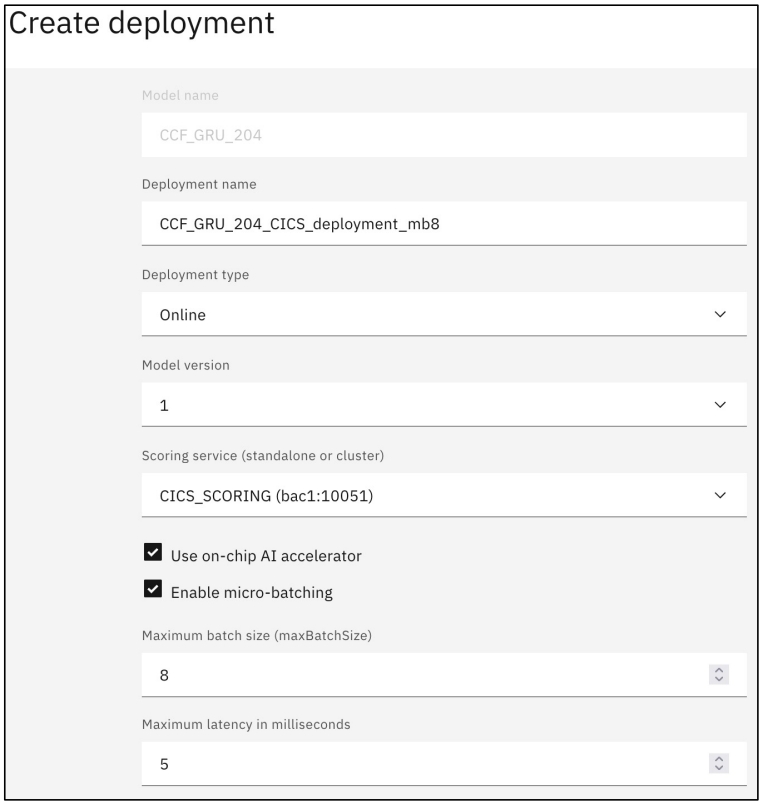


1. Simulated customer transactions are sent by JMeter to Java credit card application through a REST API call. Can control thread count.
2. A CICS transaction is initiated by the Java program, and the customers 7 previous credit card transactions are retrieved.
3. An AI prediction is performed using ALNSCORE to pass the previous transactions as input to the AI model
4. Prediction result is returned to Java application and then to JMeter

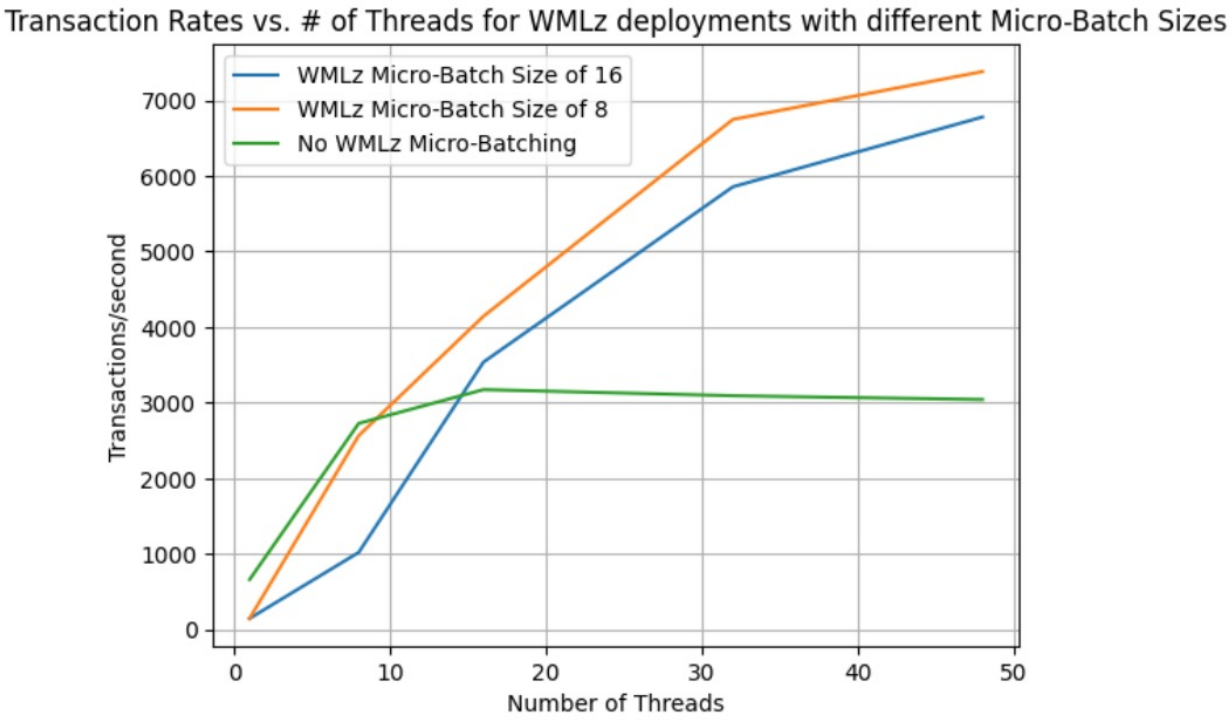
Accelerated Inferencing Demo

Deployment using Watson Machine Learning for z/OS (WMLz) and ALNSCORE

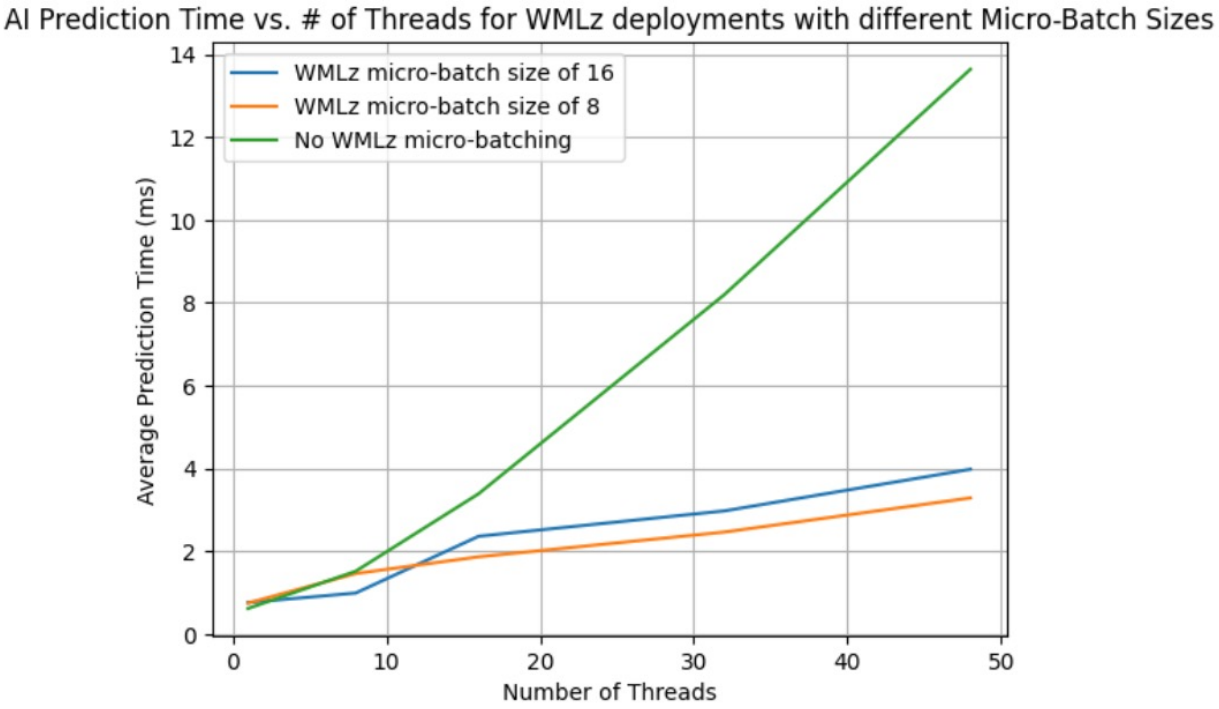
- First, model must be imported into WMLz
  - Deep Learning Compiler compiles model for efficient execution on the Integrated Accelerator for AI or CPU
- Can deploy to scoring server running in CICS region
- Can select whether deployment will use AI accelerator
- Specify deployment parameters to make online scoring more efficient
  - Micro-batch size: How many scoring requests to queue and send for processing simultaneously
  - Wait time: How long do I wait for the batch queue to fill up before processing scoring requests



Effect of WMLz Micro-Batching on Transaction Rates



Effect of WMLz Micro-Batching on Prediction Time



- Note: WMLz prediction times recorded by Java/COBOL application

Resources

- AI on IBM Z POCs
  - Contact aaminin@ibm.com
- AI on IBM Z 101
  - https://ibm.github.io/ai-on-z-101/
- IBM Z and LinuxONE Container Registry
  - https://ibm.github.io/ibm-z-oss-hub/main/main.html
- Watson Machine Learning for z/OS Overview
  - https://www.ibm.com/products/machine-learning-for-zos

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