National Science Foundation's Industry/University Cooperative Research (I/UCRC) Program

# F2-24: Development of Large AI Applications and Systems

Mission-Critical Computing
NSF CENTER FOR SPACE, HIGH-PERFORMANCE, AND RESILIENT COMPUTING (SHREC)

SHREC Annual Workshop (SAW23-24)

University of Pittsburgh
BYU BRIGHAM YOUNG UNIVERSITY
VIRGINIA TECH
UF UNIVERSITY of FLORIDA

January 17-18, 2024

**Dr. Janise McNair**
Professor of ECE

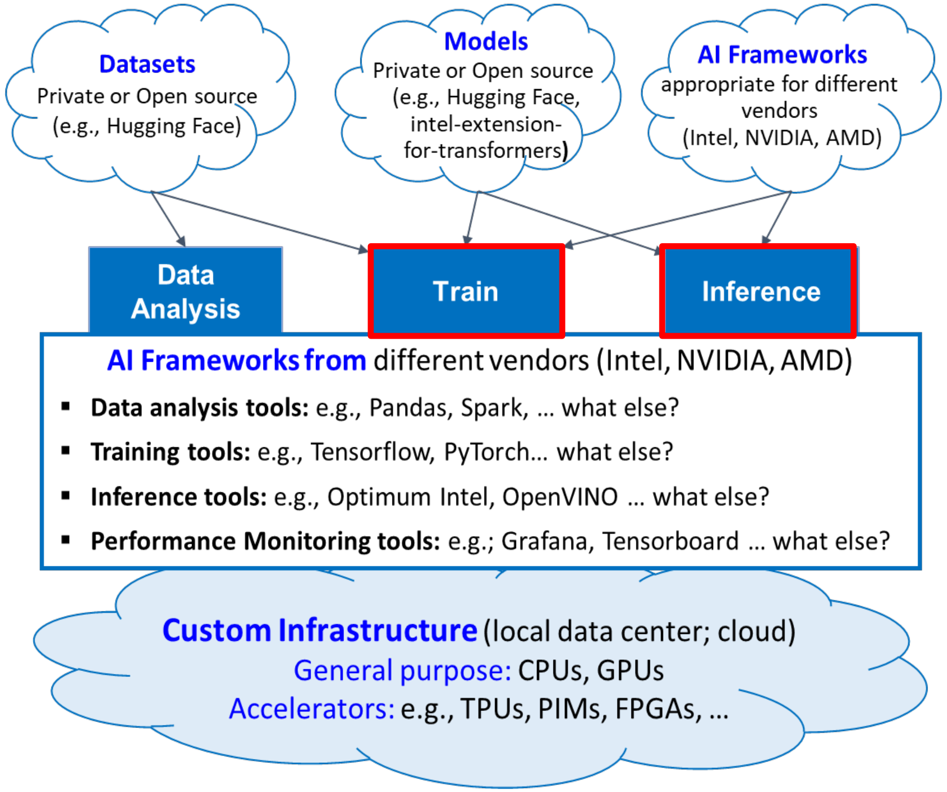
**Dr. Herman Lam**
Assoc. Professor of ECE

D. Agnew, A. Koti, A. Mandala,
P. Pinninti, A. Rice-Bladykas
Research Students
University of Florida

Number of requested memberships ≥ 4

# Motivation

- Creating hardware infrastructures for developers faced with a complex array of choices: *dataset*, *model*, AI *framework*, & hardware *infrastructure*



**Datasets**
Private or Open source
(e.g., Hugging Face)

**Models**
Private or Open source
(e.g., Hugging Face, intel-extension-for-transformers)

**AI Frameworks**
appropriate for different vendors
(Intel, NVIDIA, AMD)

Data Analysis

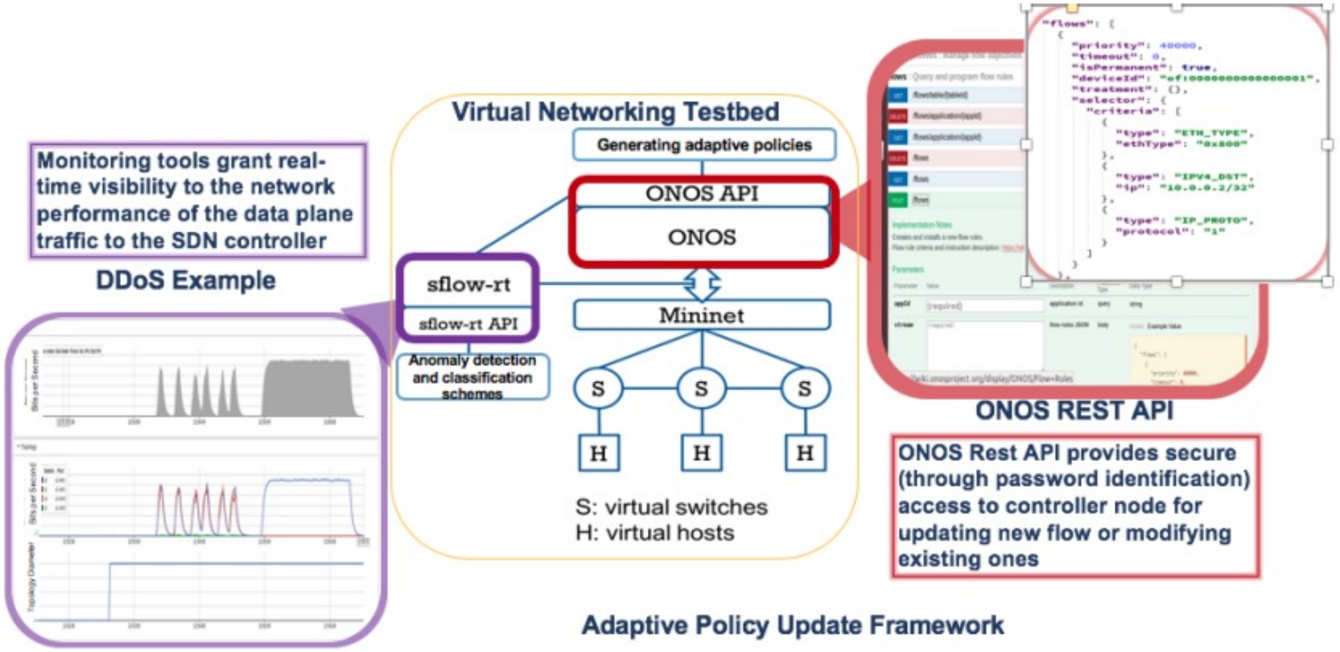
Train

Inference

**AI Frameworks from** different vendors (Intel, NVIDIA, AMD)

- **Data analysis tools:** e.g., Pandas, Spark, … what else?
- **Training tools:** e.g., Tensorflow, PyTorch… what else?
- **Inference tools:** e.g., Optimum Intel, OpenVINO … what else?
- **Performance Monitoring tools:** e.g.; Grafana, Tensorboard … what else?

**Custom Infrastructure** (local data center; cloud)
General purpose: CPUs, GPUs
Accelerators: e.g., TPUs, PIMs, FPGAs, …



Monitoring tools grant real-time visibility to the network performance of the data plane traffic to the SDN controller

DDoS Example

Virtual Networking Testbed

Generating adaptive policies
ONOS API
ONOS
sflow-rt
sflow-rt API
Mininet
Anomaly detection and classification schemes
S
S
S
H
H
H

S: virtual switches
H: virtual hosts

ONOS REST API

```
"flows": [
  {
    "priority": 40000,
    "timeout": 0,
    "isPermanent": true,
    "deviceId": "of:0000000000000001",
    "treatment": {},
    "selector": {
      "criteria": [
        {
          "type": "ETH_TYPE",
          "ethType": "0x800"
        },
        {
          "type": "IPV4_DST",
          "ip": "10.0.0.2/32"
        },
        {
          "type": "IP_PROTO",
          "protocol": "1"
        }
      ]
    }
  },
```

ONOS Rest API provides secure (through password identification) access to controller node for updating new flow or modifying existing ones

Adaptive Policy Update Framework

- Creating network architectures for applications, *such as security and quality of service,* that can generate and leverage *real-time situational awareness* through new network profile *data sets*, network *models*, and *machine learning and AI-based* protocols.

# Project Goal & Approach

**Goal**

***Optimize and advance*** key technologies that will accelerate performance of *mission-critical* systems

- *Software-based network management* for mission-critical deployments
- *Routing performance and adaptive parameters* for 5G satellite communications
- *Enhance AI integrAItor FY-2023* & focus on a *new generation of LLMs*[1]

## R&D Approach and F2 Projects

- **T1:** Develop *RAvN*[1] for adaptive and responsive SDN[2]-manages tactical network resilience
- **T2:** Develop *reinforcement learning techniques* for satellite topology reconfiguration.
- **T3:** Enhancements to *integrAItor* FY-2023
- **T4:** Focus on integration support of *federated learning,* using *domain-specific* LLM[3] and *RAG*[4]*/RCG*[5] generation of LLM[3]
  - **T4a:** *Fine-tuning* of *pre-trained* models (not training from scratch)
  - **T4b:** *RAG*[4]*/RCG*[5] generation for Large Language Models

[1] RAvN: Responsive auto-nomic and data-driven adaptive virtual networking framework

[2] SDN: Software-defined networks

[3] LLMs: Large Language Models

[4] RAG: Retrieval-Augmented Generation

[5] RCG: Retrieval-Centric Generation

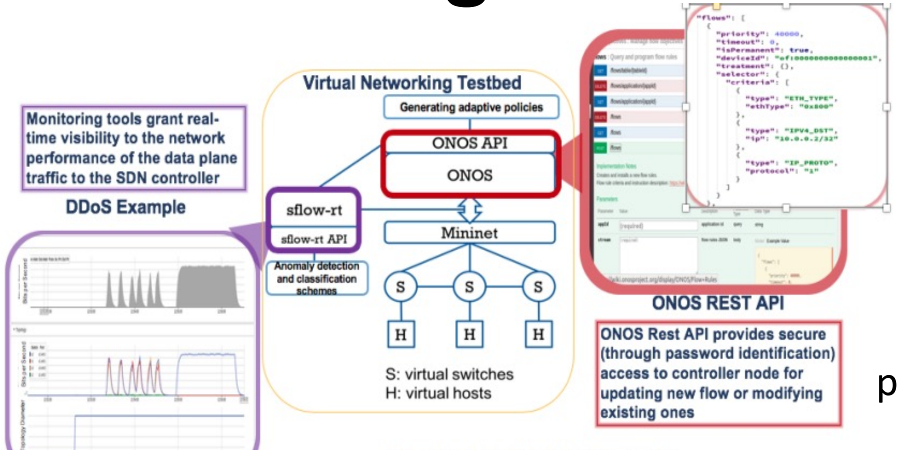
F2

# T1: RAvN[1] for SDN[2]-managed tactical network resilience

## Research Thrust 1

### Machine-Learning Approach

Determine tactical network status by analyzing a library of  observed network metrics.



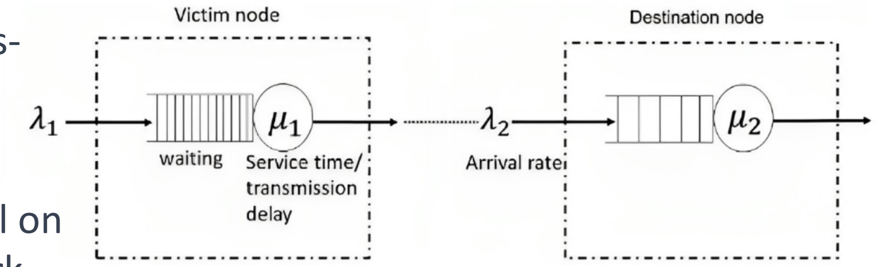
Monitoring tools grant real-time visibility to the network performance of the data plane traffic to the SDN controller

```
"flows": [
  {
    "priority": 40000,
    "timeout": 0,
    "isPermanent": true,
    "deviceId": "of:0000000000000001",
    "treatment": {},
    "selector": {
      "criteria": [
        {
          "type": "ETH_TYPE",
          "ethType": "0x800"
        },
        {
          "type": "IPV4_DST",
          "ip": "10.0.0.2/32"
        },
        {
          "type": "IP_PROTO",
          "protocol": "1"
        }
      ]
    }
  },
```

ONOS Rest API provides secure (through password identification) access to controller node for updating new flow or modifying existing ones

Adaptive Policy Update Framework

- **Data Generation:** Generate data and network samples utilizing SimComponents a network traffic simulation software developed based on the SimPY process-based discrete event simulation framework.
- **Training**: Train machine learning model on non-attack (normal) and attack network samples.
- **Analysis:** Detect, identify, and mitigate cyberattacks within a tactical network.



## Research Thrust 2

### SDN Controller Architecture

Only fine-tune a small number of (extra) model parameters while freezing most parameters of model

- **Architecture:** Comparison of decentralized, distributed, redundant, and hierarchical controllers.
- **Metrics**: Collect interarrival times, transmission delay, and packet counts received at a servers and controllers.
- **Actuation:** Metrics can be captured by the forwarding nodes and subsequently transmitted to the controller for onward transmission to the network operator. Isolate and redirect network traffic away from a compromised node.
- **Analysis:** Mininet network emulation analysis.

[1] RAvN: Responsive auto-nomic and data-driven adaptive virtual networking framework

[2] SDN: Software-defined networks

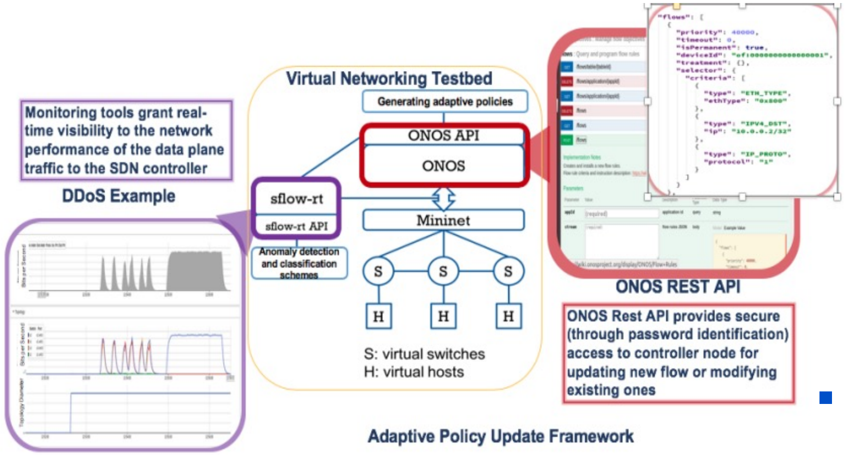
# T2: *Reinforcement learning techniques* for satellite topology reconfiguration.

## Research Thrust 1

### Machine-Learning Approaches

Explores the application of a shortest-distance reconfiguration algorithm in satellite constellations.

- **Shortest Distance Algorithm:** Address the performance disparity according to the size of the satellite constellations.
- **Training**: Train machine learning model on failure conditions, including device, link and signal failures.
- **Analysis:** Investigate using reinforcement learning or some other machine learning approach for satellite topology reconfiguration for various constellation sizes.

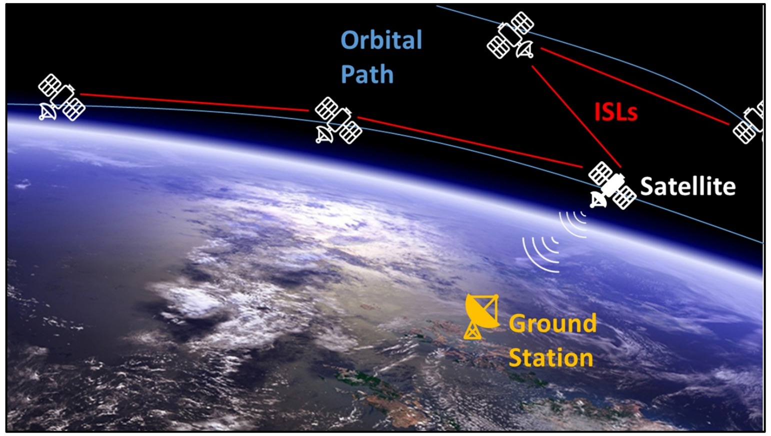


Monitoring tools grant real-time visibility to the network performance of the data plane traffic to the SDN controller

ONOS Rest API provides secure (through password identification) access to controller node for updating new flow or modifying existing ones

```
"flows": [
  {
    "priority": 40000,
    "timeout": 0,
    "isPermanent": true,
    "deviceId": "of:0000000000000001",
    "treatment": {},
    "selector": {
      "criteria": [
        {
          "type": "ETH_TYPE",
          "ethType": "0x800"
        },
        {
          "type": "IPV4_DST",
          "ip": "10.0.0.2/32"
        },
        {
          "type": "IP_PROTO",
          "protocol": "1"
        }
      ]
    }
  },
```

Adaptive Policy Update Framework



## Research Thrust 2

### Satellite Network Performance Analysis

Examine new tools for more accurate performance evaluation

- **SDN-based Approach:** Using SDN controllers to manage satellite topology.
- **Quantum Satellite Networks:** Begin an investigation of quantum networking for satellites
- **Topology:** Access to systems tool kit (formerly satellite tool kit for topology generation with connectivity data.
- **Metrics**: Collect connection times, duration, delay, transition time, from orbital dynamics.
- **Integrated Analysis:** Integrate STK output data with a network simulator, e.g., satellite network simulator 3, omnet++, or Mininet.

[1] RAG: Retrieval-Augmented Generation

[2] RCG: Retrieval-Centric Generation

F2

# T3: Enhancements to *integrAItor* FY-2023

## Research Thrust 1

### onDemand Developer Mode

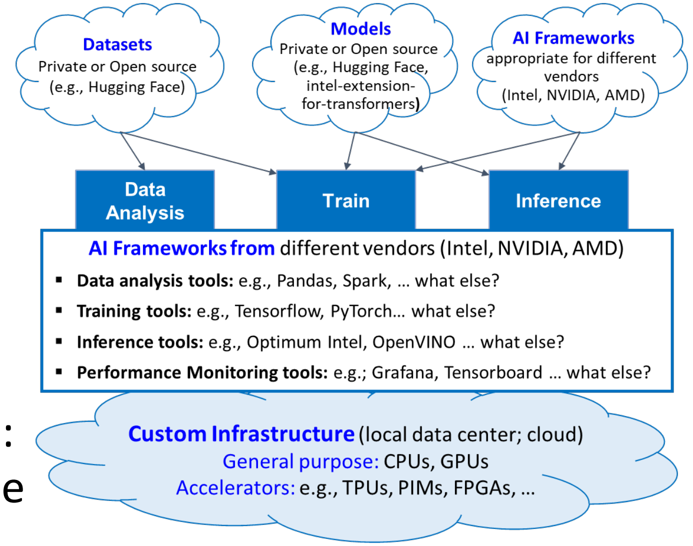
Interactive ***Jupyter Notebook*** environ. for *flexible* development, experimentation, & evaluation

A "***playground"*** that supports developers:

- to customize *existing* or write *new* code
- to flexibility explore, monitor, analyze, and optimize AI applications

Harness ***OnDemand*,***

- *an Open-source web portal* to access computer systems through the web
- enables computation from anywhere, on any device



**Datasets** Private or Open source (e.g., Hugging Face)

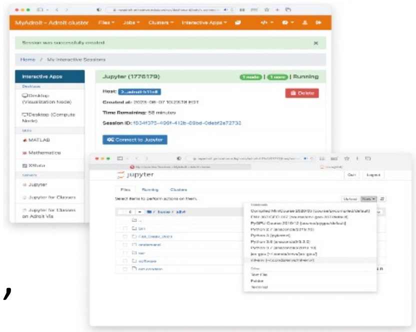
**Models** Private or Open source (e.g., Hugging Face, intel-extension-for-transformers)

**AI Frameworks** appropriate for different vendors (Intel, NVIDIA, AMD)

**AI Frameworks from** different vendors (Intel, NVIDIA, AMD)

- **Data analysis tools:** e.g., Pandas, Spark, … what else?
- **Training tools:** e.g., Tensorflow, PyTorch… what else?
- **Inference tools:** e.g., Optimum Intel, OpenVINO … what else?
- **Performance Monitoring tools:** e.g.; Grafana, Tensorboard … what else?

**Custom Infrastructure** (local data center; cloud)
General purpose: CPUs, GPUs
Accelerators: e.g., TPUs, PIMs, FPGAs, …



## Research Thrust 2

### Flexible Experiment Tracking & Monitoring

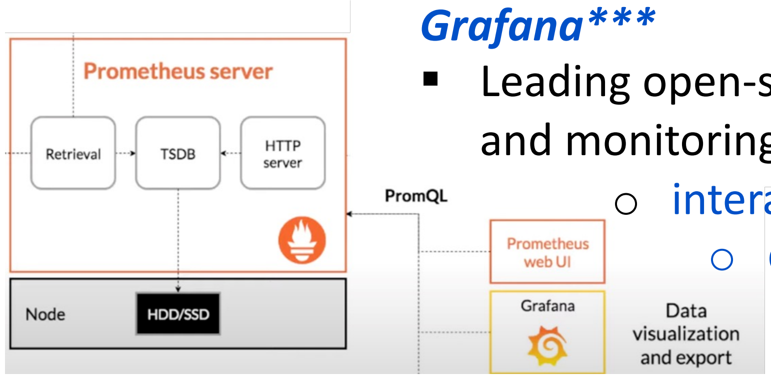
Extensive collection/presentation of evaluation metrics using ***Prometheus***** and ***Grafana******

***Prometheus*****

- Monitor/track metrics from servers, network, and applications to provide real-time insights

***Grafana******

- Leading open-source data visualization and monitoring platform:
  - interactive dashboards
  - data consolidation
  - highly customizable



* onDemand: https://openondemand.org/

** Prometheus: https://prometheus.io

*** Grafana: https://grafana.com/

# T4a: Fine-tuning of Pre-trained Models

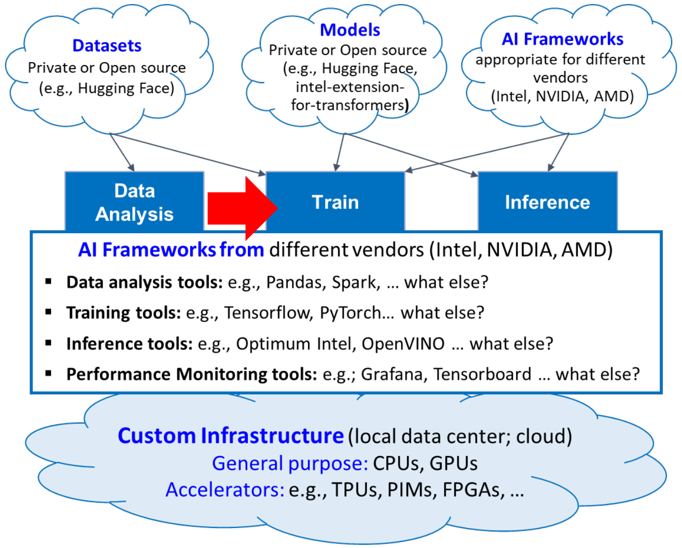
FY-2024 Focus on integration support of *federated learning* using *domain-specific LLM* & *RAG1[1]/RCG[2] generation of LLM*

## Research Thrust 1

**Train Models with *Transfer Learning***

Model trained on one task is adapted and fine-tuned for a different but related task

- **Inter-Task Compatibility:** Explore task compatibility in transfer learning to support selecting the best pre-trained models for given tasks
- **Adaptive Learning Rates**: Provide adaptive visualization tools to support experiments in varying learning rates



**Datasets** Private or Open source (e.g., Hugging Face)

**Models** Private or Open source (e.g., Hugging Face, intel-extension-for-transformers)

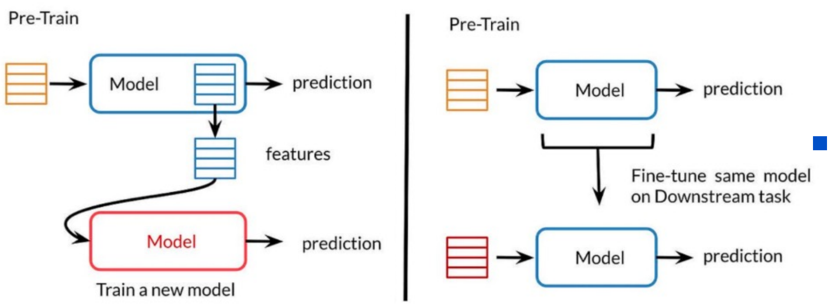
**AI Frameworks** appropriate for different vendors (Intel, NVIDIA, AMD)

Data Analysis → Train — Inference

**AI Frameworks from** different vendors (Intel, NVIDIA, AMD)

- **Data analysis tools:** e.g., Pandas, Spark, … what else?
- **Training tools:** e.g., Tensorflow, PyTorch… what else?
- **Inference tools:** e.g., Optimum Intel, OpenVINO … what else?
- **Performance Monitoring tools:** e.g.; Grafana, Tensorboard … what else?

**Custom Infrastructure** (local data center; cloud)
General purpose: CPUs, GPUs
Accelerators: e.g., TPUs, PIMs, FPGAs, …



## Research Thrust 2

**Fine-Tuning Model with *PEFT****

Only fine-tune a small number of (extra) model parameters while freezing most parameters of model

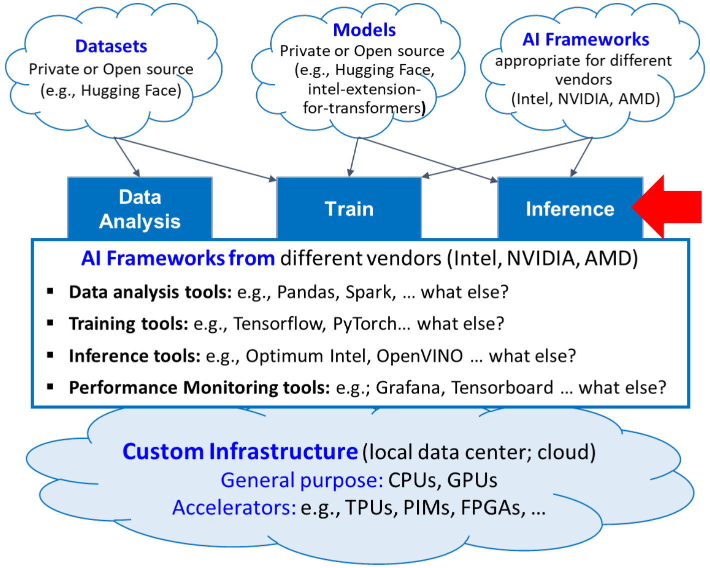
- **Incremental Learning** : Integrate existing and emerging libraries/tools for PEFT models to learn new data incrementally while retaining existing knowledge
- **Performance Monitoring:** Set up metrics and monitoring to quickly optimize PEFT

*PEFT: Parameter-Efficient Fine-Tuning

[1] RAG: Retrieval-Augmented Generation
[2] RCG: Retrieval-Centric Generation

F2

# T4b: RAG and RCG for LLM



Datasets
Private or Open source
(e.g., Hugging Face)

Models
Private or Open source
(e.g., Hugging Face, intel-extension-for-transformers)

AI Frameworks
appropriate for different vendors
(Intel, NVIDIA, AMD)

Data Analysis

Train

Inference

**AI Frameworks from** different vendors (Intel, NVIDIA, AMD)

- **Data analysis tools:** e.g., Pandas, Spark, … what else?
- **Training tools:** e.g., Tensorflow, PyTorch… what else?
- **Inference tools:** e.g., Optimum Intel, OpenVINO … what else?
- **Performance Monitoring tools:** e.g.; Grafana, Tensorboard … what else?

**Custom Infrastructure** (local data center; cloud)
General purpose: CPUs, GPUs
Accelerators: e.g., TPUs, PIMs, FPGAs, …

**Research Thrust 1:**

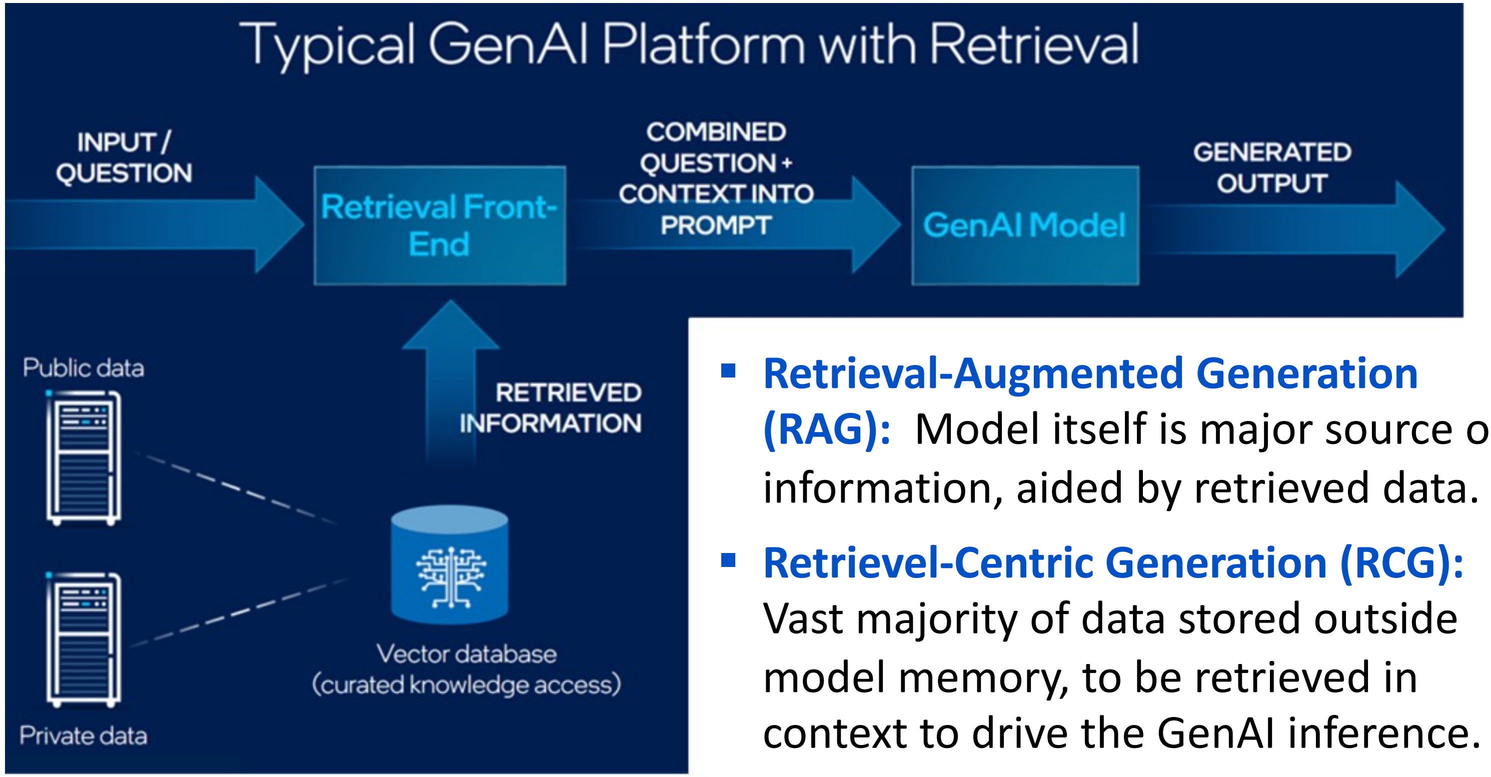
- Explore & determine capabilities of commercial and open-source RAG/RCG tools and libraries (e.g., from Intel, Hugging Face)

**Research Thrust 2:**

- Integrate RAG/RCG tools into *integrAItor*

**FY-2024** Focus on integration support of ***federated learning*** using ***domain-specific LLM*** & ***RAG/RCG generation of LLM***

## T4b: RAG and RCG for LLM*



Typical GenAI Platform with Retrieval
INPUT / QUESTION
Retrieval Front-End
COMBINED QUESTION + CONTEXT INTO PROMPT
GenAI Model
GENERATED OUTPUT
RETRIEVED INFORMATION
Public data
Private data
Vector database (curated knowledge access)

- **Retrieval-Augmented Generation (RAG):** Model itself is major source of information, aided by retrieved data.
- **Retrievel-Centric Generation (RCG):** Vast majority of data stored outside model memory, to be retrieved in context to drive the GenAI inference.

* "GenAI Architecture Shifting from RAG Toward Interpretive Retrieval-Centric Generation (RCG) Models", Gadi Singer, Director of Emergent AI Research at Intel Labs.

# Milestones, Deliverables & Budget

## Milestones

- SMW24: Showcase midway progress on framework, platform, and interconnect exploration
- SAW24-25: Present completed project results

## Deliverables

- Application source code and technology-transfer support
- Progress reports documenting research methods, progress, results, and analysis
- Several conference and/or journal publications

## Membership Budget

- Requesting ≥ 4 memberships

# Conclusions & Member Benefits

## Conclusions

- Creating network architectures for applications, *such as security and quality of service,* that can generate and leverage *real-time situational awareness* through new network profile *data sets*, network *models*, and *machine learning and AI-based* protocols.
- A developer is faced with a complex array of choices: *dataset*, *model*, AI *framework*, & hardware *infrastructure*
  - The goal is to enhance *AI integrAItor* FY-2023 & focus on a new generation of LLMs

## Member Benefits

- Direct influence over selected architecture, app, and inter-connect studies
- Technology transfer of accelerated archs/apps/techniques of interest to members
- Key insights and lessons learned from design space explorations & tradeoff analyses