F2-25: Development of Large AI Applications and Systems

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

SHREC Annual Workshop (SAW24-25)

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

January 14-15, 2025

Dr. Janise McNair
Professor of ECE

Dr. Herman Lam
Assoc. Professor of ECE

S. Boamah, C. Cheemarla, A. Koti, P. Mangipudi,
A. Rice-Bladykas
Research Students
University of Florida

Number of requested memberships 3 to 4

Introduction

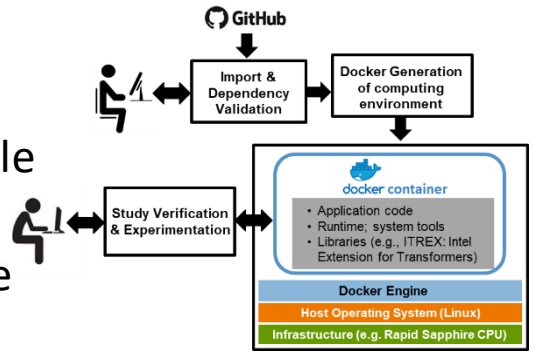
Unprecedented advances in AI and LLMs*

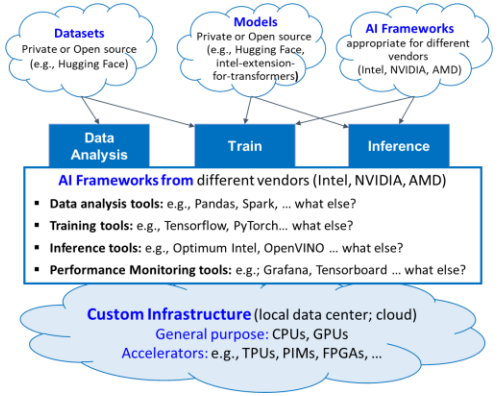
* Overwhelming availability of *datasets*, *models*, *AI tools*, and *hardware* platforms/devices

** Emergence of a large body of research papers

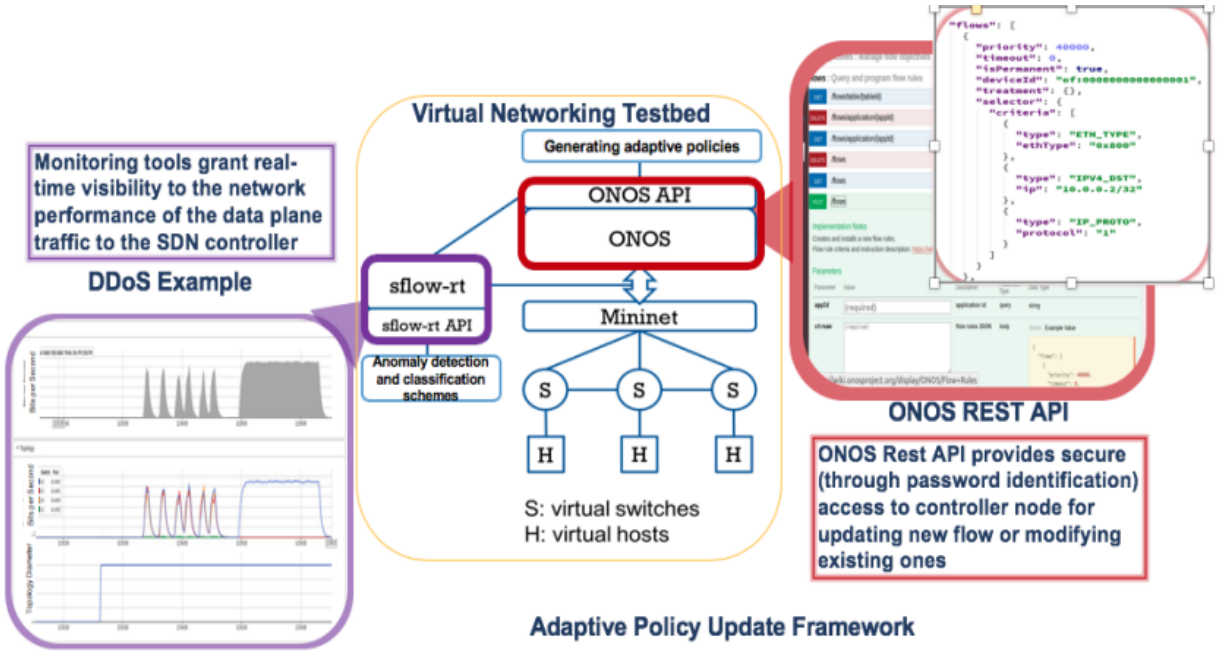
- Many published with *associated codebase*

** *AI integrAItor* ***Research*** *Mode*: Rapid-prototyping tools to enable *study* & *experimentation* of published papers w/codebase





* *AI integrAItor* ***Development*** *Mode*: Rapid-prototyping tools to manage/support the *development* of AI/LLM *applications* and *systems*



- 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.

* LLMs: Large Language Models

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
- *AI integrAItor-2025:* enhancement of *integrAItor-2024* & new *integrAItor-2025* capabilities

R&D Approach and F2 Projects

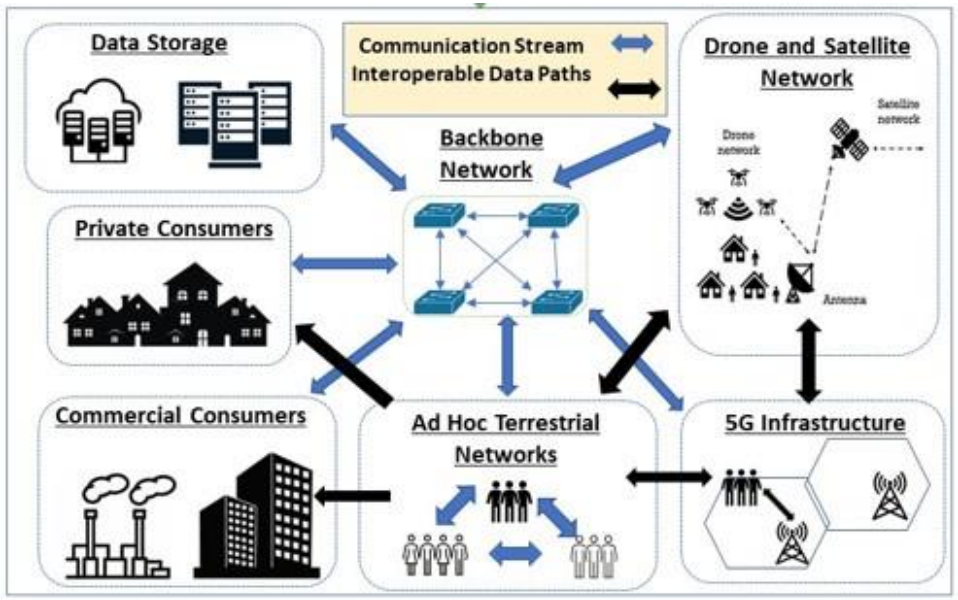
- **T1:** Develop adaptive and responsive SDN[1]-managed 5G interoperable networks.
- **T2:** Develop *reinforcement learning techniques* for satellite topology reconfiguration.
- **T3:** Enhancement of *integrAItor-2024* capabilities
  - **T3a: Development** Mode enhancements; **T3b: Research Mode** enhancements
- **T4:** New *integrAItor-2025 c*apabilities: complement 2024 user-friend *GUI support* with flexible support for *advanced users*, using full capabilities of *OnDemand* [2], *Prometheous* [3], and *Grafana* [4]

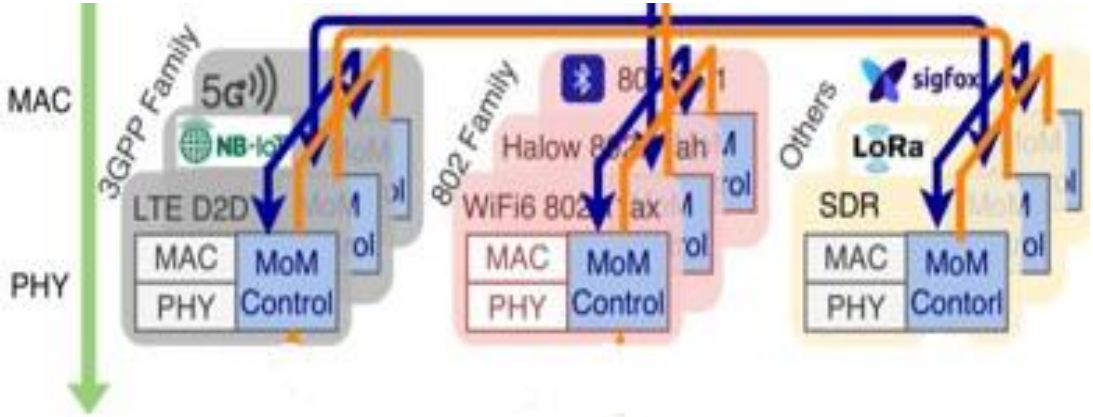
[1] SDN: Software-defined networks

[2] OnDemand (https://openondemand.org/)
[3] Prometheus Monitoring System (https://prometheus.io/)Grafana
[4] Data Visualization System (https://grafana.com/)

T1: 5G Multi Radio Access Technology (RAT) Interoperable Networks

- Multi RAT networks consist of various RATs coexisting with each other, giving the opportunity to increase connectivity for Beyond 5G networks
- Inclusion of multiple criteria from the available set of performance characteristics for quality of service of each RAT, improves overall system performance.
- Strategic offloading of users from cellular network to non cellular networks (local ad hoc, drones, satellite) within the multi-RAT network, can preserve as much of the 5G spectrum possible.





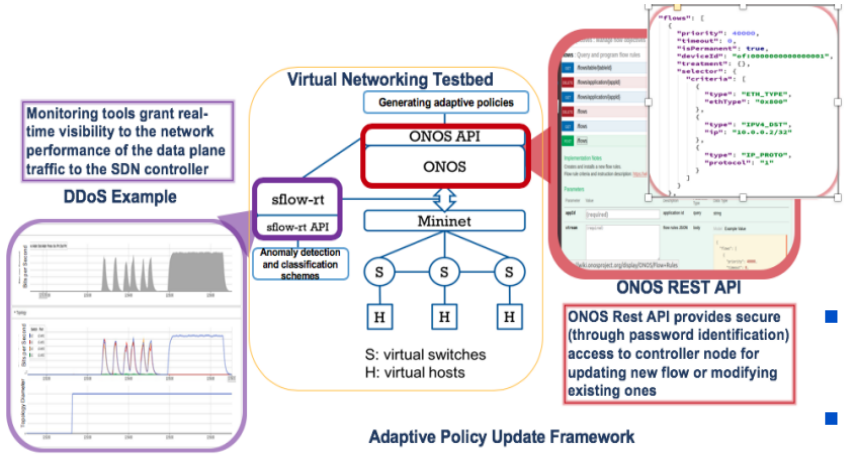
T2: *Reinforcement learning techniques* for satellite topology reconfiguration.

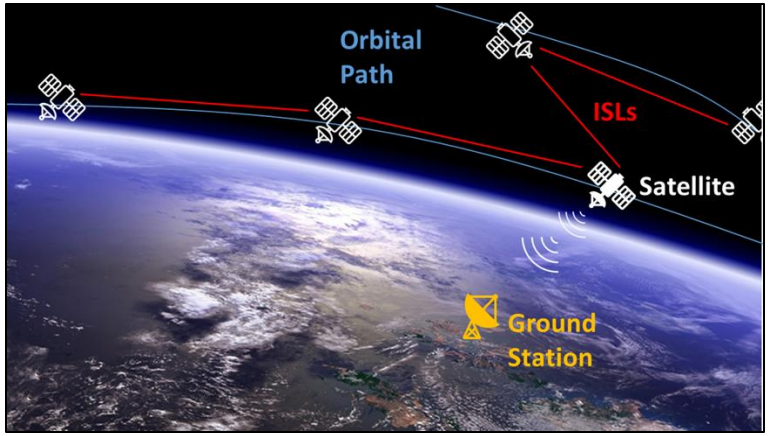
Research Thrust 1

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.





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 (Collaboration with Pitt)** Integrate STK output data with a network simulator, e.g., satellite network simulator 3, omnet++, or Mininet.

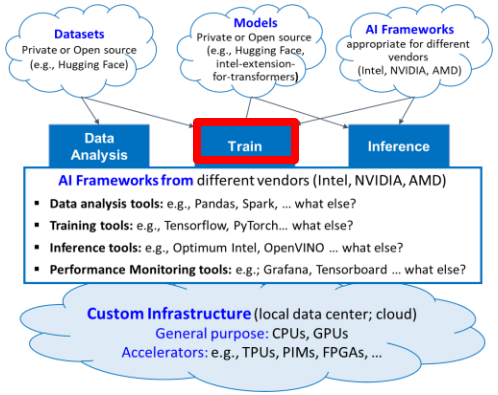
[1] RAG: Retrieval-Augmented Generation

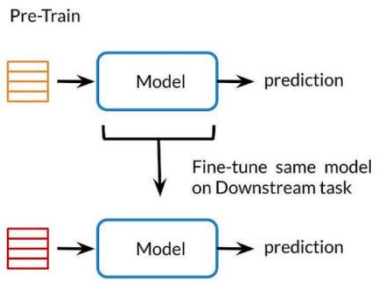
[2] RCG: Retrieval-Centric Generation

T3: Enhancement of *integrAItor-2024* Capabilities

Task T3a

AI integrAItor
Development Mode
Enhancements



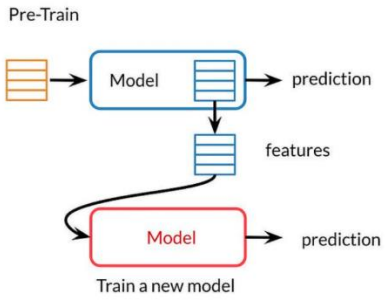


Fine-Tuning Model with *PEFT**

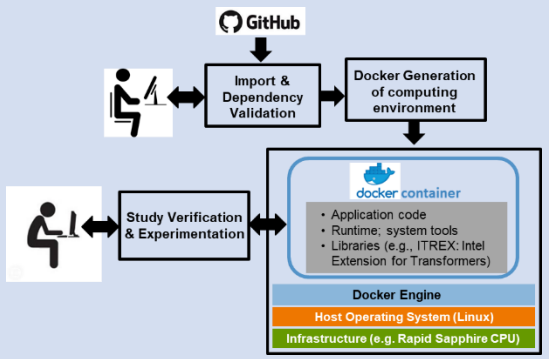
Only fine-tune a small number of (extra) model parameters while freezing most parameters of model

Train Models with *Transfer Learning*

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



Task T3b



AI integrAItor
Research Mode
Enhancements

Ability to perform *extensive experiments on import codebase* using ***different***:

- Datasets, models, AI frameworks
- Hardware platforms (with new *devices/accelerators*)

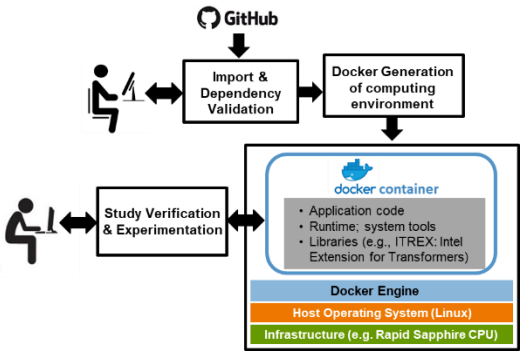
Integration of ***Prometheus**** and ***Grafana**** into *AI integrAItor* (for codebase experimentation)

- Integrates and correlates ***system performance*** metrics with ***model performance*** metrics
- Support discovery of actionable insights from integrated (system/model) performance data

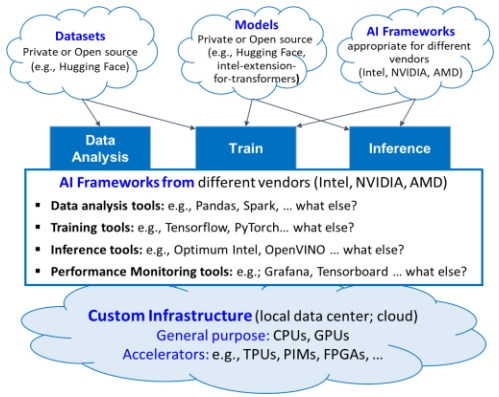
*PEFT: Parameter-Efficient Fine-Tuning (Initial version completed in 2024)

** Prometheus Monitoring System (https://prometheus.io/)
Grafana Data Visualization System (https://grafana.com/)

T4: Advanced User Support in *integrAItor-2025*



Research Mode



Development Mode

For both **Research** and **Development** modes:

- *integrAItor-2024* has user-friend ***GUI*** support
  - Easy to used, but restricted to the menu items
- Complement w/ ***flexible advanced user support***
  - Using full capabilities of *OnDemand* [1], *Prometheous* [2], *and Grafana* [3]

OnDemand[1] (Advanced) Developer Mode

Interactive ***Jupyter Notebook*** environment for *flexible* development, experimentation, & evaluation

OPEN OnDemand

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

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

Flexible Experiment Tracking & Monitoring

Support advanced user with extensive collection/presentation of evaluation metrics using full power of:

Prometheus[2]

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

Grafana[3]

Grafana

Leading open-source data visualization and monitoring platform:

[1] OnDemand (https://openondemand.org/)
[2] Prometheus Monitoring System (https://prometheus.io/)
[3] Grafana Data Visualization System (https://grafana.com/)

Milestones, Deliverables & Budget

Milestones

- SMW25: Showcase midway progress on framework, platform, and interconnect exploration
- SAW25-26: 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 3 to 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*-2024 & 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