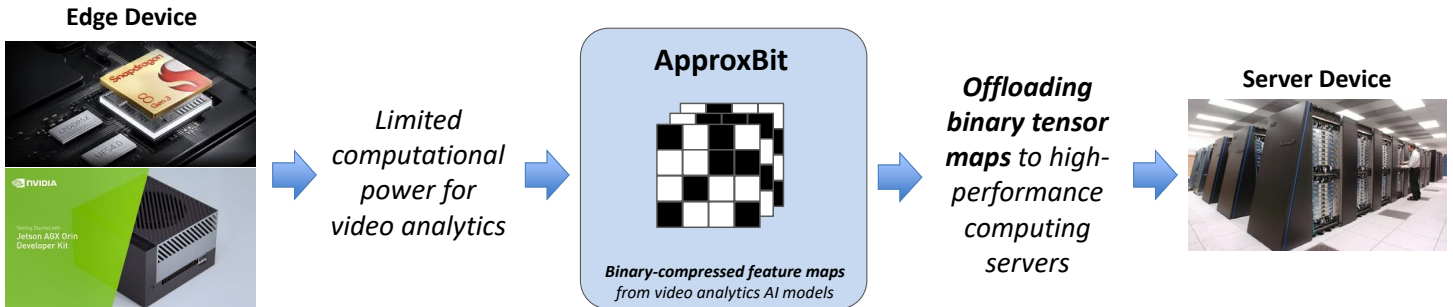# CPS Frontier: CHORUS: Resilient Distributed CPS through Rational and Dynamic Decision-Making Among Multiple Stakeholders

ApproxBit: Efficient Video Action Recognition through Latency-Aware Offloading with Learned Binary Codes

PI: Saurabh Bagchi, Purdue University

Co-PIs: Somali Chaterji (Purdue), Yin Li (Wisconsin). Graduate Researcher: Hyunseung Kim

https://choruscomputes.xyz

Edge Device → *Limited computational power for video analytics* → **ApproxBit** *Binary-compressed feature maps from video analytics AI models* → *Offloading binary tensor maps to high-performance computing servers* → **Server Device**

Problem Statement

- **Problem:** Edge devices have limited computational power and bandwidth for video analytics.
- **Significance:** High-quality video recognition demands substantial processing and data transmission.
- **Challenge:** "Efficiently offloading video features" while minimizing latency and preserving accuracy.

Scientific Impact

- Applicable to Computer and Cyber-Physical Systems (CPS) needing **efficient edge-cloud interactions**.
- Enables **autonomous driving, surveillance, and augmented reality**, where real-time processing is critical.


Our Solution: ApproxBit

[Overall framework]

[**ApproxBit** Module]
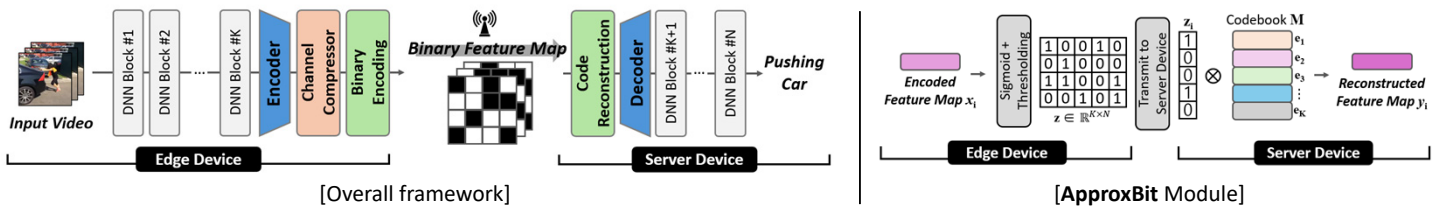
- **Binary quantization** minimizes transmission overhead.
- Implements an **adaptive offloading module** that dynamically adjusts configurations based on network conditions.

- Uses **learned binary codes** to optimize feature transmission, reducing bandwidth usage.
- **Lightweight encoder** at the edge and **heavy decoder** at the server balance computation.

Result
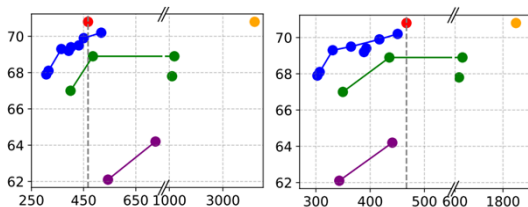
Accuracy-latency graph for MViTv2 model on A100-Jetson AGX Orin at 3Mbps (left) and 6Mbps (right)

- **Superior Performance:** Our method(•) outperforms DeepCOD (•, Sensys 20') and LimitNet (•, MobiSys 24') in accuracy-latency trade-offs.
- **Network Adaptability:** Adapts dynamically to network conditions, ensuring reliability across various bandwidths (e.g., 3Mbps ~ 50Mbps).

Potential impact

- ApproxBit can lead to **significant energy and cost savings** in large-scale edge deployments.
- Enables real-time AI processing in **resource-constrained environments** like autonomous systems, smart surveillance, and industrial IoT.