FECBench – A Framework for Measuring and Analyzing Performance Interference Effects for Latency-sensitive Applications

Yogesh Barve

Collaborators: Shashank Shekhar, Shweta Khare, Anirban Bhattacharjee, Ajay Chhokra, Aniruddha Gokhale

yogesh.d.barve@vanderbilt.edu
Institute for Software Integrated Systems,
Dept. of Electrical Engineering and Computer Science
Vanderbilt University, Nashville, TN, USA
Cloud Computing Trends for CPS

- Soft real-time Cyber-Physical Systems (CPS) and Internet of Things (IoT) applications, such as cognitive assistance, patient health monitoring, and connected vehicles are increasingly using the cloud
  - Elasticity
  - Cost benefits
- The applications have **stringent** Quality of Service (QoS) needs
- Cloud must provide:
  - predictable performance
  - lower end-to-end response time
Cloud for CPS: Key Challenges

- Performance Variability Challenges: **Interference**
- CPS tasks can experience **QoS violations**

**Machine Learning InceptionResNetV2 Model**

- **Inference Serving Performance**

- Lack of Frameworks and Methodologies to Characterize **Performance Interference**
FECBench Workflow

**OFFLINE** - One Time Profiling Step for Each Hardware

**ONLINE** - Repeat for Each New App
FECBench in Action

UPSARA: The Modeling Design Studio of FECBench

Metric Monitoring Configuration

Benchmark Deployment Configuration

Benchmarking Harness