

Mainstream: Dynamic Stem-Sharing for Multi-Tenant Video Processing

Angela Jiang, Daniel L.-K. Wong, Christopher Canel, Ishan Misra, Michael Kaminsky (Intel Labs), Michael Kozuch (Intel Labs), Padmanabhan Pillai (Intel Labs), David Andersen, Greg Ganger

Overview

Goal: Efficiently run concurrent streaming video analysis apps

Problem:

- Most video analysis apps perform DNN inference
 - Running several full DNNs becomes very slow

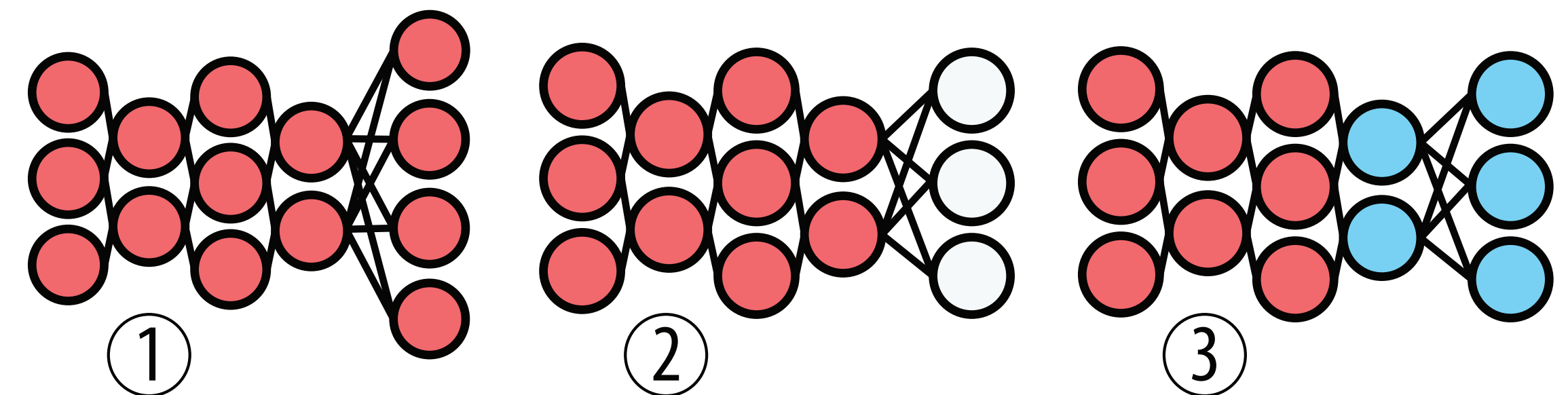
Mainstream:

- Identifies and shares redundant DNN computation
 - By exploiting nature of fine-tuned DNNs
- Decides at runtime how much to share
 - Balances specialization vs. sharing trade-off
 - Optimizes when hardware and set of apps is known

Transfer Learning

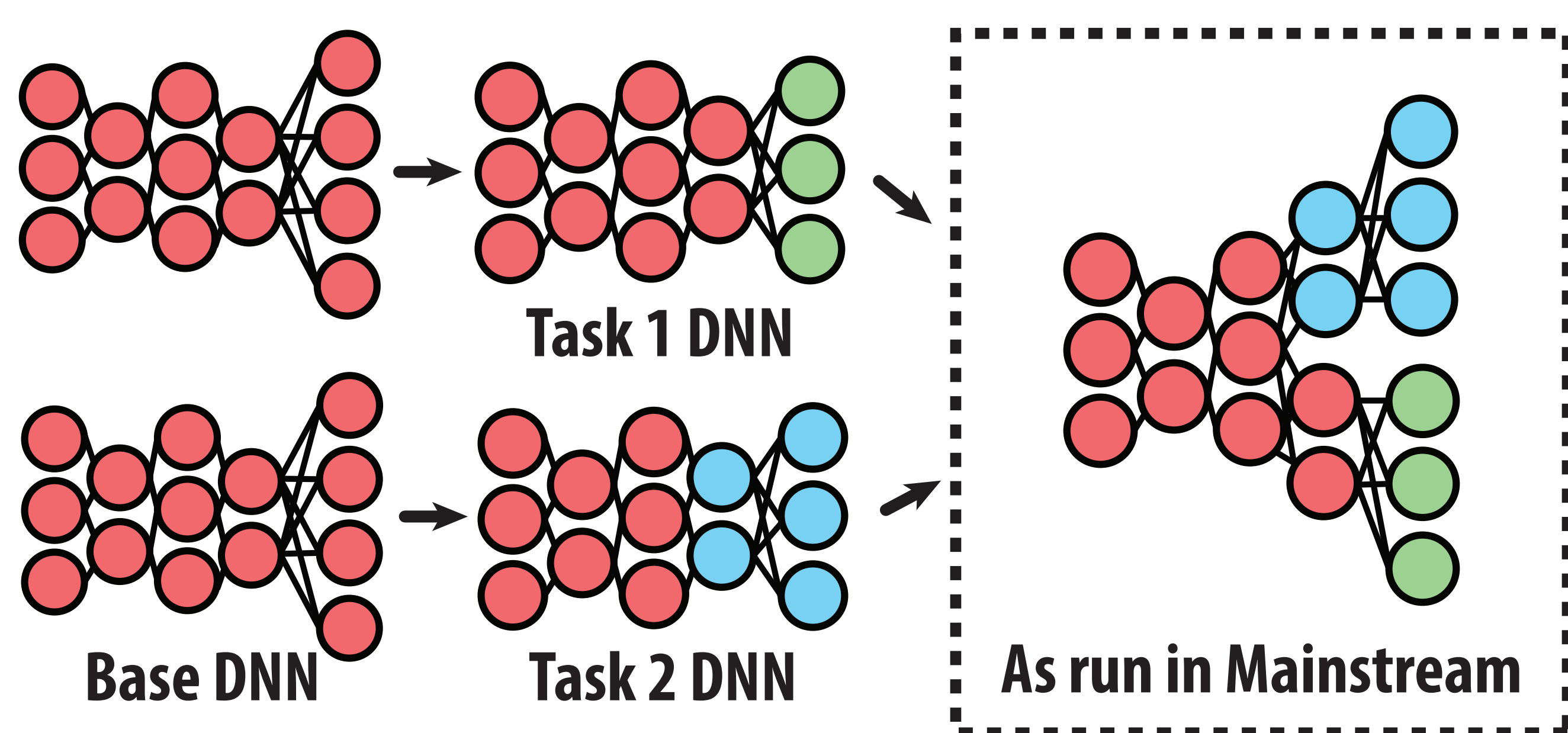
When training task B, start with DNN pre-trained for task A

- Common practice for training networks

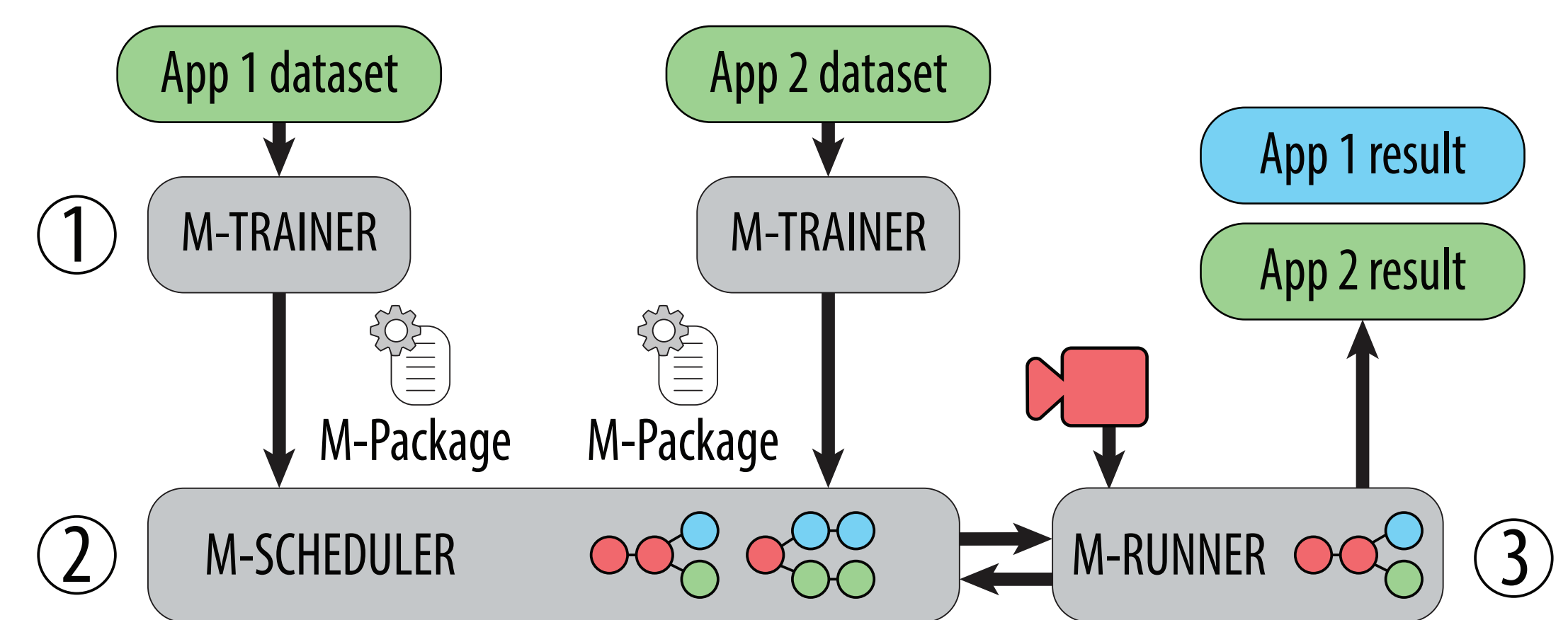


- Network is trained from scratch for task A (e.g., ImageNet)
- Replace A-specific final layer with B-specific final layer
- Fine-tune part of network for task B, other layers held frozen

Sharing Computation



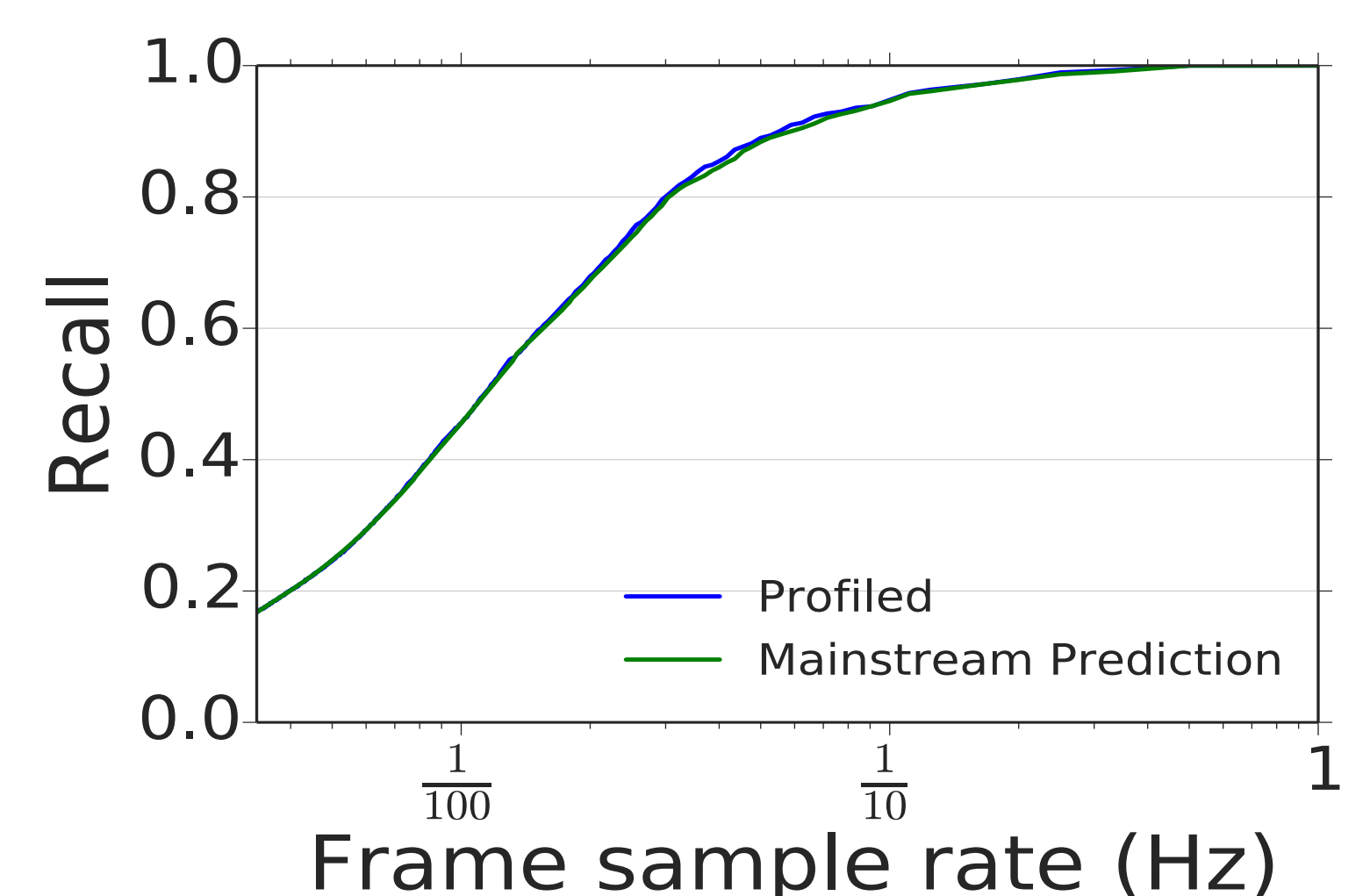
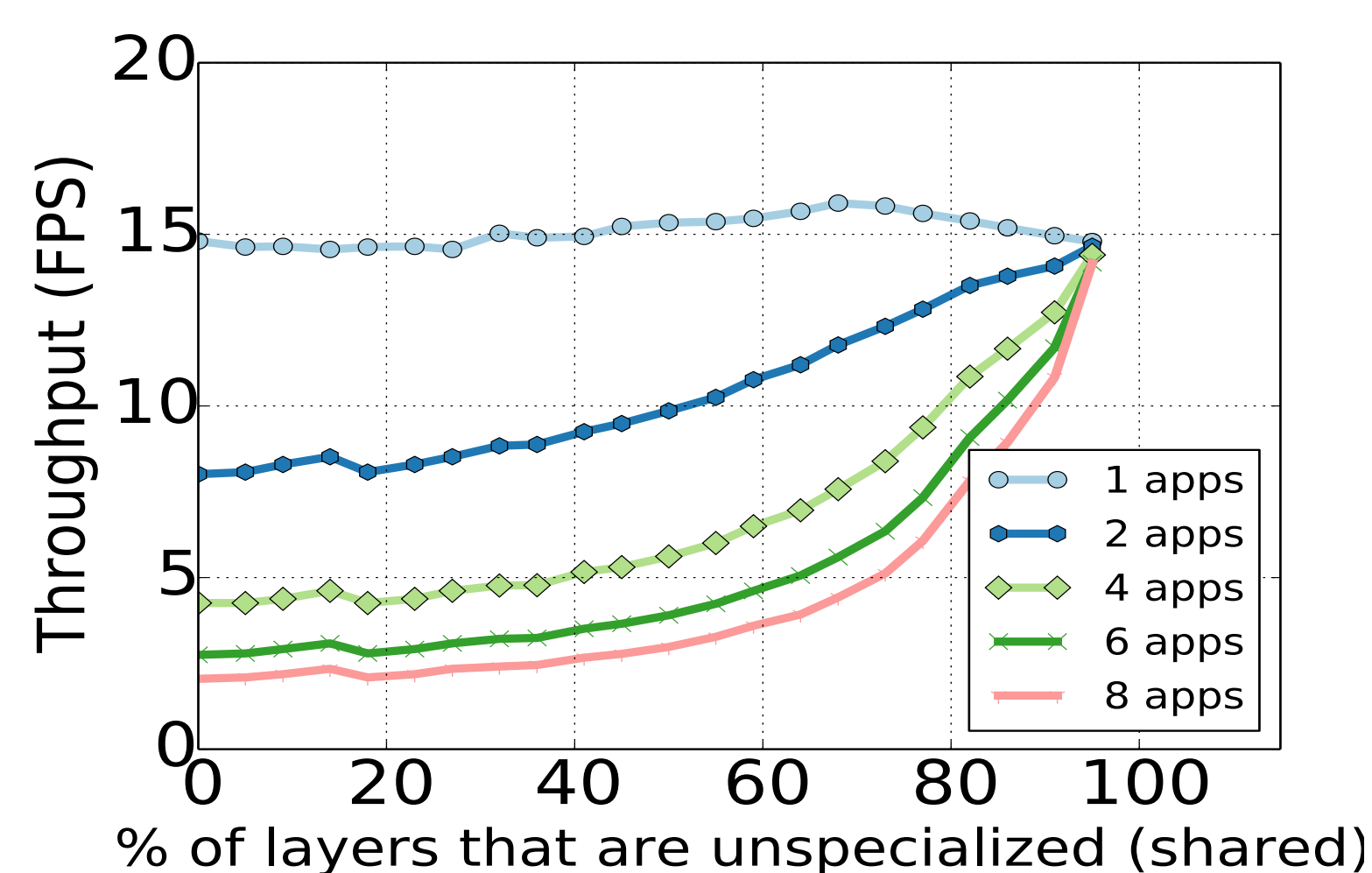
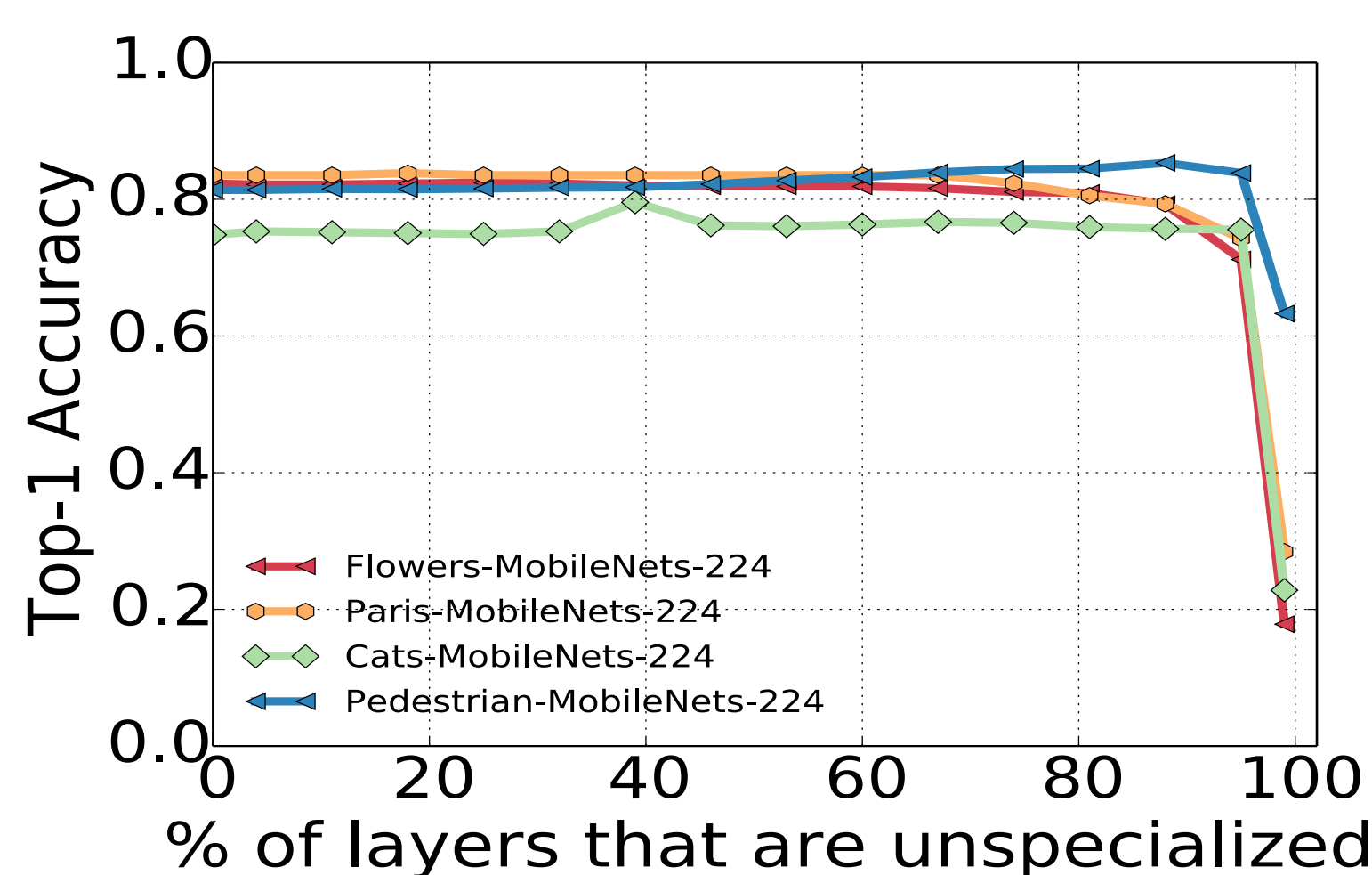
Mainstream Architecture



- M-trainer trains DNNs w/ varying % of network held frozen
- M-Scheduler determines amt. of DNN to share for each app
- M-Runner processes video stream using deployed DNNs

Specialization vs. Sharing Trade-off

Experimental setup: Train image classifiers using MobileNet-224. Run simultaneous classification pipelines on an Intel NUC.



Less specialization → Lower per-frame acc.

Less specialization → Higher throughput

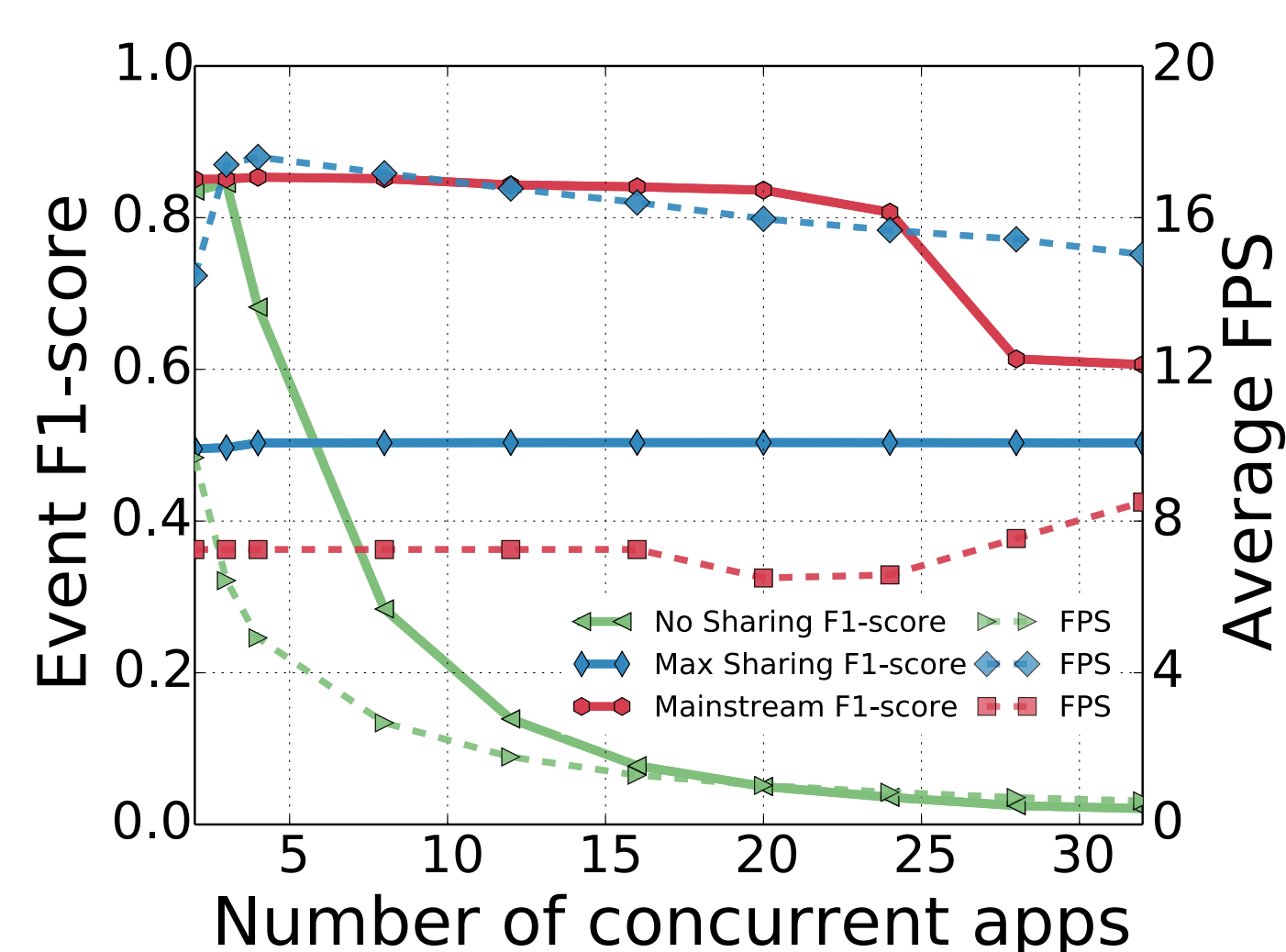
Use analytical model to find/navigate trade-offs

Application Performance

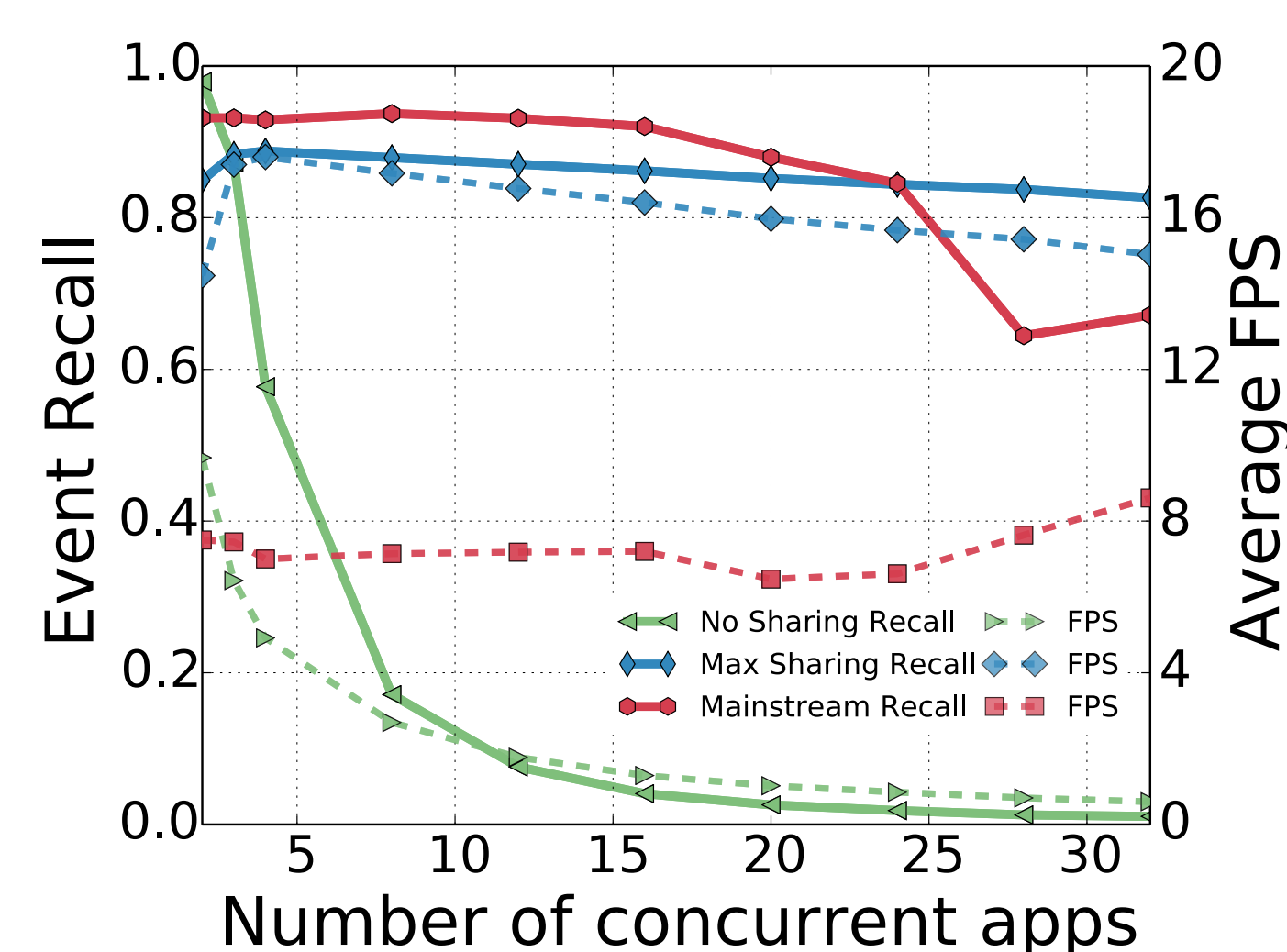
Recall: % of events detected;

Precision: % of detected events that are correct;

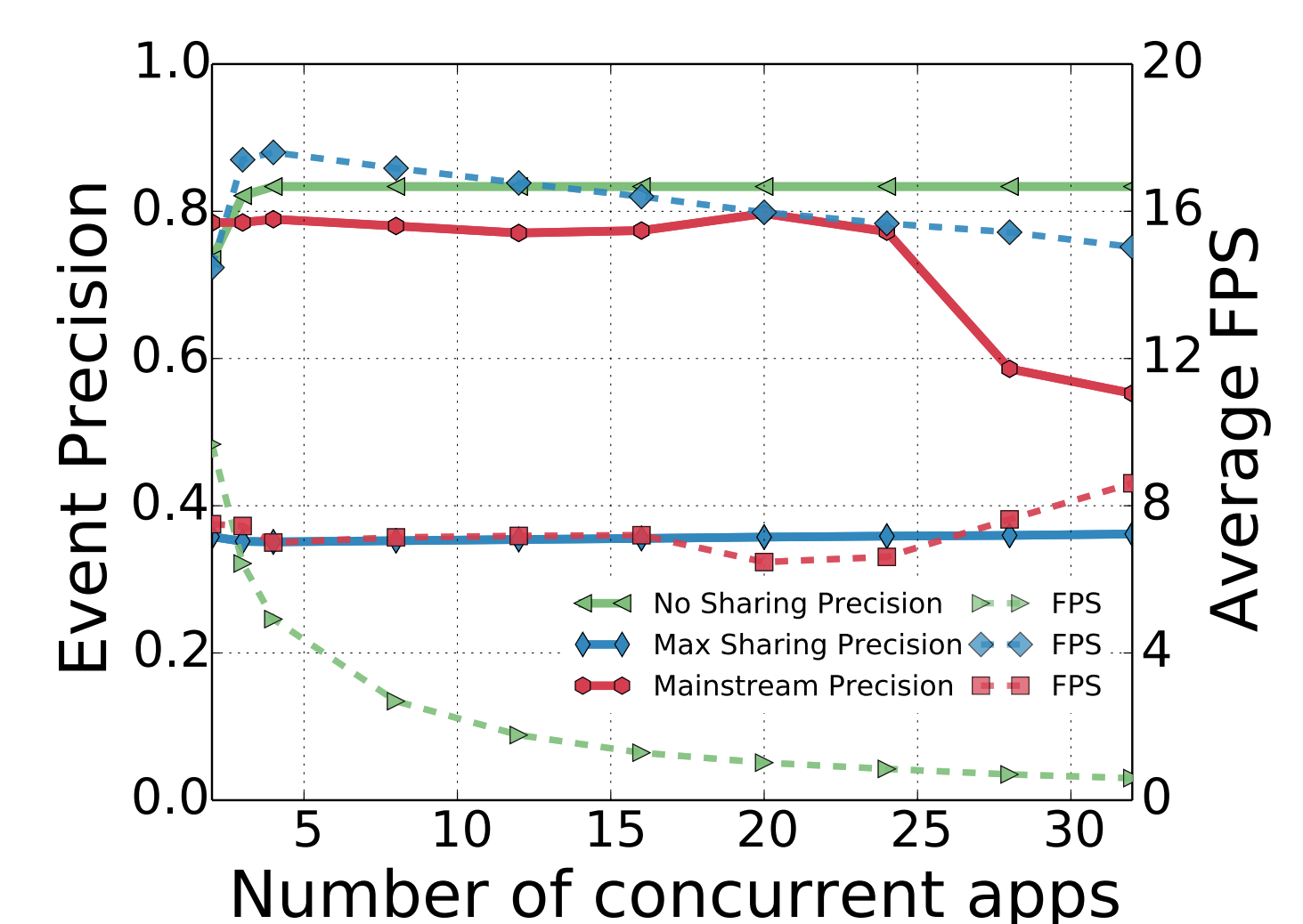
F1 score: Harmonic mean of precision and recall



- “No Sharing” deploys full DNN for each app
- “Max Sharing” shares all but final layer
- Mainstream achieves up to 28X higher F1



- “No Sharing” (NS) has low FPS, high acc.
- NS misses events, incurring low recall
- Mainstream balances FPS and acc.



- “Max Sharing” (Max) has high FPS, low acc.
- Max’s false positives cause low precision
- Mainstream balances precision and recall