

Scaling Video Analytics on Constrained Edge Nodes

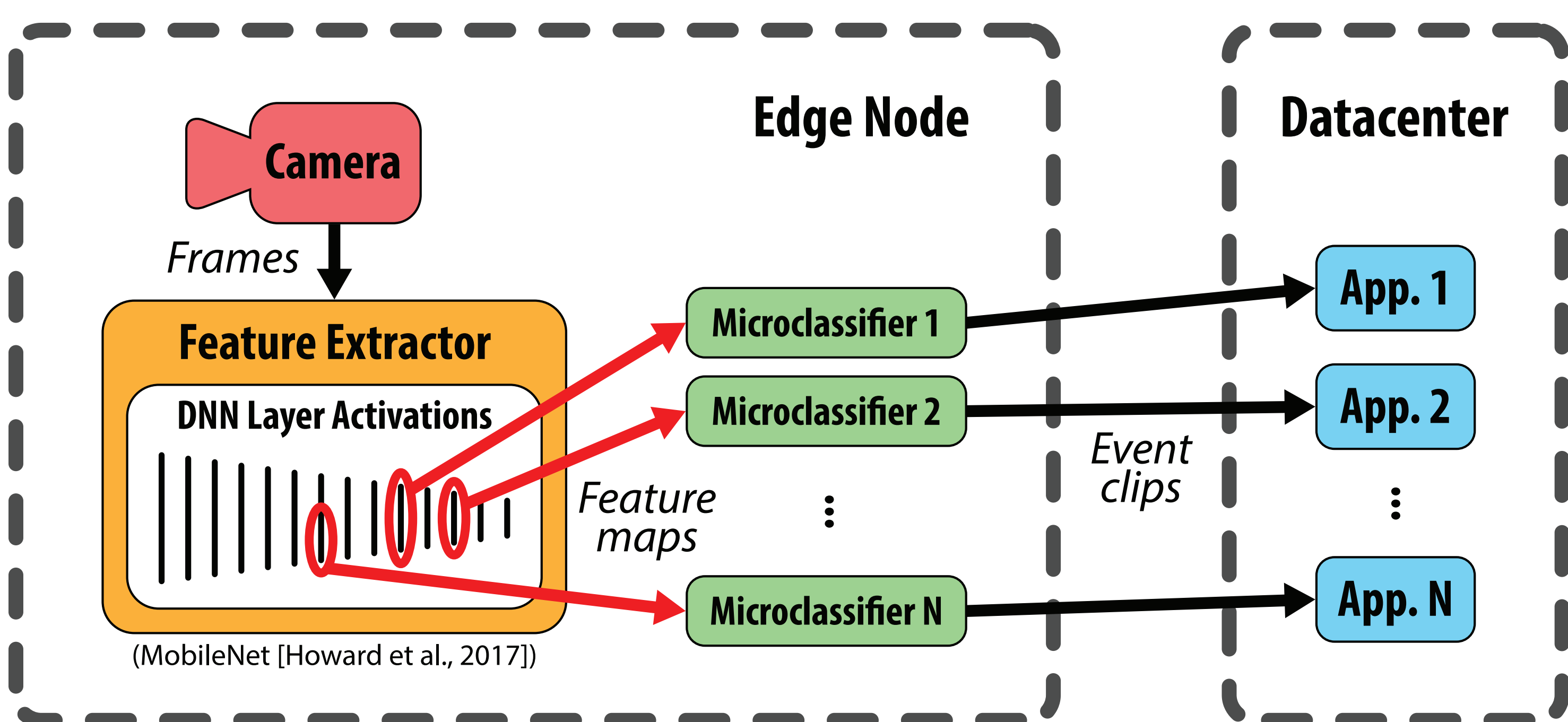
Christopher Canel, Thomas Kim, Giulio Zhou, Conglong Li, Hyeontaek Lim,
David G. Andersen, Michael Kaminsky[†], Subramanya R. Dulloor[‡]
Carnegie Mellon University; [†]Intel Labs; [‡]ThoughtSpot

MOTIVATION

- City-wide camera deployments are ubiquitous
- Apps running in the cloud demand high-quality video
- Data volume is cost-prohibitive for wide-area networks

OUR WORK

- Saves bandwidth by dropping irrelevant frames
- A scalable framework that enables edge nodes to determine which frames are relevant to cloud apps



FilterForward (FF): Accurate and Scalable Video Filtering on the Edge

Compression obscures critical, subtle details



1000 Kb/s 100 Kb/s

KEY IDEAS

- Apps receive high-quality data at a fraction of the bandwidth cost
- Exposing fine frame details enables precise filtering
- Computation sharing allows filtering to scale to many apps

Datacenter:

- Communicates filter parameters to edge nodes
- Receives relevant frames, distributes them to apps

Edge node (the core of FilterForward):

- Runs a base Deep Neural Network (e.g., MobileNet) on each frame to extract semantic features
- Uses features as input to many *microclassifiers* (MCs), thereby amortizing expensive pixel processing
- Forwards frames filtered by MCs back to cloud

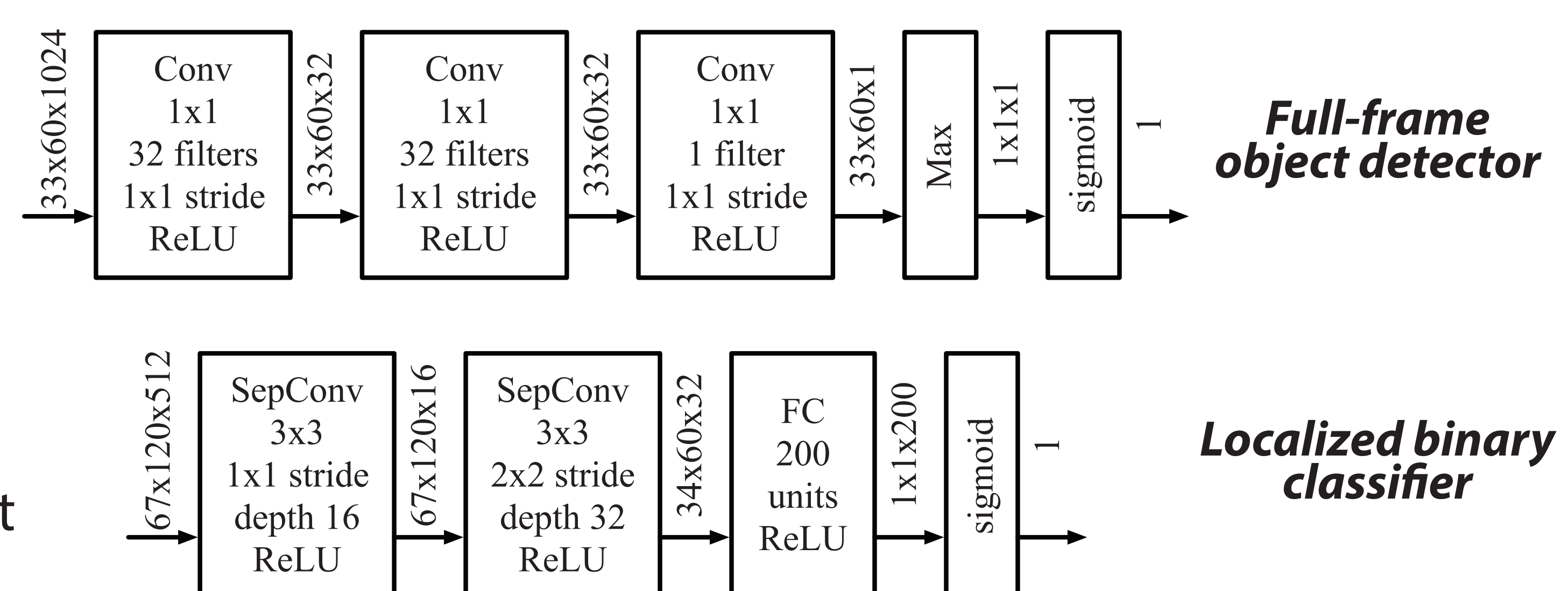
Microclassifiers: Small Neural Networks that Filter Content Relevant to Cloud Applications

Sensitive to fine-grained details

- Base DNN processes full-resolution (e.g., 1080p) frames
- Feature maps are drawn from any layer in the base DNN
- Spatially cropping feature maps enables MCs to localize

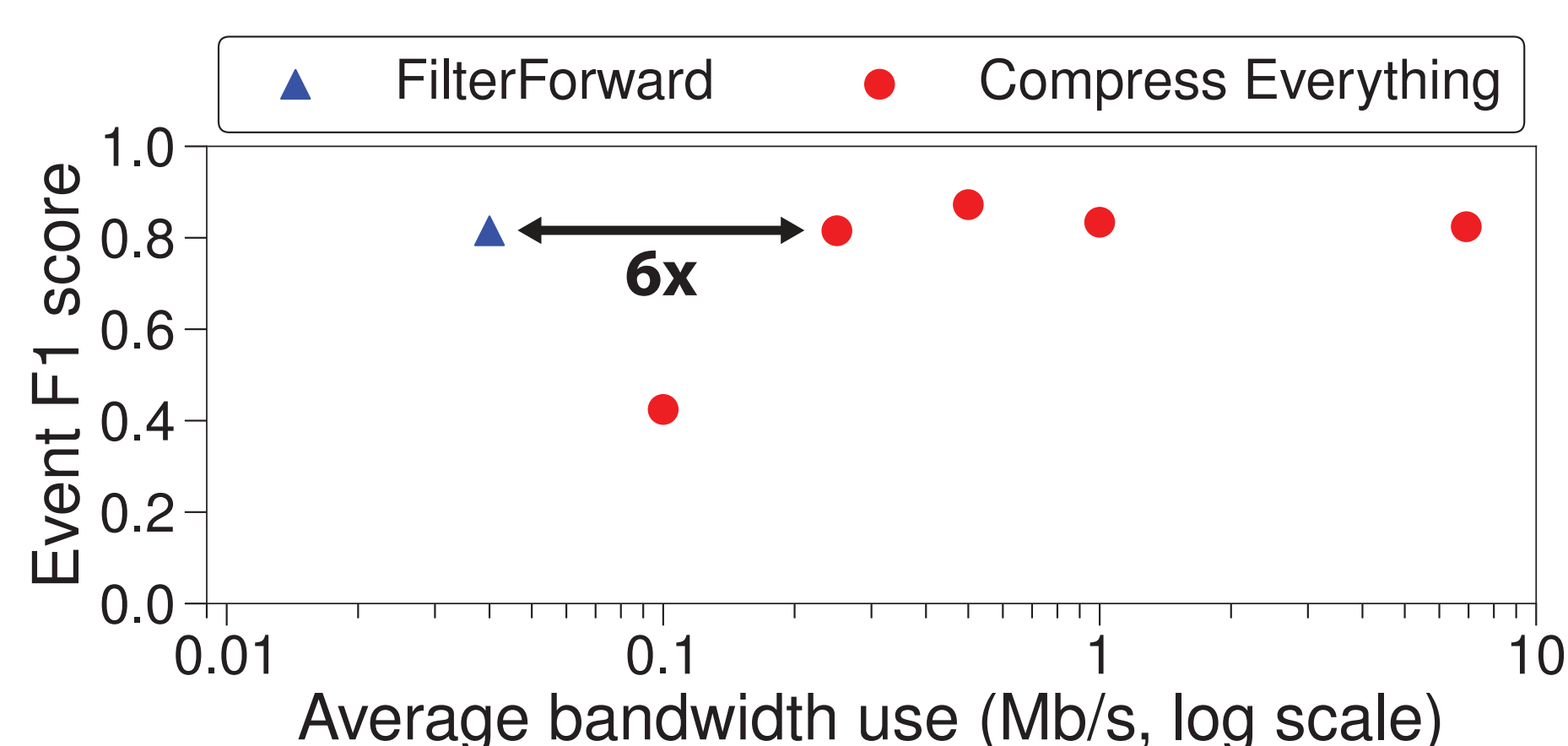
Computationally scalable

- Using a base DNN shares the pixel processing across all MCs
- Operating directly on semantic features makes MCs lightweight
- Adding more MCs amortizes the expensive base DNN

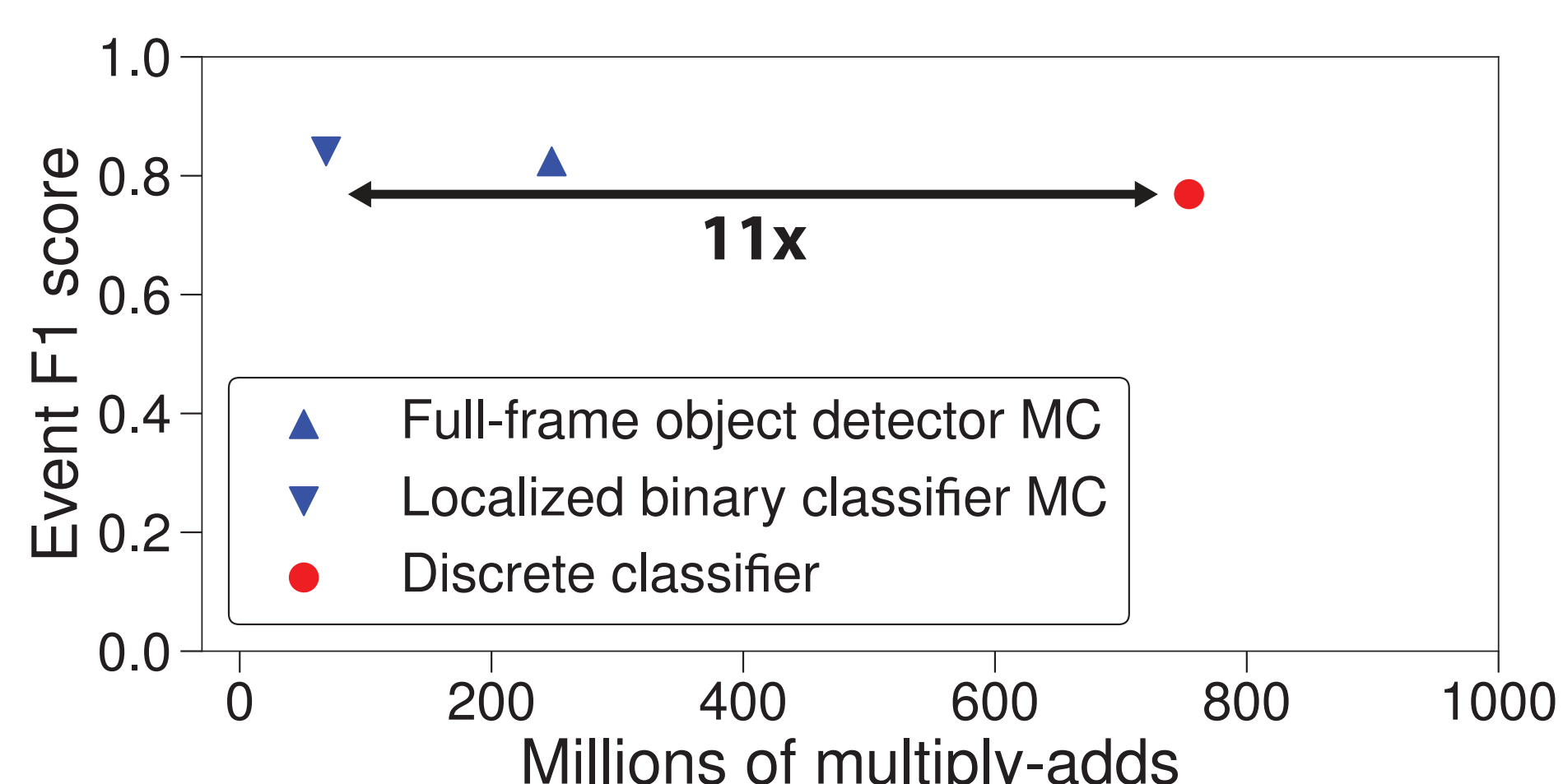


Evaluation: FilterForward Uses Less Bandwidth for Same Accuracy; Scales to Many Apps

FilterForward uses **6x** less bandwidth compared to compressing the entire video, without sacrificing accuracy



Microclassifiers have up to **11x** lower marginal cost, yet similar accuracy, compared to pixel-level classifiers (NoScope [Kang et al., 2017])



Computation sharing makes FF more efficient than pixel-level classifiers when running ≥ 4 , with **4x** higher throughput with 20 classifiers

