Introduction to Facebook AI Performance Evaluation Platform

Fei Sun

Software Engineer, AI Infra Developer Platform, Facebook

Key Components of Edge ML Benchmarking

- Relevant edge ML Models
- Representative edge ML workload
- *Unified and standardized methodology of evaluating the metrics in different scenarios*

What is FAI-PEP

- Backend & Framework agnostic benchmarking platform
- Normalize the benchmarking metrics and conditions
- Automate the benchmarking process
- Honest measurement on performance

- https://github.com/facebook/FAI-PEP

The Use Cases of FAI-PEP

- Compare performance/quality of the models
- Compare performance of the software stack
  - Different commits
  - Different software kernels
- Compare performance of the hardware implementations

- Facebook uses FAI-PEP internally the same way as OSS

Design Philosophy

- Generalizability
- Explicitness
- Composability
- Extensibility
- Centralization

The Runtime

- One runtime per model or one runtime for all models?
- FAI-PEP recommends to use one runtime for all models
- Pytorch/Caffe2 will provide one runtime to measure both quality and performance
  - ONNX models supported

Who Owns the Validation?

- Performance metrics are measured by the runtime
  - Too much overhead if measured outside
  - One runtime for all models to reduce uncertainty
- Accuracy metrics are not measured by the runtime
  - Runtime takes inputs and generates outputs
  - Harness/shared repo determines the quality of the outputs

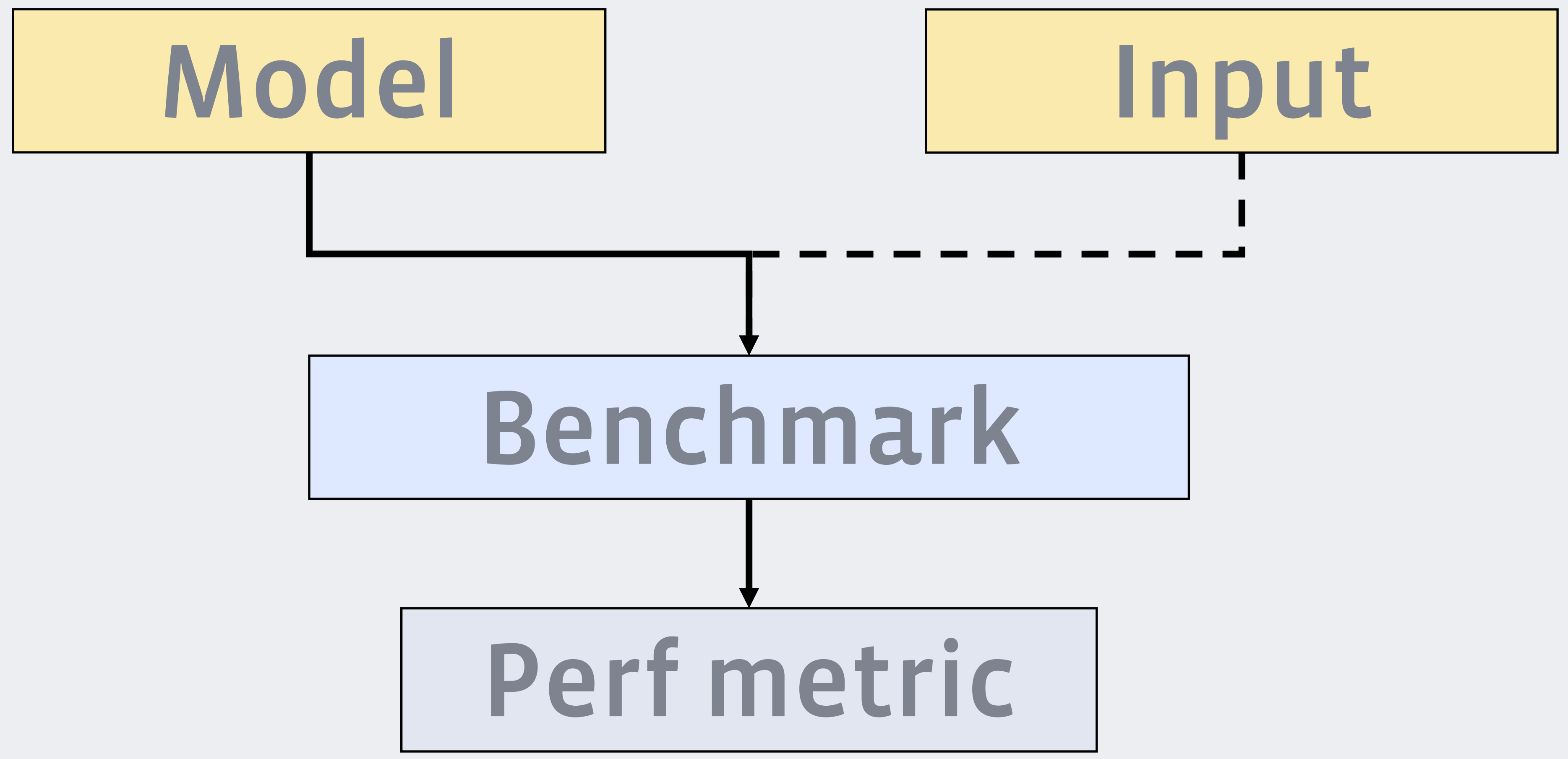
Run Command

- Setup sensible defaults in the first time invocation
- Follow-up run commands:
    - benchmarking/run_bench.py -b specifications/models/caffe2/shufflenet/shufflenet.json --platforms android

The Build

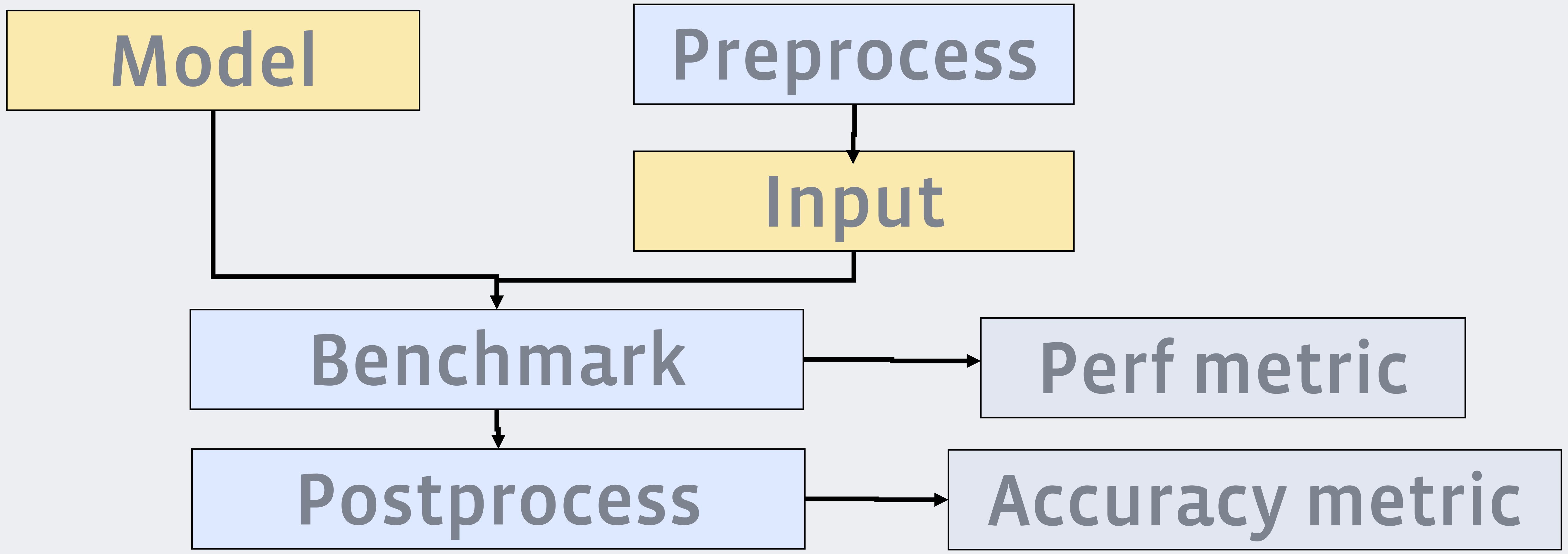
- Shell script build.sh
- Script is selected based on the `--platforms` argument
- Example:
  - https://github.com/facebook/FAI-PEP/blob/master/specifications/frameworks/caffe2/android/build.sh

The Run - Performance



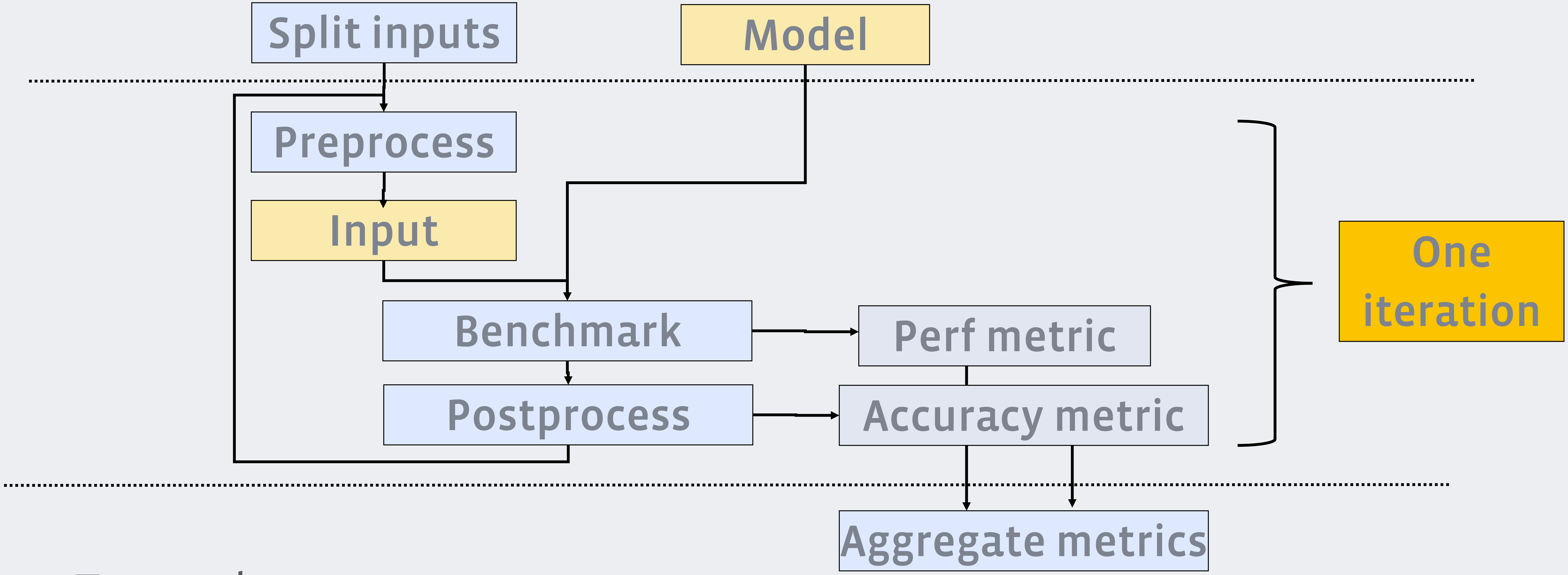
- Example:
  - https://github.com/facebook/FAI-PEP/blob/master/specifications/models/caffe2/squeezenet/squeezenet.json

The Run – Accuracy with Performance



- Example:
  - https://github.com/facebook/FAI-PEP/blob/master/specifications/models/caffe2/squeezenet/squeezenet_accuracy_input_file.json

The Run – Complicated Flow



- Example:
  - https://github.com/facebook/FAI-PEP/blob/master/specifications/models/caffe2/squeezenet/squeezenet_accuracy_imagenet.json
  - https://github.com/facebook/FAI-PEP/wiki/Run-Imagenet-validate-dataset

Deep Dive into Code