

Kazuaki Ishizaki, Tuan Hoang Trong, Umamaheswari Devi, Anubhav Jana, Ashok Pon Kumar Sree Prakash
IBM Research

1. What problem are we solving?

Op-level tests built from synthetic inputs pass while real models fail

- Which of PyTorch's 2000+ operators must a backend implement to run a given model?
- Which shapes, strides, dtypes, broadcast patterns, dtype promotions must each operator handle?
- How do we validate one operator before the full model runs end to end?

Visit here



2. What is our solution?

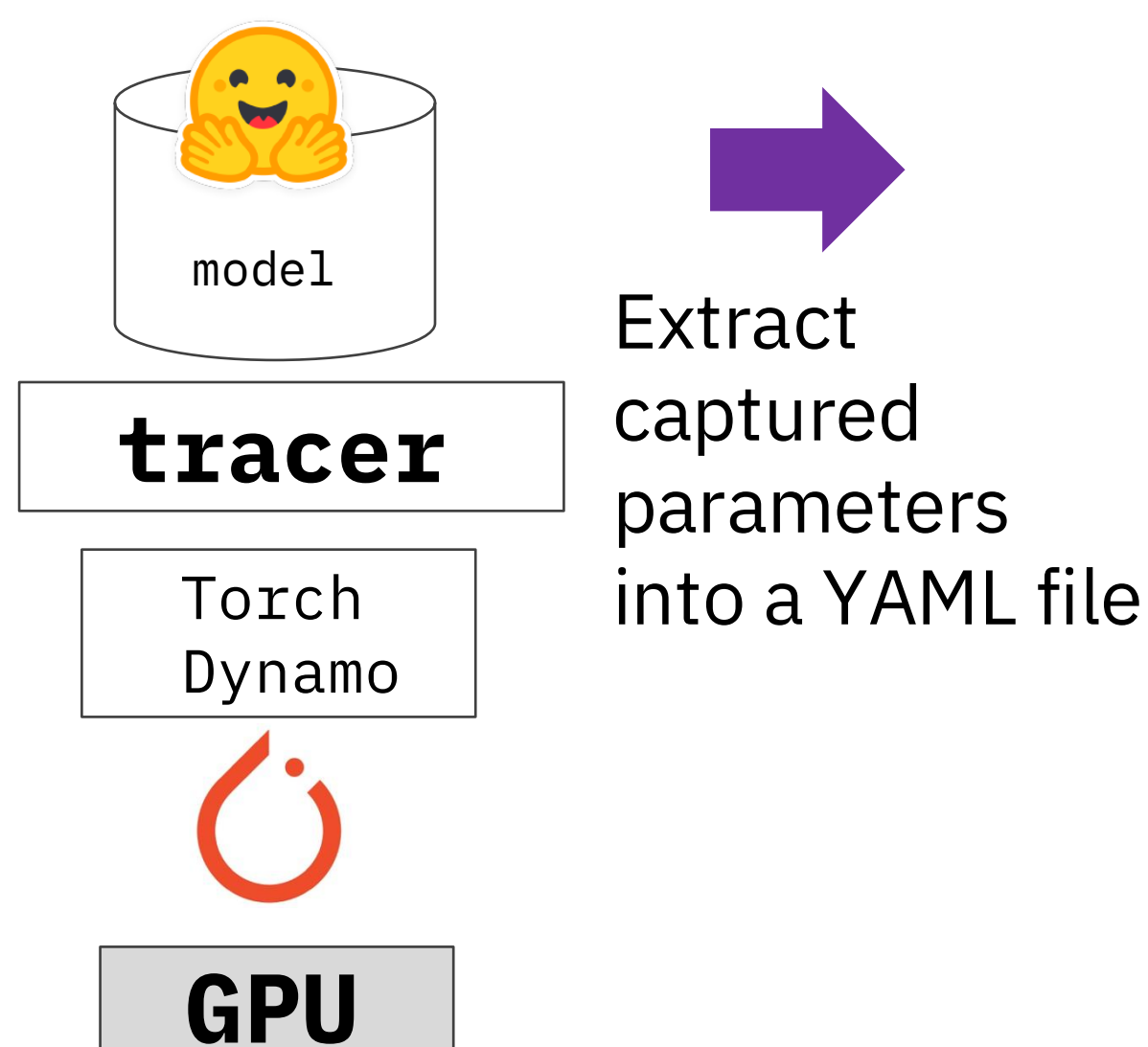
Generate and validate test cases from models

- Tracer: A TorchDynamo records every operator on GPU and emits per-op YAML with observed shapes, dtypes, strides, and offsets

The same YAML files can be reused across accelerators

- Runner: A pytest runner, which is built on PyTorch's test ecosystem, replays each recorded operator in isolation on the target accelerator, and validate its result against a CPU reference

No HF Transformers dependency



```
model.yaml
name: softmax.1
op: torch.softmax
inputs:
  - tensor:
      shape: [2, 128]
      stride: [128, 1]
      dtype: bfloat16
  kwdim: -1
name: add.1
op: torch.add
inputs:
  - tensor:
      shape: [64, 128]
      stride: [1, 64]
      dtype: bfloat16
  - tensor:
      shape: [128]
      stride: [1]
      dtype: float32
  marks: [float32]
```

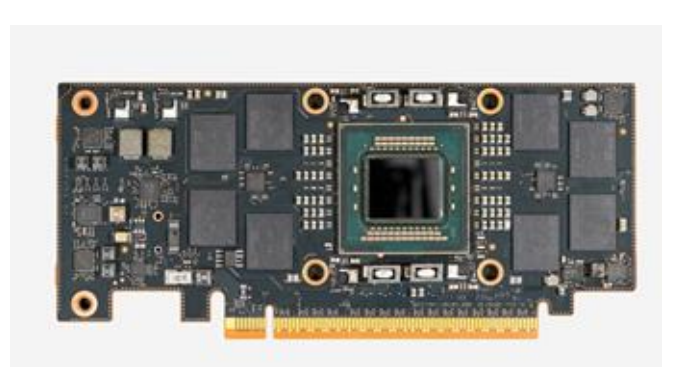
Conditions (e.g. float32) can be marked as skipped until the software stack supports them.

YAML decouples tracer and runner

Replay each operator on the device with captured parameters

Runner

```
import pytest
def read_yaml()
...
@ops(op, ...)
def test_...
...
op(shape, ...)
assert ...
```



IBM Spyre

3. Results

- Collected 2,261 test cases covering 79 operators across 10 decoder models
- Some model patterns suited our accelerator poorly, so an AI-agent-generated adapters hf-adapters that patch the model at load time.
- Improved bug detection and operator validation during the bring-up of IBM Granite 3.3 (a Llama-family architecture) on IBM Spyre using the torch-spyre stack

This presentation is based on results obtained from the project, "Research and Development Project of the Enhanced Infrastructures for Post-5G Information and Communication Systems" (JPNP20017), subsidized by the New Energy and Industrial Technology Development Organization (NEDO).