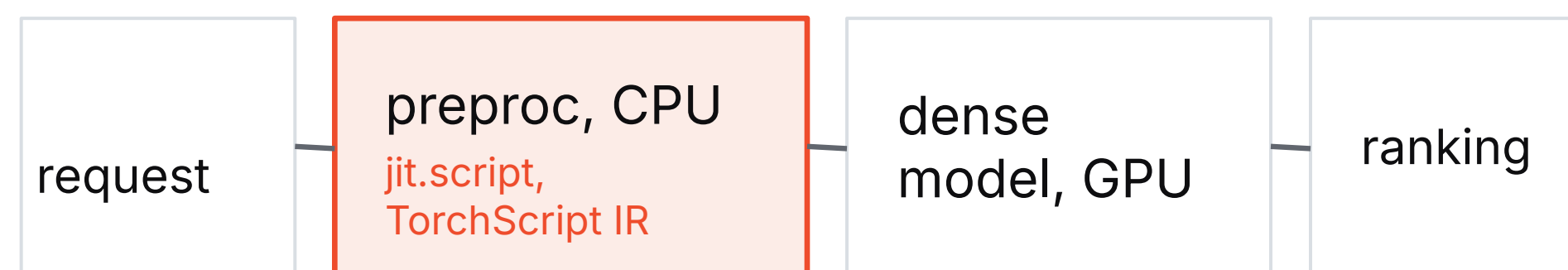


Upgrading CPU Preprocessing from TorchScript to PT2

IR design, developer experience, and production gains on recommendation serving

Georgia Phillips · AI Inference Acceleration, Meta

- 1 CPU preproc runs ahead of every ranking inference, on a deprecated stack

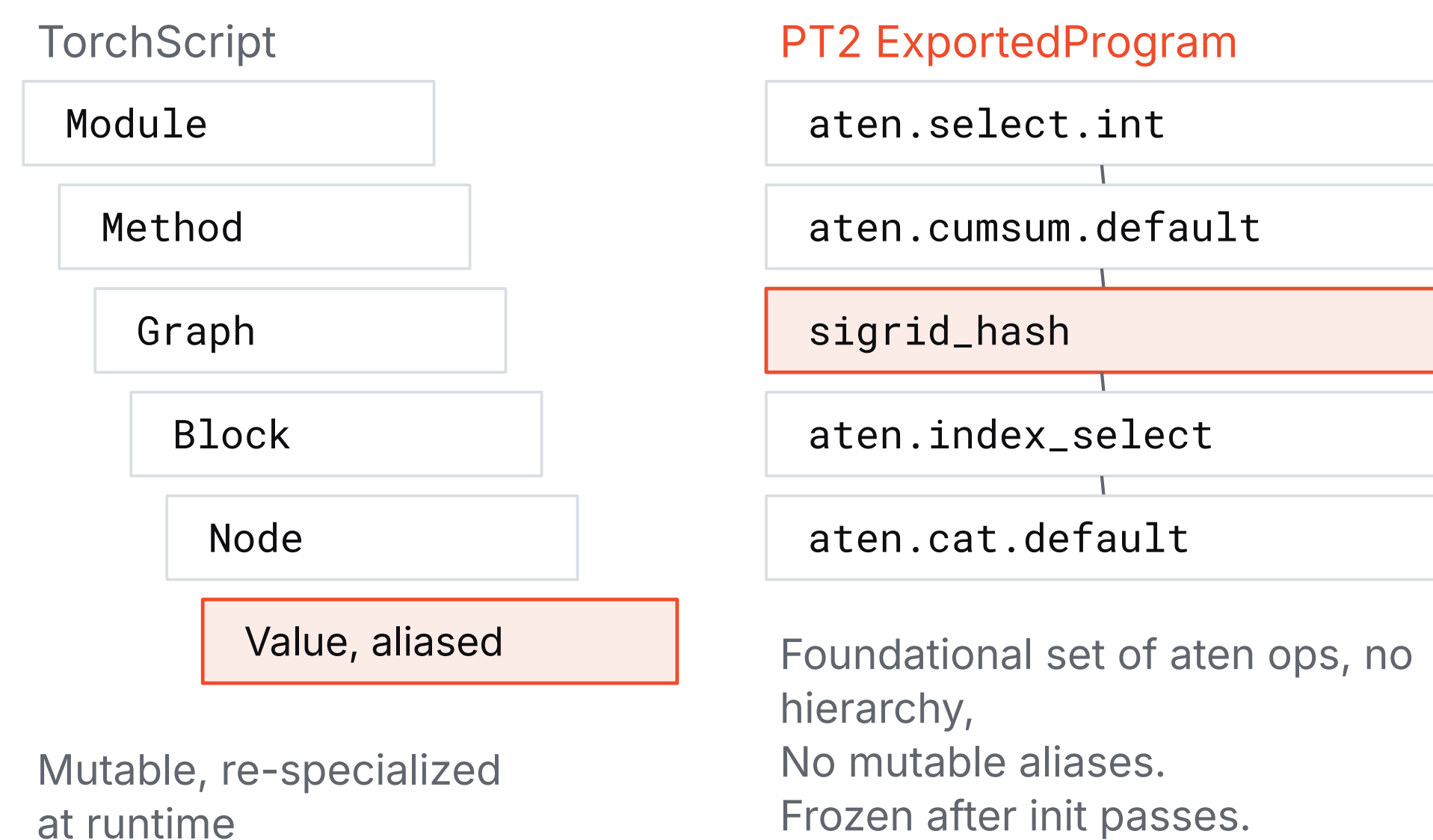


2.6 MW
Preproc power consumption

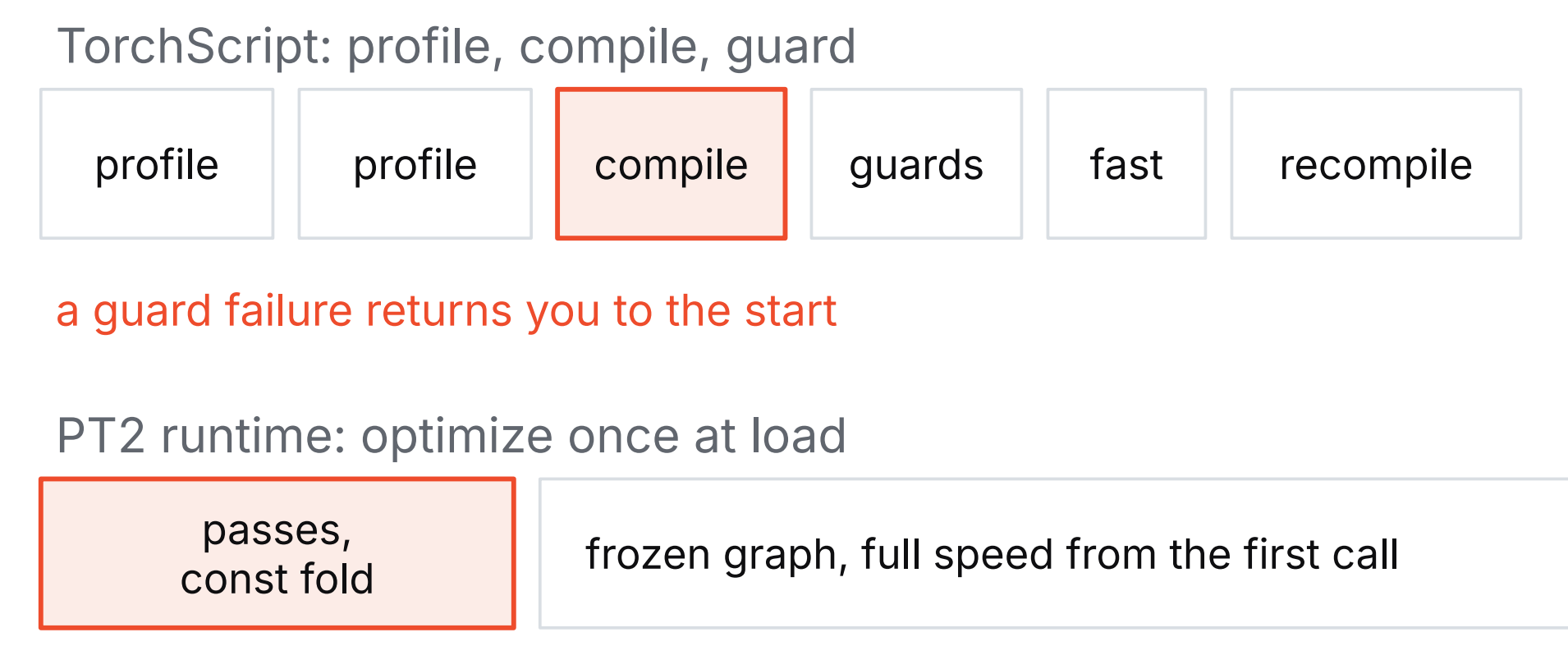
SEV1
capacity incidents, host overload

14.9 s
JIT warm-up per replica

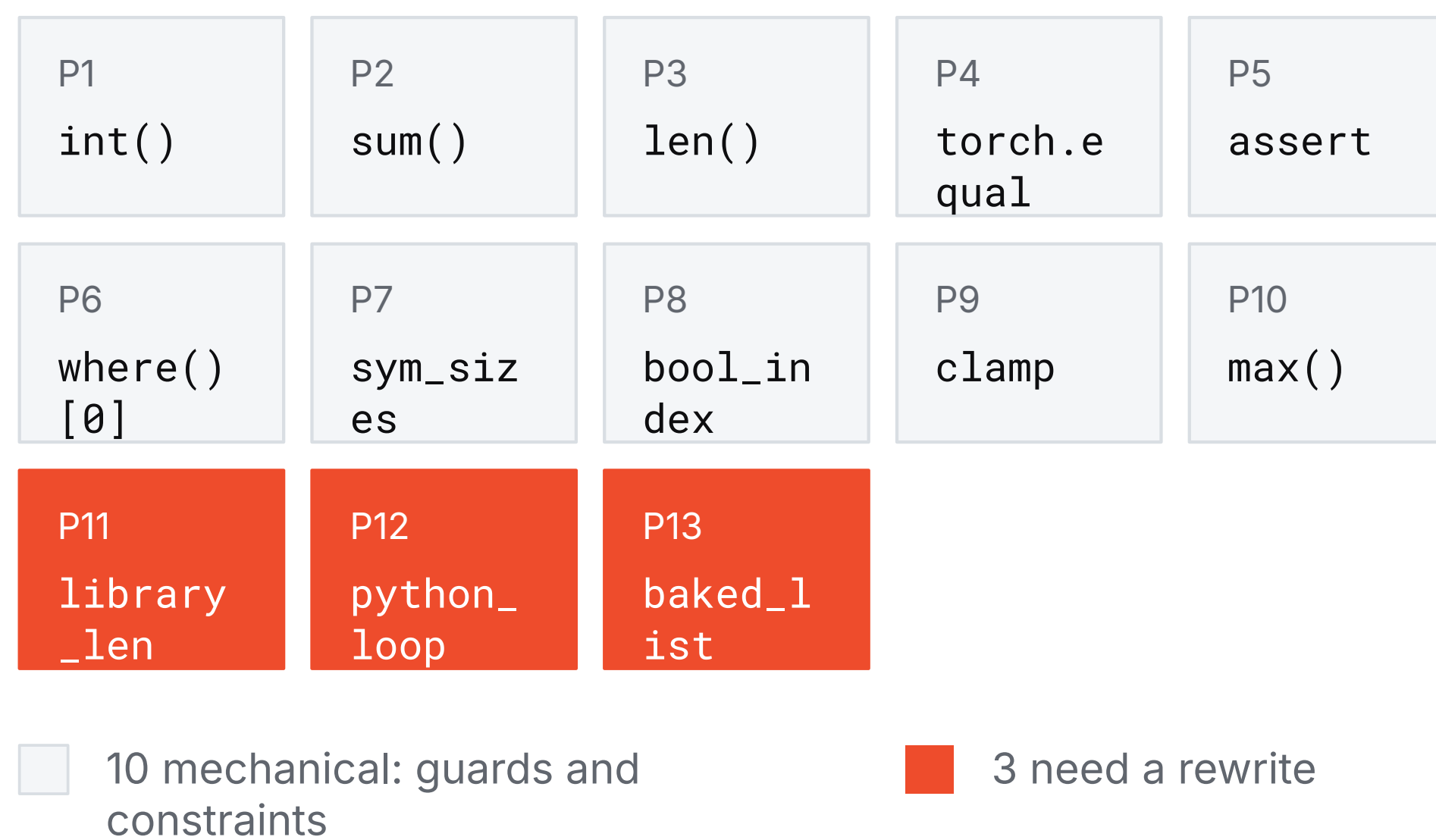
- 2 Hierarchical and aliased, against flat and functional



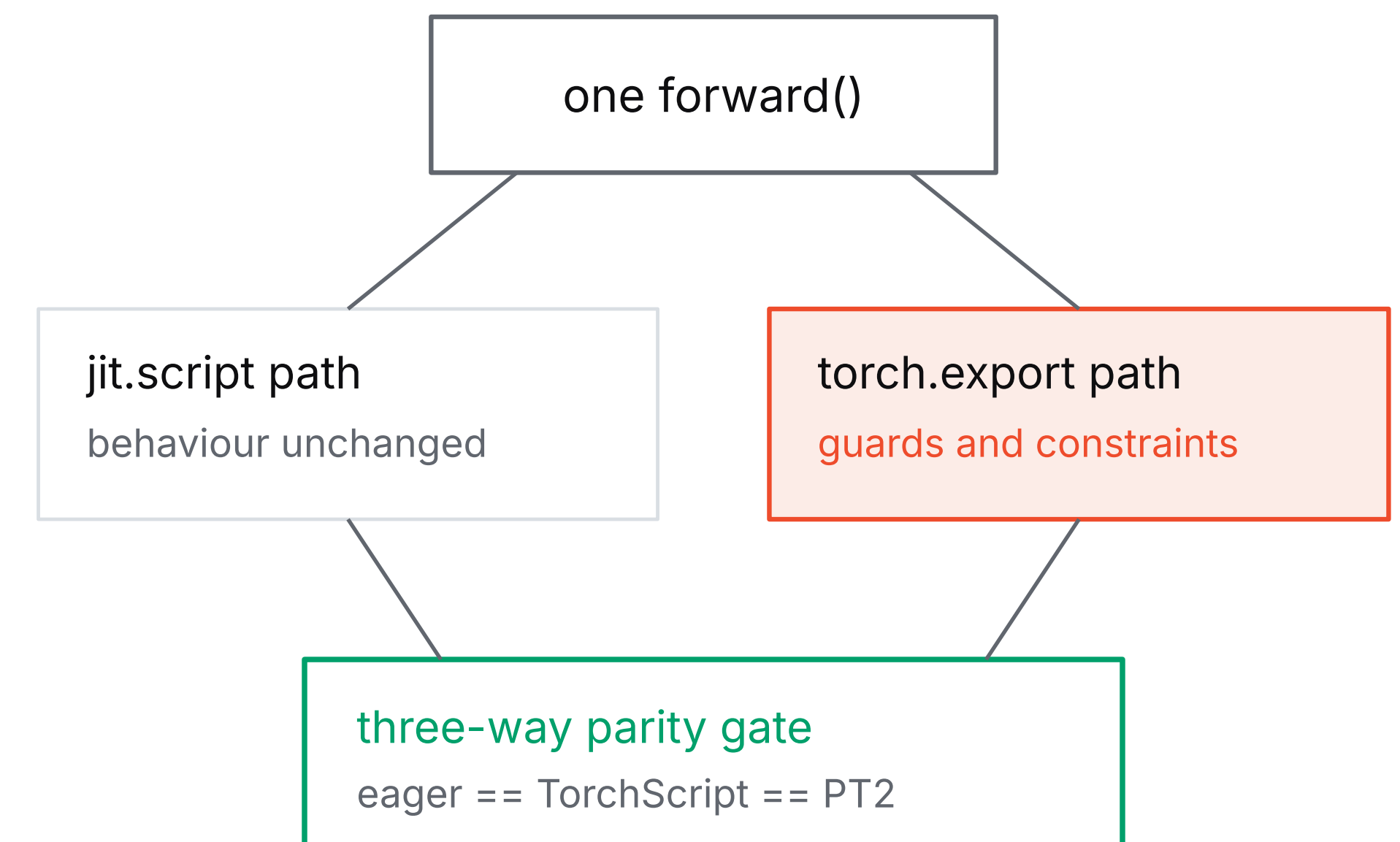
- 3 A flat, functional graph moves optimization from run time to load time



- 4 Data-dependent expressions in sparse preproc are a finite, nameable set



- 5 One source file serves both frontends, so migration lands incrementally



- 6 Consistent latency and memory wins, with no accuracy change

