Dynamic Control Flow in Large-Scale Machine Learning, Yu et al. (2018)[1] Review

Presented by Gabriel Mahler

Presentation Overview

- Motivation
- TensorFlow: a Data-Flow System
- Implementation Overview
- Evaluation
- Criticisms
- References

Motivation

Motivation

- Training and running of recurrence relations models (RNNs, Reinforcement Learning)

  - Static vs. Dynamic unrolling

Motivation

- Recurrence relations models (RNNs, Reinforcement Learning)

  - Static vs. Dynamic unrolling

- Training on distributed computation units

  - Parallelism and asynchrony

Motivation

- Recurrence relations models (RNNs, Reinforcement Learning)

  - Static vs. Dynamic unrolling

- Training on distributed computation units

  - Parallelism and asynchrony

- Dynamic control flow

  - Ability to define models as general data flow constructs

Motivation

- Recurrence relations models (RNNs, Reinforcement Learning)

  - Static vs. Dynamic unrolling

- Training on distributed computation units

  - Parallelism and asynchrony

- Dynamic control flow

  - Ability to define models as general data flow constructs

- No existing dynamic control flow system supporting automatic differentiation

TensorFlow

- Data-flow system, not exclusively for machine-learning

TensorFlow

- Data-flow system, not exclusively for machine-learning
- Computations represented as directed dataflow graphs [2]

TensorFlow

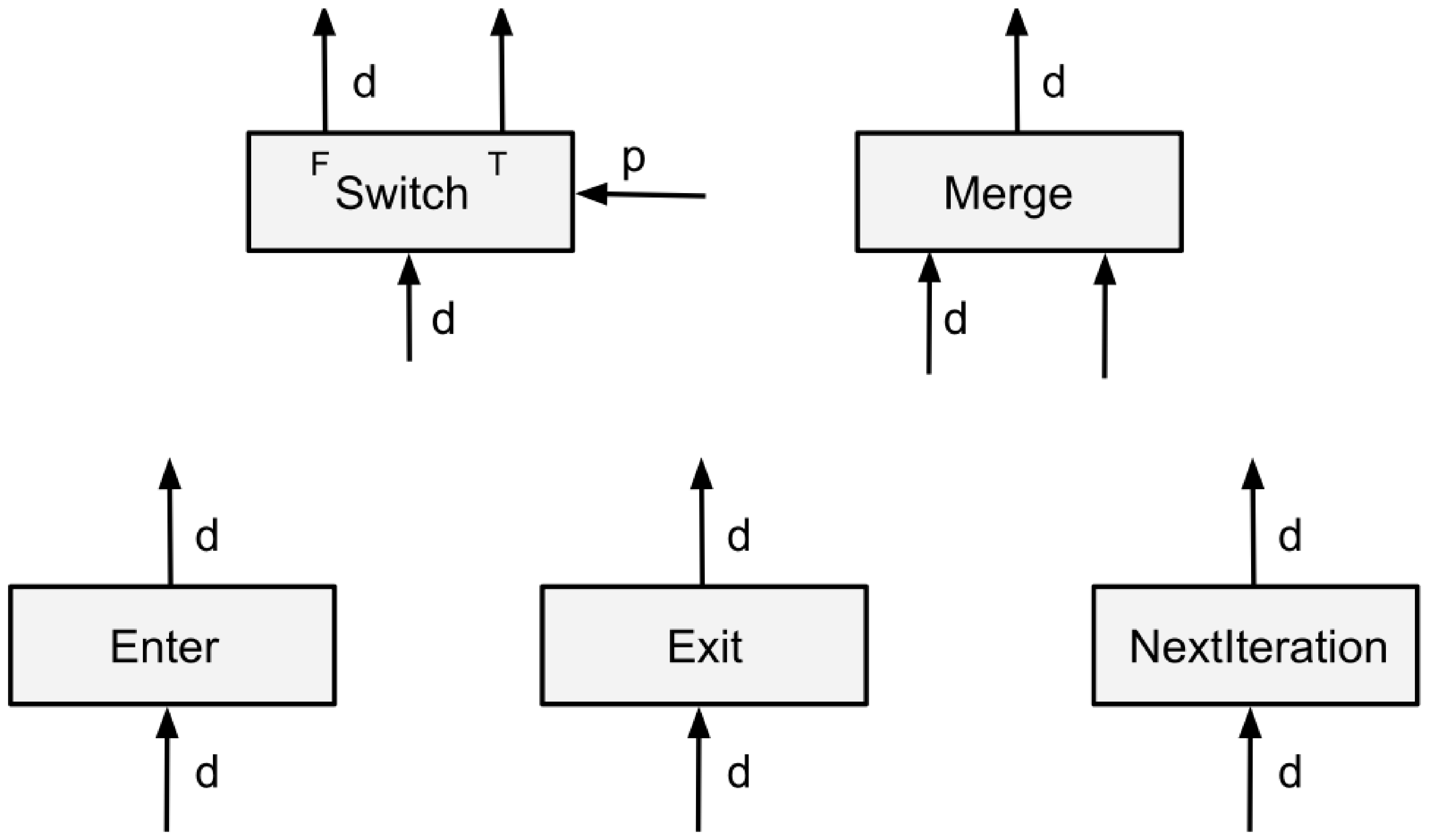
- Data-flow system, not exclusively for machine-learning
- Computations represented as directed dataflow graphs [2]
- Built to support a wide variety of hardware (first major system to support computation mapping to multiple devices) [2]

Implementation Overview

Implementation Overview

Control Flow Operations

- Dynamic flow with basic graph primitives



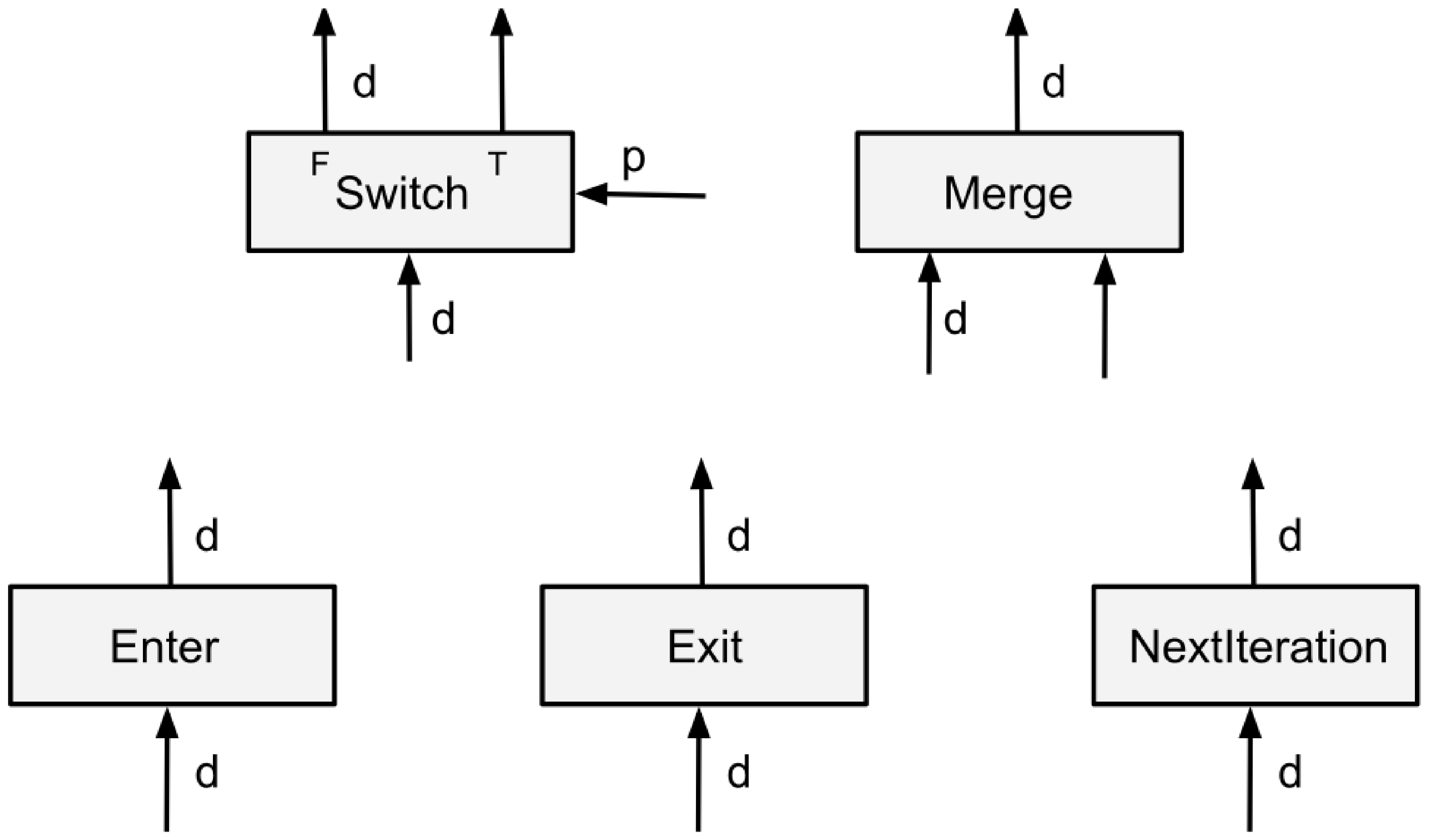
"The Control Flow Primitives" [1]

Implementation Overview

Control Flow Operations

- Dynamic flow with basic TensorFlow primitives

  - while-loops, conditionals



"The Control Flow Primitives" [1]

Implementation Overview

Graph Partitioning

- Ability to split a graph into subgraphs

Implementation Overview

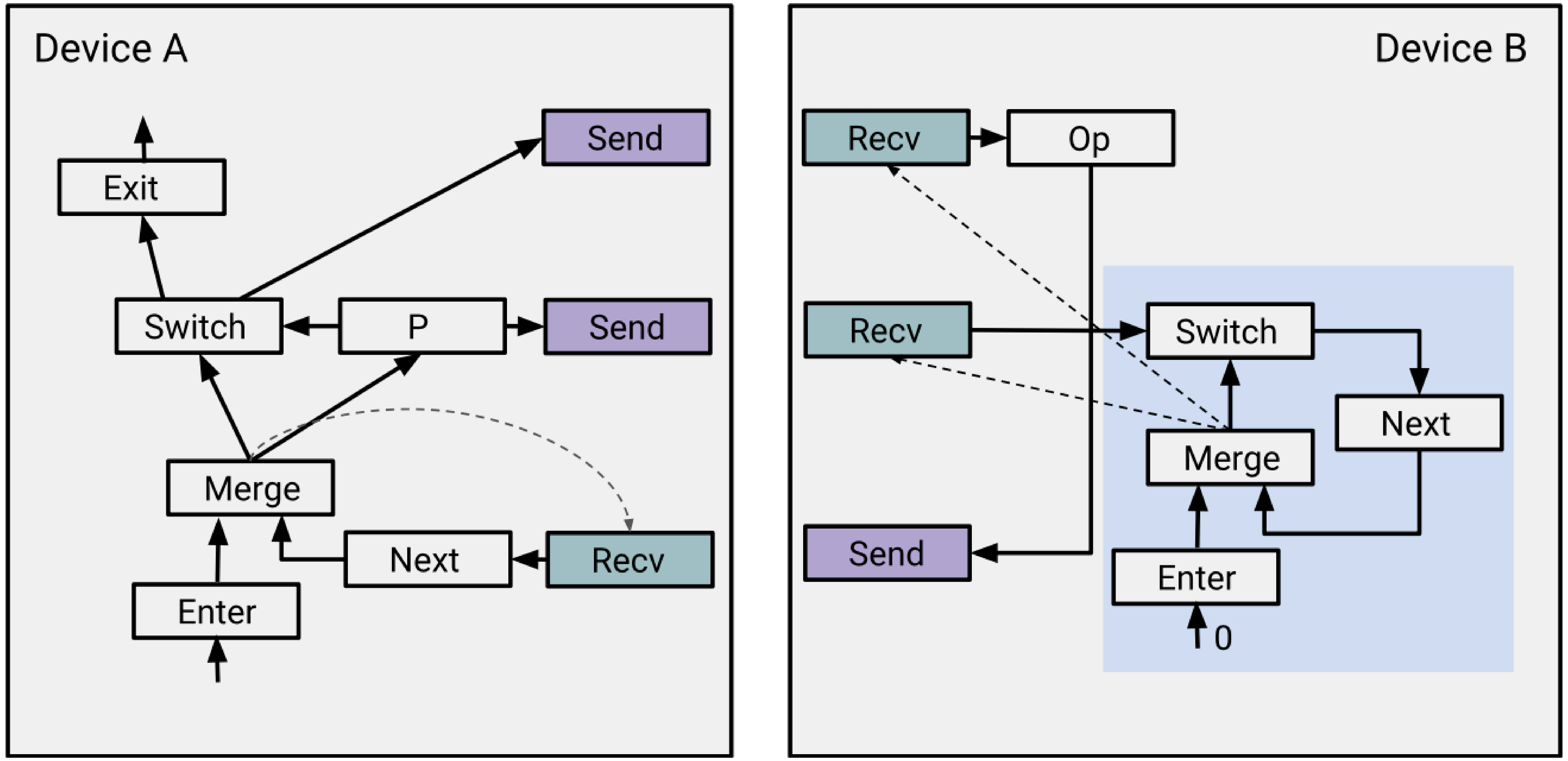
Graph Partitioning

- Ability to split a graph into subgraphs
- TensorFlow: running subgraphs on various devices

Implementation Overview

Direct cross-device communication

- Send and Recv operations



"Distributed execution of a while-loop" [1]

Implementation Overview

Memory swapping

- Temporary use of abundant memory (GPU ↔ CPU)
- Dependent primarily on the parallel execution

Implementation Overview

Automatic Differentiation

- TensorFlow mechanisms
- Saving intermediate values (while-loops)
- Memory management (memory swapping)

Evaluation

Evaluation

- Good memory performance (memory swapping, distributed systems)

Evaluation

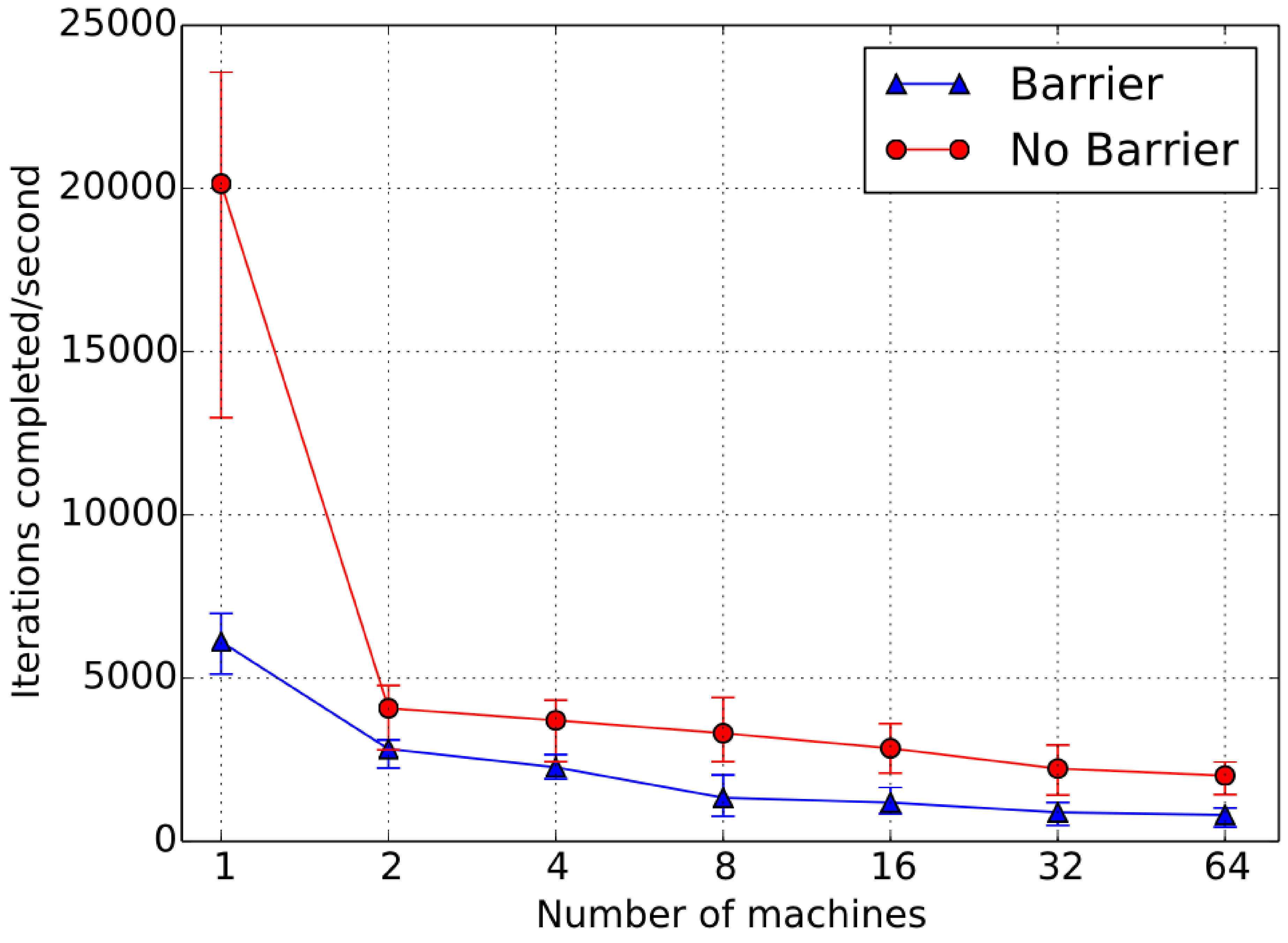
	Training time per loop iteration (ms), by sequence length						
Swap	100	200	500	600	700	900	1000
Disabled	5.81	5.78	5.75	OOM	OOM	OOM	OOM
Enabled	5.76	5.76	5.73	5.72	5.77	5.74	5.74

"Training time per loop iteration for an LSTM model with increasing sequence lengths." [1]

Evaluation

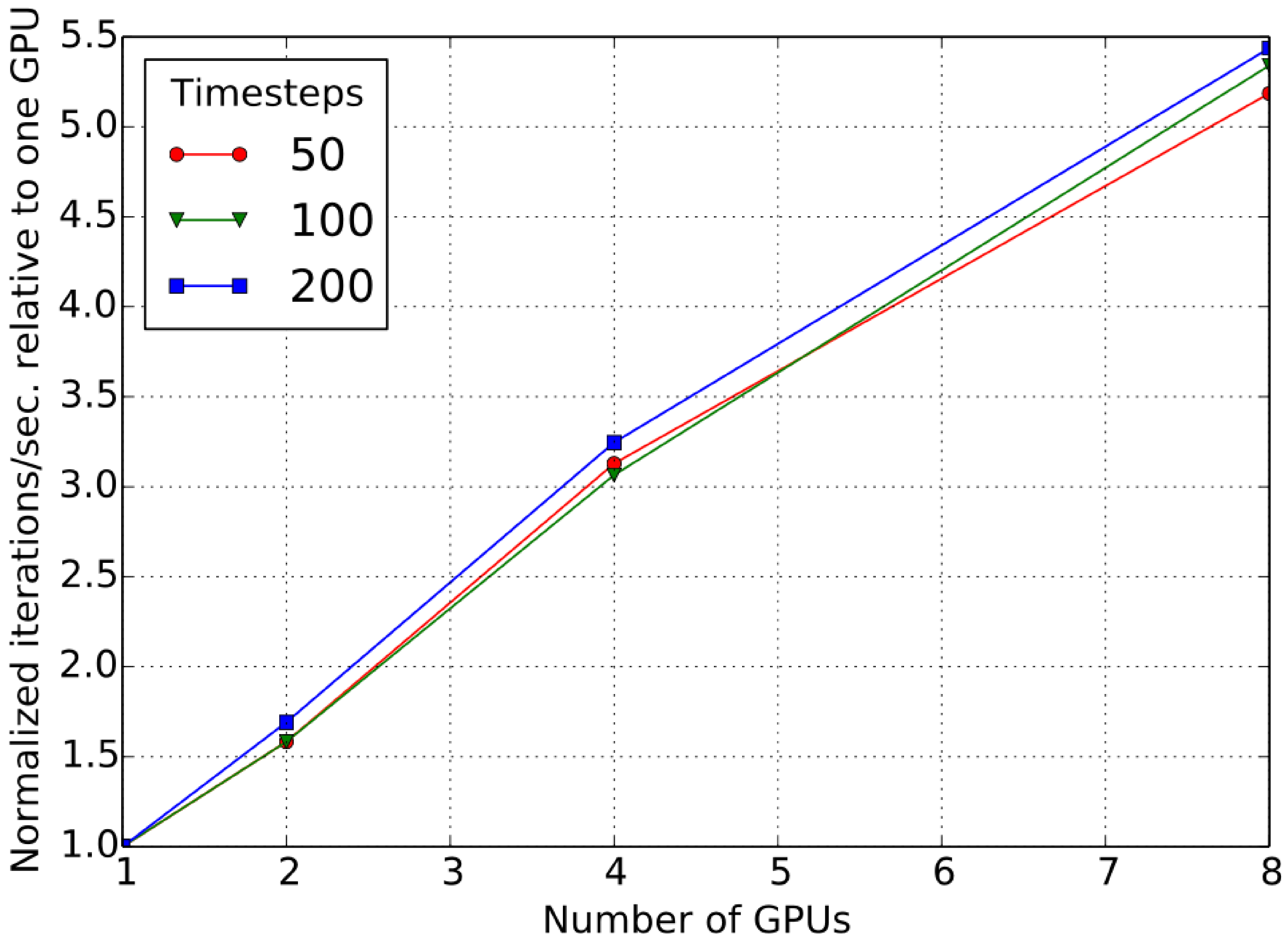
- Good memory performance (memory swapping, distributed systems)
- Mixed speed performance (overheads)

Evaluation



"Performance of a distributed while-loop with a trivial body on a GPU cluster" [1]

Evaluation



"Parallel speedup for an 8-layer LSTM as we vary the number of GPUs from 1 to 8."
(training) [1]

Evaluation

Analysis of 11.7 million machine-learning graphs:

- 65% contain conditional computation
- 5% contain one or more loop

Criticisms

Criticisms

- Limited discussion of alternative approaches

Criticisms

- Limited discussion of alternative approaches
- Limited testing approaches (only performance)

Criticisms

- Limited discussion of alternative approaches
- Limited testing approaches (only performance)
- Tested on limited and fairly homogenous distributed systems
- Limited discussion of implementation (graph semantics)

References

[1] - Yu, Yuan, et al. "Dynamic control flow in large-scale machine learning." Proceedings of the Thirteenth EuroSys Conference. 2018.

[2] - Abadi, Martín, et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint arXiv:1603.04467.* 2016.