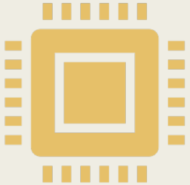
PLACETO: LEARNING GENERALIZABLE DEVICE PLACEMENT ALGORITHMS FOR DISTRIBUTED MACHINE LEARNING

Ravichandra Addanki, Shaileshh Bojja Venkatakrishnan, Shreyan Gupta, Hongzi Mao, Mohammad Alizadeh

Presented by: Obodoekwe Nnaemeka

Problem



Distributed training (GPU and CPU)

Human experts?

Reinforcement learning?

Problem

Sometimes tolerable.

Solutions do not generalize

The optimization is done for a single graph.

Single computational graph vs Class of computational graph

Placeto

Efficiency

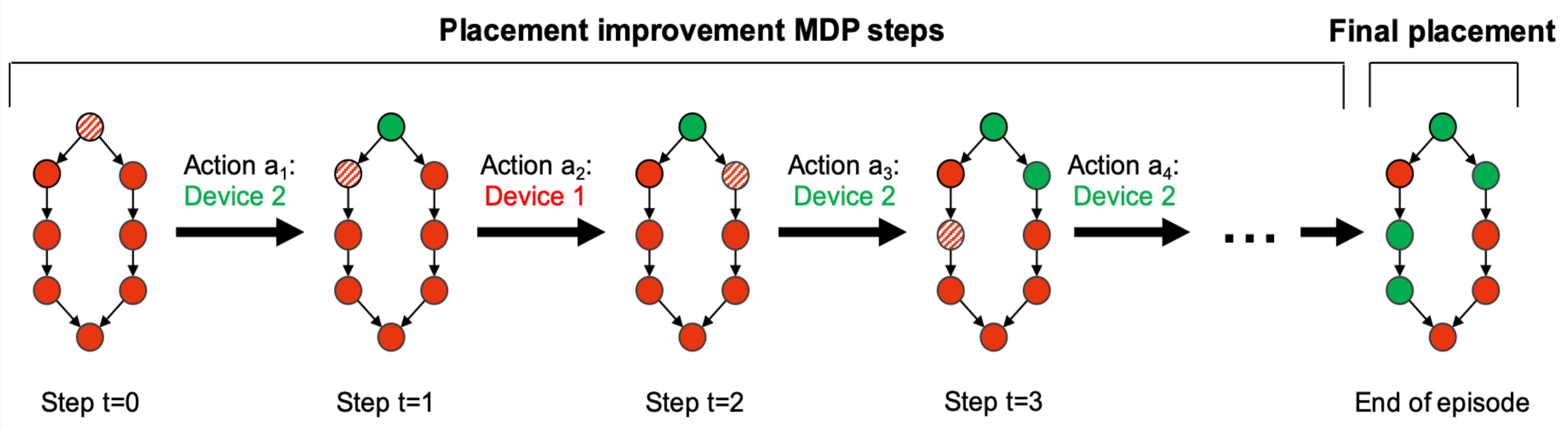
Sequence of iterative placement improvements

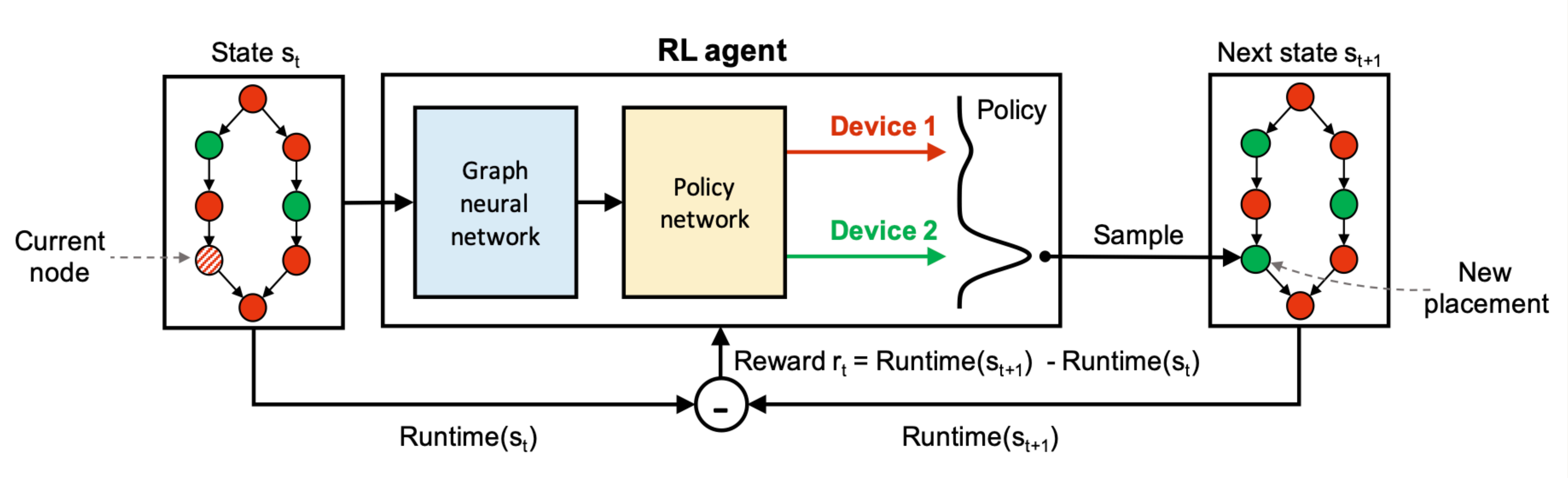
Generalizability

NN architecture that uses graph embedding to encode the computation of graph structure in the placement policy.

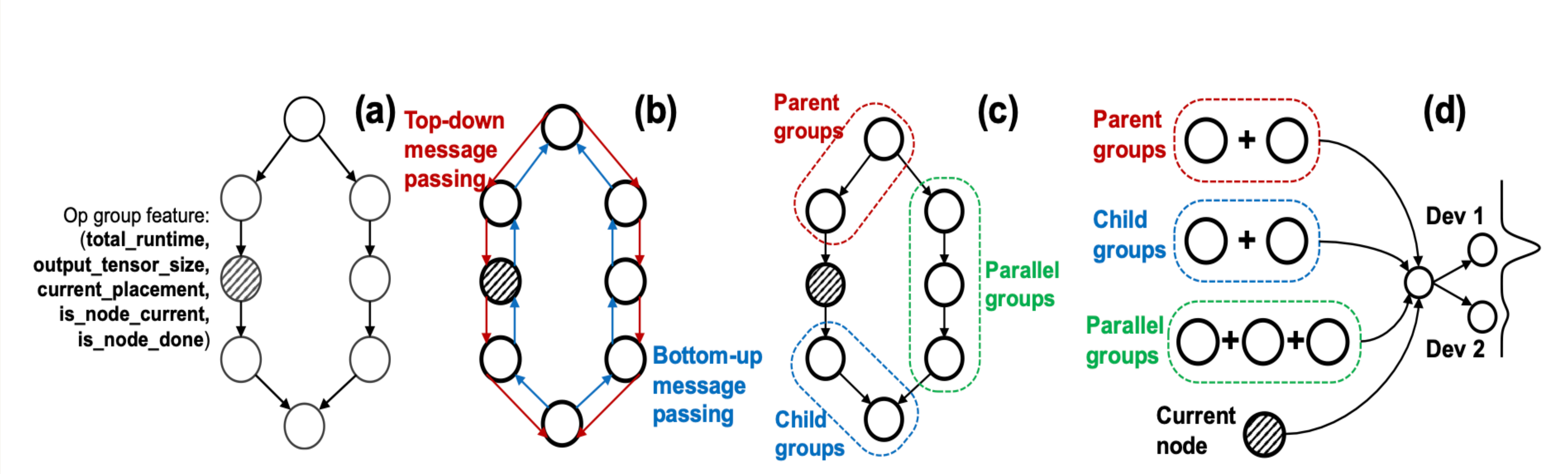
Learning method

- Markov Decision Process



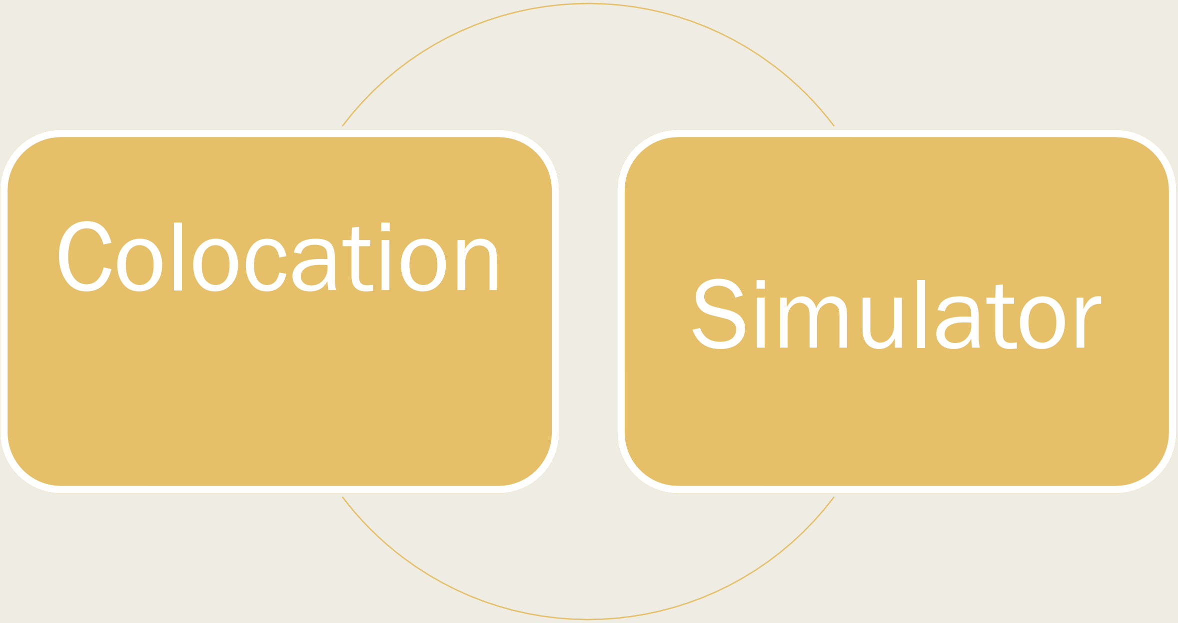


POLICY NETWORK ARCHITECTURE



GRAPH EMBEDDING

Training Details



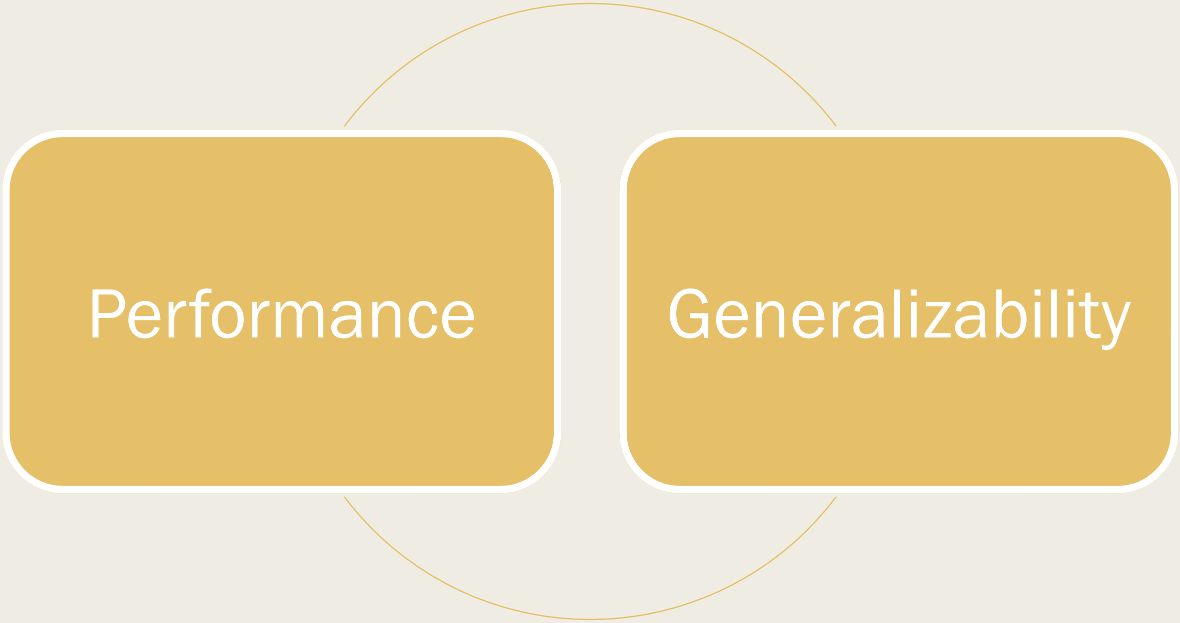
Experimentation

Deep learning models (**Inception-V3, NASNet, NMT**)

Synthetic data (cifar10, ptb, nmt)

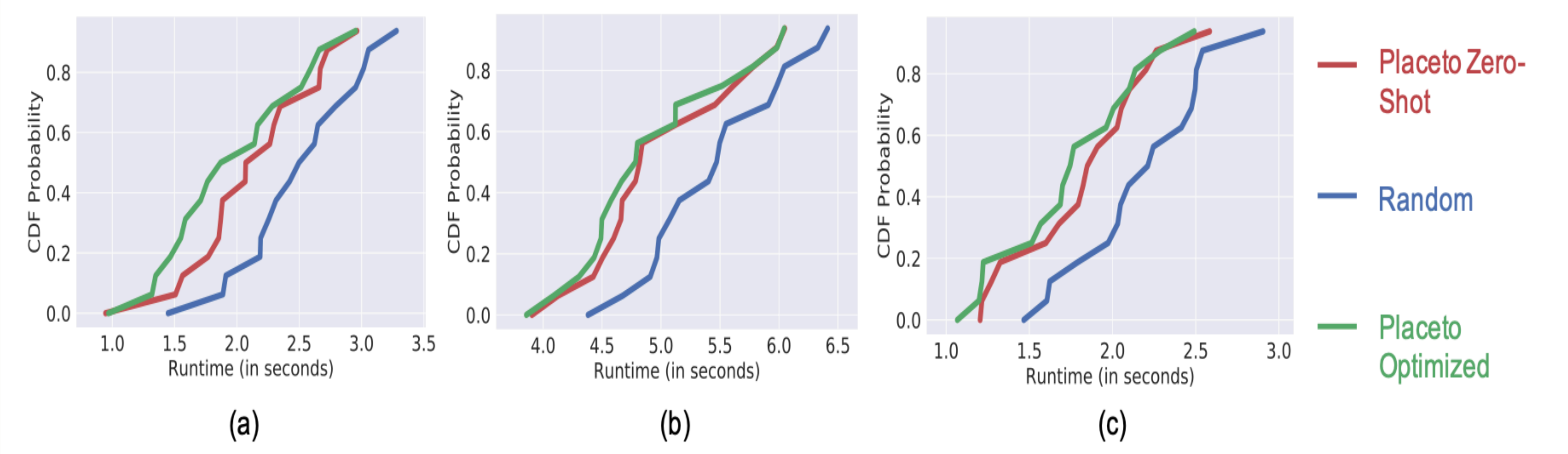
Single GPU, Scotch, Human Expert, RNN based approach.

Result

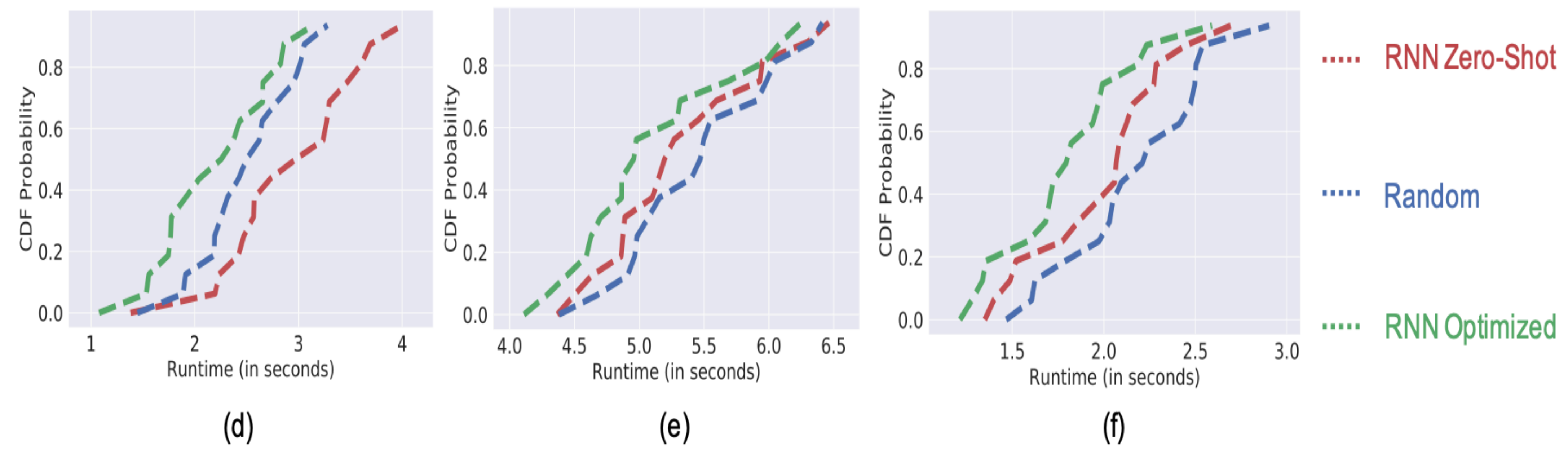


Model	Placement runtime (sec)							Training time (# placements sampled)		Improvement	
	CPU only	Single GPU	#GPUs	Expert	Scotch	Placeto	RNN-based	Placeto	RNN-based	Runtime Reduction	Speedup factor
Inception-V3	12.54	1.56	2	1.28	1.54	1.18	1.17	1.6 K	7.8 K	- 0.85%	**4.8** ×
			4	1.15	1.74	1.13	1.19	5.8 K	35.8 K	5%	**6.1** ×
NMT	33.5	OOM	2	OOM	OOM	2.32	2.35	20.4 K	73 K	1.3 %	**3.5** ×
			4	OOM	OOM	2.63	3.15	94 K	51.7 K	**16.5 %**	0.55 ×
NASNet	37.5	1.28	2	0.86	1.28	0.86	0.89	3.5 K	16.3 K	3.4%	**4.7** ×
			4	0.84	1.22	0.74	0.76	29 K	37 K	2.6%	**1.3** ×

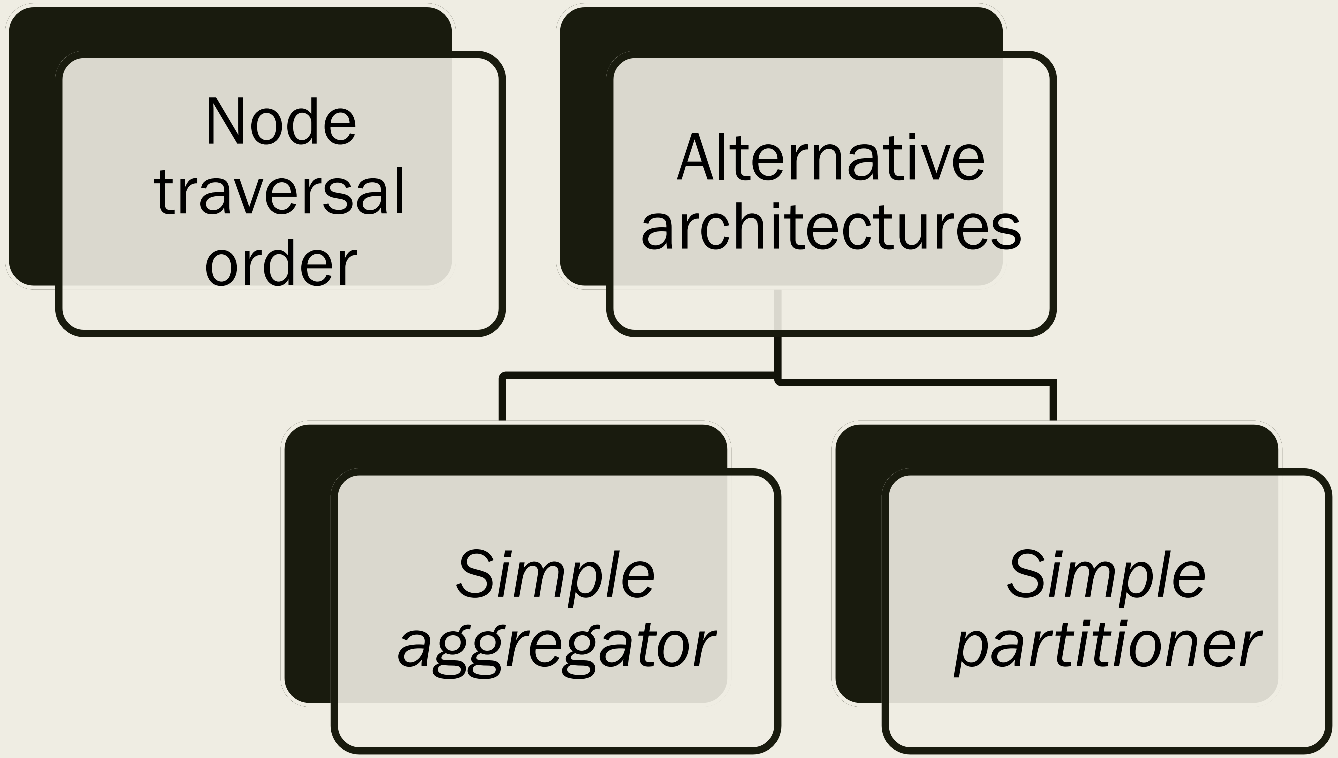
PLACETO VS RNN



GENERALIZABILITY



GENERALIZABILITY



Place deep dive

Critic

+ First attempt to generalize device placement using a graph embedding network

+ Really Impressive performance

- Only optimizes placement decisions

- It shows generalization to unseen graphs, but they are generated artificially by architecture search for a single learning task and dataset.

How does the framework handle failure. Evaluation protocol needs to be more explicit.