Berkeley
UNIVERSITY OF CALIFORNIA

riselab
UC Berkeley

Checkmate

Checkmate: Breaking the Memory Wall with Optimal Tensor Rematerialization

Paras Jain

Joint work with: Ajay Jain, Ani Nrusimha, Amir Gholami, Pieter Abbeel, Kurt Keutzer, Ion Stoica, Joseph Gonzalez

Artifacts Available
Artifacts Evaluated Functional
Results Replicated

BigGAN (2018)

Image generation

Brock et al. 2019

VideoBERT (2019)

Video generation

Sun et al. 2019

GPT-2 (2019)

Text generation

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

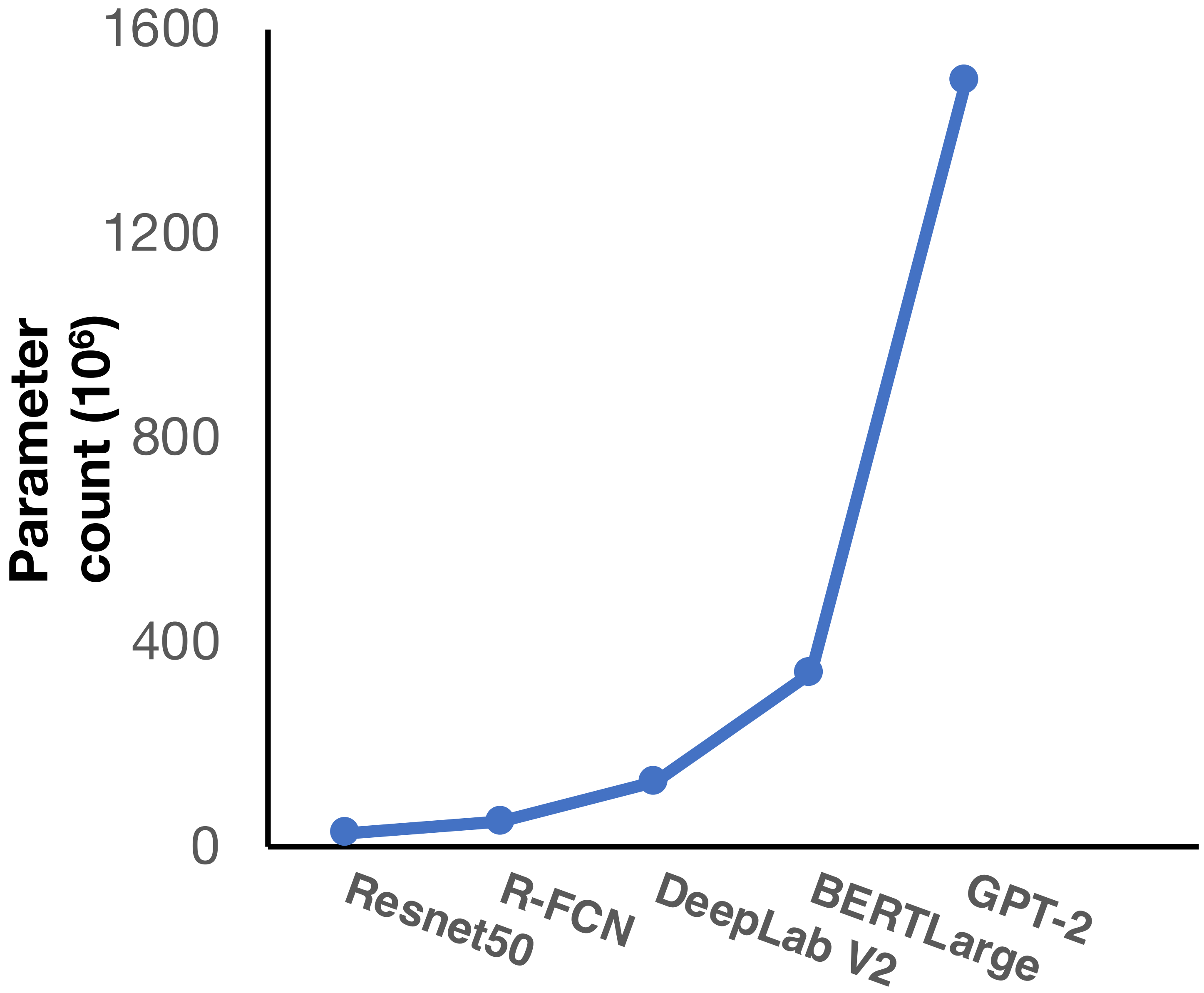
MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several

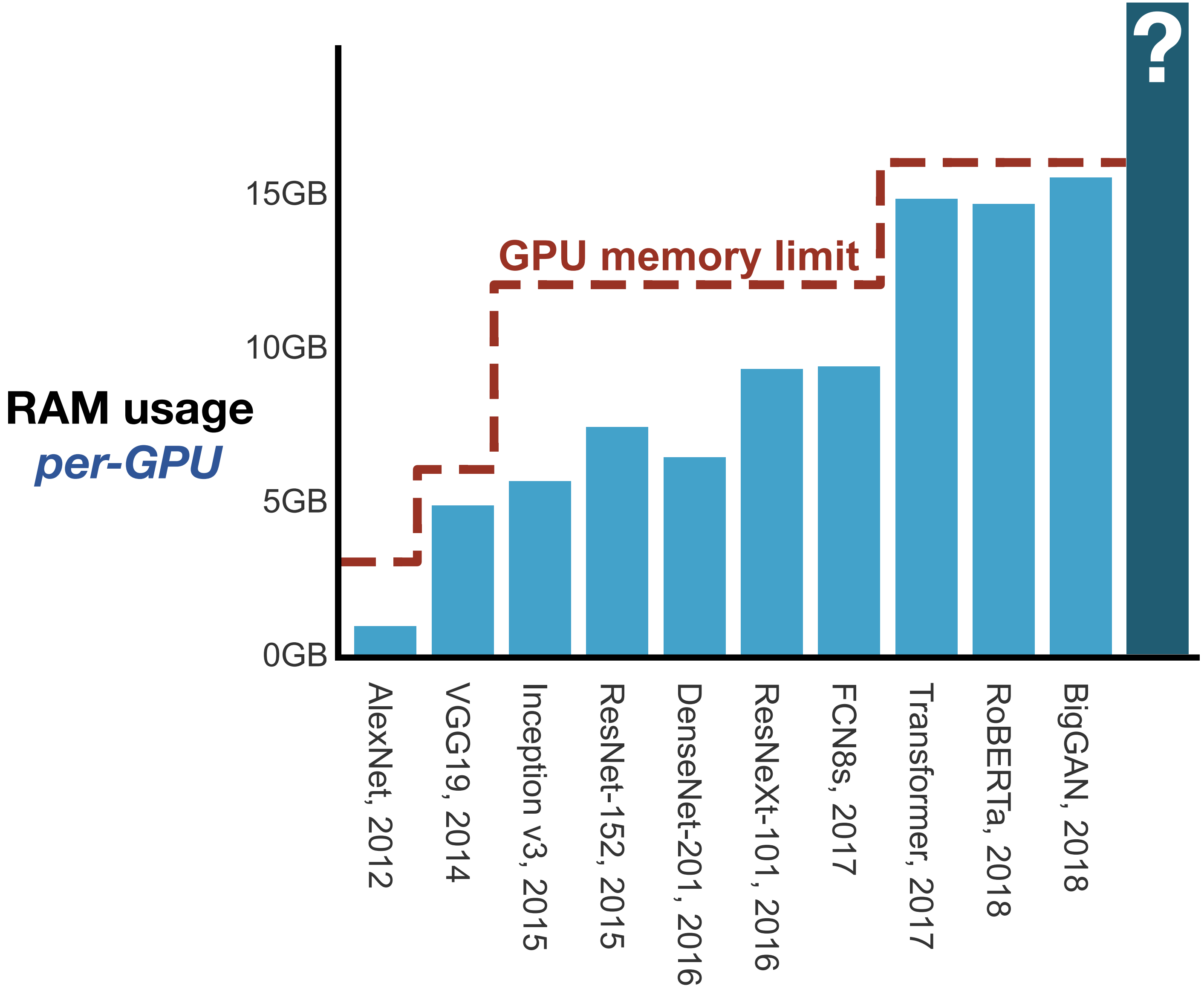
Radford et al. 2019



Emerging trend:

Rapid **growth in model size**

Figure adapted from NVIDIA



State-of-the-art models have hit a **memory capacity wall**.

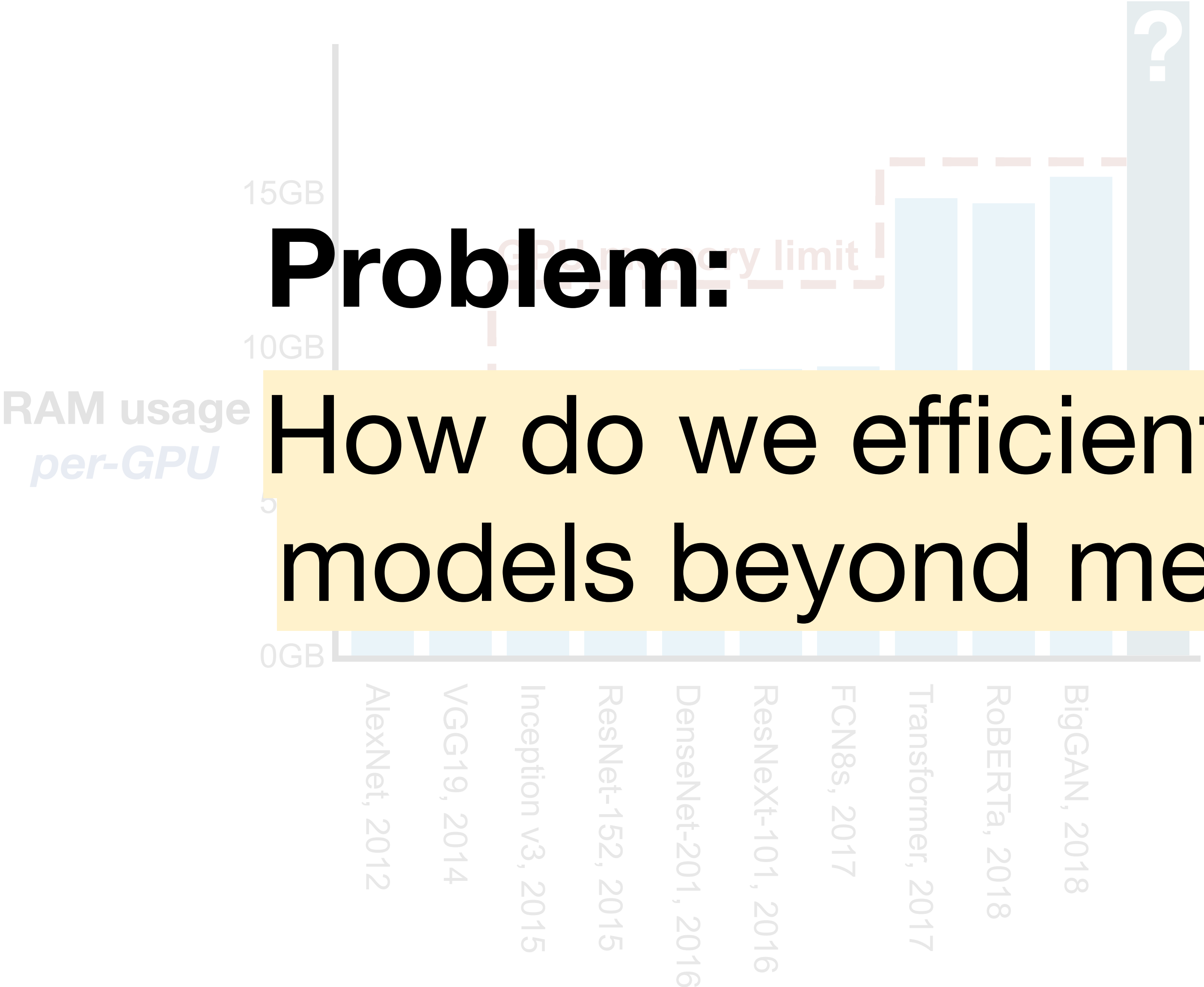
Cited memory as limiting factor

Chen et al. 2016	Liu et al. 2019
Gomez et al. 2017	Dai et al. 2019
Pohlen et al. 2017	Child et al. 2019

Limited GPU memory is slowing progress in new deep learning models!

Problem:

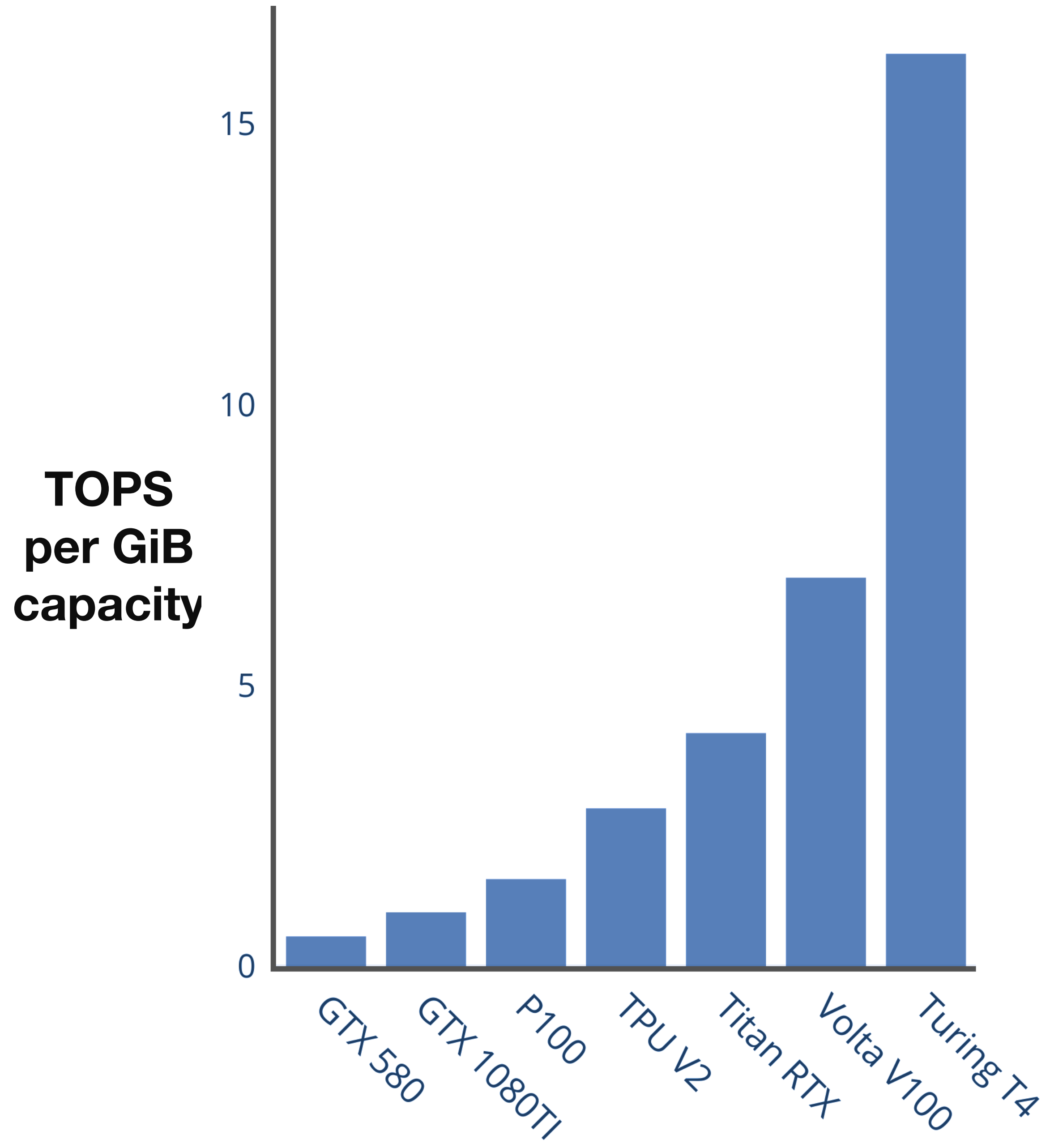
How do we efficiently train large models beyond memory limits?



State-of-the-art models have hit a **memory capacity wall**.

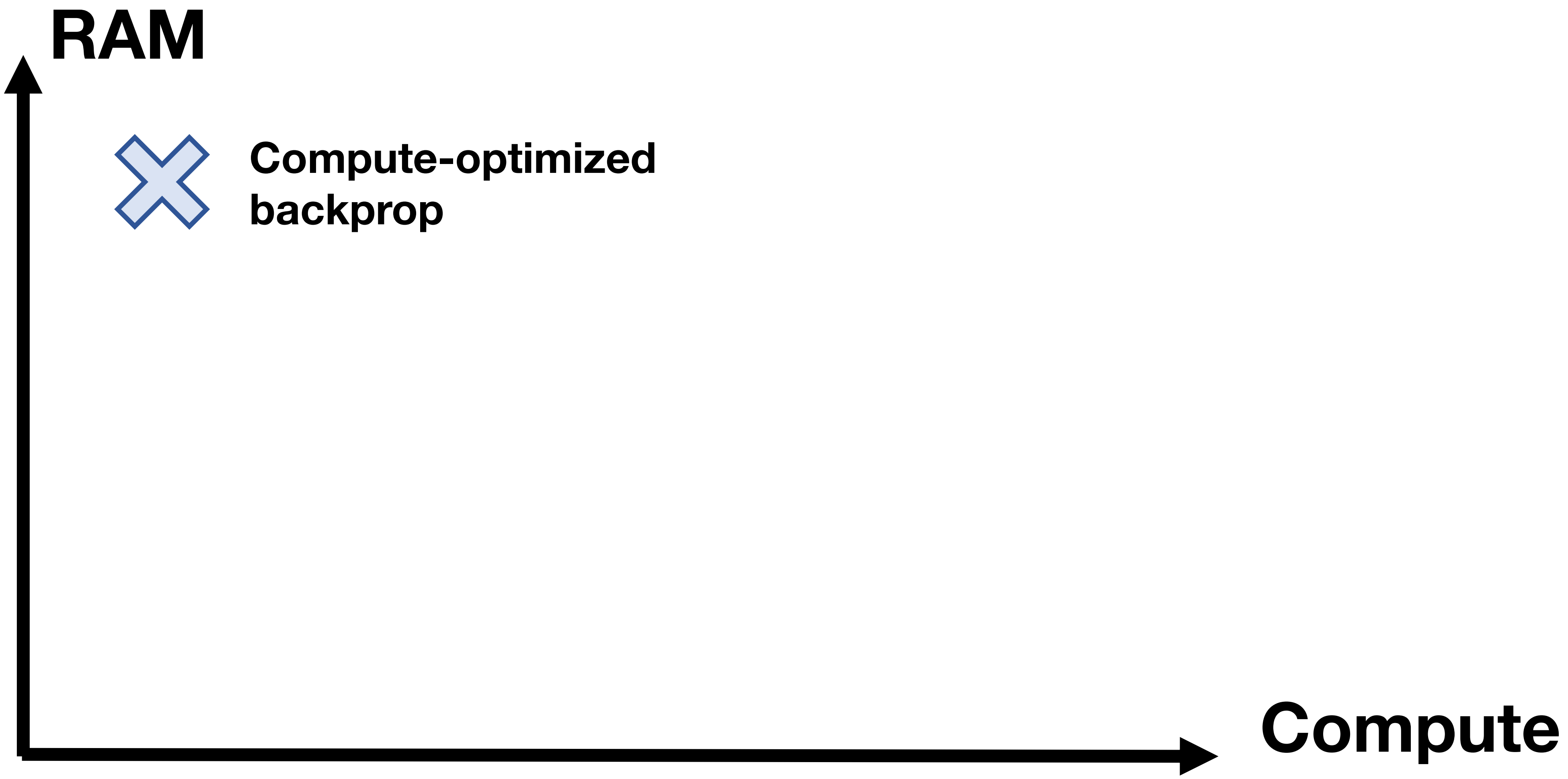
Cited memory as limiting factor

Limited GPU memory is slowing progress in new deep learning models!

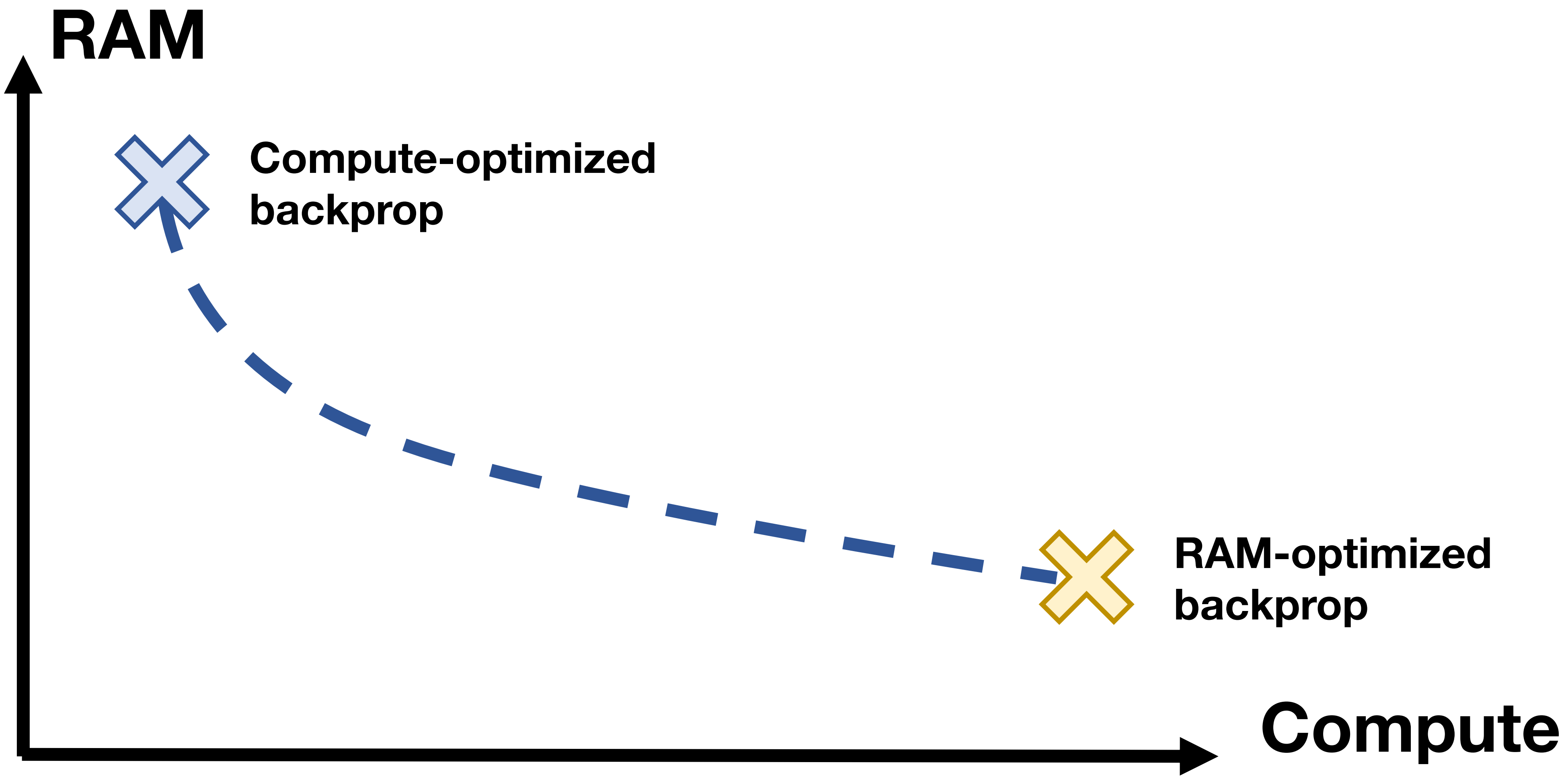


Compute is outstripping DRAM capacity growth

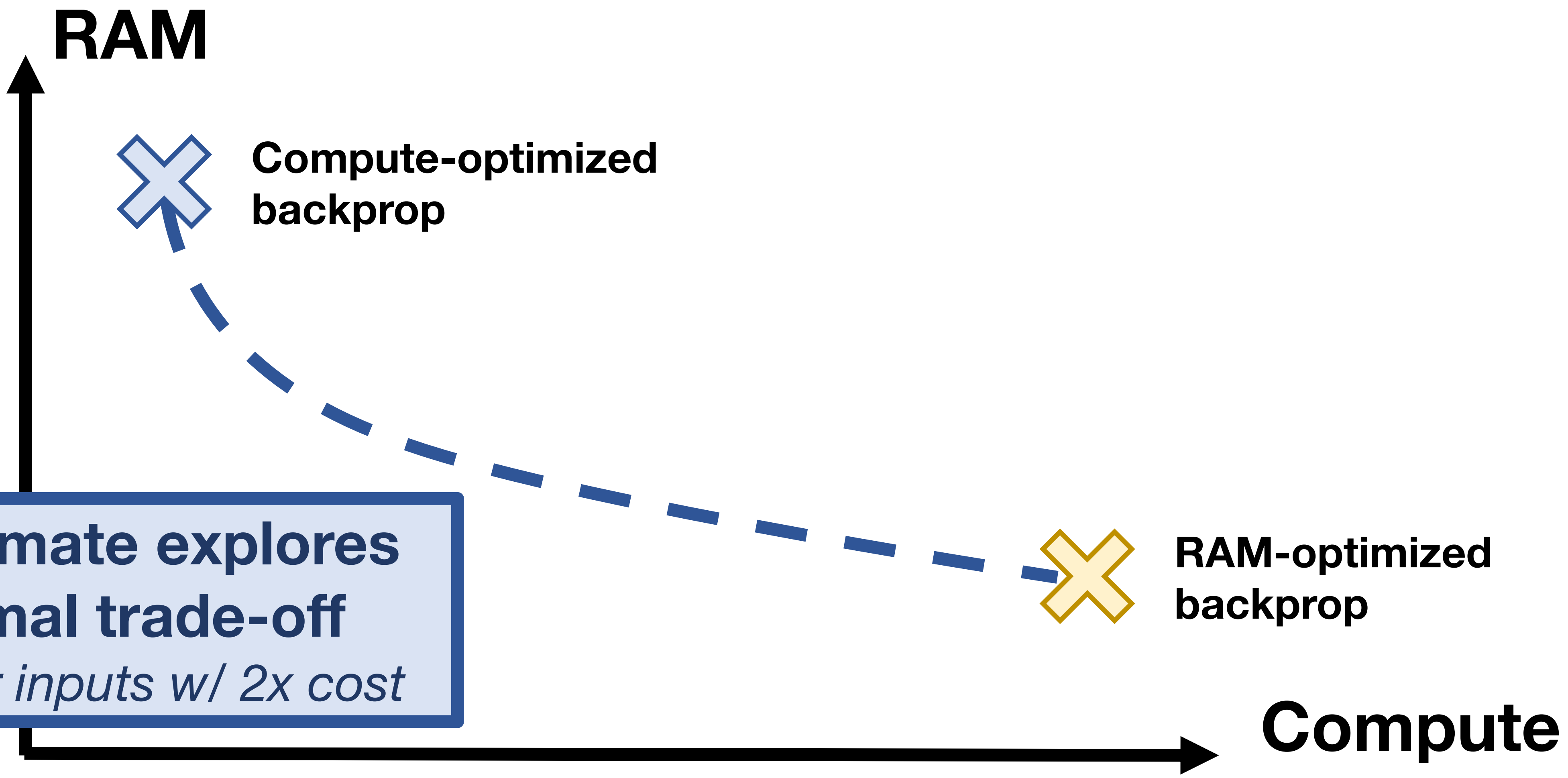
Backprop is optimized for compute efficiency, not RAM usage



Ideal: scalable algorithm for backprop that adapts to RAM constraints



This work: optimal space-time tradeoff for backpropagation

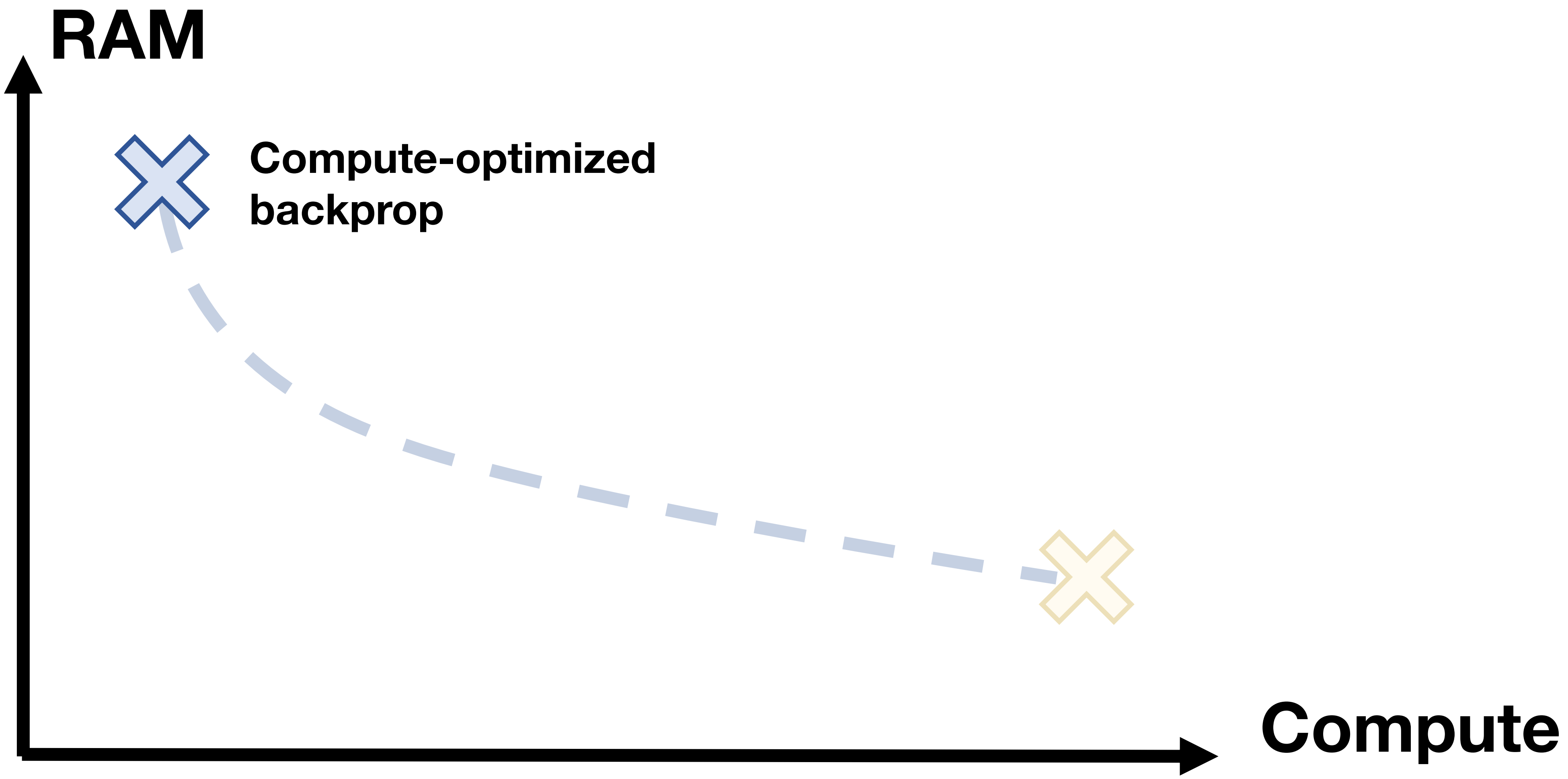


Checkmate explores optimal trade-off

5x larger inputs w/ 2x cost

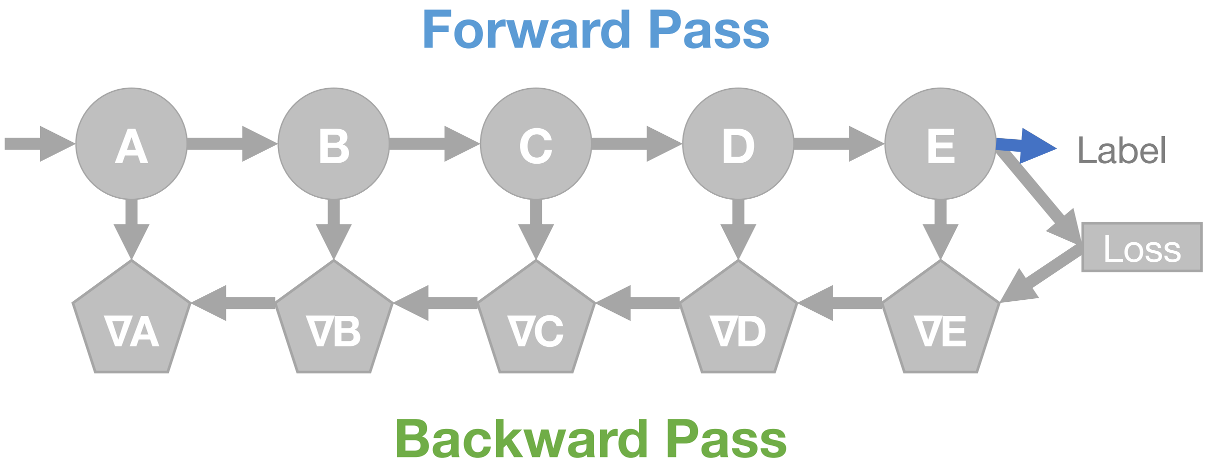
RAM-hungry backprop policy

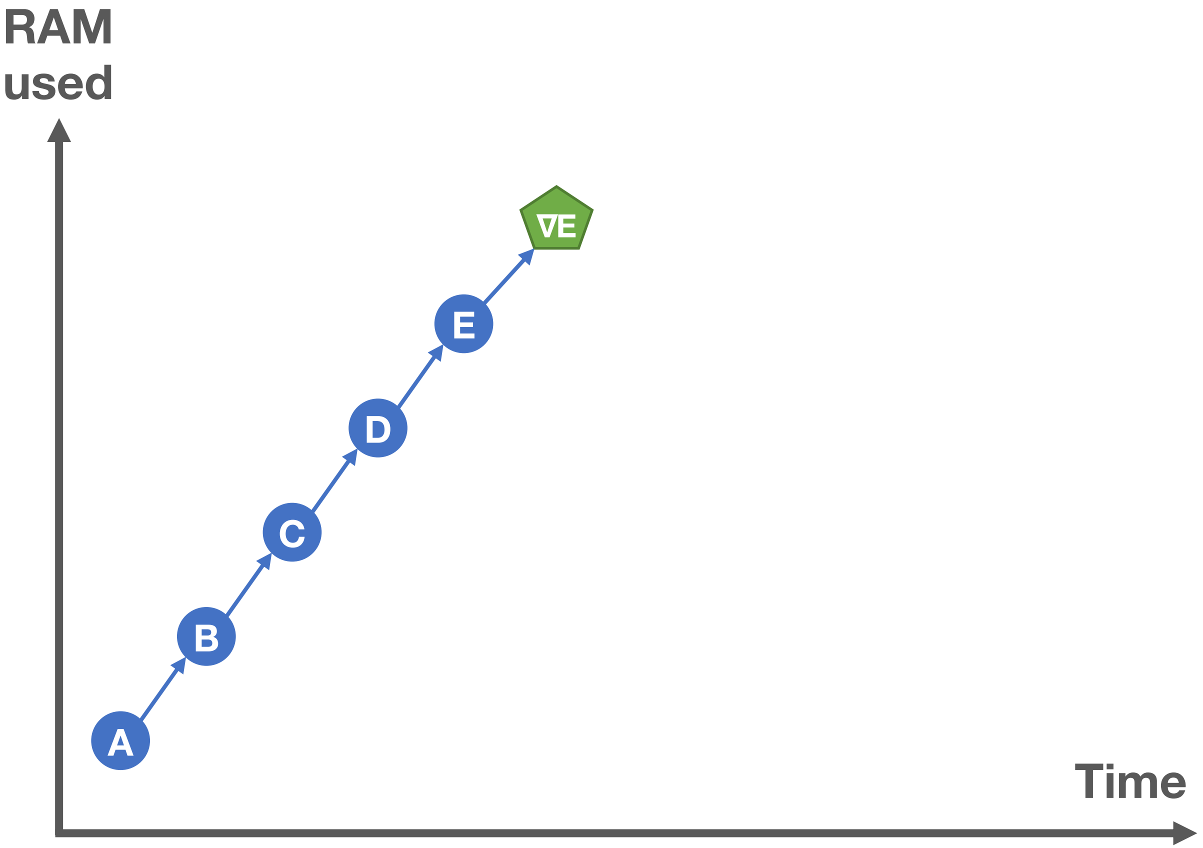
Keep all layers in RAM



RAM-hungry backpropagation policy

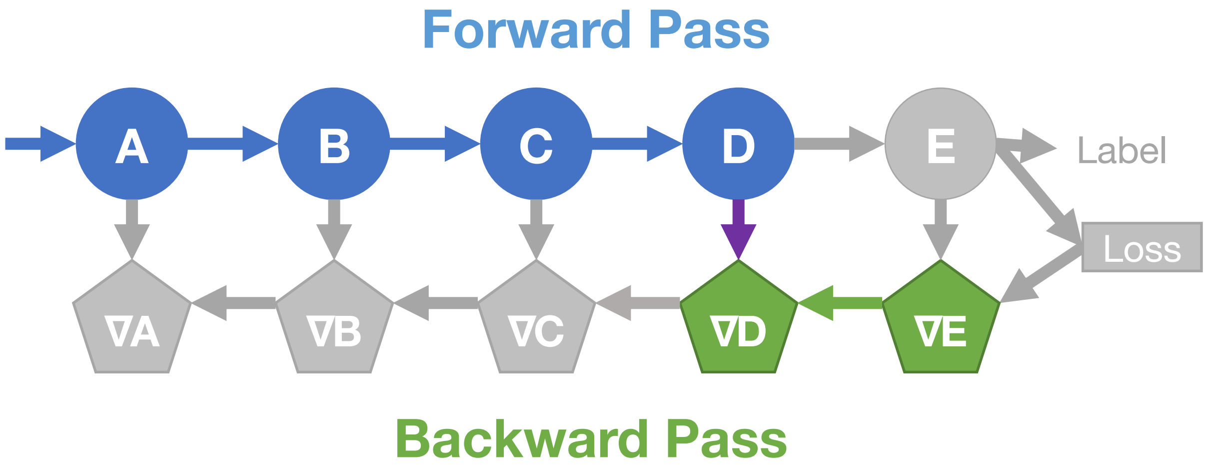
Keep all layers in RAM

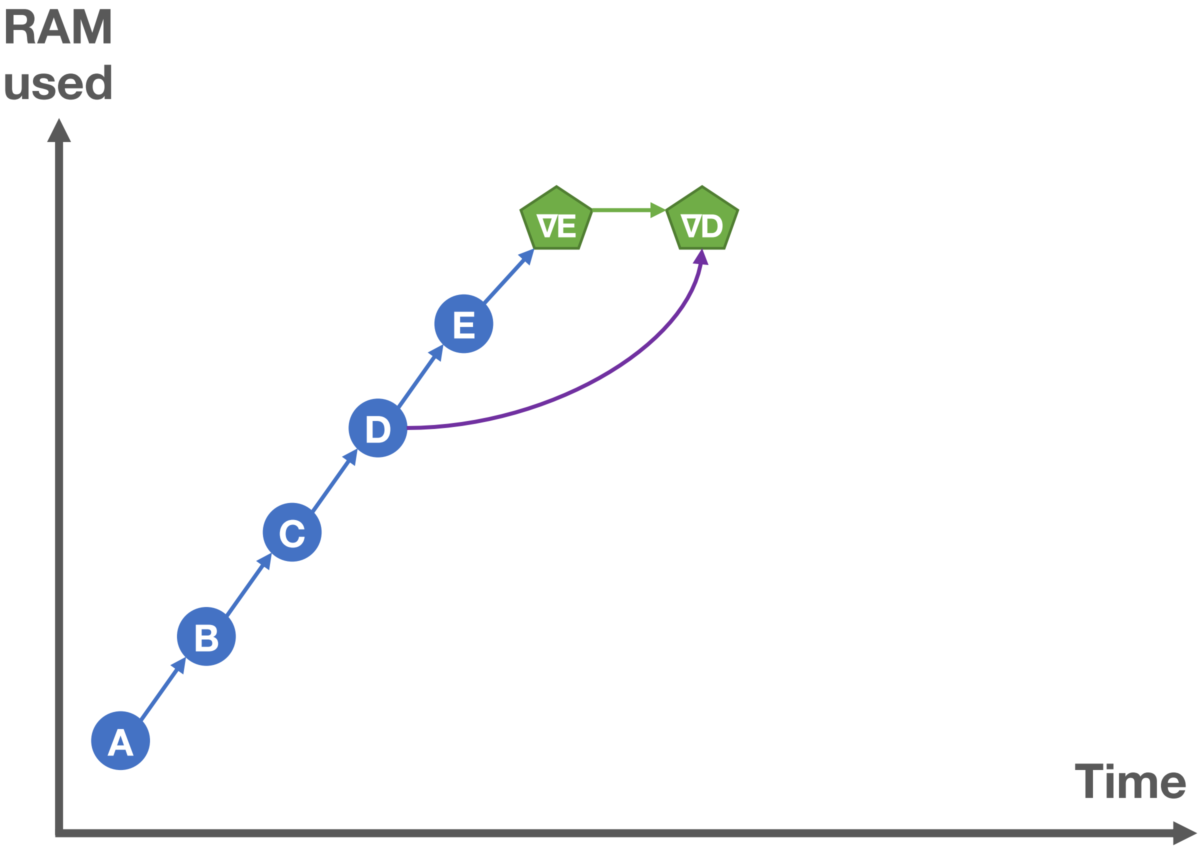




RAM-hungry backpropagation policy

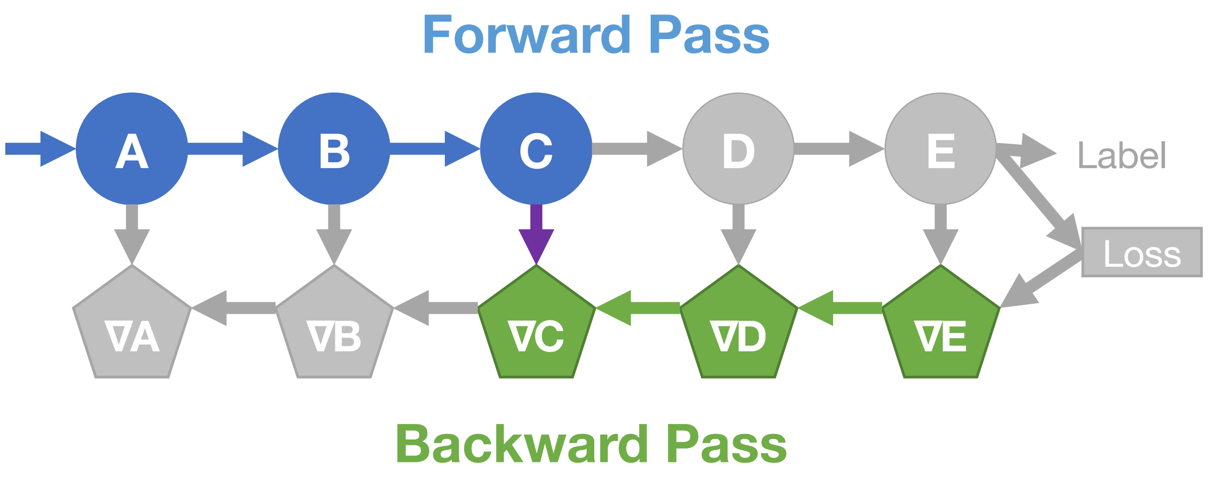
Keep all layers in RAM

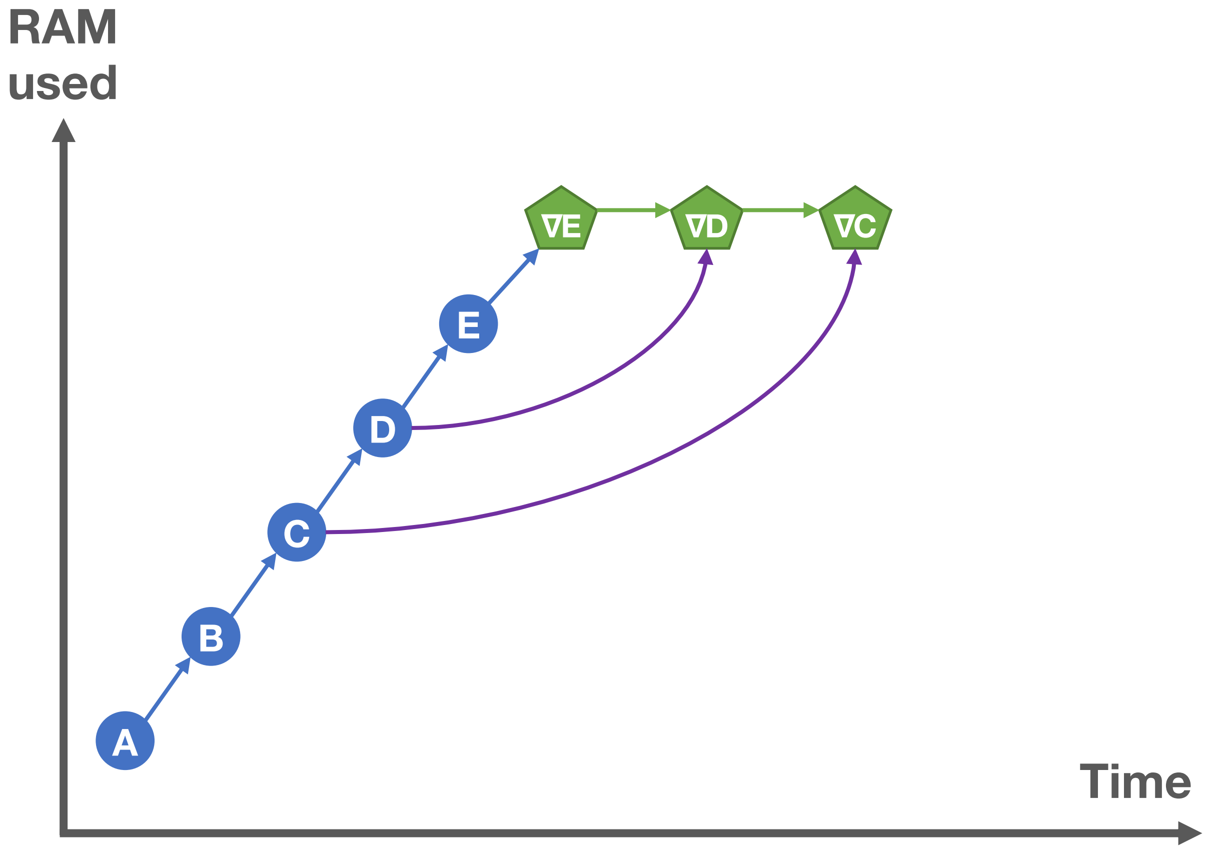




RAM-hungry backpropagation policy

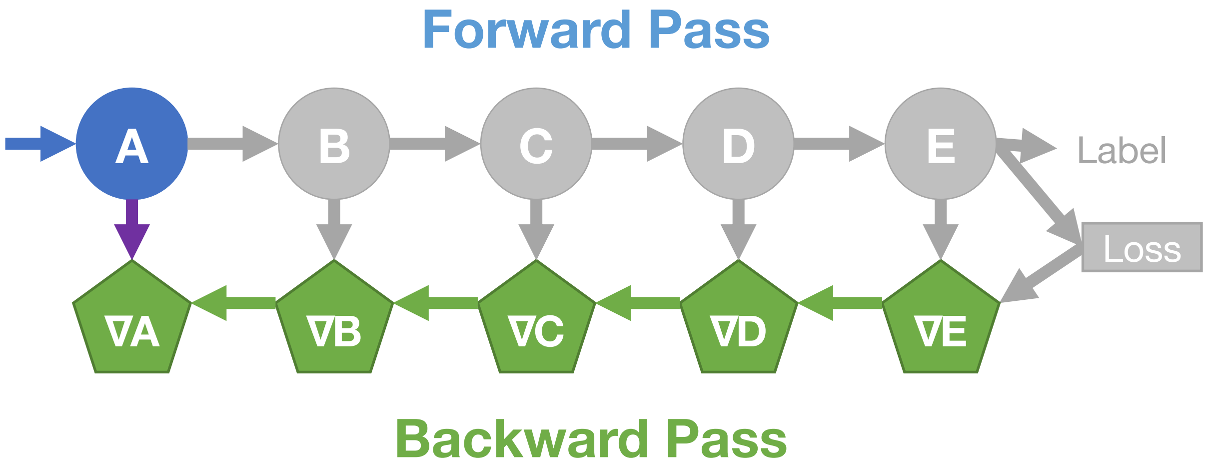
Keep all layers in RAM

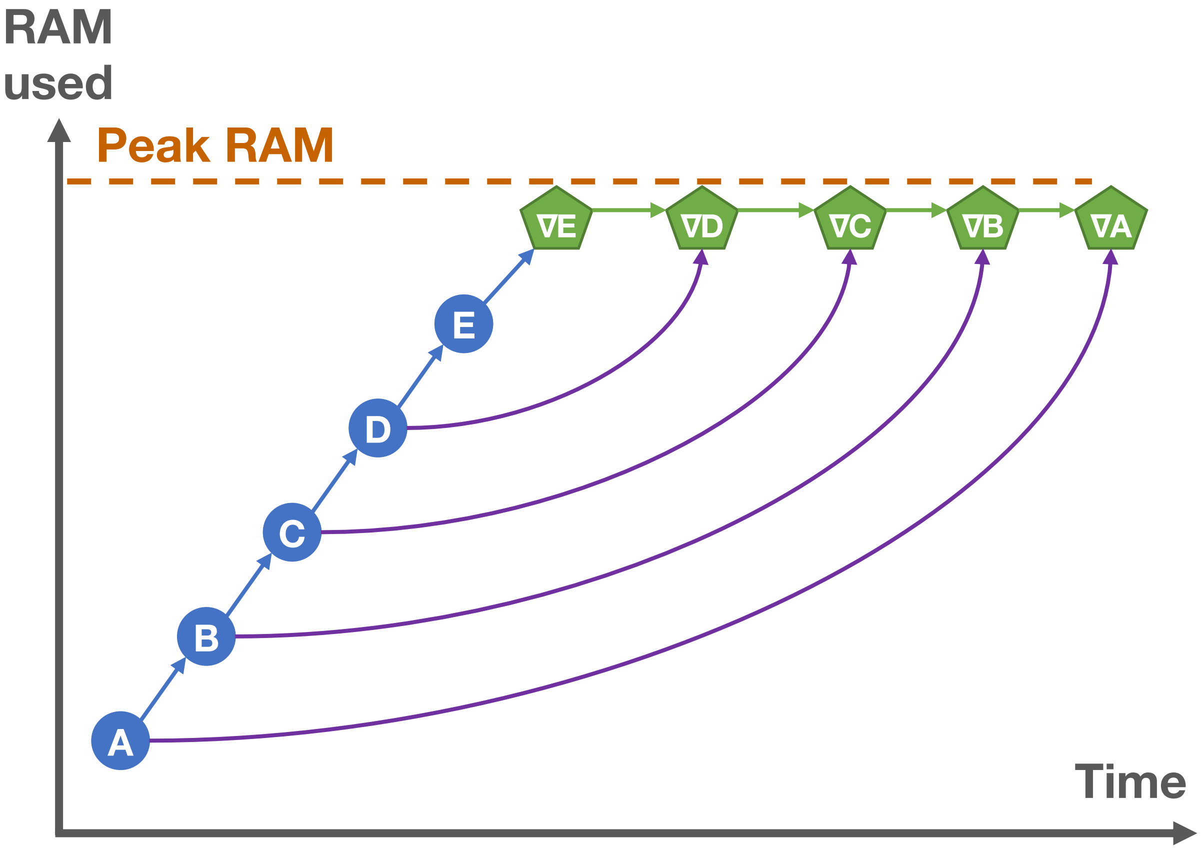




RAM-hungry backpropagation policy

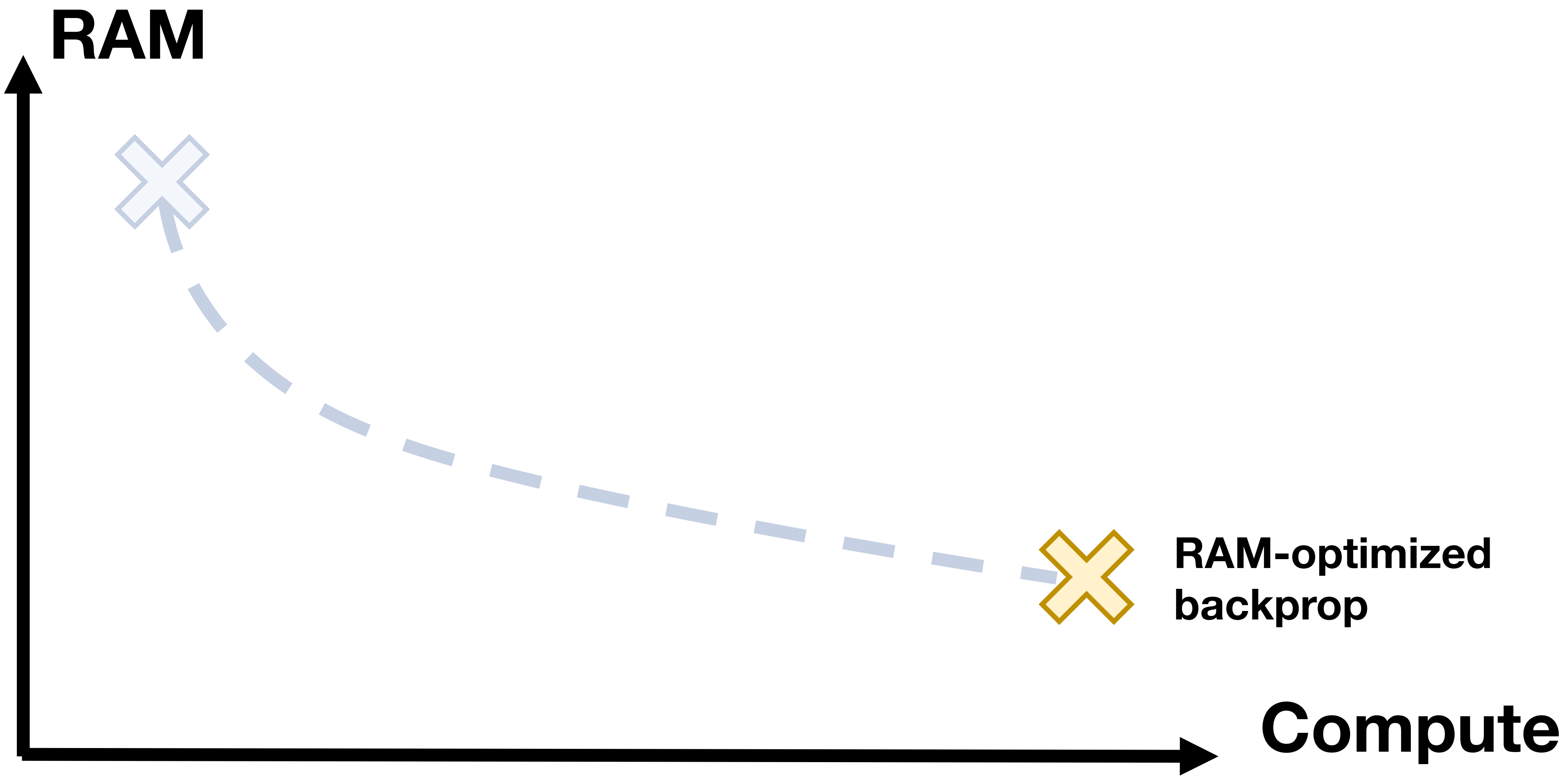
Keep all layers in RAM





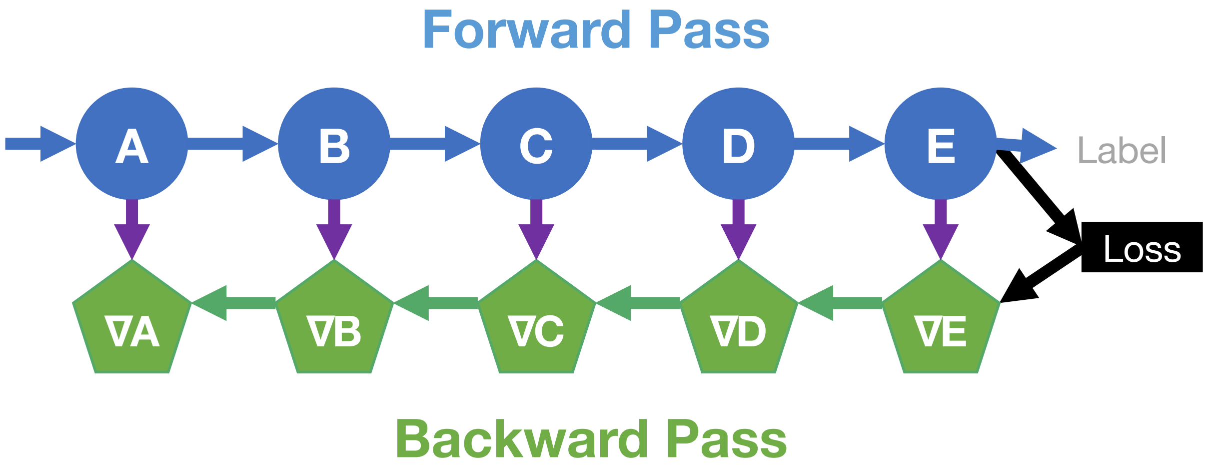
RAM-optimized backpropagation policy

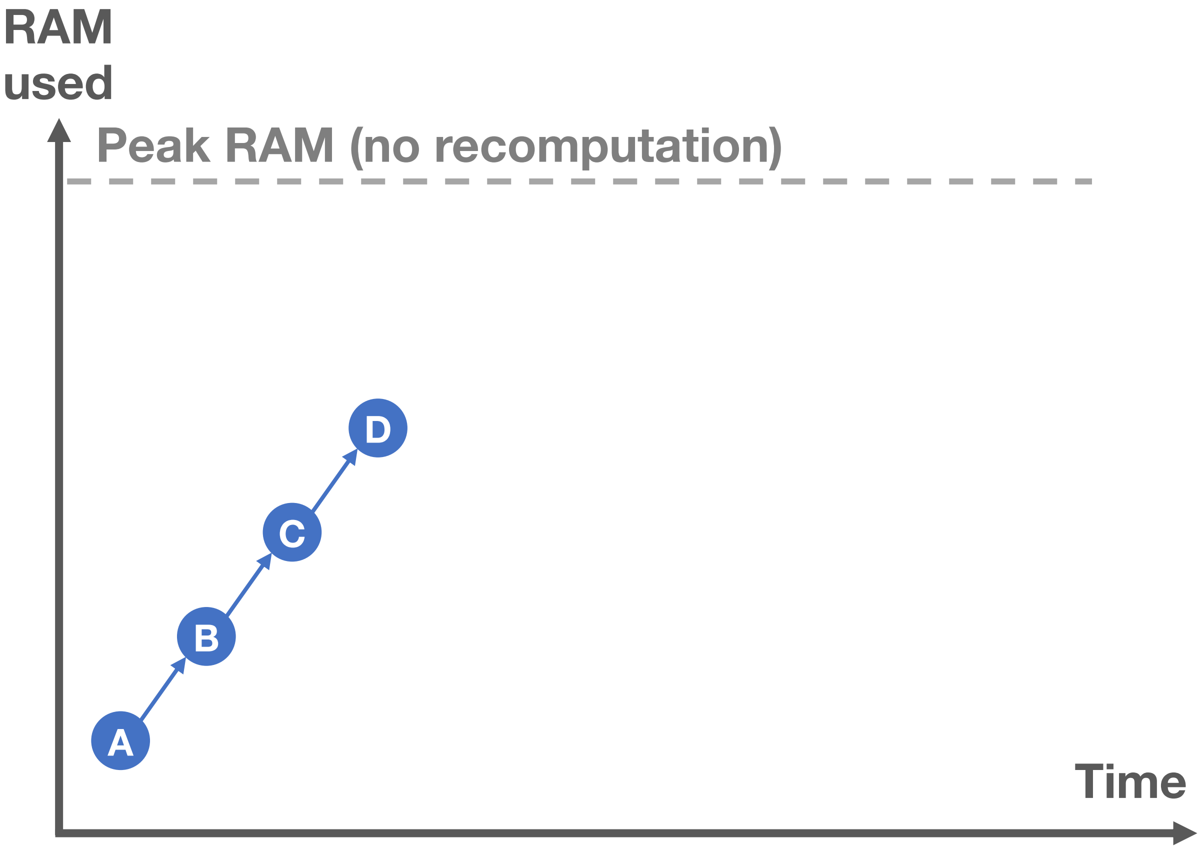
Recompute all layers as needed



RAM-optimized backpropagation policy

Recompute all layers



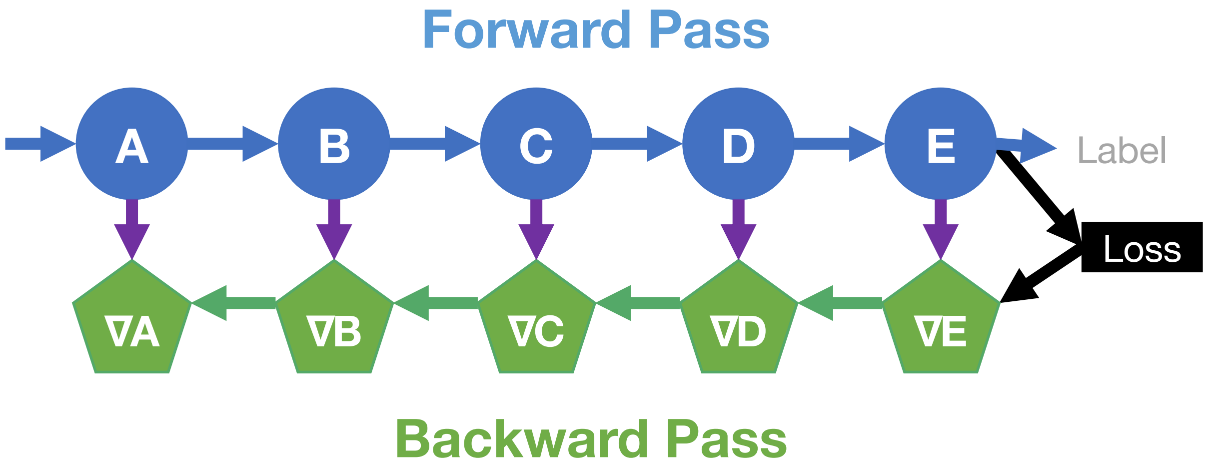


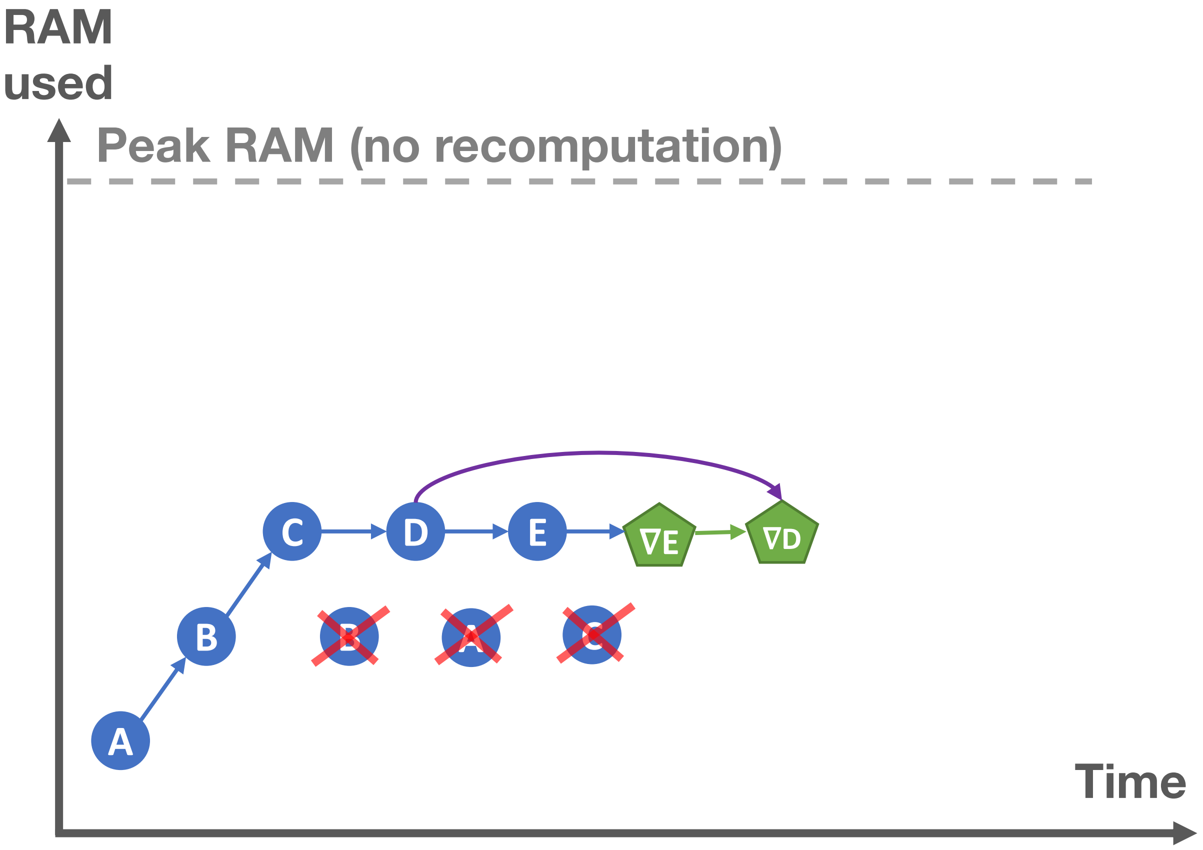
How can we use less memory?

Free early & recompute

RAM-optimized backpropagation policy

Recompute all layers



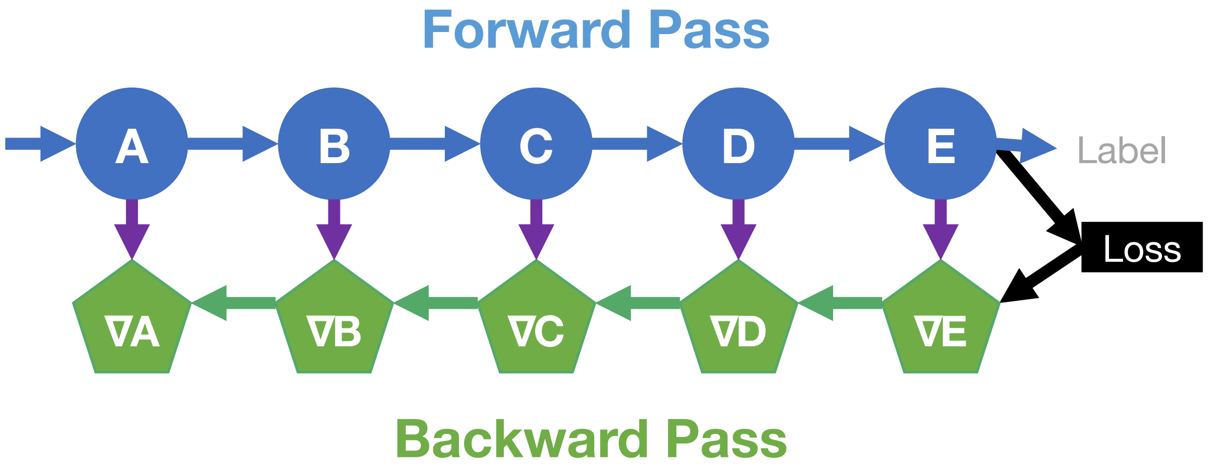


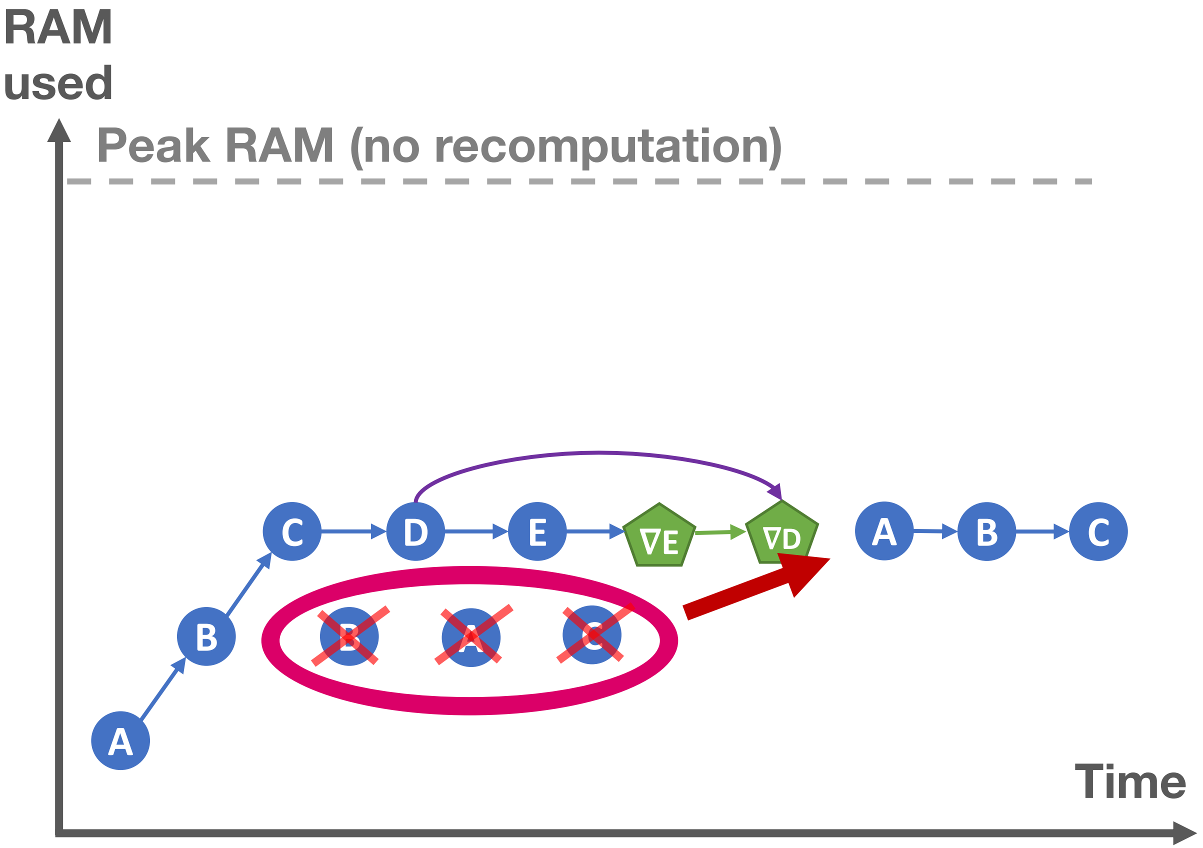
How can we use less memory?

Free early & recompute

RAM-optimized backpropagation policy

Recompute all layers



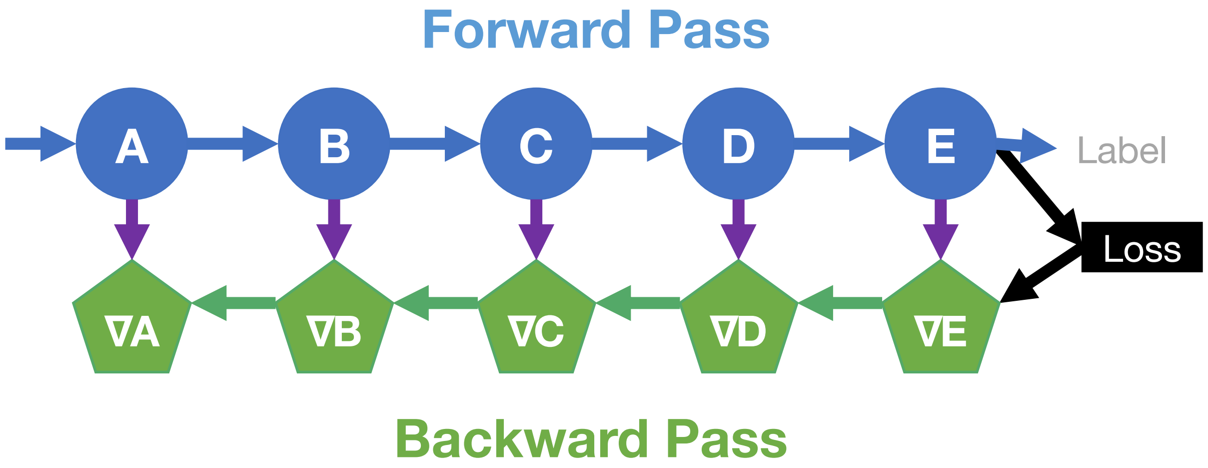


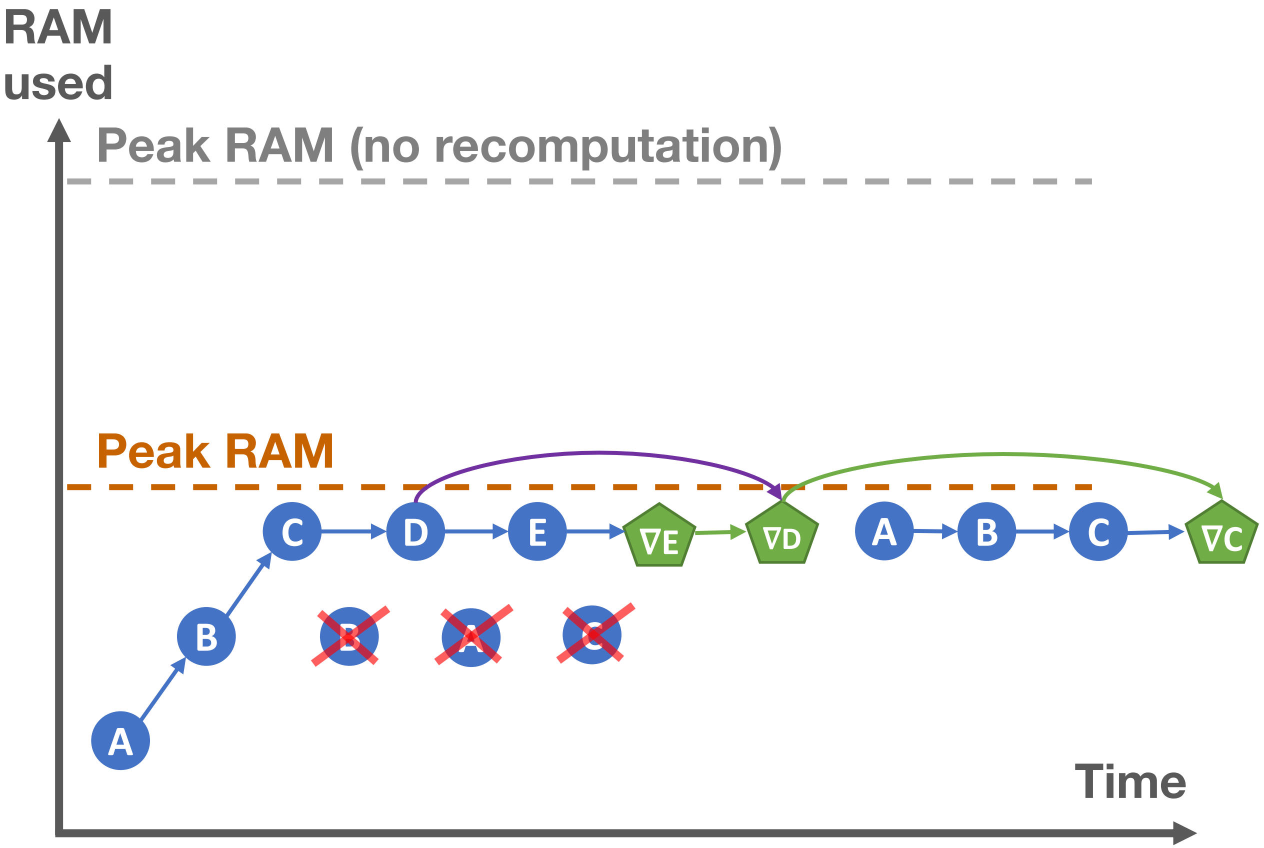
How can we use less memory?

Free early & recompute

RAM-optimized backpropagation policy

Recompute all layers

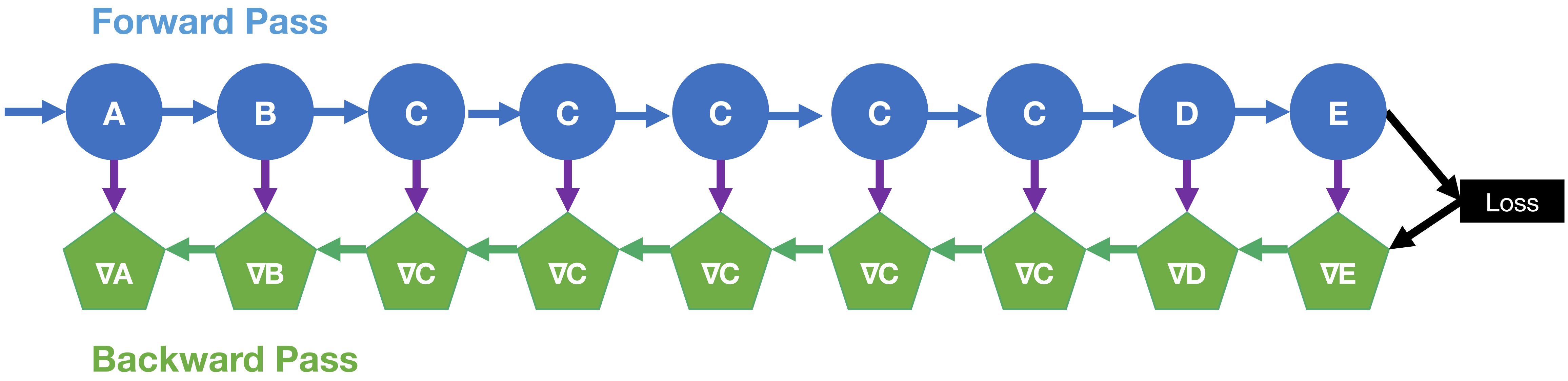




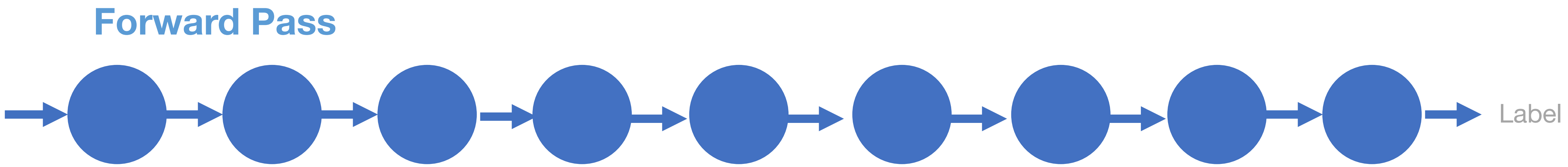
How can we use less memory?

Free early & recompute

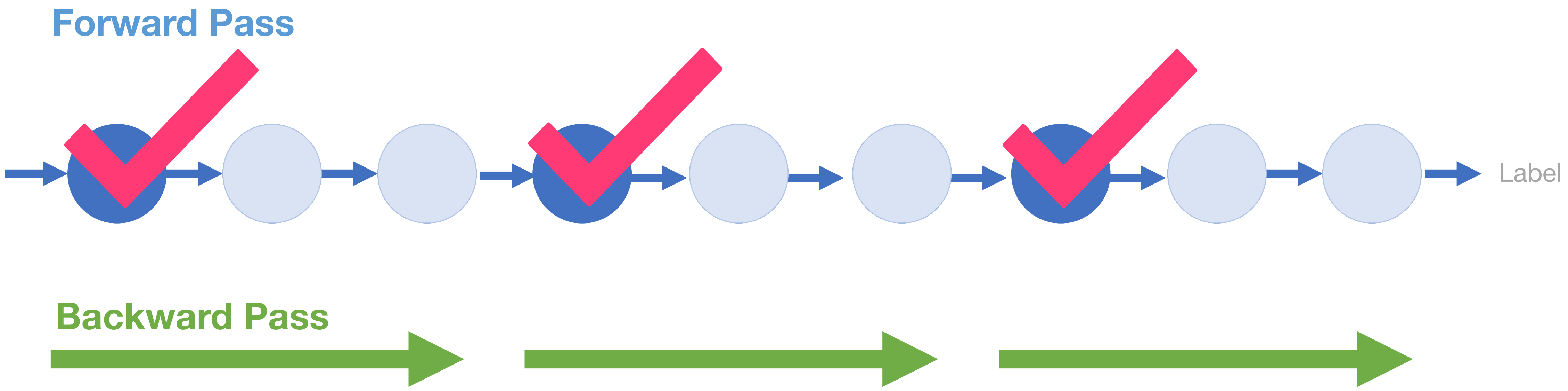
How to choose which layers to recompute?



How to choose which layers to recompute?

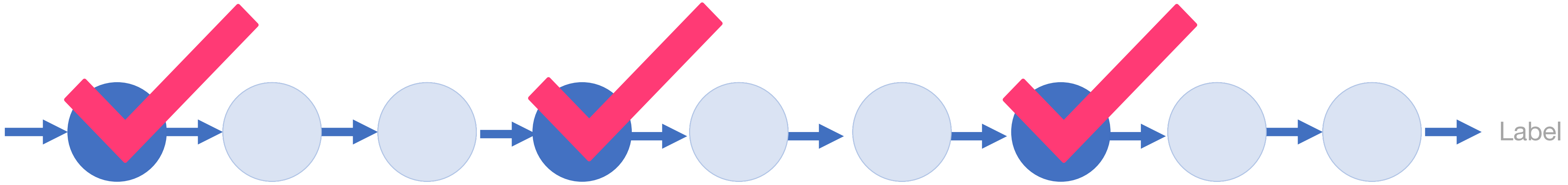


How to choose which layers to recompute?



$$\textbf{Compute: } O(n) \text{ additional overhead}$$

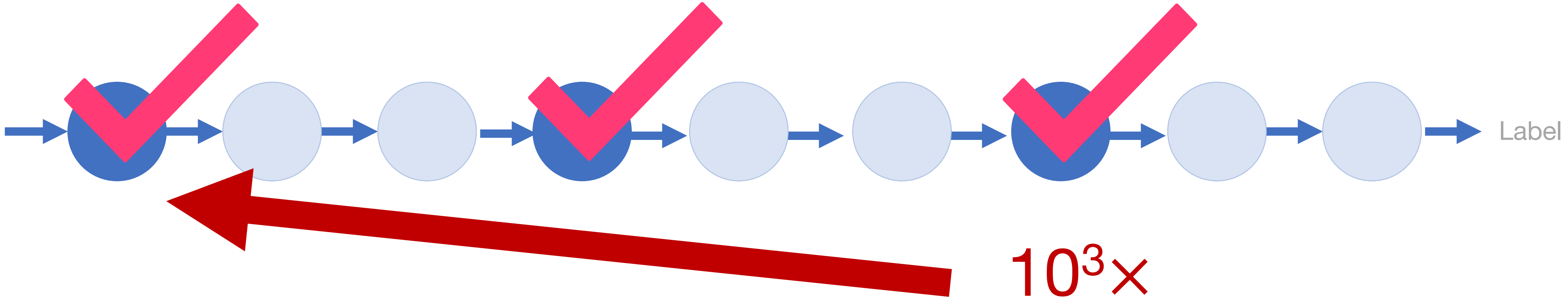
$$\textbf{RAM: } O(\sqrt{n}) \text{ RAM usage}$$



Challenges of heuristics:

1. Variable runtime per layer

$10^6\times$
slower

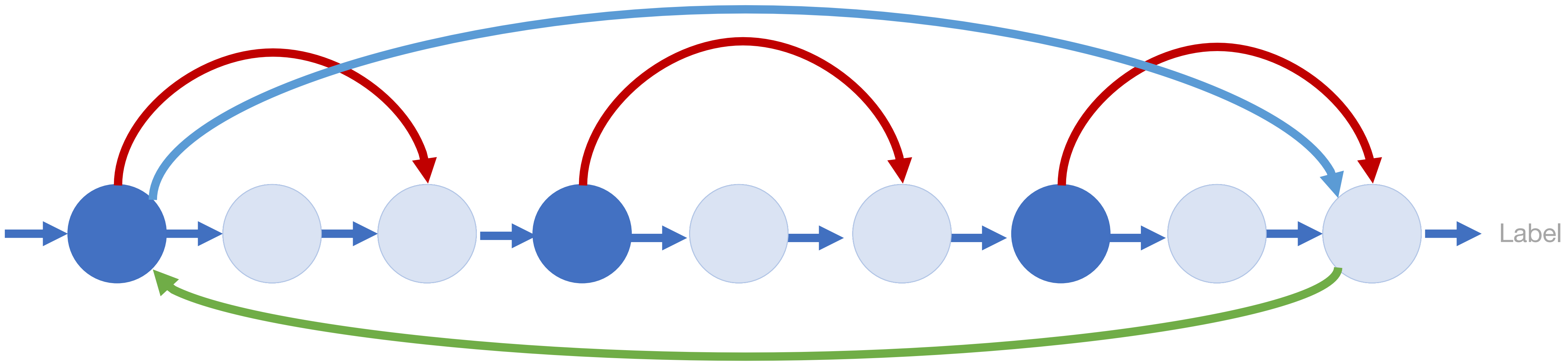


$10^3\times$
more RAM

Challenges of heuristics:

1. Variable runtime per layer

2. Variable RAM usage per layer



Challenges of heuristics:

1. Variable runtime per layer

2. Variable RAM usage per layer

3. Real DNNs are non-linear

Prior work is suboptimal in general setting!

Greedy heuristic

[Chen 2016]

[XLA authors 2017, 2020]

Divide-and-conquer heuristic

[Griewank 2000]

[Kowarz 2006]

[Siskind 2018]

[Kumar 2019]

Optimal for specific architecture

[Gruslys 2016]

[Feng 2018]

[Beaumont 2019]

Challenges:

1. Variable runtime per layer

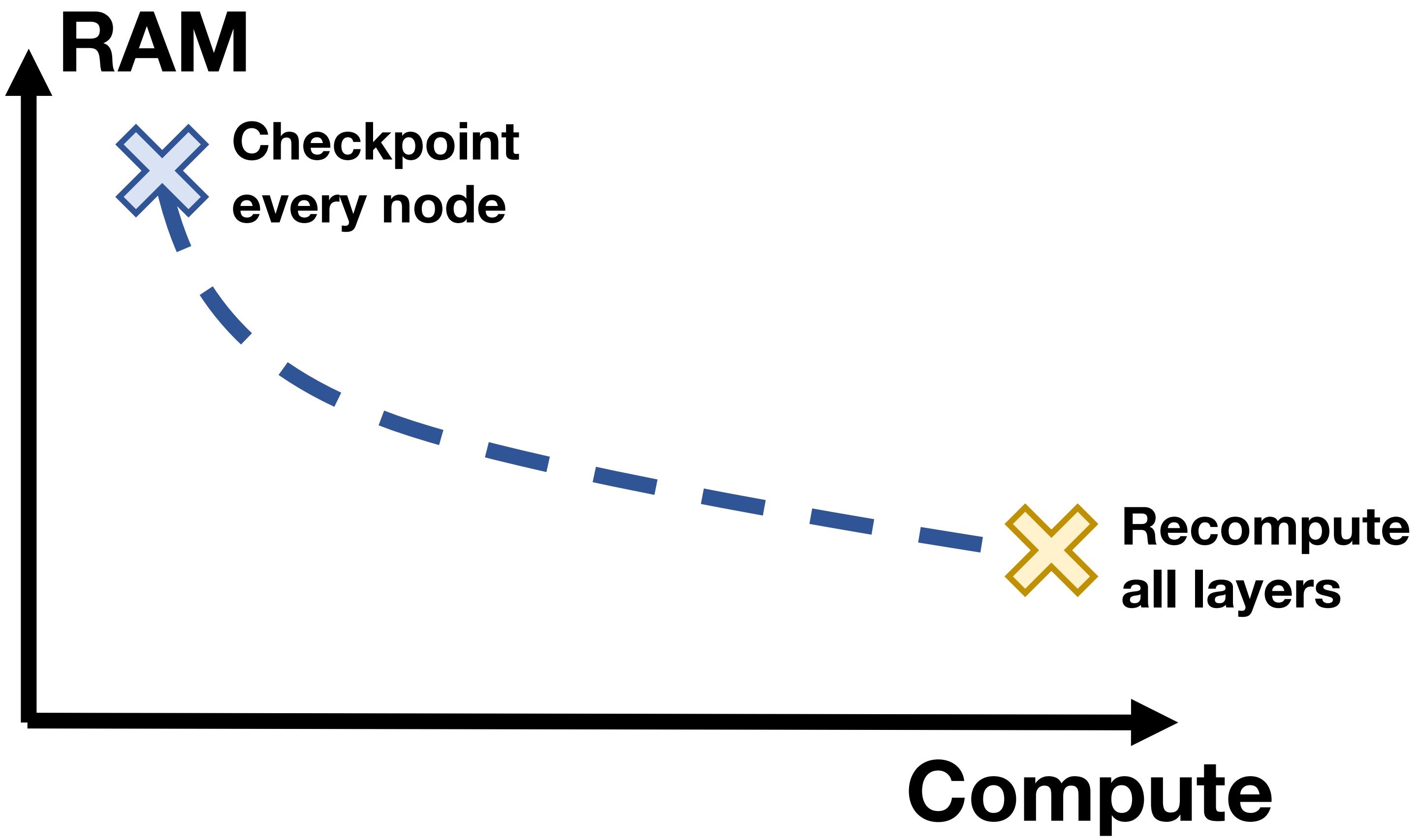
2. Variable RAM usage per layer

3. Real DNNs are non-linear

Can we optimally trade-off RAM for compute?

Let's be:

1. Hardware-aware
2. RAM-aware
3. DAG flexibility



Checkmate

A system for optimal tensor rematerialization

Hardware + RAM aware

Solve for 10s-1hr
Train for 1mo

GPU, CPU, TPU support

Accurate cost model

Flexible search space

Optimal solver
Integer Linear Program

Near-optimal solver
Two phase rounding

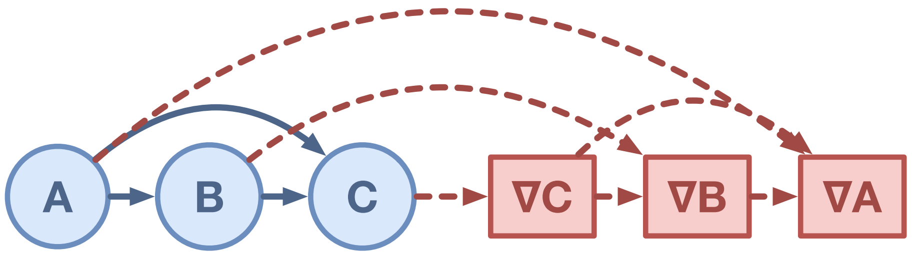
Graph rewrite

Checkmate

A system for optimal tensor rematerialization

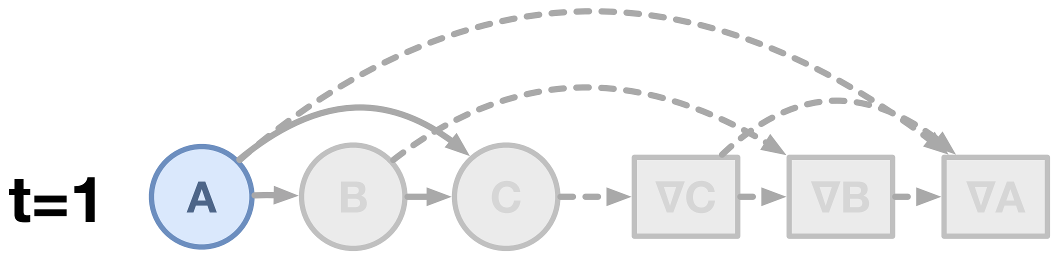
Checkmate

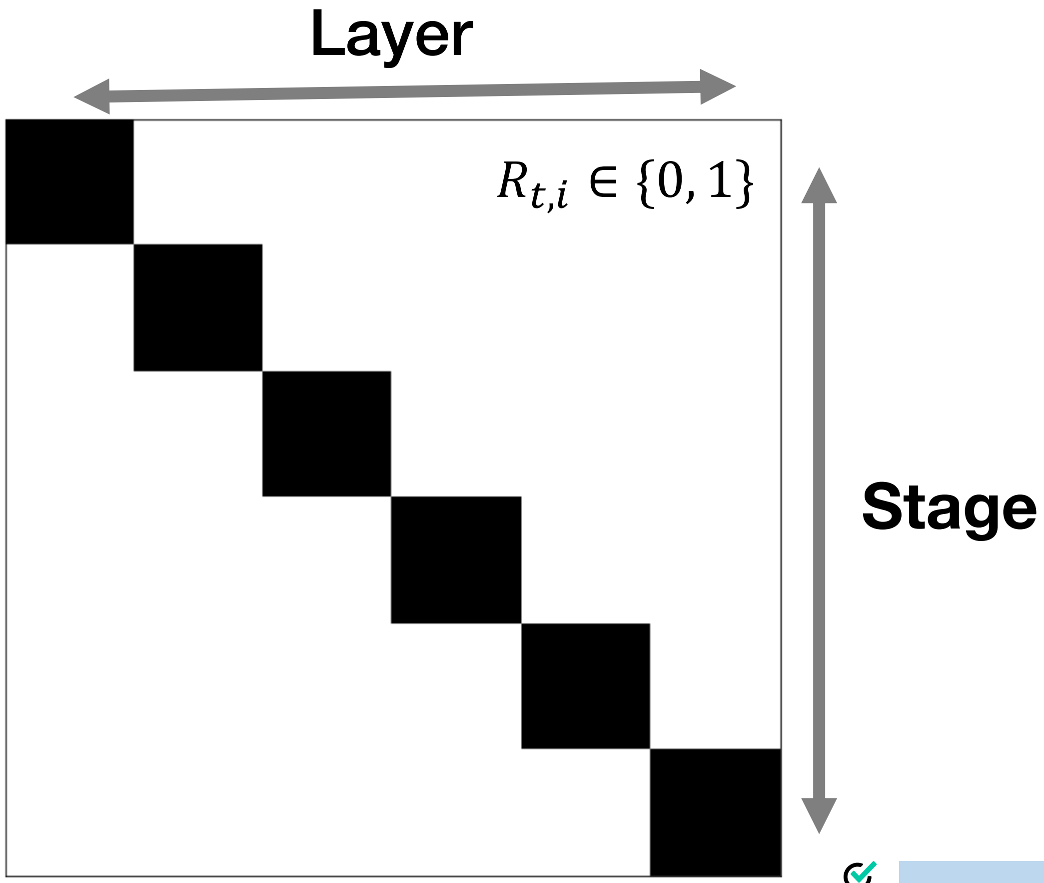
A system for optimal tensor rematerialization



Checkmate

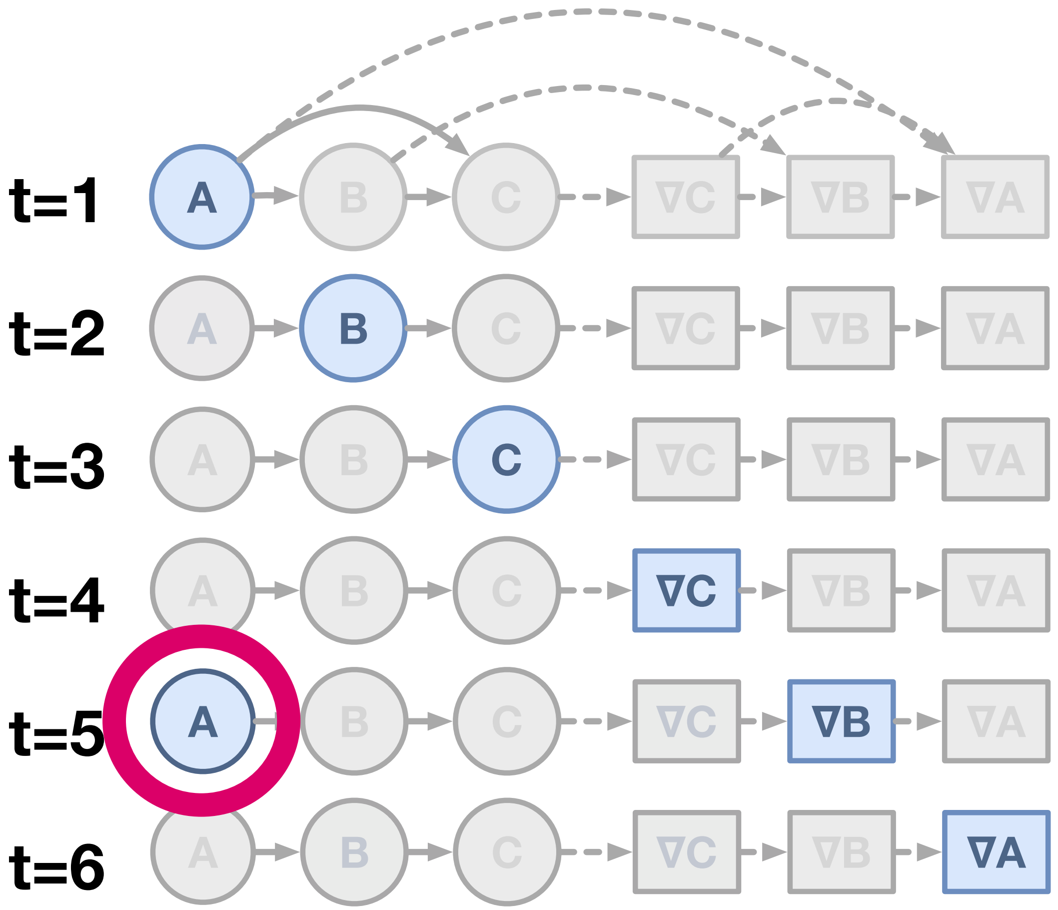
A system for optimal tensor rematerialization

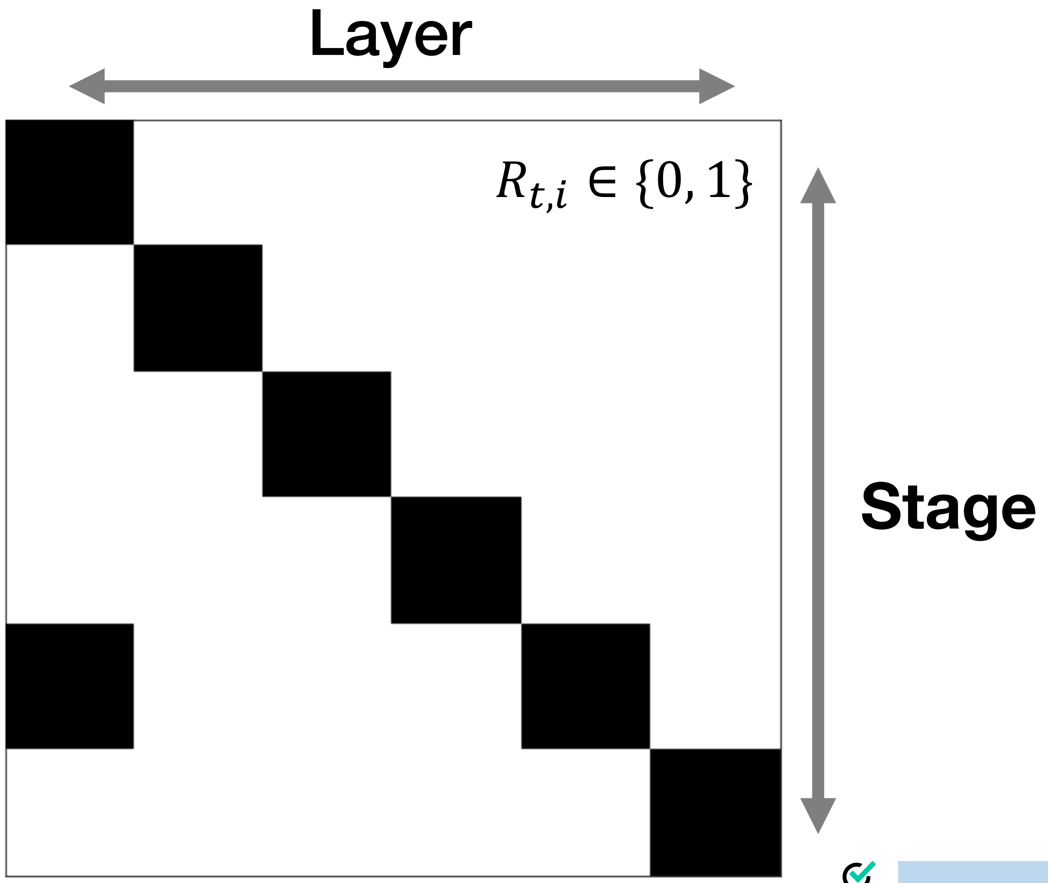




Checkmate

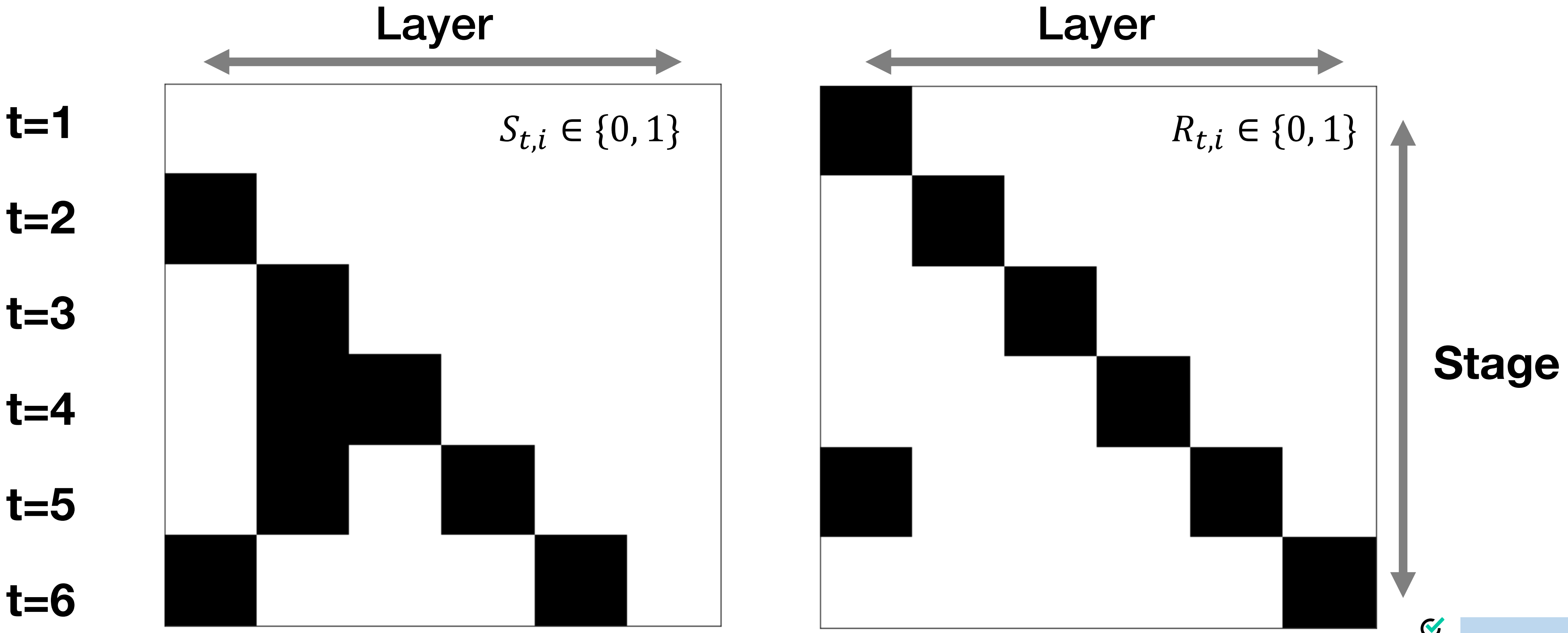
A system for optimal tensor rematerialization





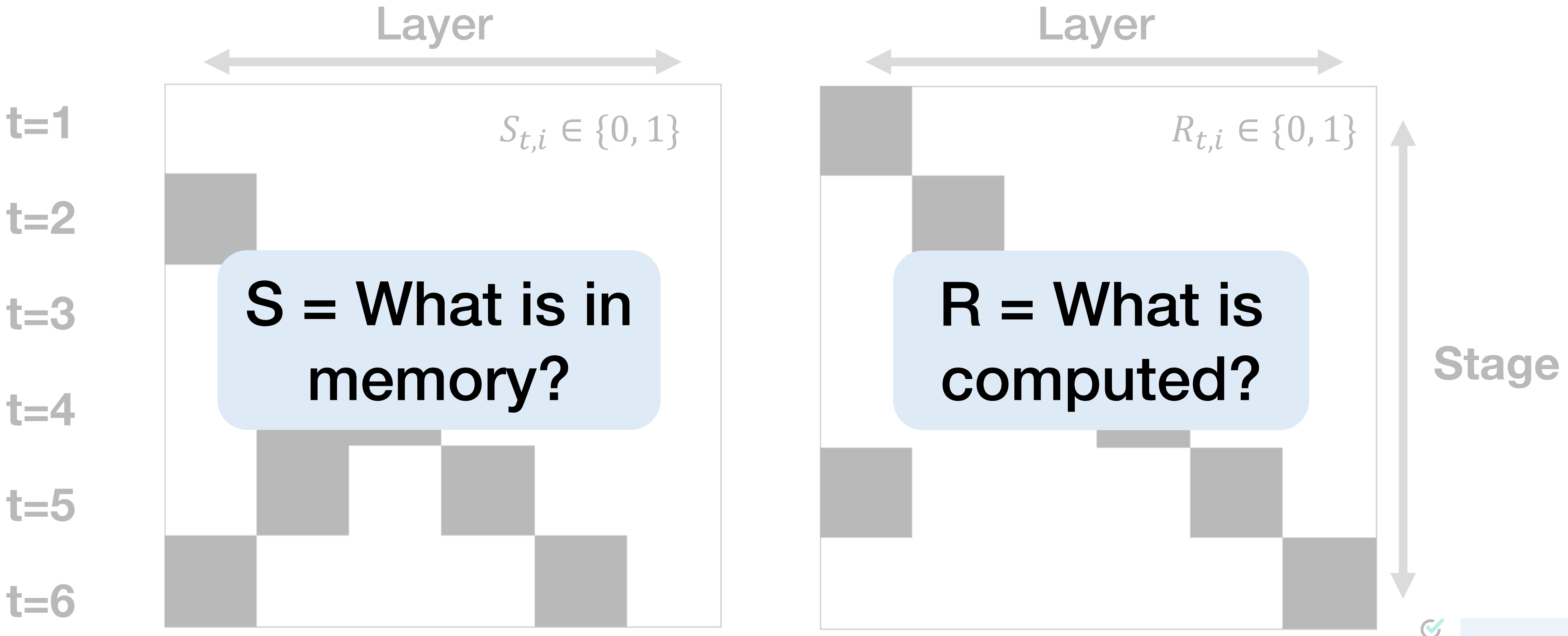
Checkmate

A system for optimal tensor rematerialization

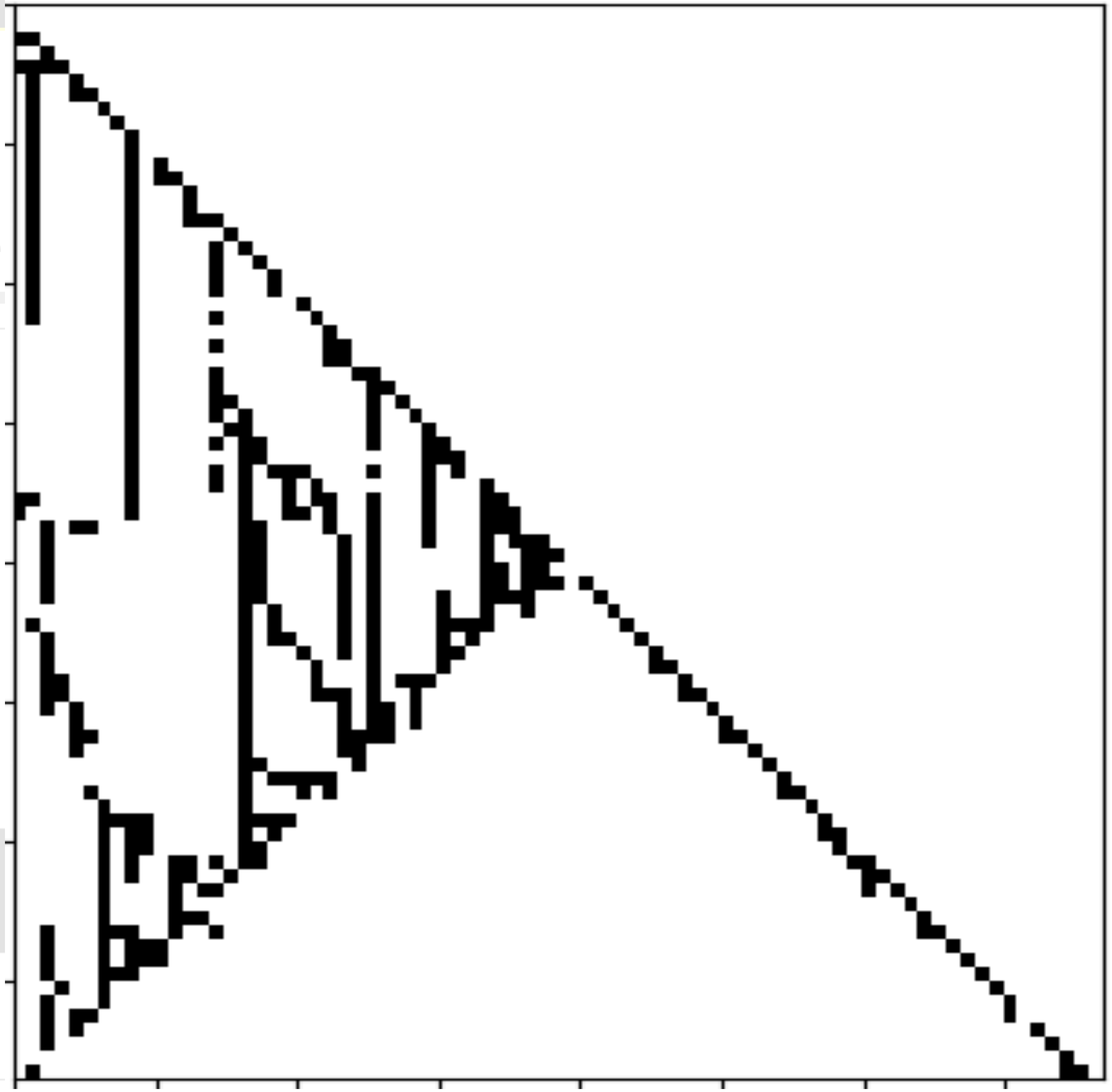


Checkmate

A system for optimal tensor rematerialization



Example of optimal "S" (SegNet)



Checkmate

A system for optimal tensor rematerialization

Layers

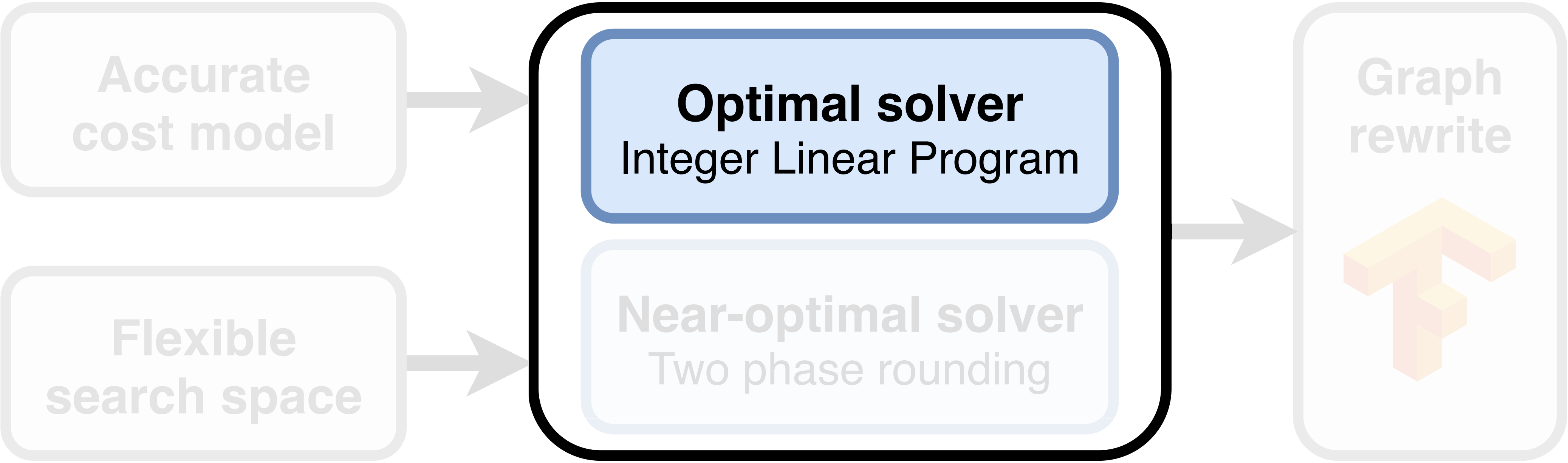
t=1
t=2
t=3
t=4
t=5
t=6

{0, 1}

Stage

Checkmate

A system for optimal tensor rematerialization



Checkmate

A system for optimal tensor rematerialization

Minimize forward + backward cost

$$\min_{S,R,U} \sum \sum C_i \, R_{t,i}$$

Use R matrix to create linear objective

Decision variables

$S_{t,i} \in \{0,1\}$ Layer i stored for stage t

$R_{t,i} \in \{0,1\}$ Layer i (re)computed in stage t

Checkmate

A system for optimal tensor rematerialization

Minimize forward + backward cost

$$\min_{S,R,U} \sum \sum C_i \, R_{t,i}$$

Decision variables

$S_{t,i} \in \{0, 1\}$ Layer i stored for stage t

$R_{t,i} \in \{0, 1\}$ Layer i (re)computed in stage t

Correctness

$$R_{t,j} \leq R_{t,i} + S_{t,i}$$

$$S_{t,i} \leq R_{t-1,i} + S_{t-1,i}$$

"A layer's dependencies must be computed before evaluation"

"A layer must be computed before it can be stored in RAM"

Checkmate

A system for optimal tensor rematerialization

Minimize forward + backward cost

$$\min_{S,R,U} \sum \sum C_i \, R_{t,i}$$

Correctness

$$R_{t,j} \leq R_{t,i} + S_{t,i}$$

$$S_{t,i} \leq R_{t-1,i} + S_{t-1,i}$$

Memory limit

$U_{t,k} \leq$ budget, …

Constrain memory via an implicit variable to model memory usage at each stage

Decision variables

$S_{t,i} \in \{0, 1\}$	Layer i stored for stage t
$R_{t,i} \in \{0, 1\}$	Layer i (re)computed in stage t
$U_{t,i} \in \mathbb{R}_+$	Memory usage in stage t

Checkmate

A system for optimal tens

Minimize forward + backward cost

$$\min_{S,R,U} \sum \sum C_i \, R_{t,i}$$

Correctness

$$R_{t,j} \le R_{t,i} + S_{t,i}$$

$$S_{t,i} \le R_{t-1,i} + S_{t-1,i}$$

Memory limit

$U_{t,k} \le$ budget, ...

Constrain me to model memory usage at each stage

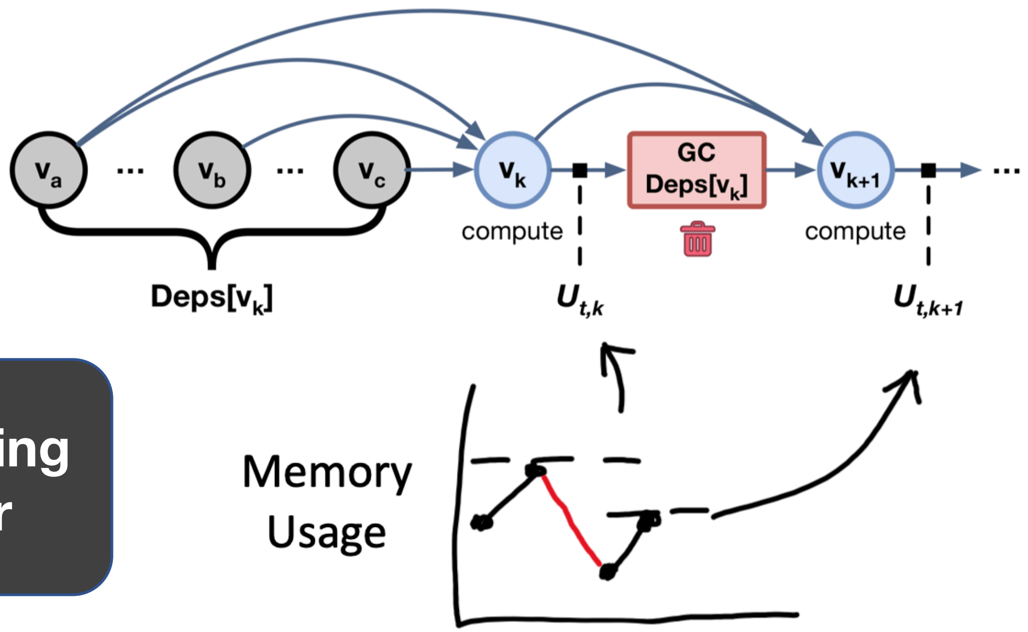
Constrain memory at all execution steps

$$U_{t,k} \le \text{ budget}$$

More checkpoints → more initial memory

$$U_{t,0} = \sum M_i S_{t,i}$$

$$U_{t,k+1} = U_{t,k} - \boxed{\text{GC Deps}[v_k]} + \boxed{M_{k+1} R_{t,k+1}}$$



Memory accounting details in paper

Checkmate

A system for optimal tensor rematerialization

Minimize forward + backward cost

$$\min_{S,R,U} \sum \sum C_i \, R_{t,i}$$

Correctness

$$R_{t,j} \leq R_{t,i} + S_{t,i}$$

$$S_{t,i} \leq R_{t-1,i} + S_{t-1,i}$$

Memory limit

$U_{t,k} \leq$ budget, ...

How long is the solve time?

9 hours 😳

Checkmate

A system for optimal tensor rematerialization

Minimize forward + backward cost

$$\min_{S,R,U} \sum \sum C_i \, R_{t,i}$$

Correctness

$$R_{t,j} \leq R_{t,i} + S_{t,i}$$

$$S_{t,i} \leq R_{t-1,i} + S_{t-1,i}$$

Memory limit

$U_{t,k} \leq$ budget, ...

Partition schedule into frontier-advancing stages

9 hours → 0.2 seconds

Tractability

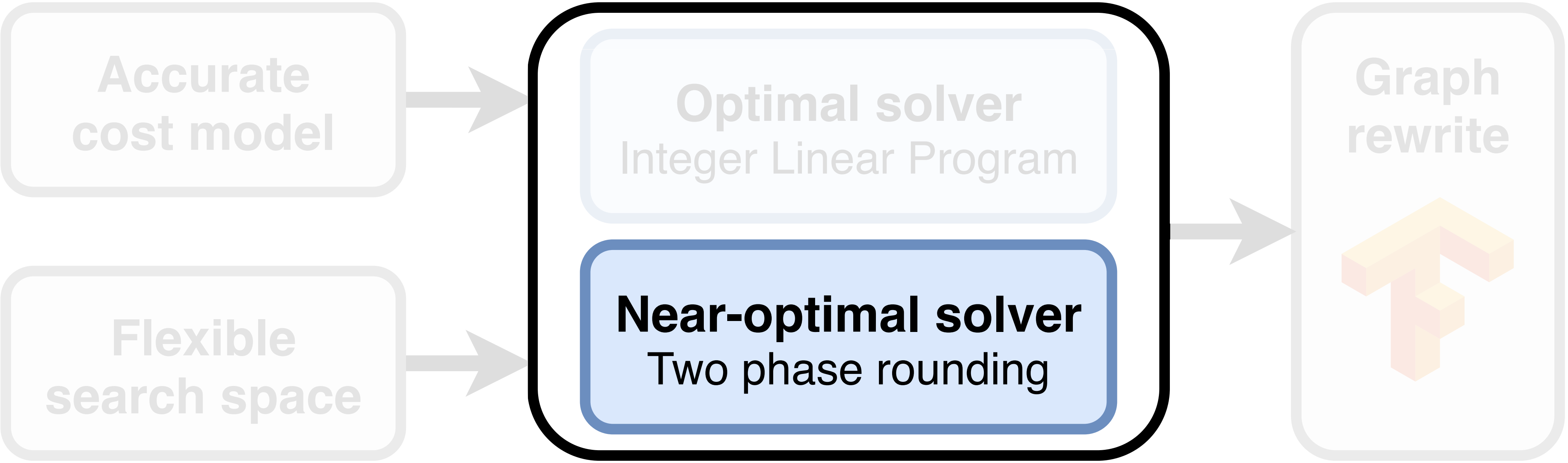
$R_{t,t} = 1$

R, S, U lower triangular

Prunes n! permutations of nodes

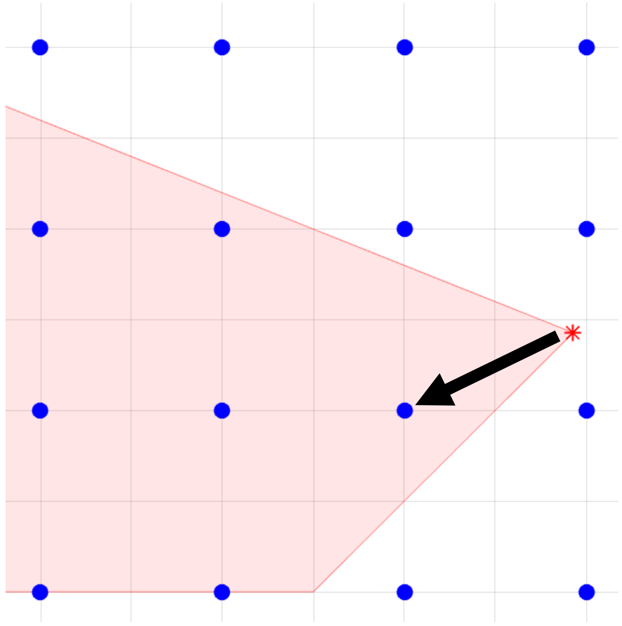
Checkmate

A system for optimal tensor rematerialization



ILP optimization is NP-hard (combinatorial search)

Polynomial-time approximation?



1. Relax boolean constraints
2. Solve LP
3. Round solution

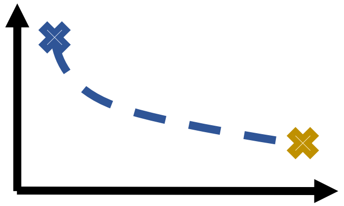
How to maintain feasibility?

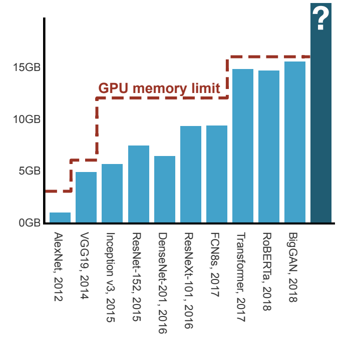
Insight: Given S, optimal R easy to compute

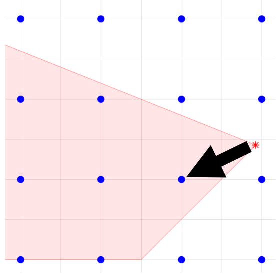
Proposed method: Two-Phase Rounding

Round S, solve other variables optimally

Evaluation: Questions

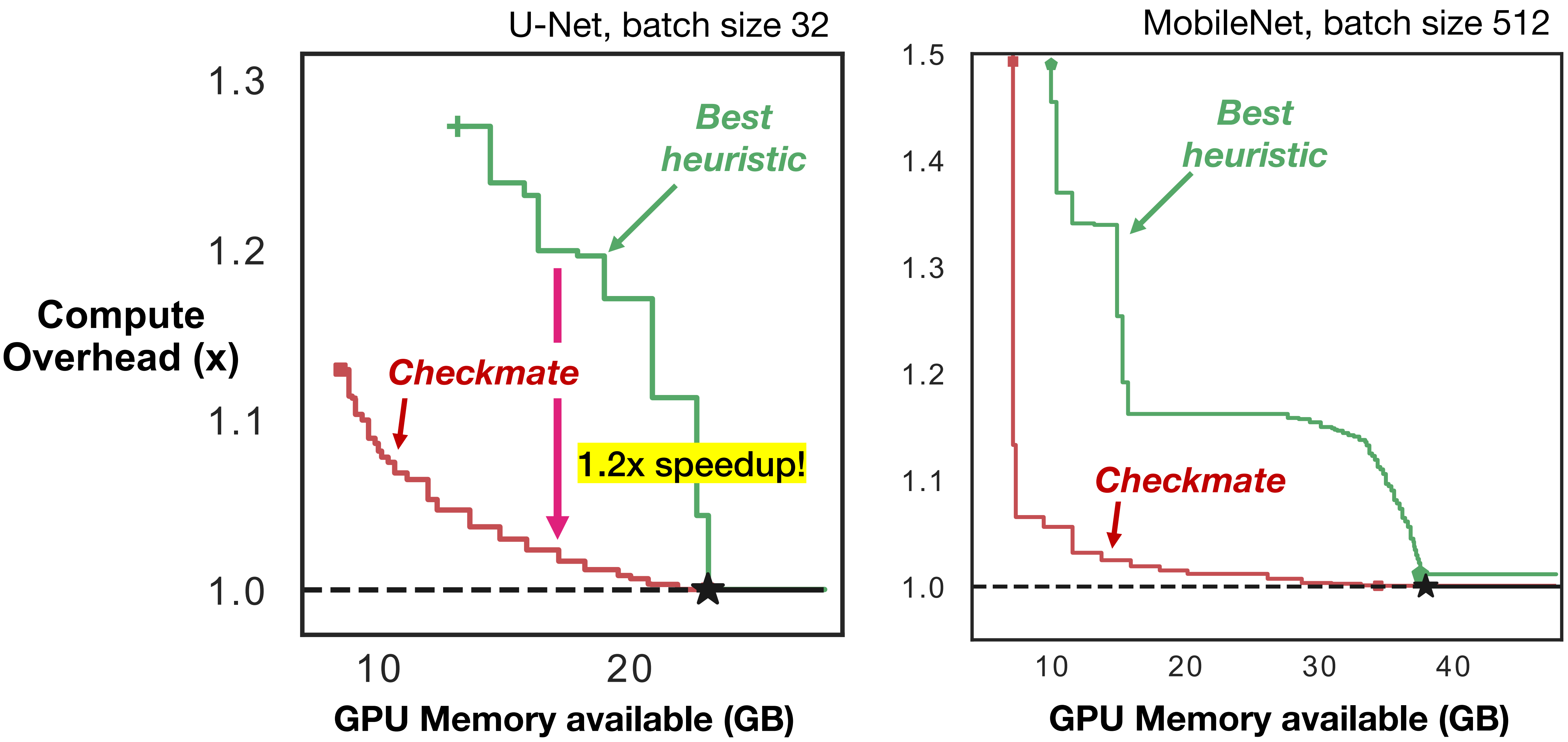
1. What is the memory vs compute trade-off? 

2. How much can we increase batch/model size? 

3. How well does two-phase rounding do? 

checkmateai.github.io

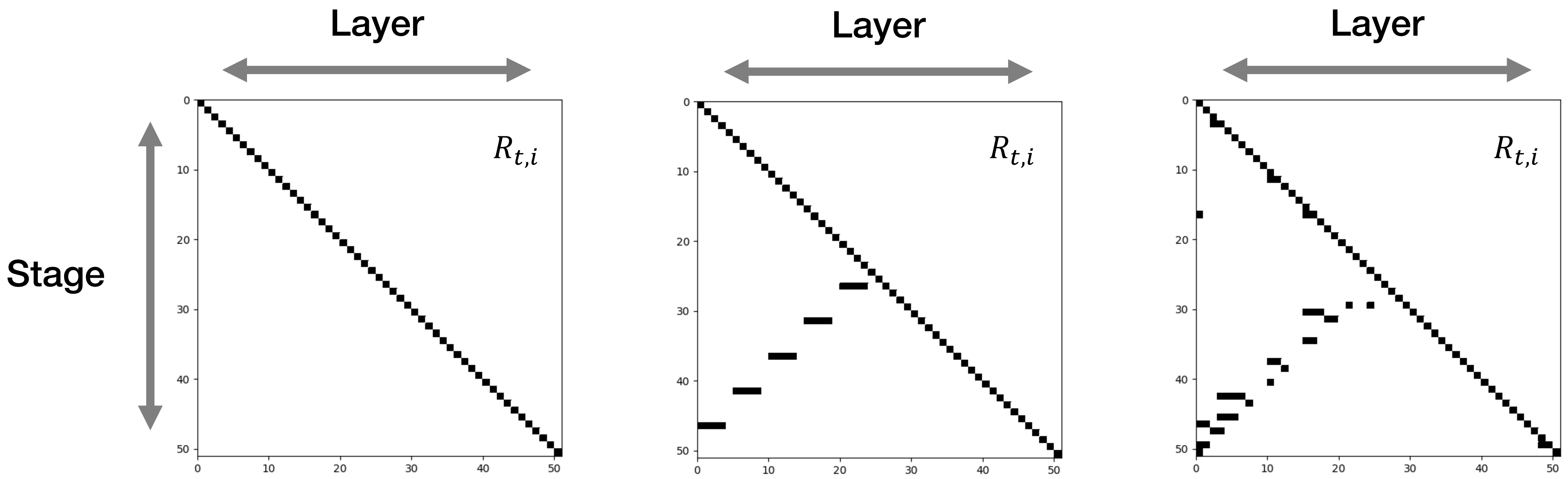
Evaluation: What is the memory vs compute trade-off?



checkmateai.github.io

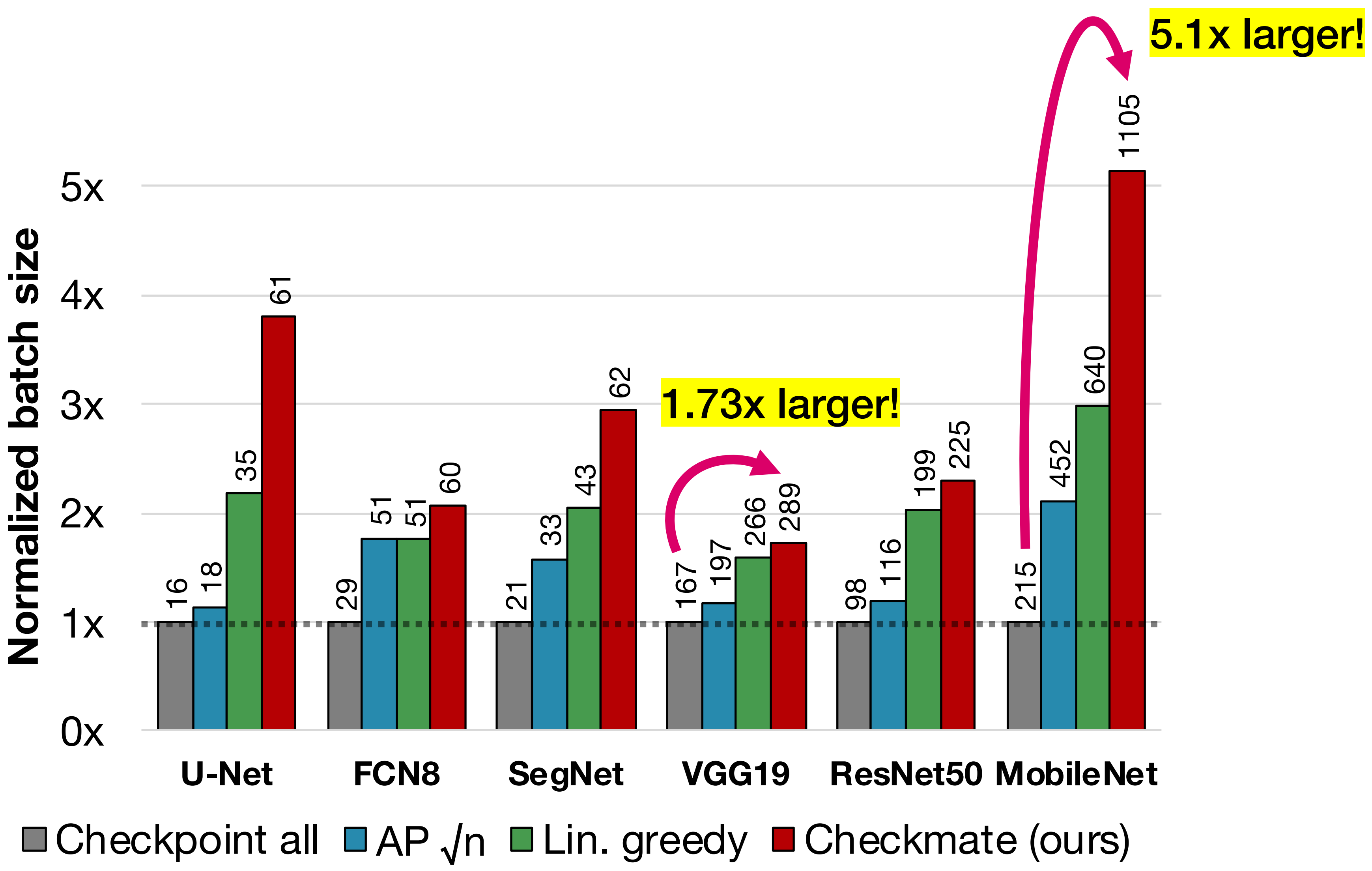
Evaluation: How much can we increase batch size?

VGG19
224x224 images

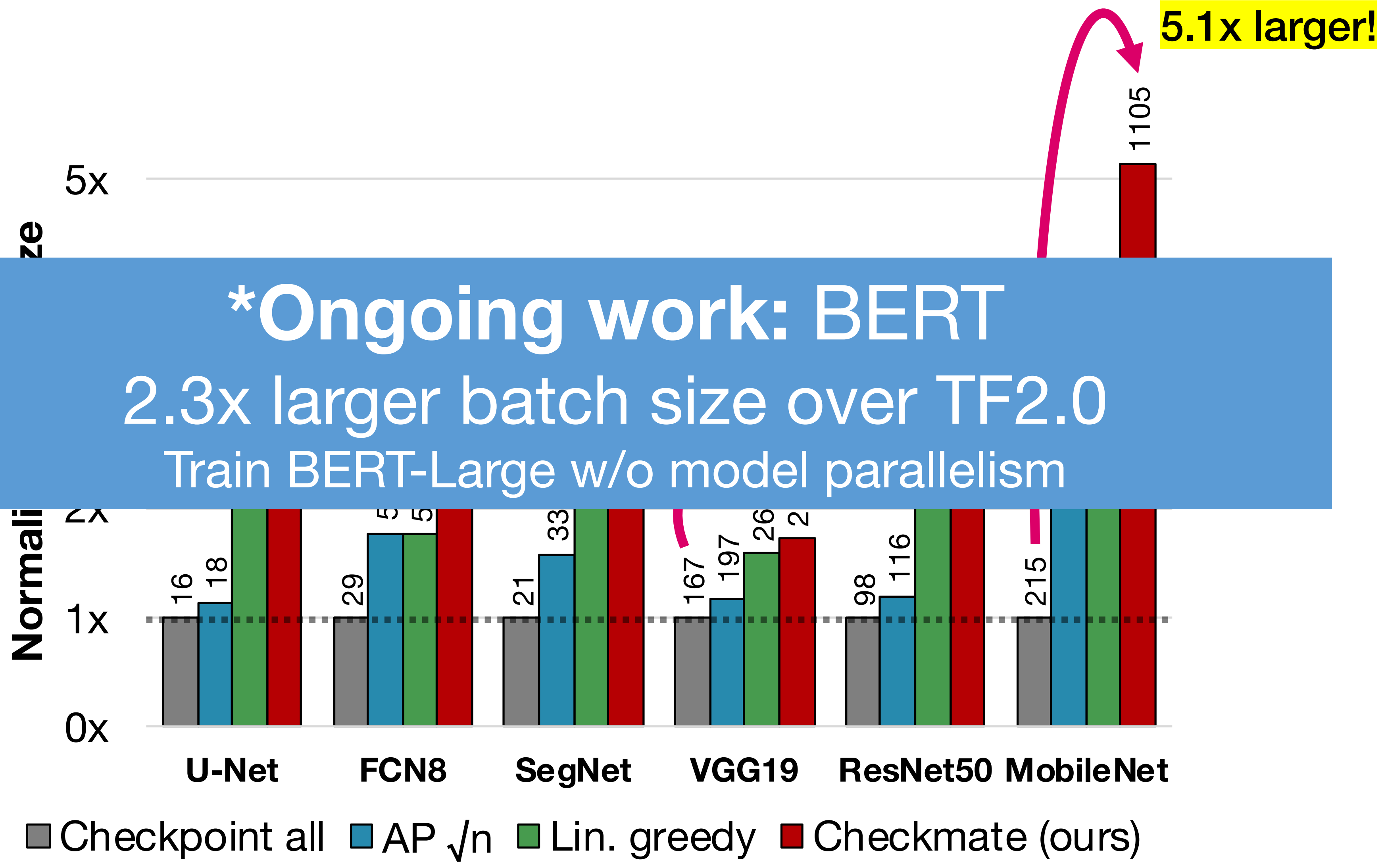


No rematerialization	Square root heuristic	Checkmate
Batch size 167	Batch size 197	Batch size 289
	1.18x larger!	1.73x larger! 10 sec solve

Evaluation: How much can we increase batch size?



Evaluation: How much can we increase batch size?



***Ongoing work:** BERT

2.3x larger batch size over TF2.0

Train BERT-Large w/o model parallelism

Evaluation: How well does 2P rounding approximate ILP?

	AP $\sqrt{n}$	AP greedy	Griewank $\log n$	Two-phase LP rounding
MobileNet	$1.14\times$	$1.07\times$	$7.07\times$	**1.06**$\times$
VGG16	$1.28\times$	$1.06\times$	$1.44\times$	**1.01**$\times$
VGG19	$1.54\times$	$1.39\times$	$1.75\times$	**1.00**$\times$
U-Net	$1.27\times$	$1.23\times$	-	**1.03**$\times$
ResNet50	$1.20\times$	$1.25\times$	-	**1.05**$\times$

Within 6% of optimal cost (geomean)

43x speedup for ResNet50
440x speedup for MobileNet

Checkmate

Key ideas:

- GPU memory limits are preventing the development of new deep learning models.
- We present the first general solution for optimal & near-optimal graph rematerialization.
- Formulation supports **arbitrary DAGs** and is both **hardware-aware** and **memory-aware**
- Integration with just **one line of code**

```
train_iteration = checkmate.compile_tf2(
    model, loss, optimizer, input_shape, label_shape)
```

Code and paper:
checkmateai.github.io

Email me:
parasj@berkeley.edu

Paras Jain, Ajay Jain, Ani Nrusimha, Amir Gholami, Pieter Abbeel, Kurt Keutzer, Ion Stoica, Joseph Gonzalez

Berkeley
UNIVERSITY OF CALIFORNIA