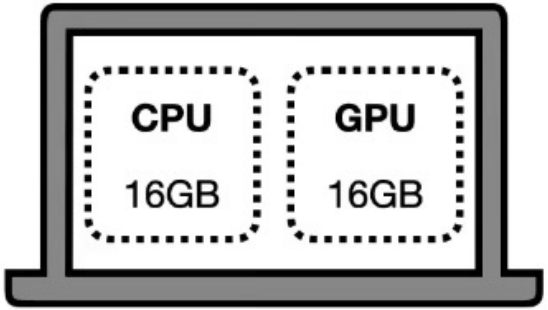
# QLoRA

LLM fine-tuning made accessible

# The Problem

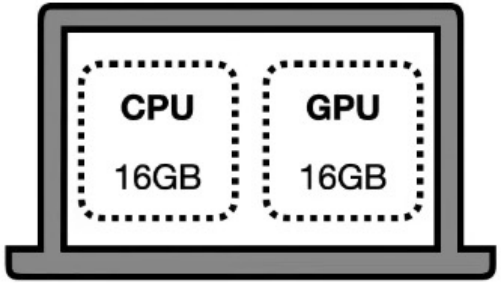
LLMs are (computationally) expensive

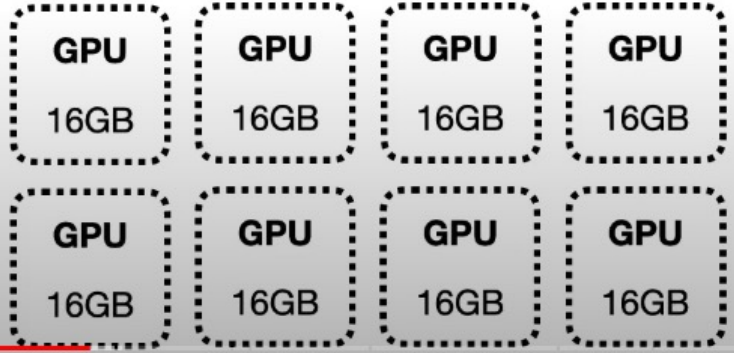
**10B Parameter Model**



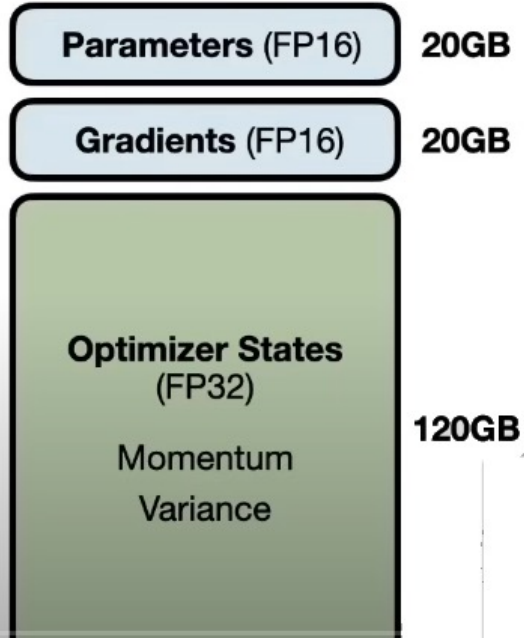
# The Problem

LLMs are (computationally) expensive



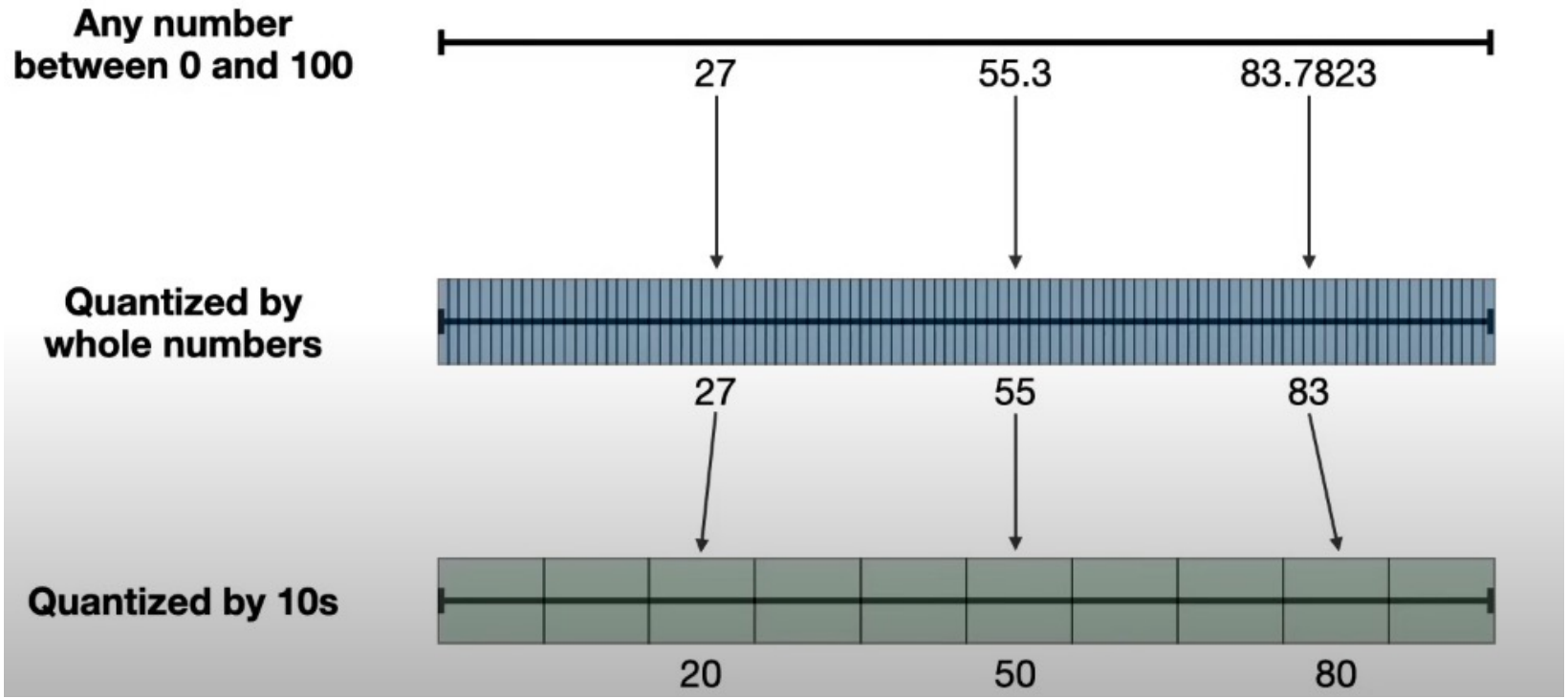




**10B Parameter Model = 160GB!**



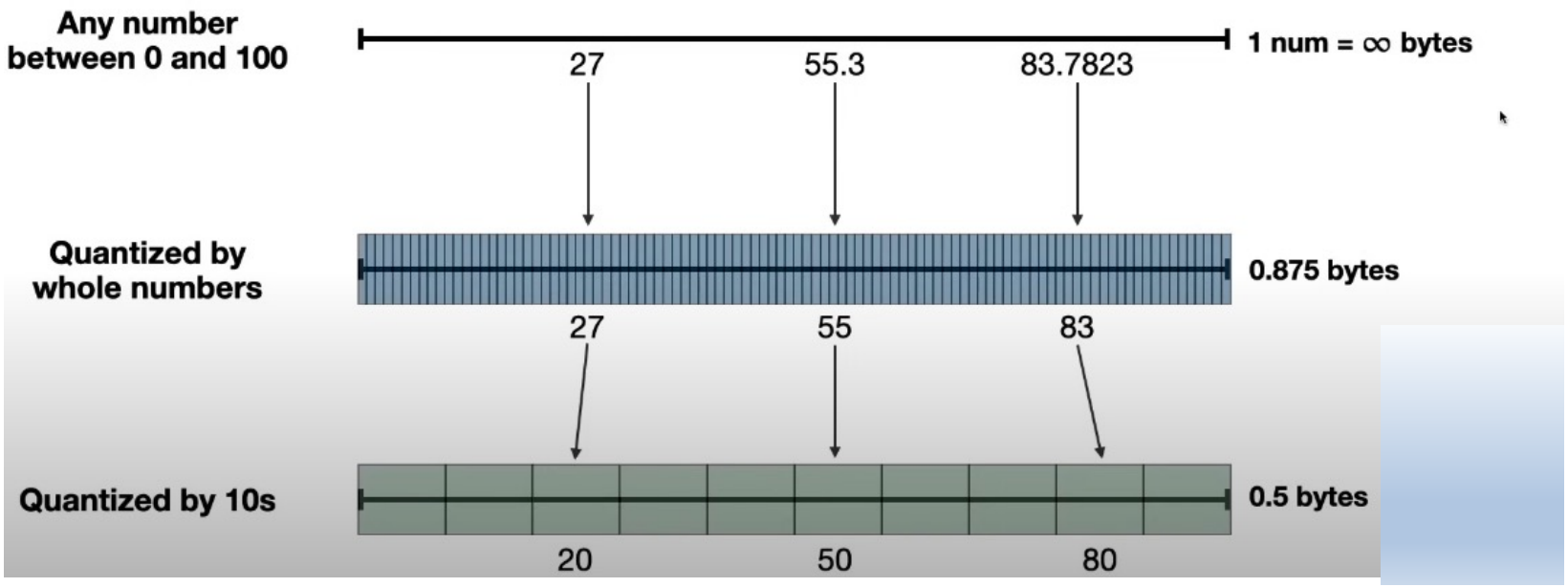
# What is Quantization?

Quantization = splitting range into buckets



# What is Quantization?

Quantization = splitting range into buckets



# Ingredient 1: 4-bit NormalFloat

A better way to bucket numbers

**4-bit** e.g. 0101

$\implies 2^4 = 16$ unique combinations

$\implies 16$ buckets for quantizations

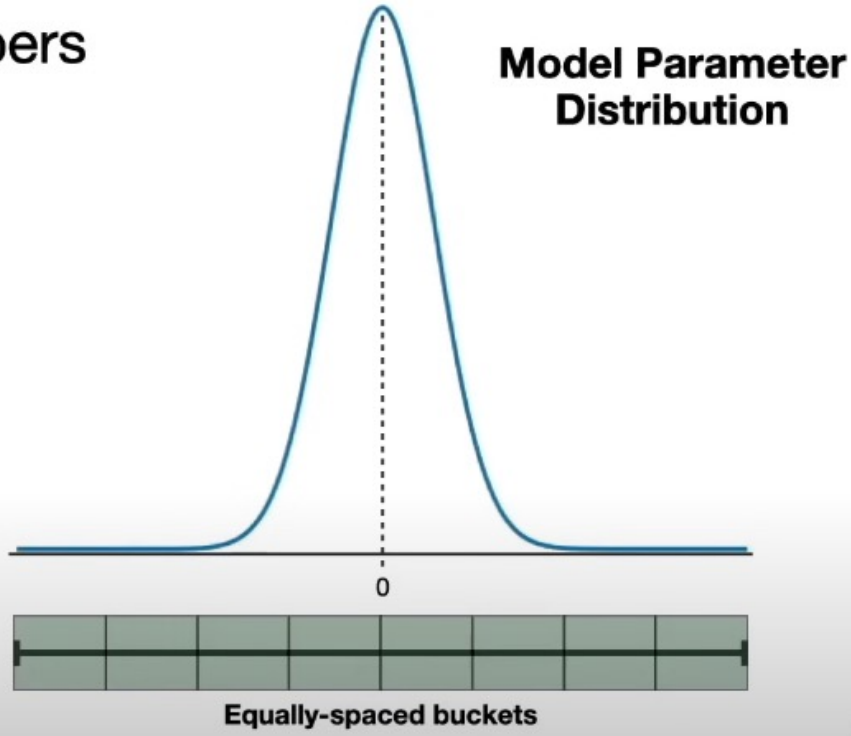
# Ingredient 1: 4-bit NormalFloat

A better way to bucket numbers

**4-bit** e.g. 0101

$\implies 2^4 = 16$ unique combinations

$\implies 16$ buckets for quantizations



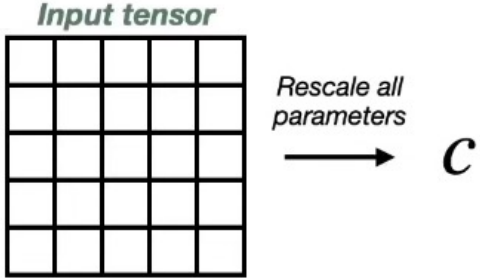
# Ingredient 2: Double Quantization

## Quantizing the Quantization Constants

$$X^{Int8} = \text{round}\left(\frac{127}{\text{absmax}(X^{FP32})} X^{FP32}\right)$$

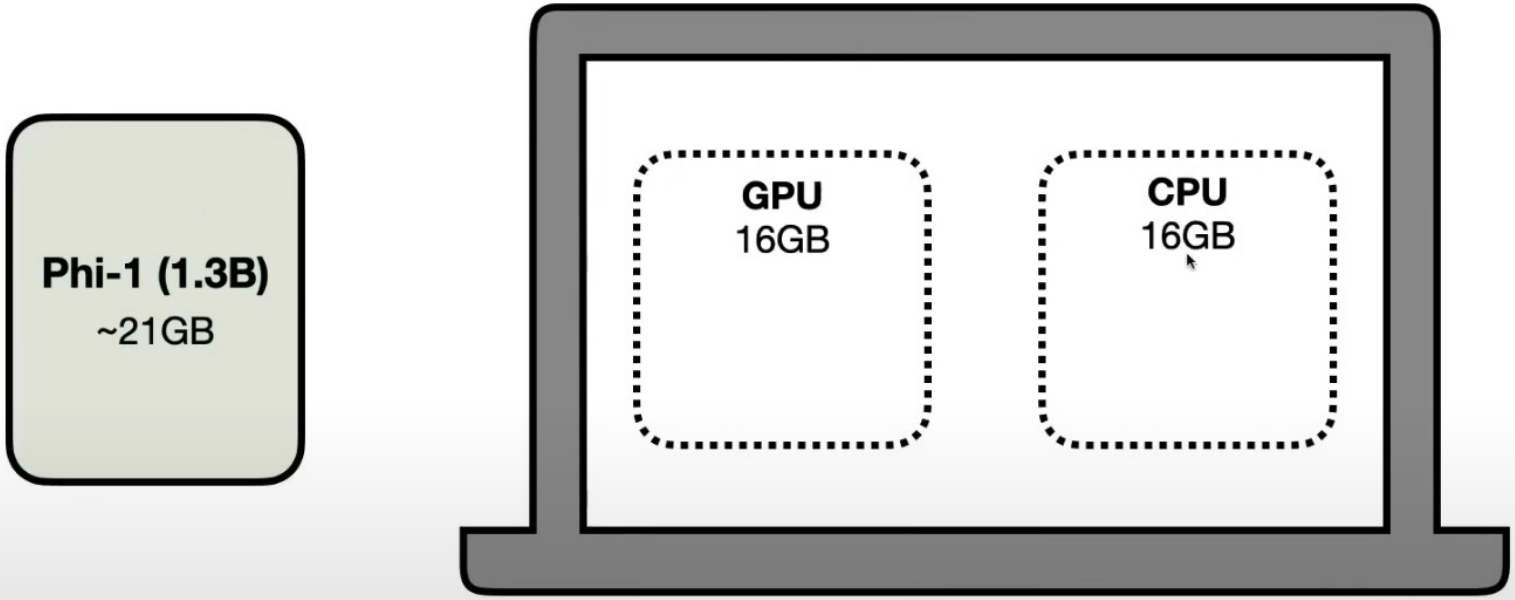
$$= \text{round}\left(c^{FP32} . X^{FP32}\right)$$

**Takes up precious memory**



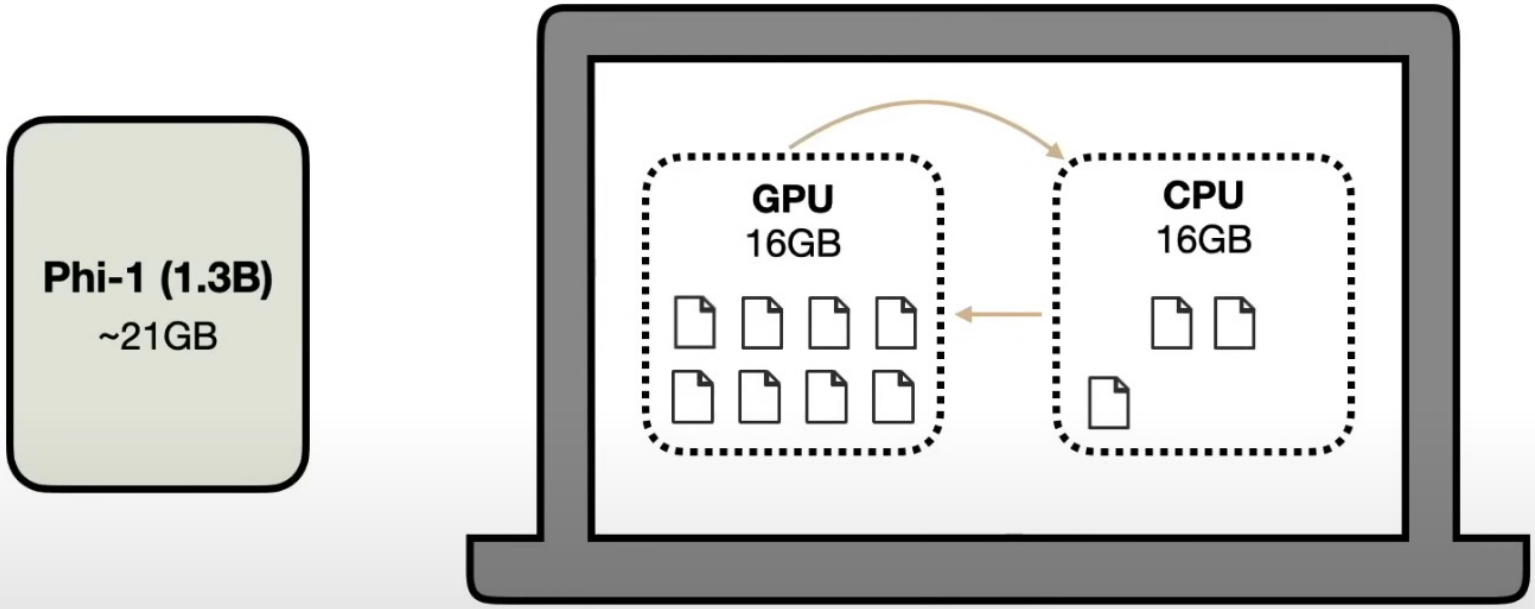
# Ingredient 3: Paged Optimizer

Looping in your CPU



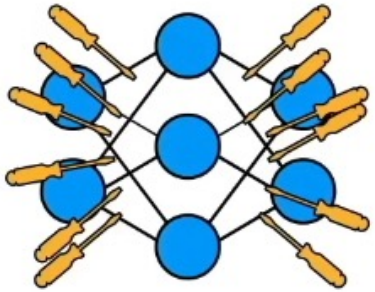
# Ingredient 3: Paged Optimizer

Looping in your CPU

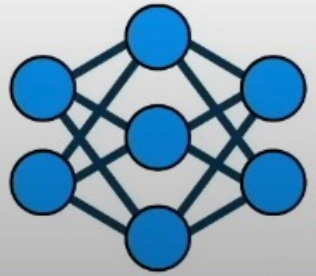


# Ingredient 4: LoRA

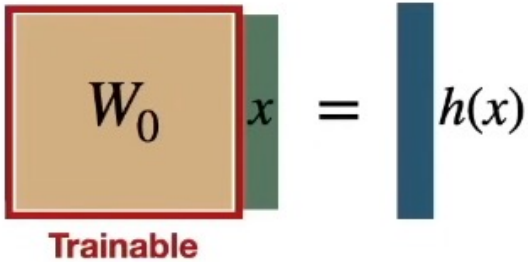
Fine-tunes model by adding **small set** of trainable parameters



$x \quad h(x) \quad y$

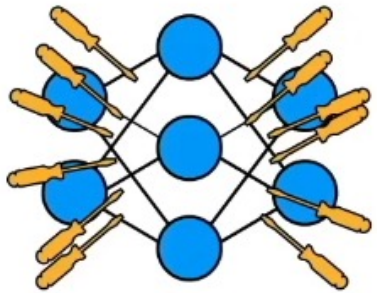


**Full Fine-tuning:** $h(x) = W_0 x$

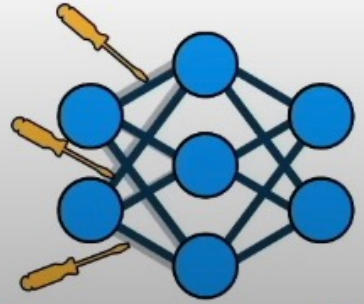


# Ingredient 4: LoRA

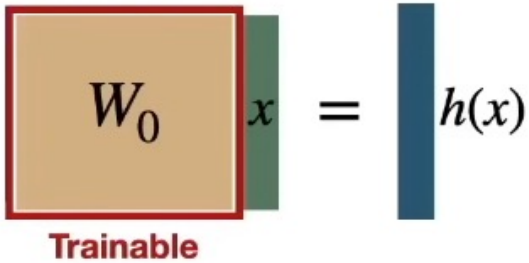
Fine-tunes model by adding **small set** of trainable parameters



$x \quad h(x) \quad y$



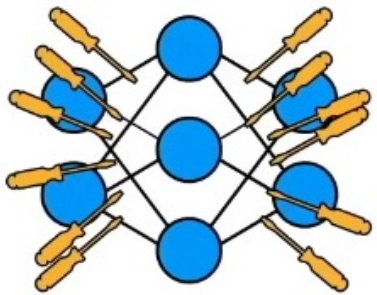
**Full Fine-tuning:** $h(x) = W_0 x$



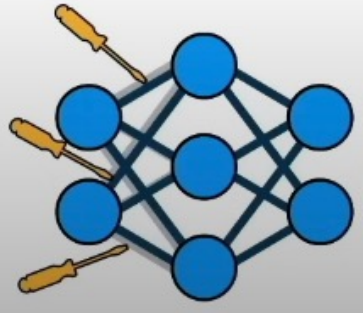

**LoRA:** $h(x) = W_0 x + \Delta W x = W_0 x + BAx$

# Ingredient 4: LoRA

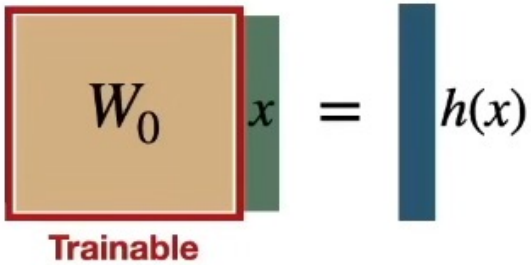
Fine-tunes model by adding **small set** of trainable parameters



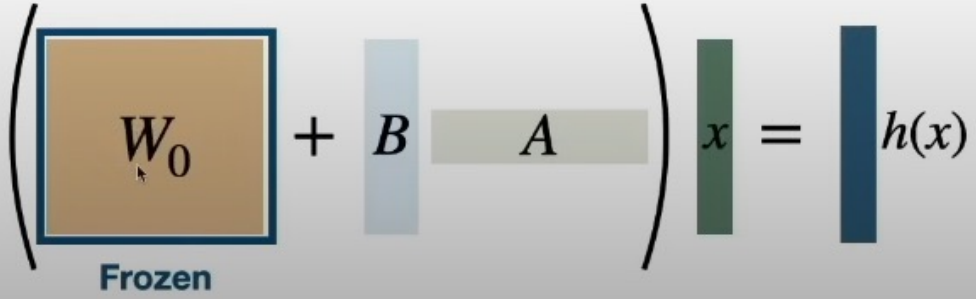
$x \quad h(x) \quad y$

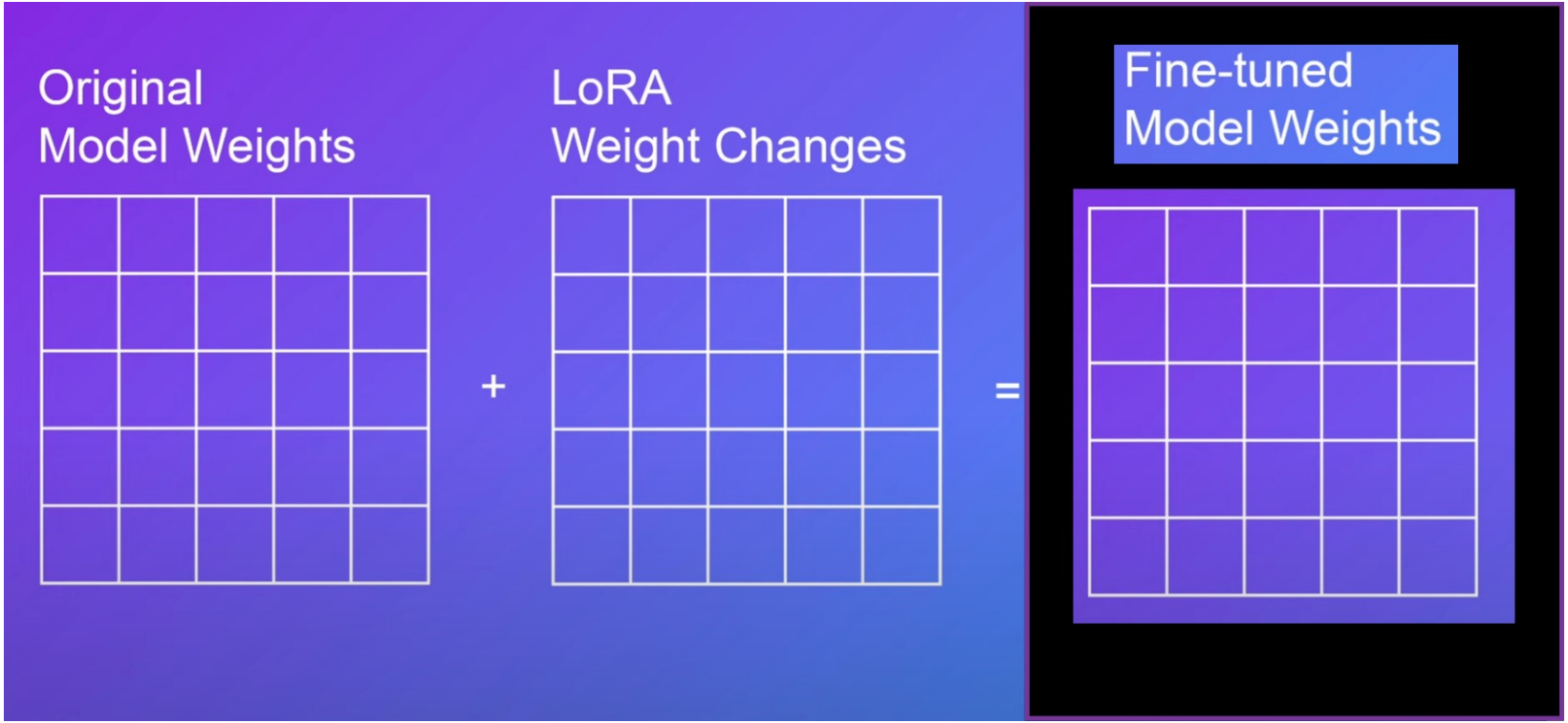


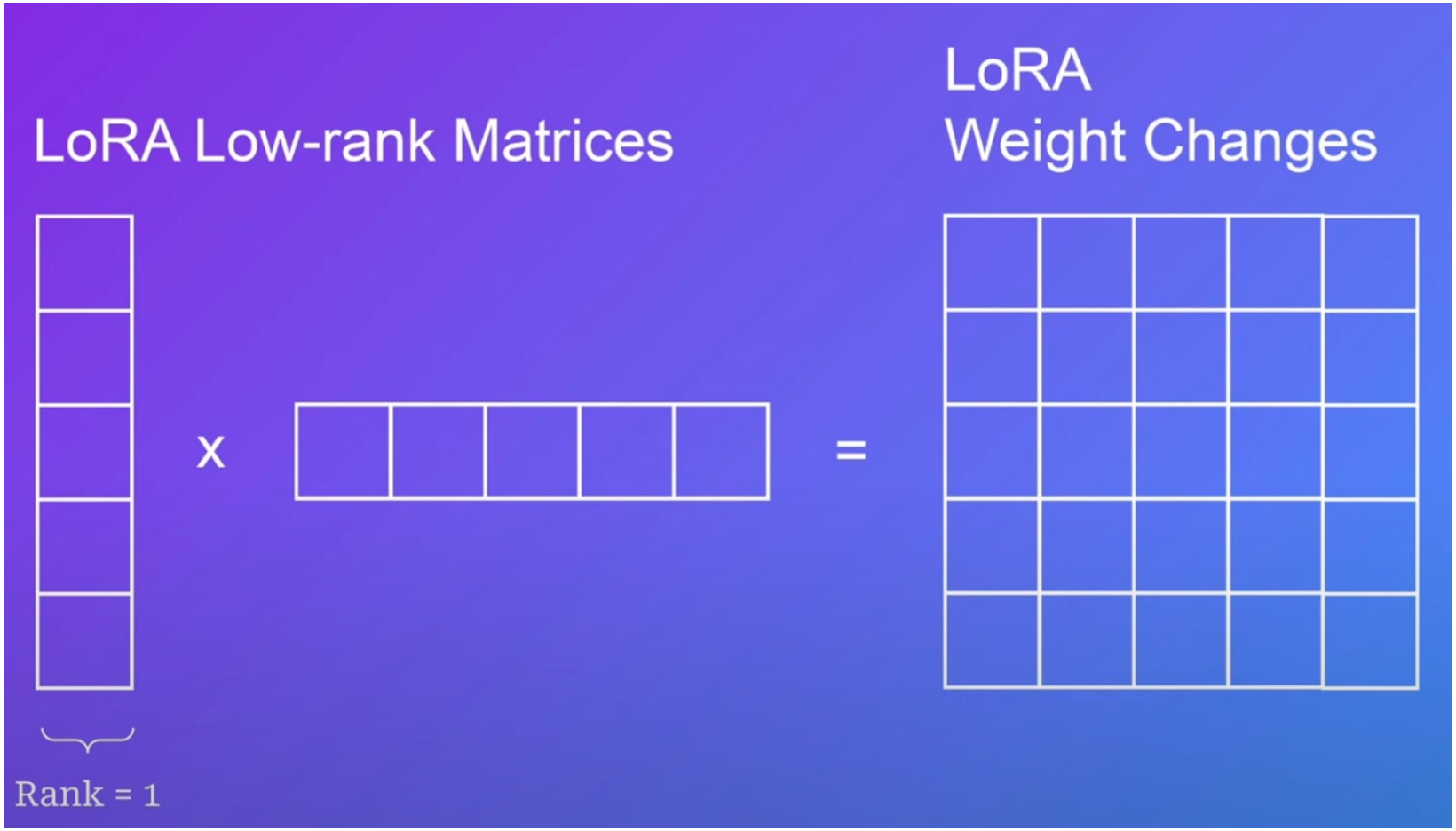
**Full Fine-tuning:** $h(x) = W_0 x$




**LoRA:** $h(x) = W_0 x + \Delta W x = W_0 x + BAx$

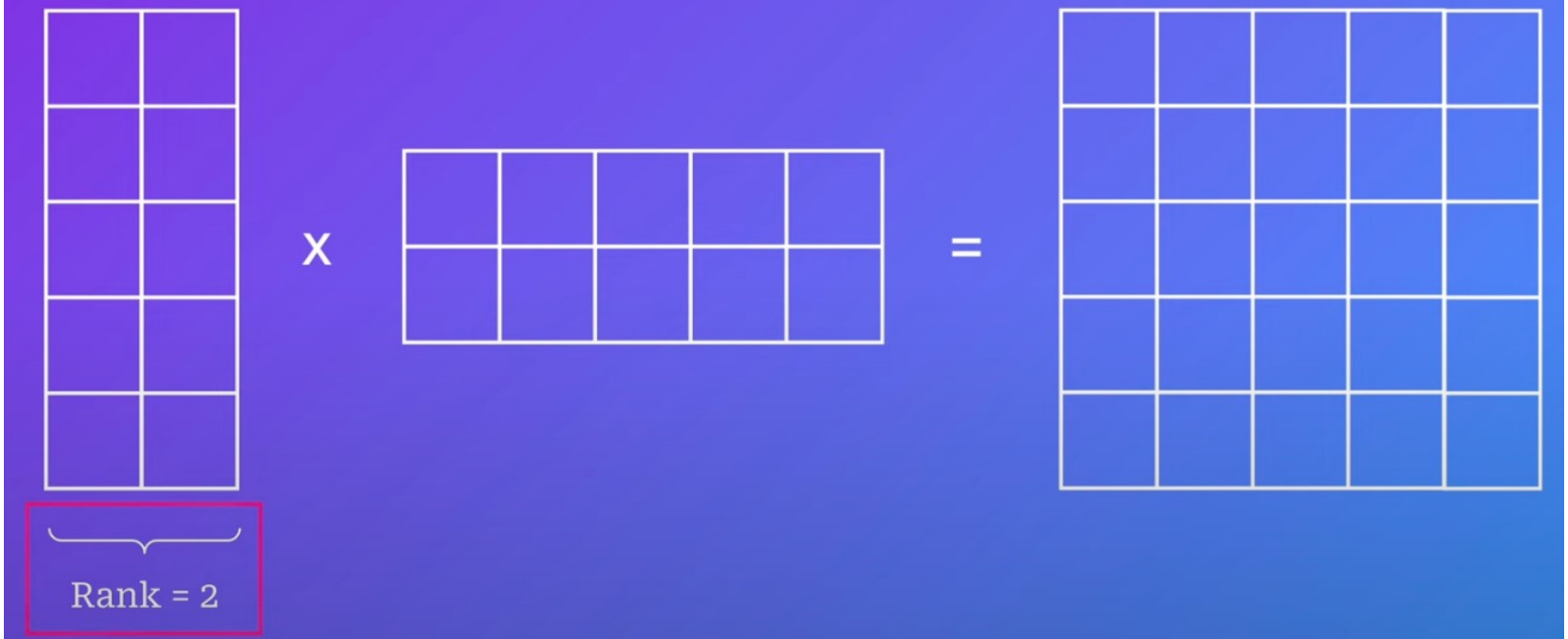



Original
Model Weights
+
LoRA
Weight Changes
=
Fine-tuned
Model Weights


LoRA Low-rank Matrices
LoRA
Weight Changes
x
=
Rank = 1

# Increasing Precision by Increasing Rank

LoRA Matrices, Rank 2



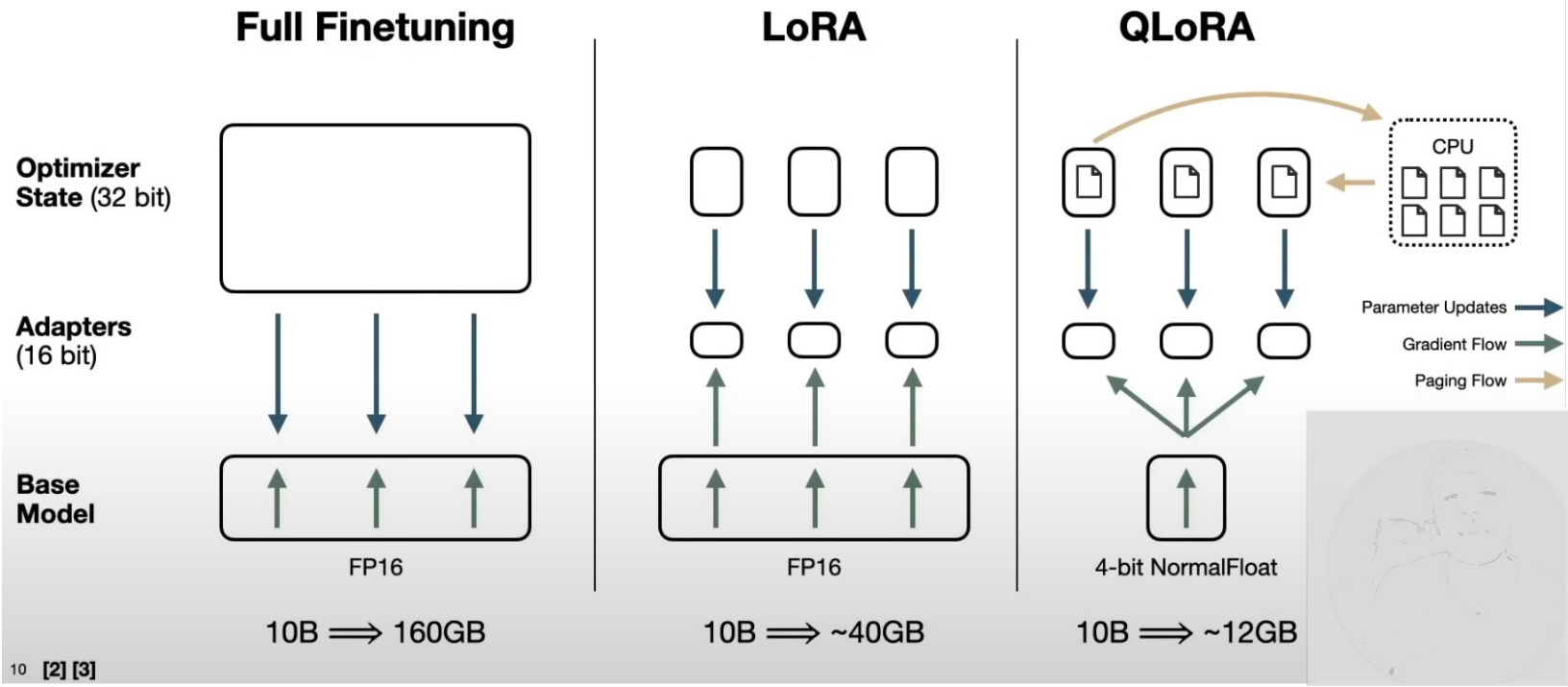

Higher Precision Weight Changes

## Number of Trainable Parameters

| Rank | 7B | 13B | 70B | 180B |
|---|---|---|---|---|
| 1 | 167k | 228k | 529k | 849k |
| 2 | 334k | 456k | 1M | 2M |
| 8 | 1M | 2M | 4M | 7M |
| 16 | 3M | 4M | 8M | 14M |
| 512 | 86M | 117M | 270M | 434M |
| 1,024 | 171M | 233M | 542M | 869M |
| 8,192 | 1.4B | 1.8B | 4.3B | 7.0B |

*In reality, LLMs are made up of multiple layers of differing sizes. This is a generalization as if the model were a single layer.*

# Bringing it all together




 [2] [3]