

Hot Chips Keynote

Noam Shazeer, Google
August 2025

**Thank you
for the
supercomputers!**

Hot Chips Keynote - Noam Shazeer

- What LLMs want
- My LLM Journey: 2015-2025
- What LLMs want from Hardware
- Q&A

What LLMs Want

Logarithmic? gains from each of:

- Computation (multiplies, adds, etc.)

What LLMs Want

Logarithmic? gains from each of:

- Computation (multiplies, adds, etc.)
- Parameters

What LLMs Want

Logarithmic? gains from each of:

- Computation (multiplies, adds, etc.)
- Parameters
- Depth
- Nonlinearities?

What LLMs Want

Logarithmic? gains from each of:

- Computation (multiplies, adds, etc.)
- Parameters
- Depth
- Nonlinearities?
- Information Flow (sequence-wise, depthwise, etc)

What LLMs Want

Logarithmic? gains from each of:

- Computation (multiplies, adds, etc.)
- Parameters
- Depth
- Nonlinearities?
- Information Flow (sequence-wise, depthwise, etc)
- Training Examples
- Unique Training Examples

2015: Got Hooked on Language Modeling

Exploring the Limits of Language Modeling

Rafal Jozefowicz

Oriol Vinyals

Mike Schuster

Noam Shazeer

Yonghui Wu

RAFALJ@GOOGLE.COM

VINYALS@GOOGLE.COM

SCHUSTER@GOOGLE.COM

NOAM@GOOGLE.COM

YONGHUI@GOOGLE.COM

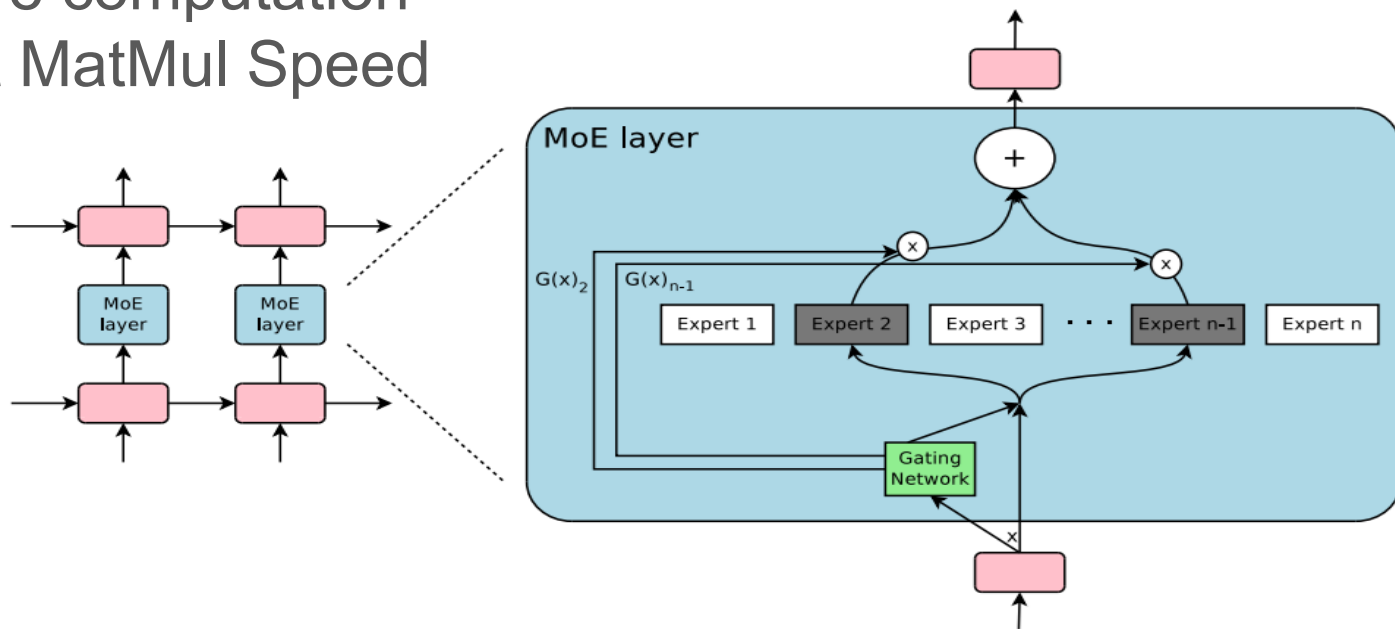
LSTM-512-512	54.1
LSTM-1024-512	48.2
LSTM-2048-512	43.7
LSTM-8192-2048 (NO DROPOUT)	37.9
LSTM-8192-2048 (50% DROPOUT)	32.2
2-LAYER LSTM-8192-1024 (BIG LSTM)	30.6
BIG LSTM+CNN INPUTS	30.0

Language Modeling - Best Problem Ever

- Text = Distilled Abstract Thought
 - A picture is only worth 1000 words
- Simple Problem Statement:
 - Produce a probability distribution on next token
- Abundant training data
- AI Complete

2016: Sparsely-Gated Mixture of Experts (MoE)

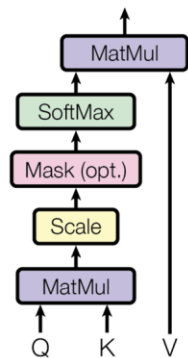
- More Parameters
- Not more computation
- Runs at MatMul Speed



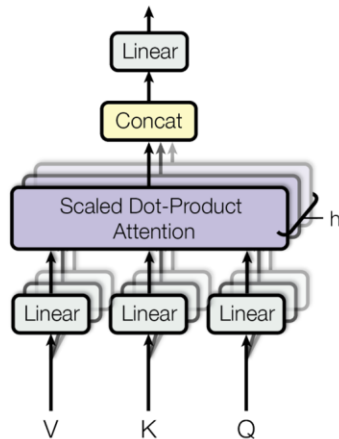
2017: Transformer (self-attention)

- Better sequence-wise information flow
- Runs at MatMul Speed

Scaled Dot-Product Attention



Multi-Head Attention



2018: Supercomputers + SPMD -> LLMs



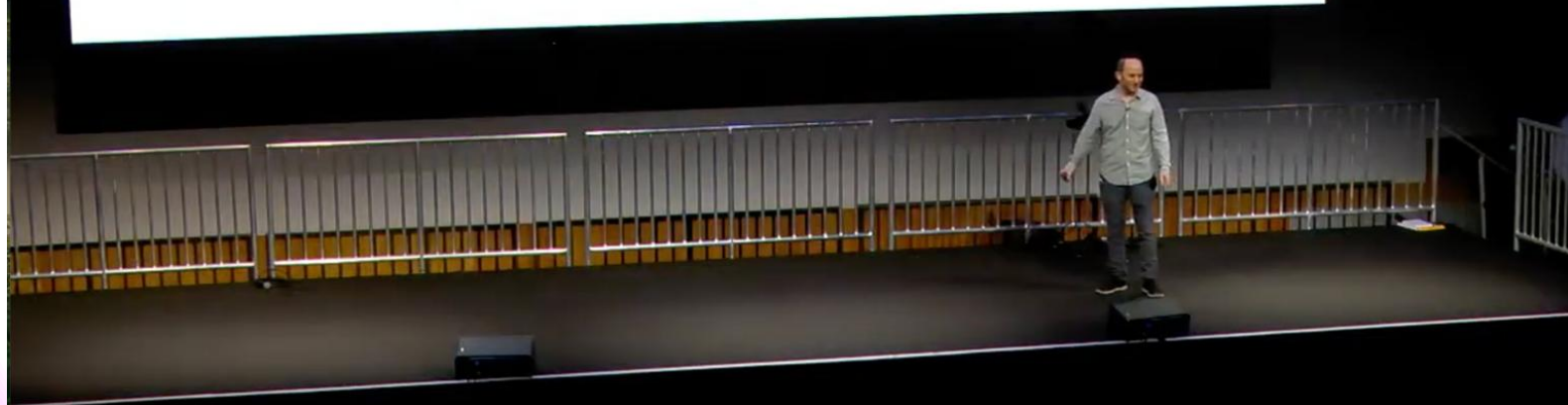
Mesh-TensorFlow: Deep Learning for Supercomputers

Noam Shazeer, Youlong Cheng, Niki Parmar,
Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee
Mingsheng Hong, Cliff Young, Ryan Sepassi, Blake Hechtman
Google Brain
{noam, ylc, nikip, trandustin, avaswani, penporn, phawkins,
hyouklee, hongm, cliffy, rsepassi, blakehechtman}@google.com

Flashback to 2018

Neural Language Models: **Bigger is Better**

Noam Shazeer, Google AI
September 2018



Samples from “Transformer” LMs.

- Transformer Language Models trained on text of Wikipedia
- Random sampling, starting from title “Abraham Lincoln”
- True “Abraham Lincoln” article was in the held-out set
- Mildly cherry-picked

“Abraham Lincoln” (33M parameters)

Abraham Lincoln was an American voodoo activist and former alcoholic who liked to have been the chief for a heroic, misguided he attended—never to represent him. He was later noted as, "one of the most triumphal scholars of the 20th century."

Lincoln's essay "Crypt of the Dark", titled "Prospecting his. The participation of icons in fusion with modern regime, is neglect", was published in 1976. Over the course of his life, Lincoln lived. Ford, who was a Loyalist, was a translator of Russian Orthodox Jewish sources, his book "Ami el repertory" which began in substitution in the 1950s.

===Study===

Lincoln was a professor at the University of Michigan and a U.S. Special ...

“Abraham Lincoln” (5B parameters)

Abraham Lincoln (March 1, 1809{{spaced ndash}}March 15, 1865) was an American lawyer, jurist, and politician. He served as the 19th President of the United States, from March 1861 until his assassination in April 1865.

Born in Newark, New Jersey, Lincoln attended Knox College of New York, studied law, and was admitted to the bar Missouri Territory in 1826. He relocated to Illinois, and was elected to the Illinois House of Representatives in 1827, 1822, 1823, and 1826. He was elected Speaker of the Illinois House in 1825, served as Attorney General of the Illinois Territory in 1829, and was elected to the Illinois Senate in 1831. Lincoln was elected to the United States House of Representatives as a Democrat in 1833, and he was elected as a representative in the United States House of Representatives, serving two terms from 1839 to 1843. He was elected as a Whig in 1842, but he only served one term in Congress, and returned to the House of Representatives, serving ...

Why Small Neural Language Models

- Generalization
- Speed
- Memory

Something is wrong with you.

- ~~Generalization~~

Your training set is too small.

- ~~Speed~~

Your computer is too slow.

- ~~Memory~~

You don't have Mesh-TensorFlow.

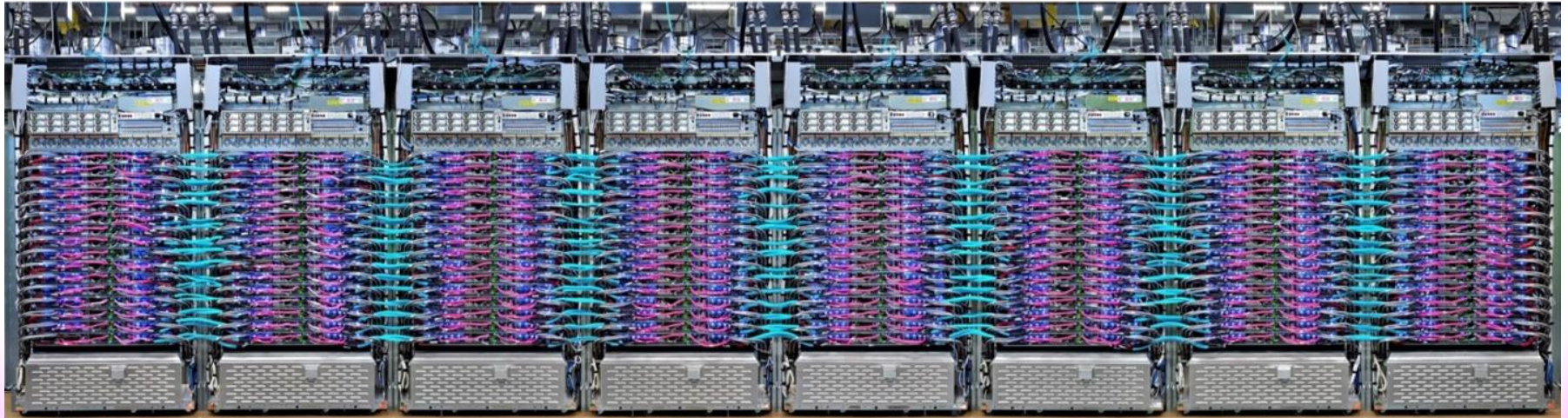
Your training set is too small.

- Trillions of words of text available
(Common Crawl, etc)
- Generative pre-training, then fine-tune on many tasks
(Radford et al, OpenAI)

Your computer is too slow.

- Training cost is quadratic in #parameters:
(more operations/token) * (more training to fill brain)
- 10^9 params * 10^9 training tokens * 6 ops / 1 day
= 10^{14} ops/s (1 cloud TPUv2 @ \$1.35/hr)
- 10^{11} params * 10^{11} training tokens * 6 ops / 1 week
= 10^{17} ops/s (need a supercomputer)

So build a supercomputer



Memory Constraints

- Multi-Billion-Parameter models must be split between many processors. (model-parallelism)
- Doing this efficiently is very tricky

Mesh-TensorFlow

Model-parallelism, using what Synchronous Data-parallelism does well.

- Every operation and every Tensor is evenly split and/or replicated across all processors.
- Same “program” on every core
- Collective communication (e.g. allreduce)

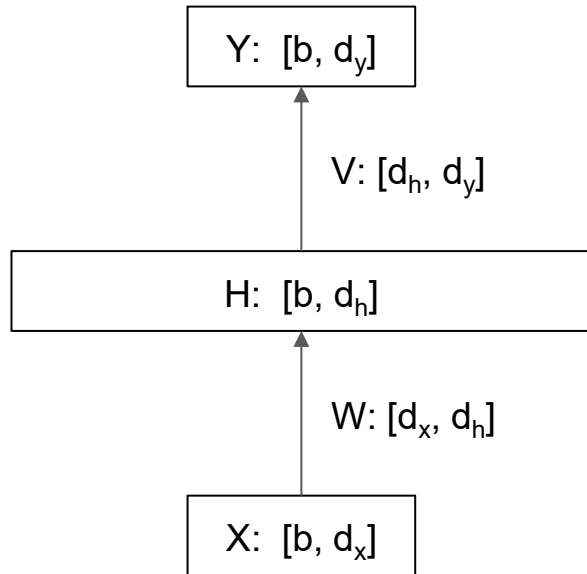
Mesh-TensorFlow

Model-parallelism, using what Synchronous Data-parallelism does well.

- User defines which dimensions to split.
- Data-Parallelism - split “batch” dimension.
 - Tensors with a “batch” dimension (activations, etc) are split.
 - Tensors with no “batch” dimension (parameters) are replicated.
- Model-Parallelism - split some layer depths.

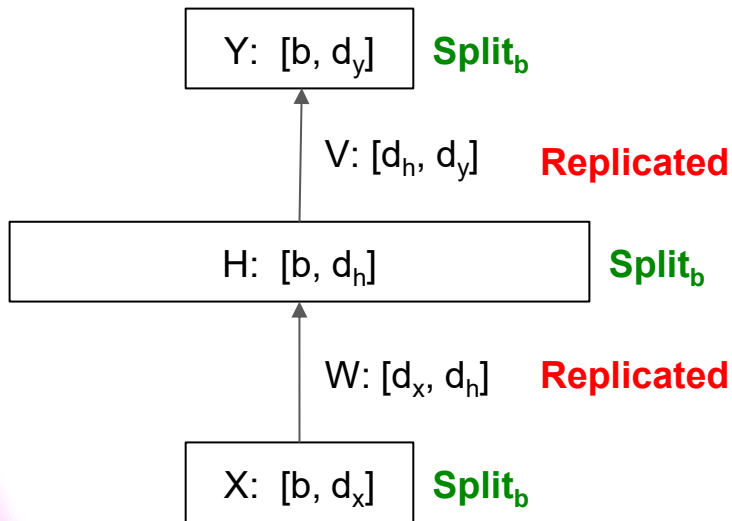
Example Perceptron

$$Y = (H = \text{Relu}(XW))V$$

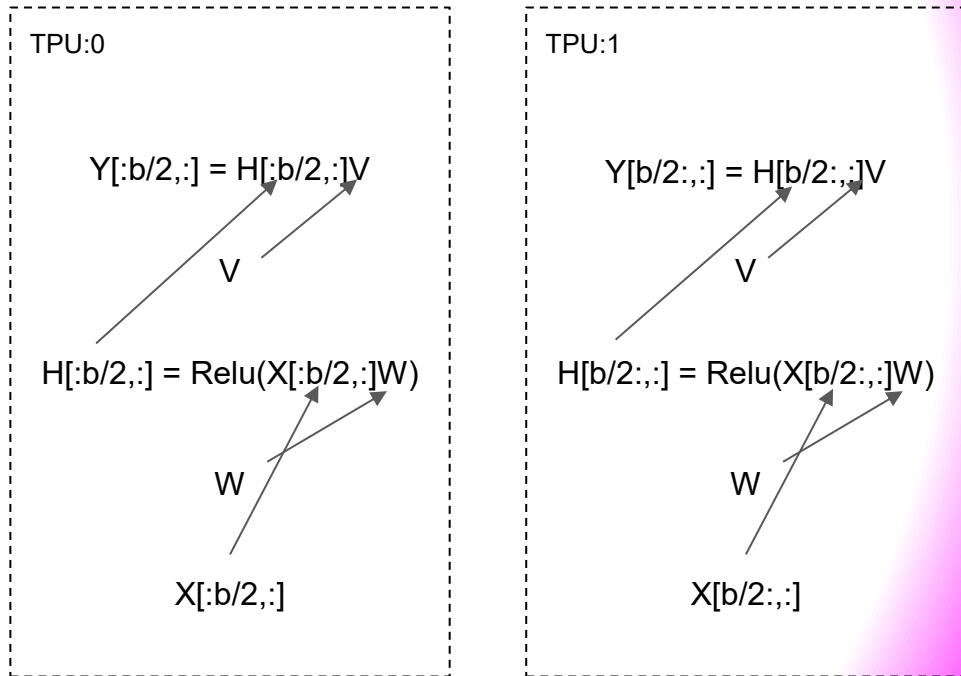


Example Perceptron

$$Y = (H = \text{Relu}(XW_1))V$$

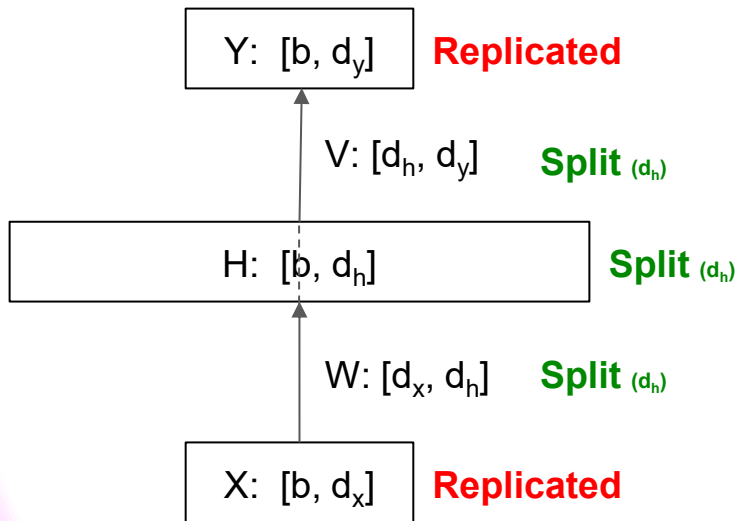


Data-Parallelism: Split dimension “b”

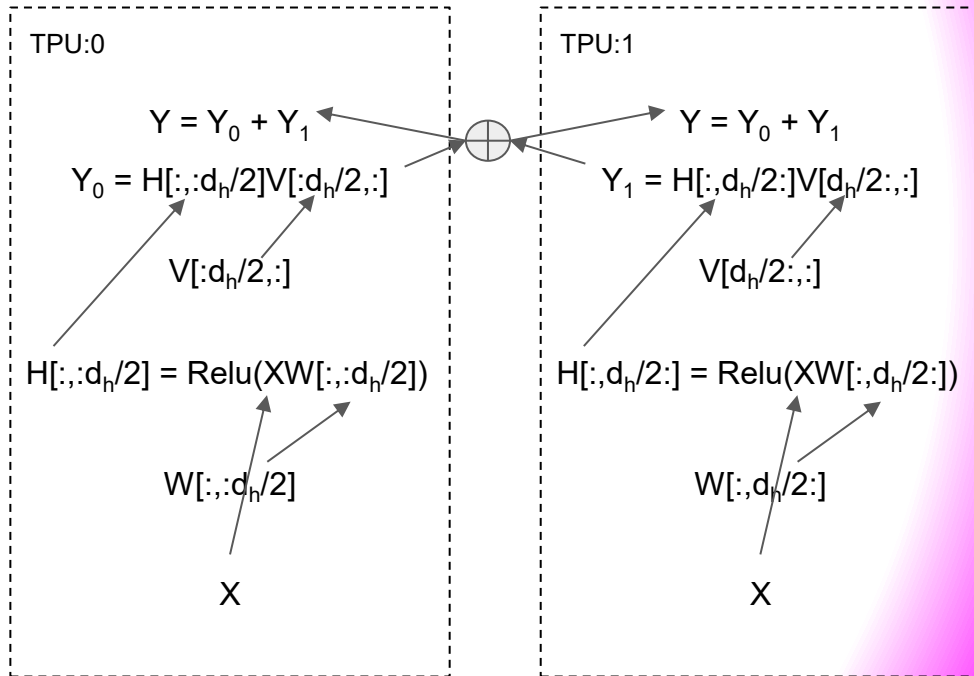


Example Perceptron

$$Y = (H = \text{Relu}(XW_1))W_2$$

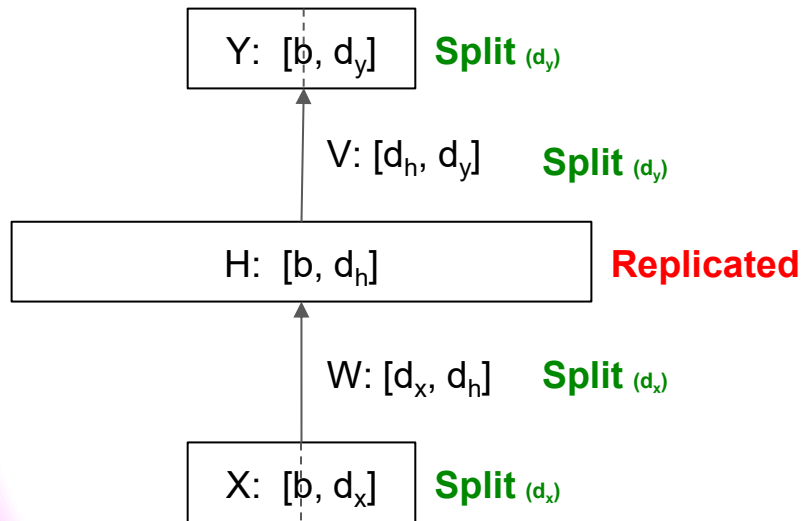


Model-Parallelism: Split dimension " d_h "

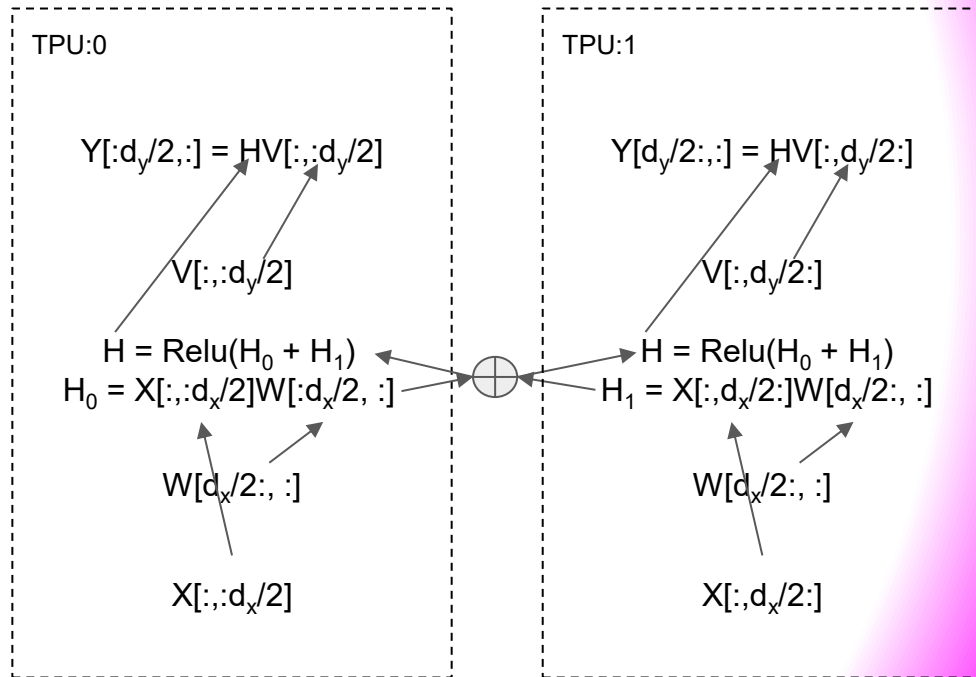


Example Perceptron

$$Y = (H = \text{Relu}(XW_1))W_2$$



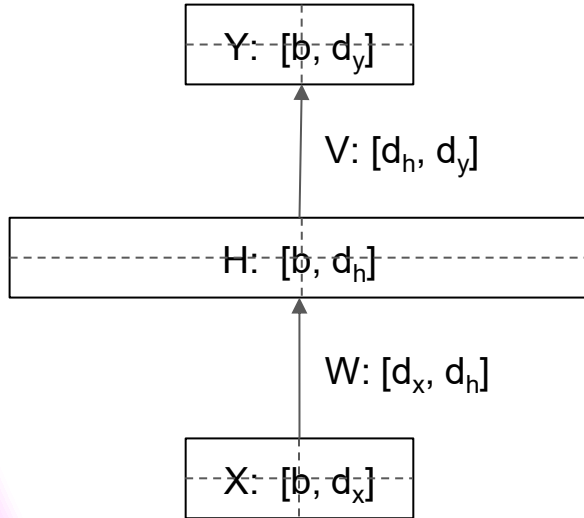
Model-Parallelism: Split dimensions d_x, d_y



**GET READY FOR
SOME SERIOUS
SLICING AND
DICING**

Example Perceptron

$$Y = (H = \text{Relu}(XW_1))W_2$$



Use a 3-dimensional mesh of processors with dimensions $[m_0, m_1, m_2]$.

- Split “ b ” across mesh dimension m_0
- Split “ d_h ” across mesh dimension m_1
- Split “ d_x ”, “ d_y ” across mesh dimension m_2
- Each matrix is tiled across two mesh dimensions and replicated across the third.
- Each matmul consists of local matmuls, followed by allreduce across one mesh dimension.

Mesh-TF Transformer Model

- Transformer models working in MeshTF

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

- Example layouts for “Transformer” model

For a 2D Mesh, split across dimensions
 (“batch”, “heads”, “vocab”, “d_ff”)

```
layout_data_parallel="batch:m0"
```

```
layout_model_parallel="vocab:m0,heads:m0,d_ff:m0"
```

```
layout_dp_mp="batch:m0,vocab:m1,heads:m1,d_ff:m1"
```

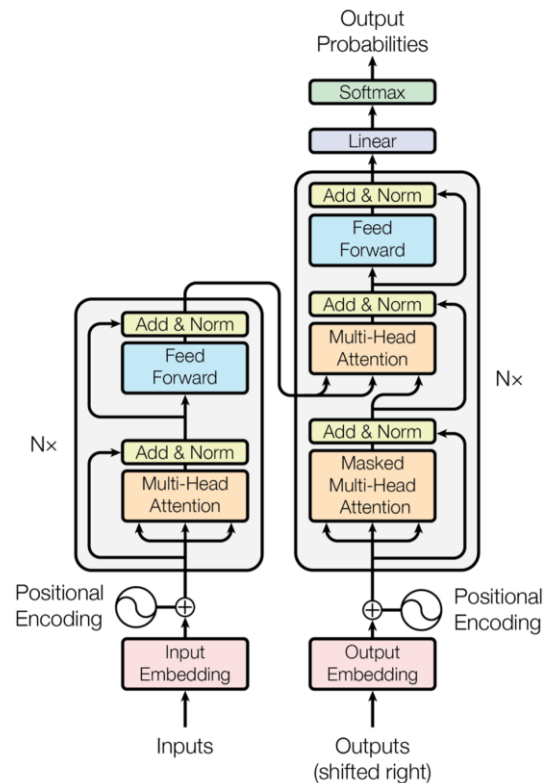


Figure 1: The Transformer - model architecture.

Mesh-TensorFlow Language

- “mesh”: n-dimensional array of processors.
- TF-like computational graph with named tensor-dimensions
- “Layout” specifies which tensor-dimensions are split across which mesh-dimensions.
- Python API, Compiles into TensorFlow

Flash-forward to 2025

I Told You So

Since 2018, I have been busy with..

- Incremental Inference Performance
 - Multiquery Attention
 - Speculative decoding
- Misc. projects on architecture / training / thinking, etc.
- General LLM Applications
 - T5, Meena/LaMDA
 - Character.AI
- Back at **Gemini** as Co-TL since 2024

What LLMs Want

Logarithmic? gains from each of:

- Computation (multiplies, adds, etc.)
- Parameters
- Depth
- Nonlinearities?
- Information Flow (sequence-wise, depthwise, etc)
- Training Examples
- Unique Training Examples

What LLMs Want from Hardware

- Logarithmic? gains from each of:
 - PFLOPs /\$

What LLMs Want from Hardware

- Logarithmic? gains from each of:
 - PFLOPs
 - Memory Capacity (all levels of hierarchy)
 - Memory Bandwidth (all levels of hierarchy)

What LLMs Want from Hardware

- Logarithmic? gains from each of:
 - PFLOPs
 - Memory Capacity (all levels of hierarchy)
 - Memory Bandwidth (all levels of hierarchy)
 - Network Bandwidth (all levels of hierarchy)

What LLMs Want from Hardware

- Logarithmic? gains from each of:
 - PFLOPs
 - Memory Capacity (all levels of hierarchy)
 - Memory Bandwidth (all levels of hierarchy)
 - Network Bandwidth (all levels of hierarchy)
- Low precision - often OK

What LLMs Want from Hardware

- Logarithmic? gains from each of:
 - PFLOPs
 - Memory Capacity (all levels of hierarchy)
 - Memory Bandwidth (all levels of hierarchy)
 - Network Bandwidth (all levels of hierarchy)
- Low precision - often OK
- Determinism - REALLY nice to have

**Thank you
for the
supercomputers!**

Q & A