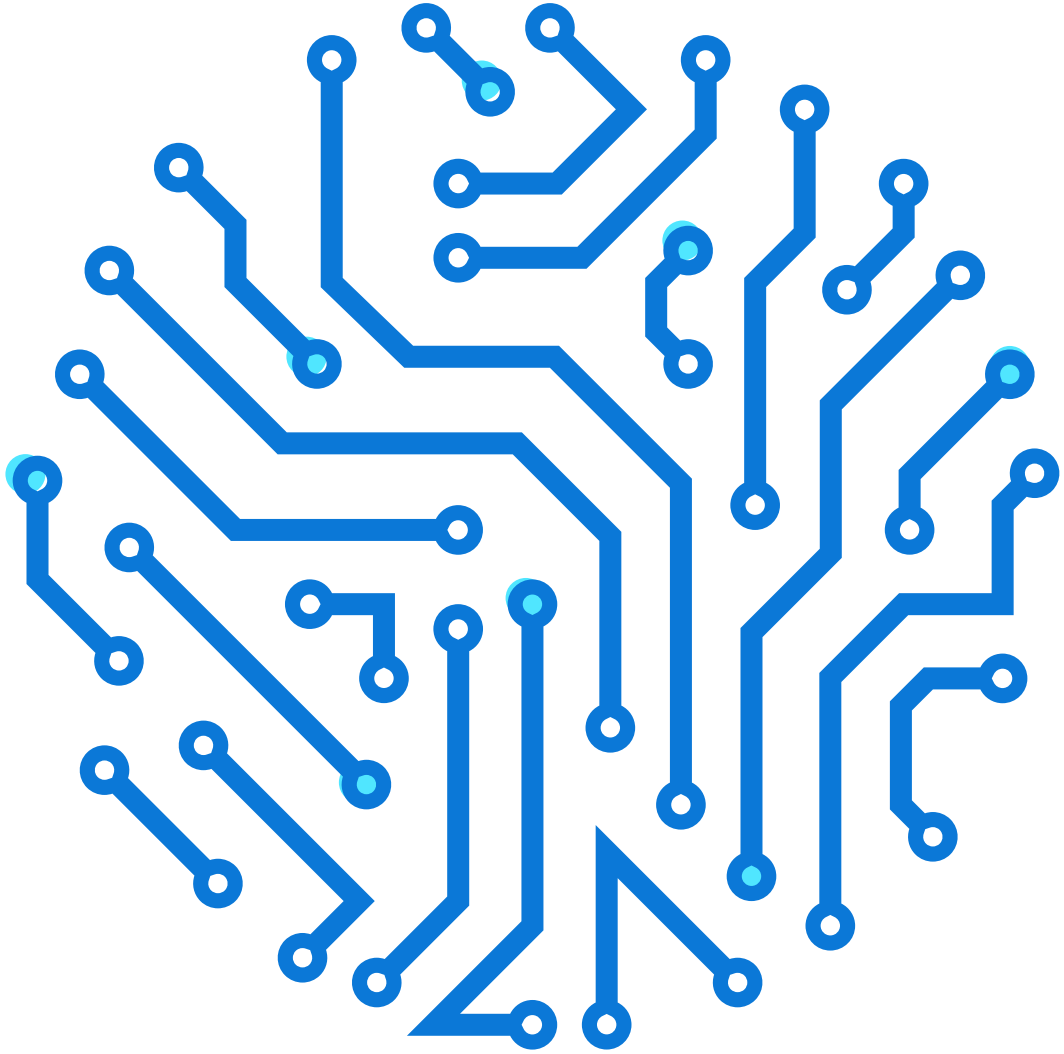
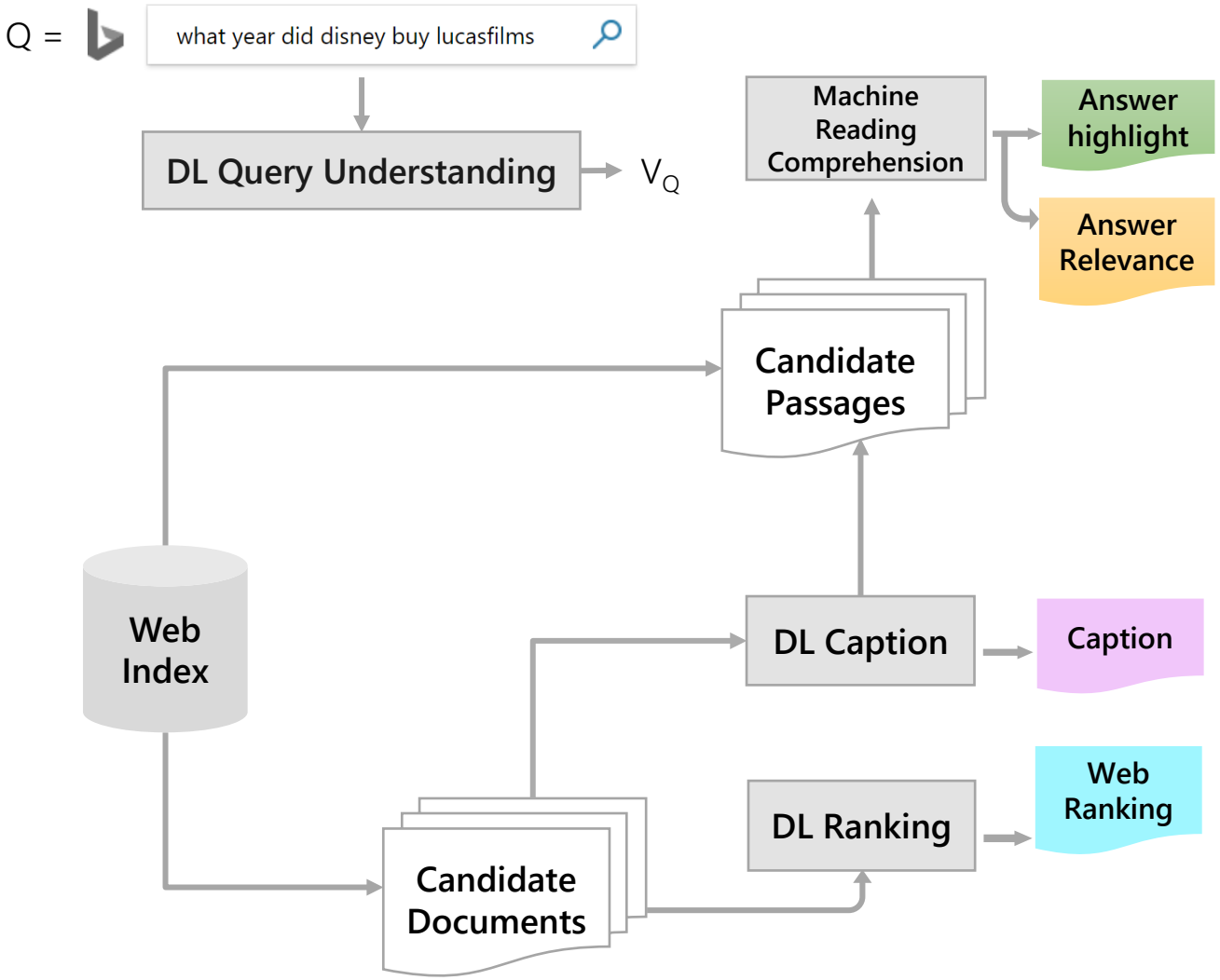
Microsoft Azure

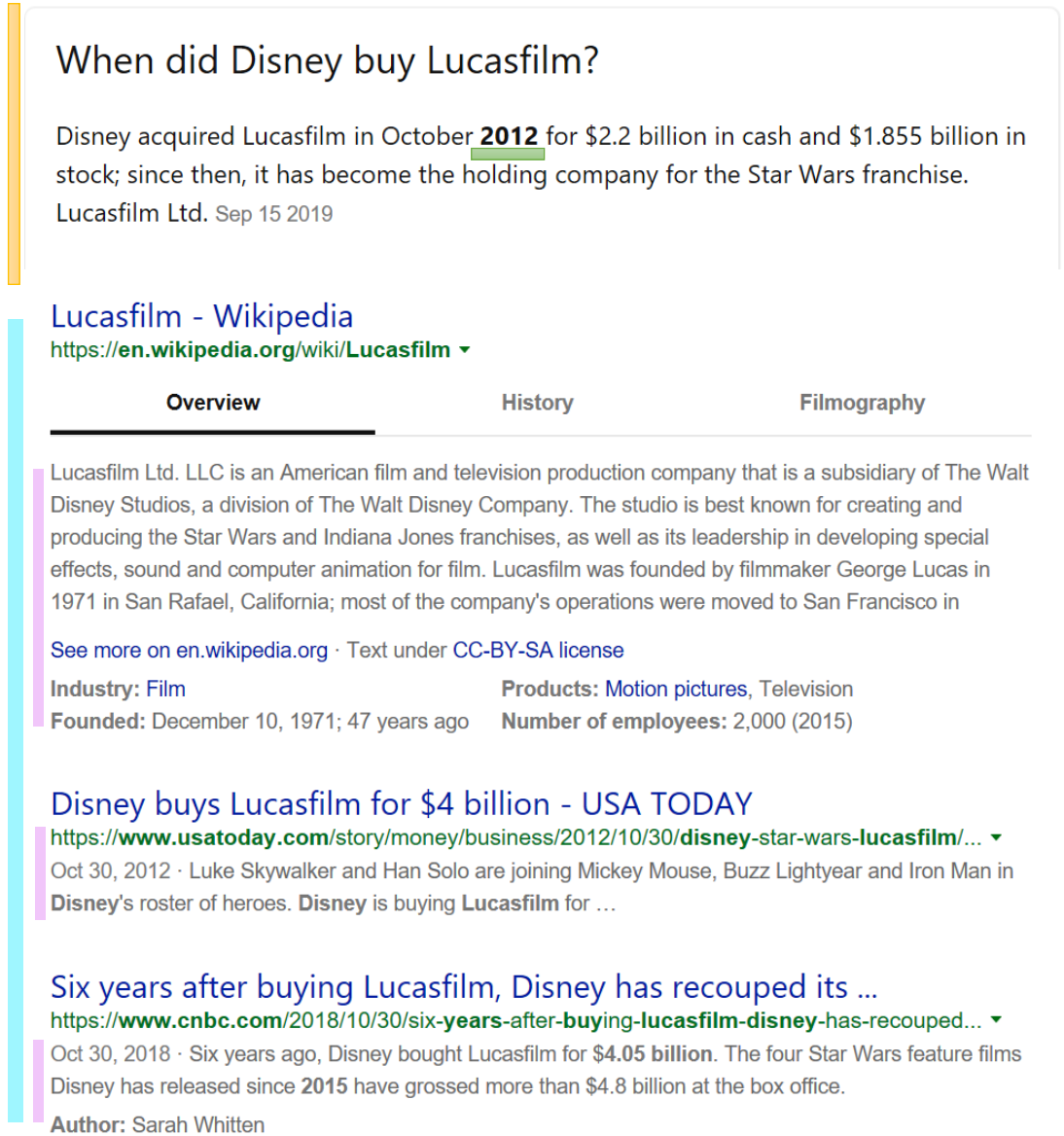
Improving Bing Web Search using Azure GPUs

Jeffrey Zhu, Program Manager, Bing
Mingqin Li, Engineering Manager, Bing



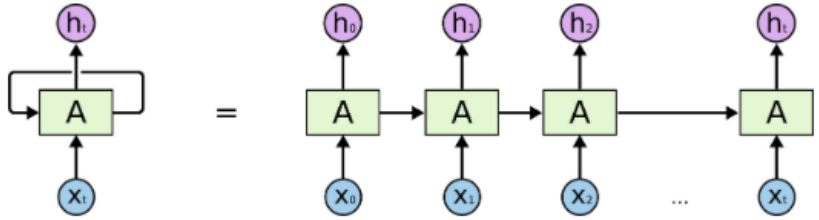
Bing is powered by Deep Learning



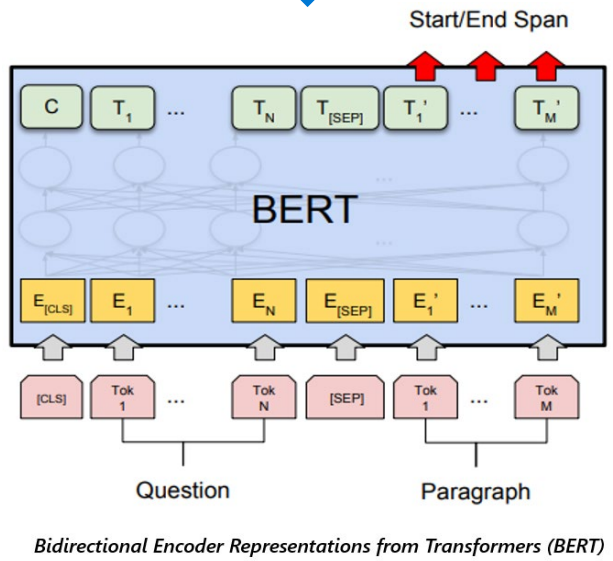


Trends in Natural Language Processing

- Most natural language processing (NLP) models were based on RNN and LSTM
  - Processes one word at a time in one direction
  - Loss of relevant information in long sentences



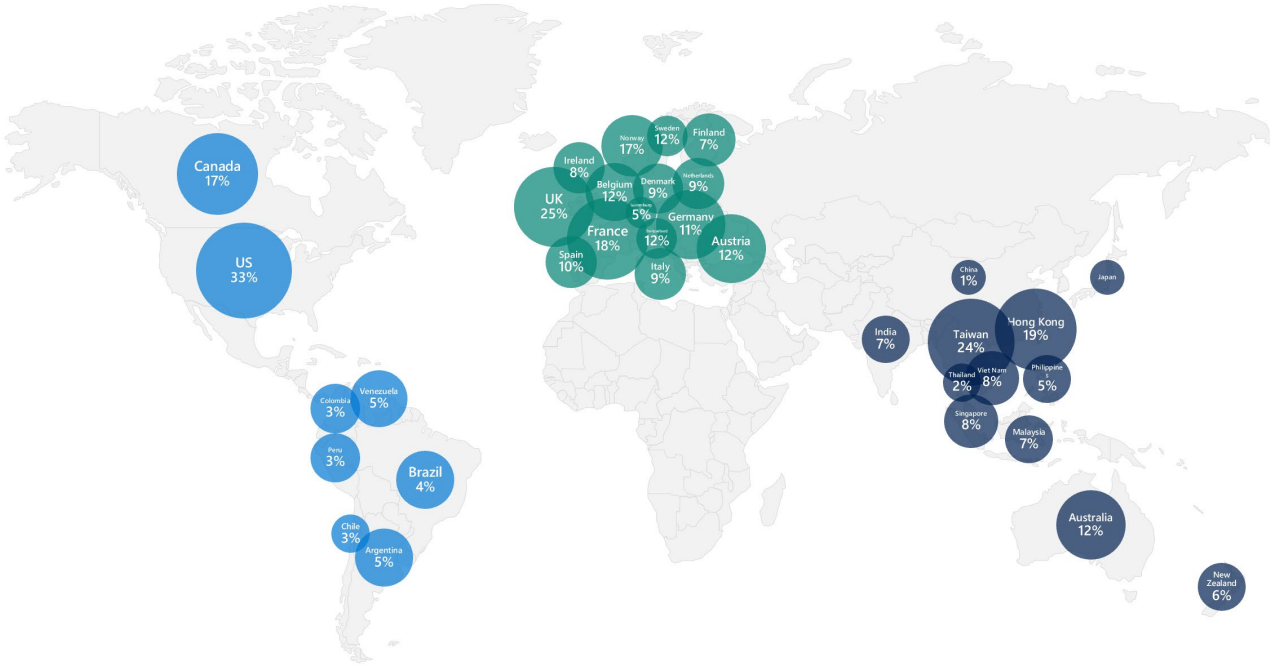
- Latest NLP models focus on large pre-trained transformer models (ex. BERT)
  - Processes each word in context with all other words at the same time
  - Highly parallelizable but significantly larger parameter size (100s of millions) and computation
  - Powered the largest search quality improvement in Bing this past year



Bidirectional Encoder Representations from Transformers (BERT)

Web Search Online Inference Challenges

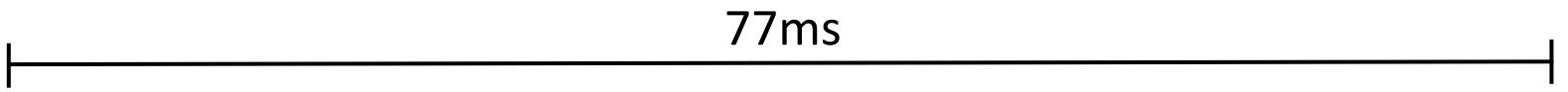
	Challenges
Performance	Large models with 100s of millions parameters need to run in single digit milliseconds
Scalability	Millions of inferences per second over hundreds of billions documents worldwide
Agility	Experiment and provision different hardware SKU at production scale within an hour



Case Study: BERT Model Optimization

Original Model:

3–layer BERT on CPU:



Case Study: BERT Model Optimization

Original Model:
3–layer BERT on CPU:

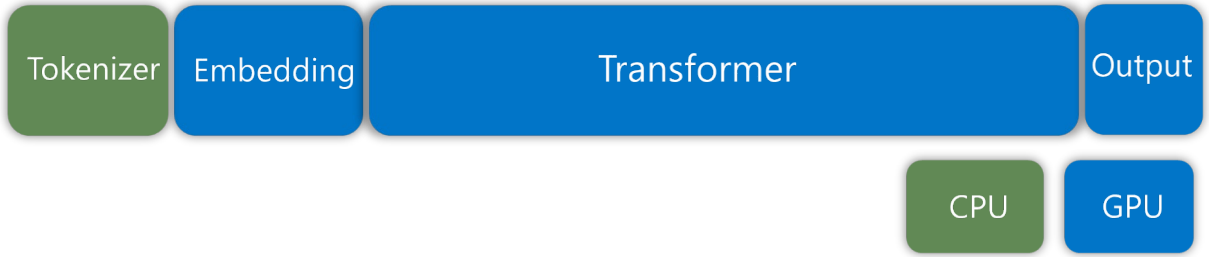
77ms

Hardware Acceleration
Model on **M60 GPU** VM:

20ms

System Optimization

- Hardware acceleration alone is not enough, system optimization at the software level is required
- GPU optimization by using TensorRT and CUDA/CUBLAS libraries
  - Operator fusion
  - Parallel execution
  - Mixed precision from Tensor Core on selected GPU
  - Partnership with NVidia
- CPU optimization on tokenizer
  - Parallel execution



Case Study: BERT Model Optimization

Original Model:
3–layer BERT on CPU:

77ms

Hardware Acceleration
Model on **M60 GPU** NV6 VM:

20ms

System Optimization
Model on **M60 GPU** NV6 VM:

9ms

Batch Size 4

Operator fusion and parallel execution
Same accuracy

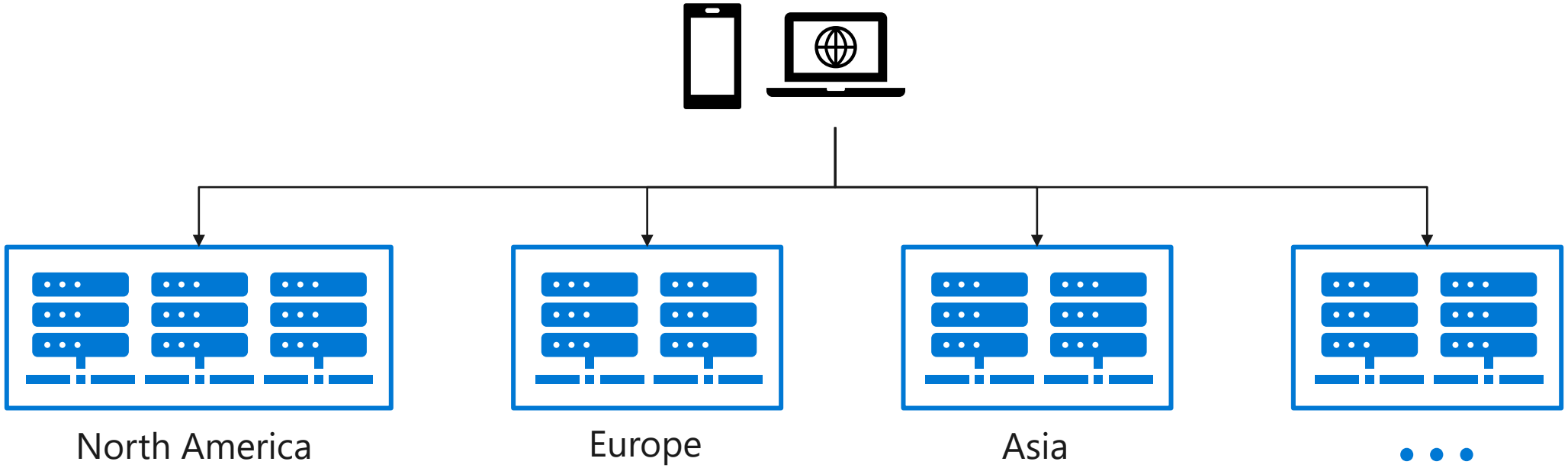
Hardware Acceleration
+ System Optimization
Model on **V100 GPU** NCSv3 VM:

6ms

Batch Size 64

Tensor Core with mixed precision
Same accuracy

Distributed serving on Azure GPU VMs worldwide



Inferences per second worldwide	1M
# of Azure N-series GPU VMs	2000+
Flops per inference	6.4 Gflops

Summary

- Bing's most significant search quality improvements in the last year were delivered by large NLP models
- System optimization and GPU hardware acceleration were necessary to meet strict performance and efficiency requirements as models grow in size
- Azure GPU VMs were critical in enabling Bing to ship large NLP models at global scale

Blog: https://azure.microsoft.com/en-us/blog/bing-delivers-its-largest-improvement-in-search-experience-using-azure-gpus/

Microsoft

Thank you for taking the time to attend our session today

We would love to hear from you! Please take a short, anonymous survey, and you will receive a gift from us. Use your phone to scan the QR code, or go to:

aka.ms/SC19survey

Once complete, visit the reception counter, show that you have finished the survey, and claim your gift.

Don’t forget to come back to see if you will win the Surface Go

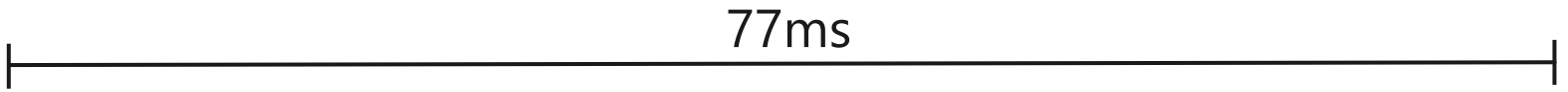
Drawings for the Surface Go devices will be held at the Microsoft booth (#633) on Tuesday (11/19) at 5:45 PM, Wednesday (11/20) at 5:45 PM, and Thursday (11/21) at 1:45 PM. Must be present to win.

Q&A

- Are these optimizations available anywhere?
  - Available ONNX in today but announcement coming shortly
- Azure ML
  - Not using AML at this point

Case Study: BERT Model Optimization

Original Model:
3–layer BERT on CPU:



Case Study: BERT Model Optimization

Original Model:
3–layer BERT on CPU:

77ms

Hardware Acceleration
Model on **M60 GPU** VM:

20ms