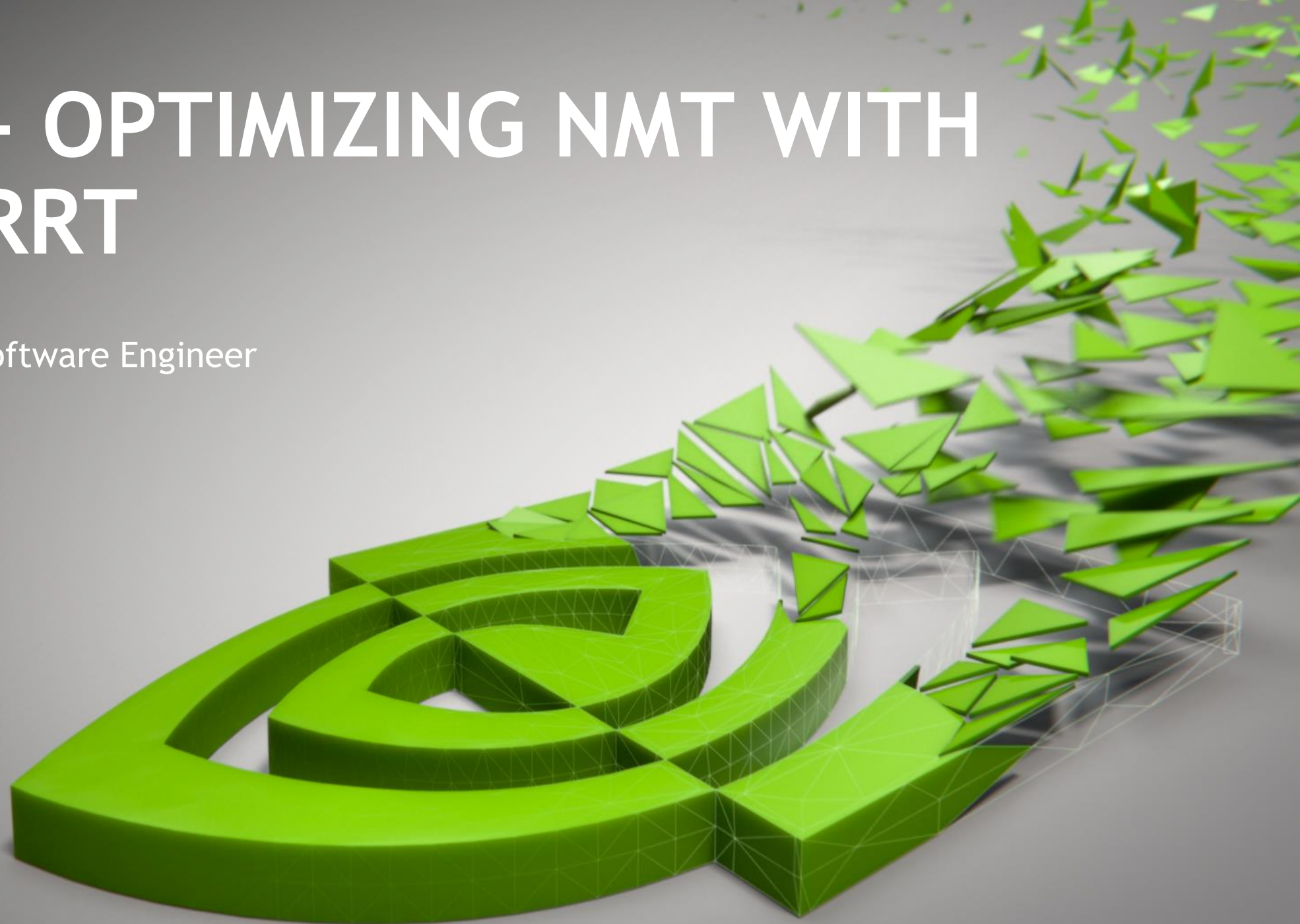


# S8822 - OPTIMIZING NMT WITH TENSORRT

Micah Villmow

Senior TensorRT Software Engineer

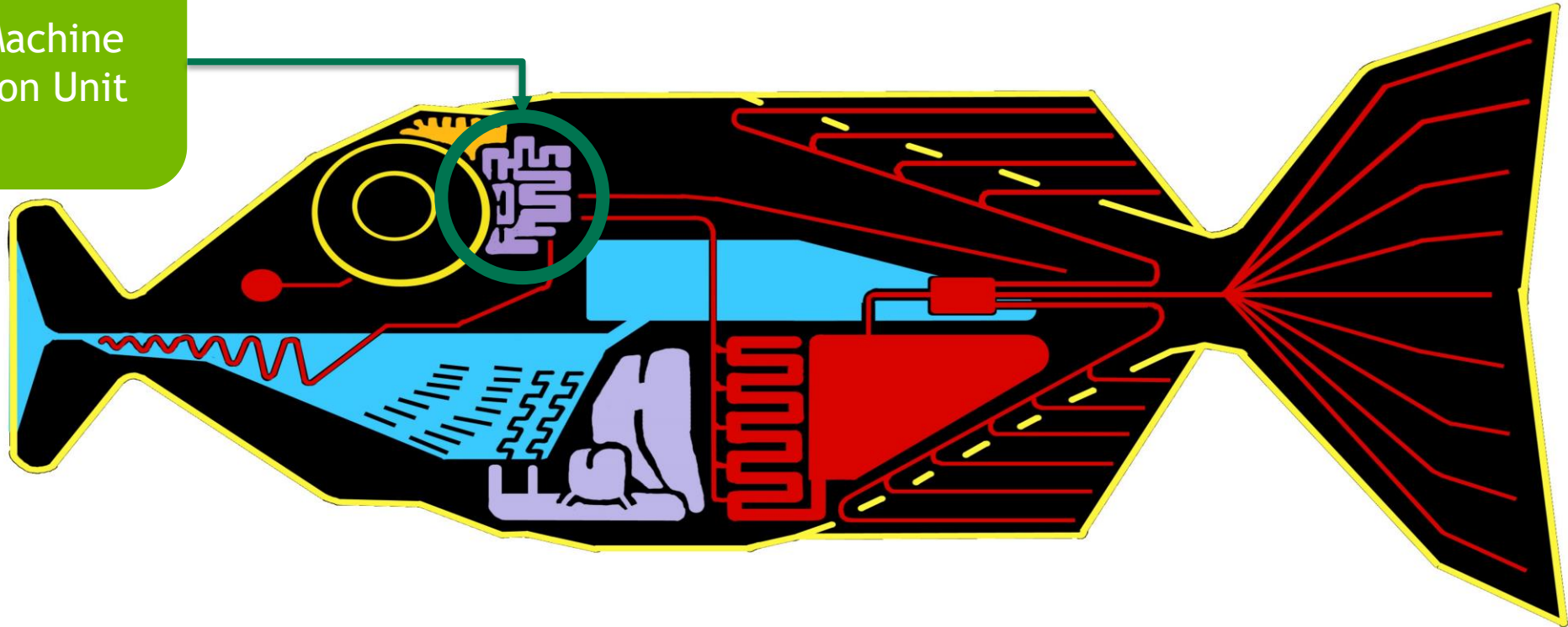


100倍以上速く、  
本当に可能ですか？



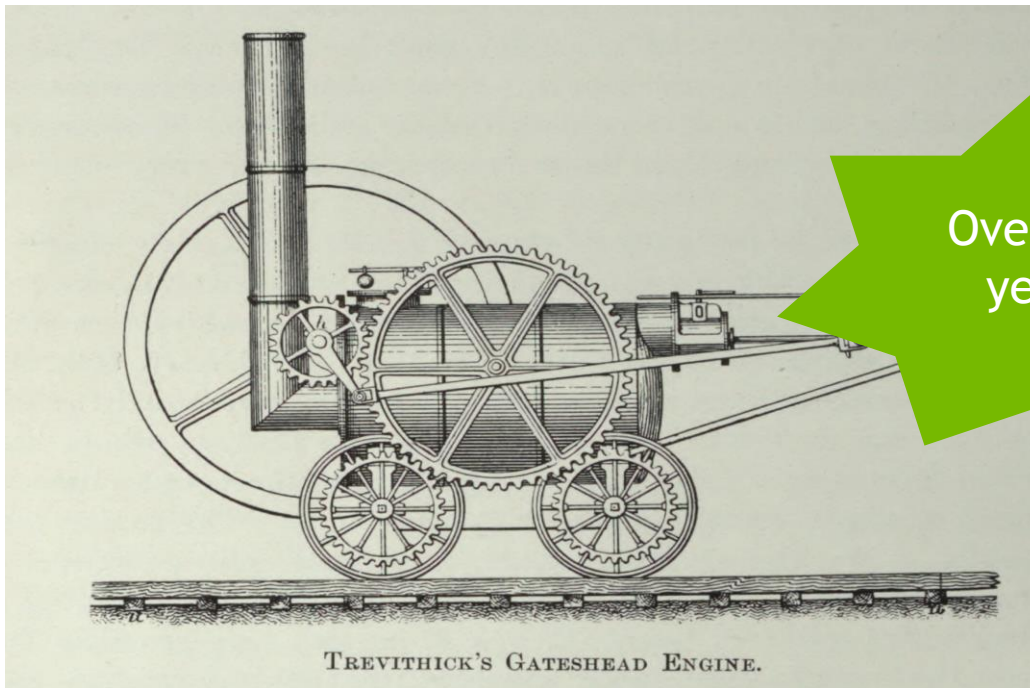
# DOUGLAS ADAMS - BABEL FISH

Neural Machine  
Translation Unit





# OVER 100X FASTER, IS IT REALLY POSSIBLE?

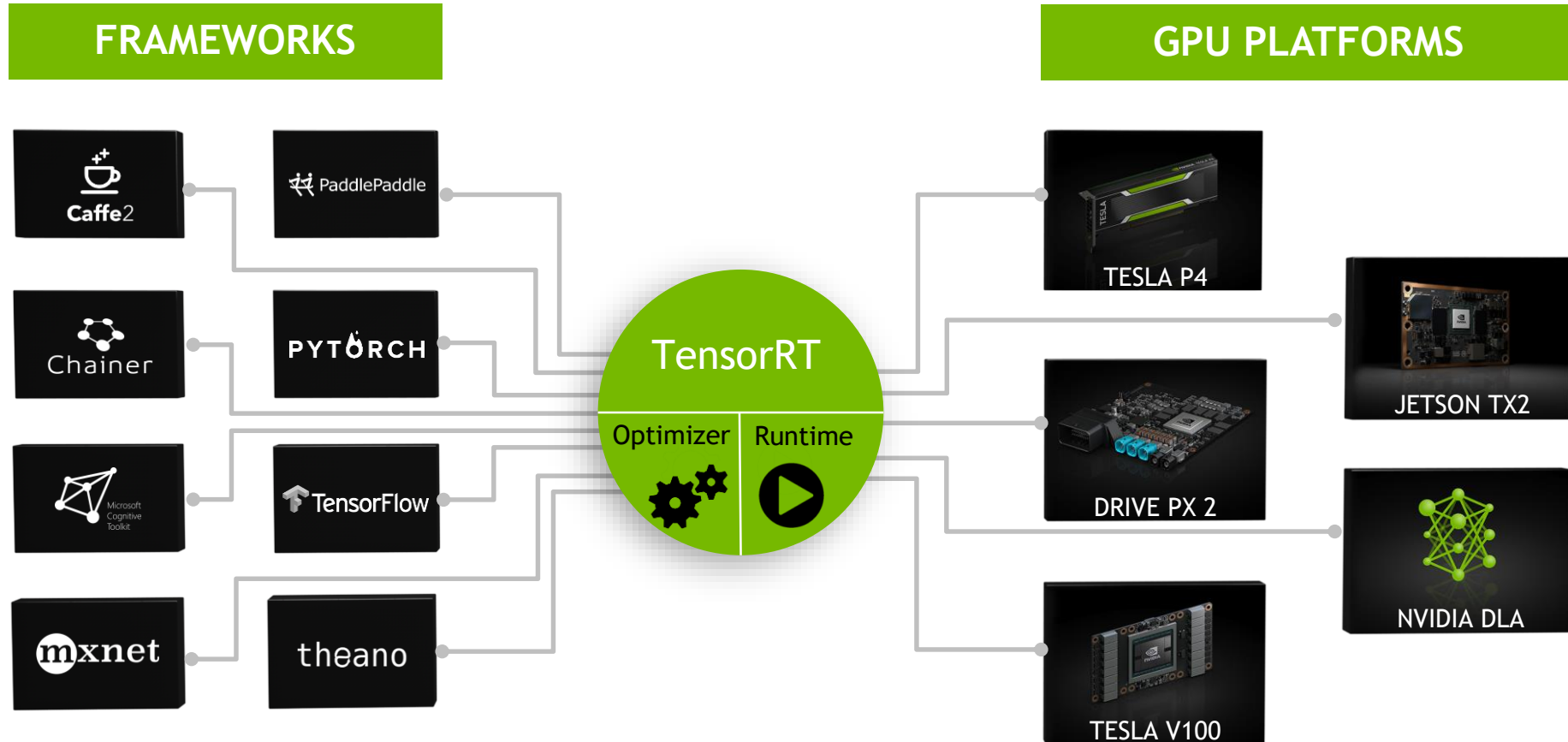


Over 200  
years



# NVIDIA TENSORRT

Programmable Inference Accelerator

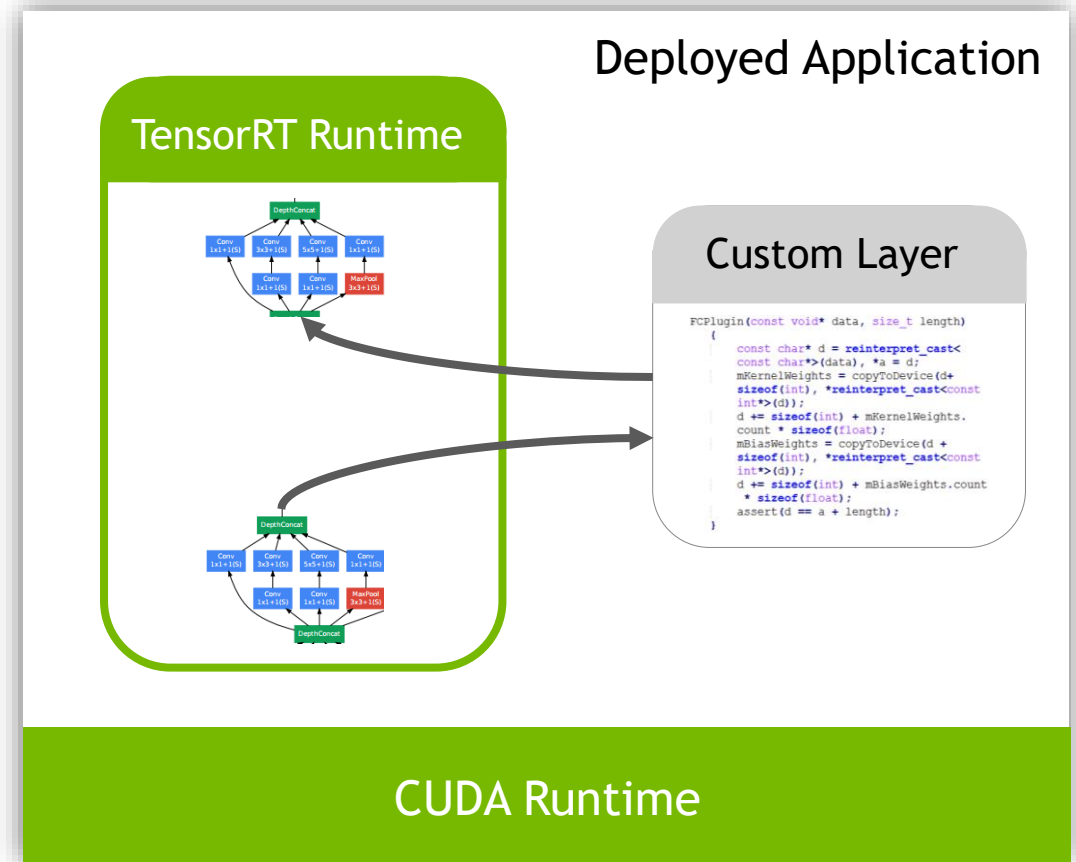


# TENSORRT LAYERS

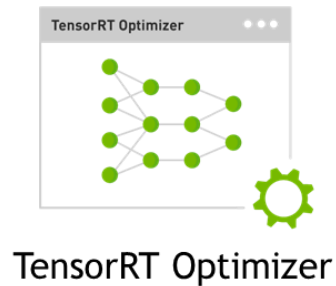
## Built-in Layer Support


- Convolution
- LSTM and GRU
- Activation: ReLU, tanh, sigmoid
- Pooling: max and average
- Scaling
- Element wise operations
- LRN
- Fully-connected
- SoftMax
- Deconvolution

## Custom Layer API





# TENSORRT OPTIMIZATIONS



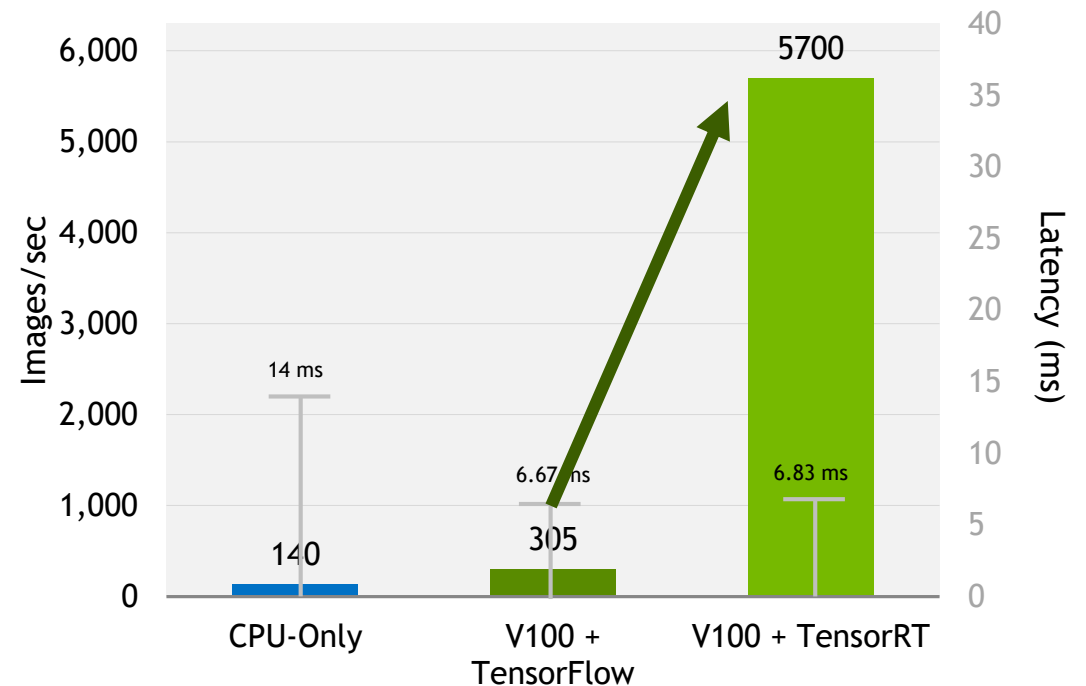
  
**Layer & Tensor Fusion**

  
**Weights & Activation  
Precision Calibration**

  
**Kernel Auto-Tuning**

  
**Dynamic Tensor  
Memory**

**40x Faster CNNs on V100 vs. CPU-Only  
Under 7ms Latency (ResNet50)**



Inference throughput (images/sec) on ResNet50. **V100 + TensorRT**: NVIDIA TensorRT (FP16), batch size 39, Tesla V100-SXM2-16GB, E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On **V100 + TensorFlow**: Preview of volta optimized TensorFlow (FP16), batch size 2, Tesla V100-PCI-E-16GB, E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. **CPU-Only**: Intel Xeon-D 1587 Broadwell-E CPU and Intel DL SDK. Score doubled to comprehend Intel's stated claim of 2x performance improvement on Skylake with AVX512.

# Agenda

- **What is NMT?**
- What is current state?
- What are the problems?
- How did we solve it?
- What perf is possible?



# ACRONYMS AND DEFINITIONS

NMT: Neural Machine Translation

OpenNMT: Open source NMT project for academia and industry

Token: The minimum representation used for encoding(symbol, word, character, subword)

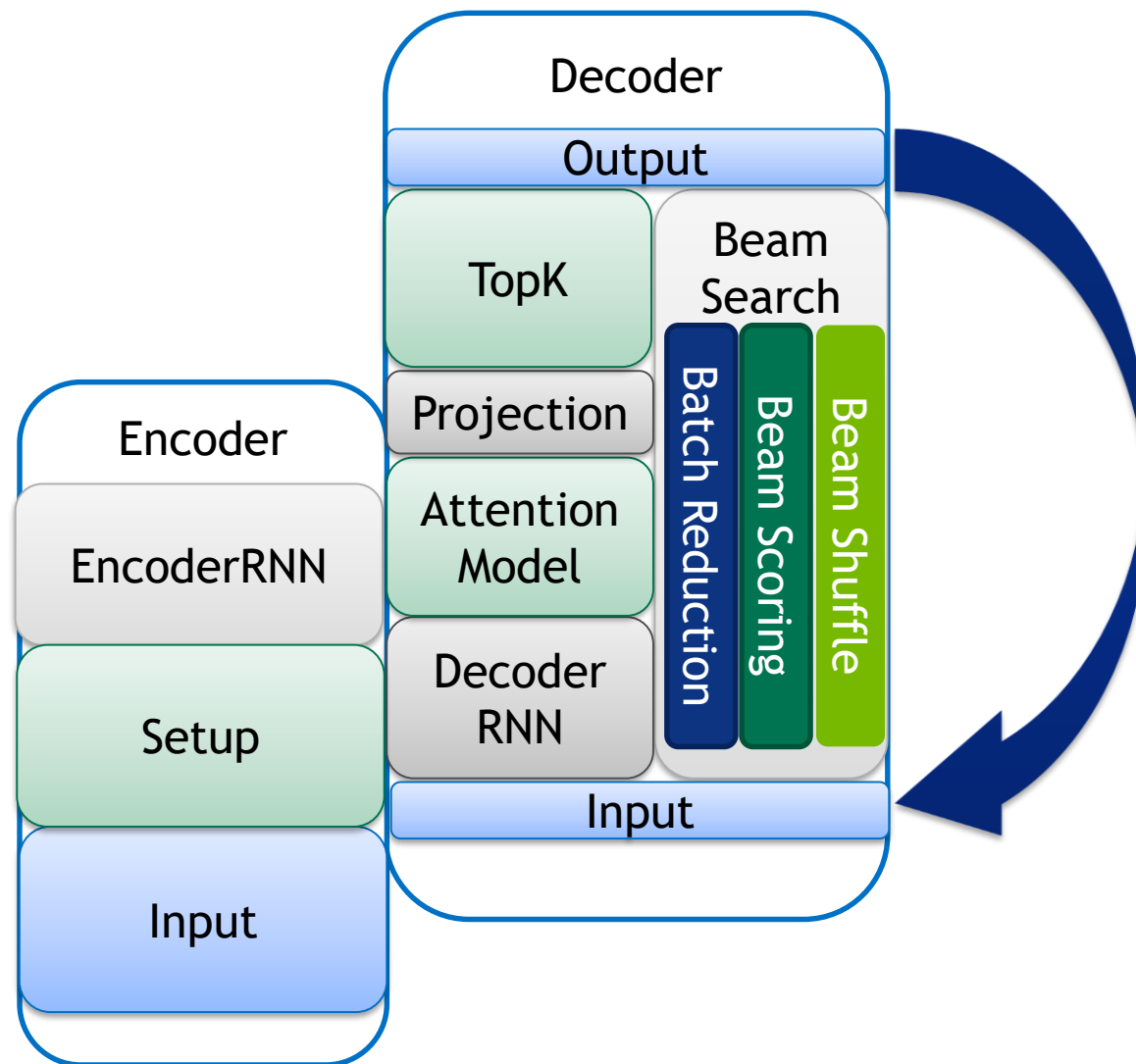
Sequence: A number of tokens wrapped by special start and end sequence tokens.

Beam Search: directed partial breadth-first tree search algorithm

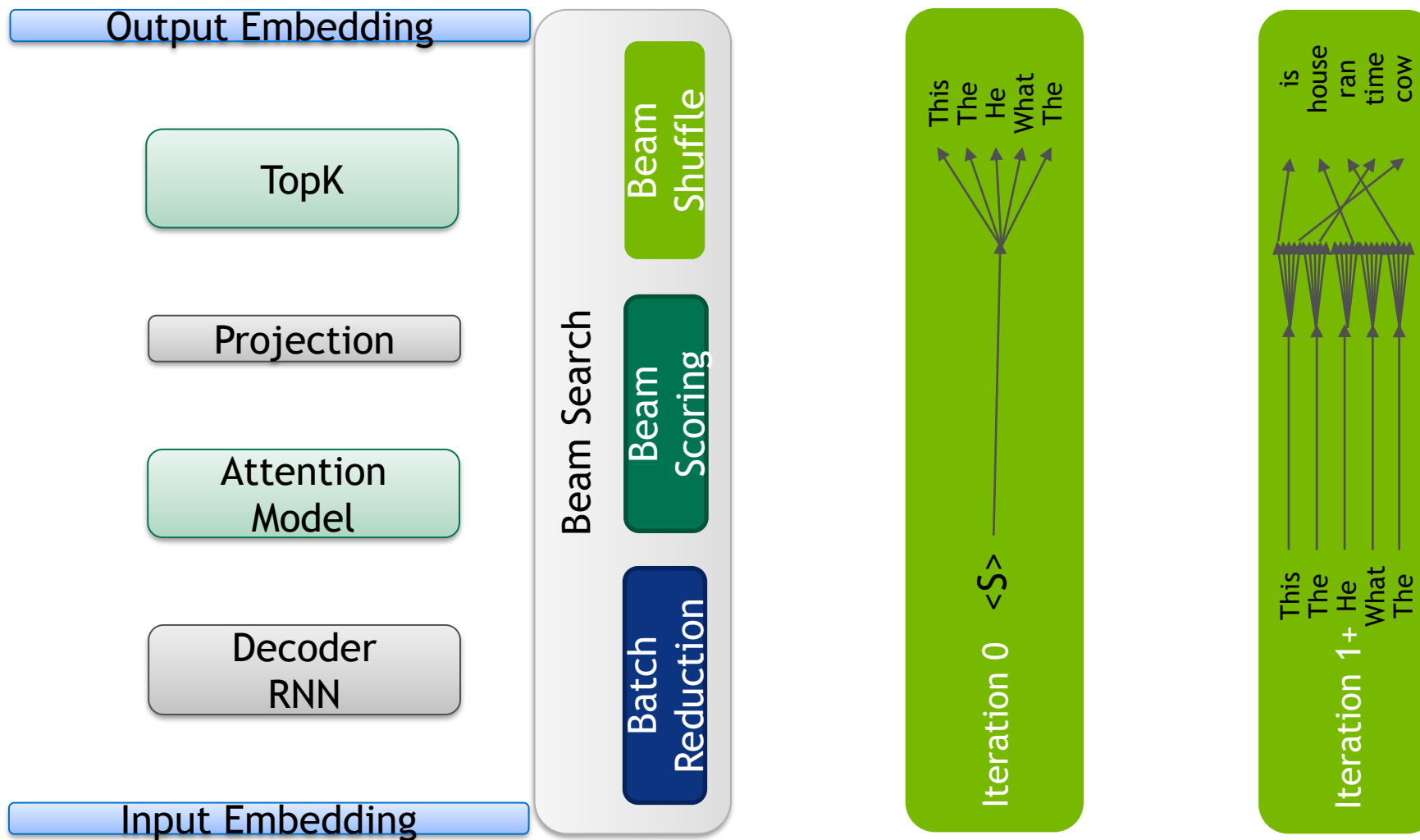
TopK: Partial sort resulting in N min/max elements

Unk: Special token that represents unknown translations.

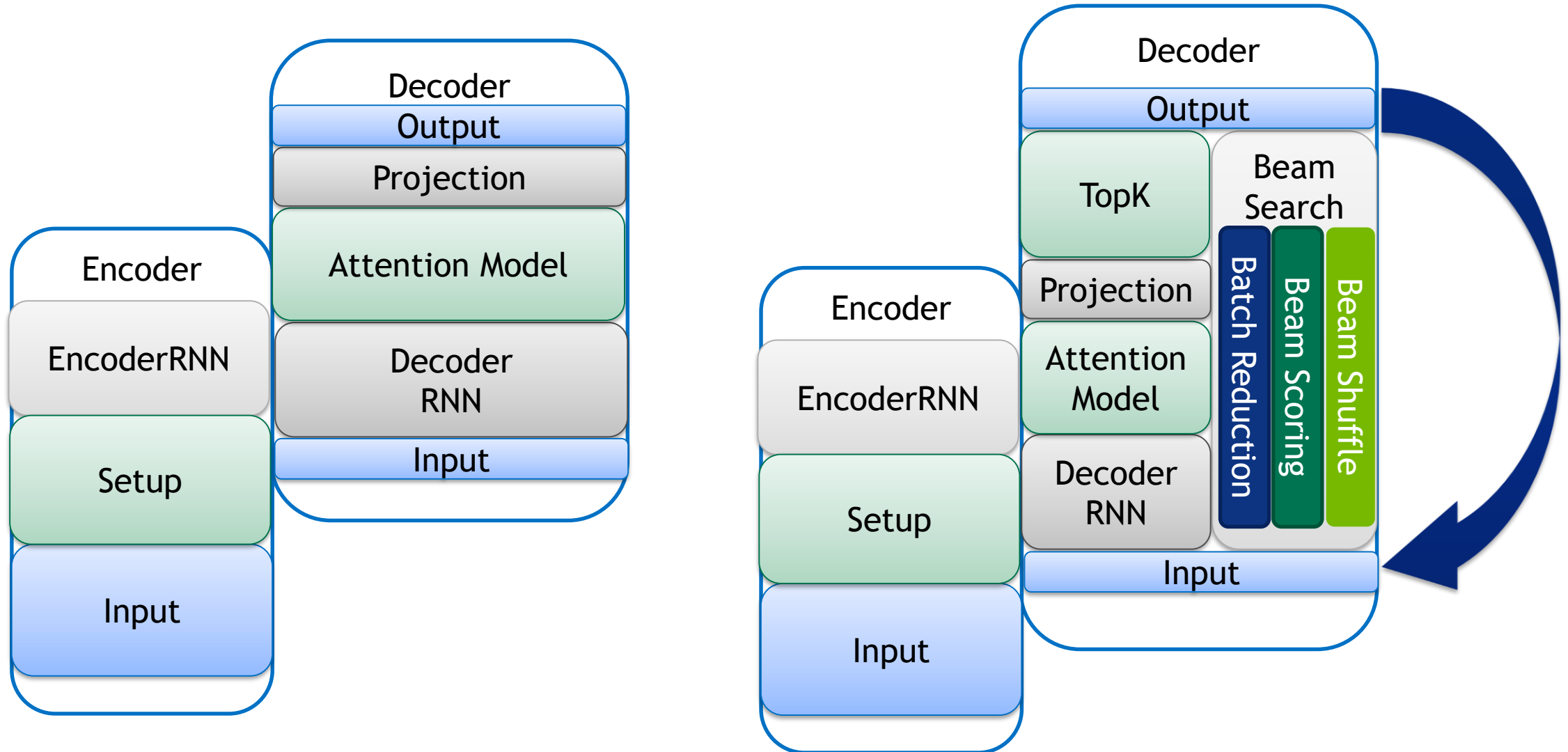
# OPENNMT INFERENCE



# DECODER EXAMPLE



# TRAINING VS INFERENCE



# Agenda

- What is NMT?
- **What is current state?**
- What are the problems?
- How did we solve it?
- What perf is possible?



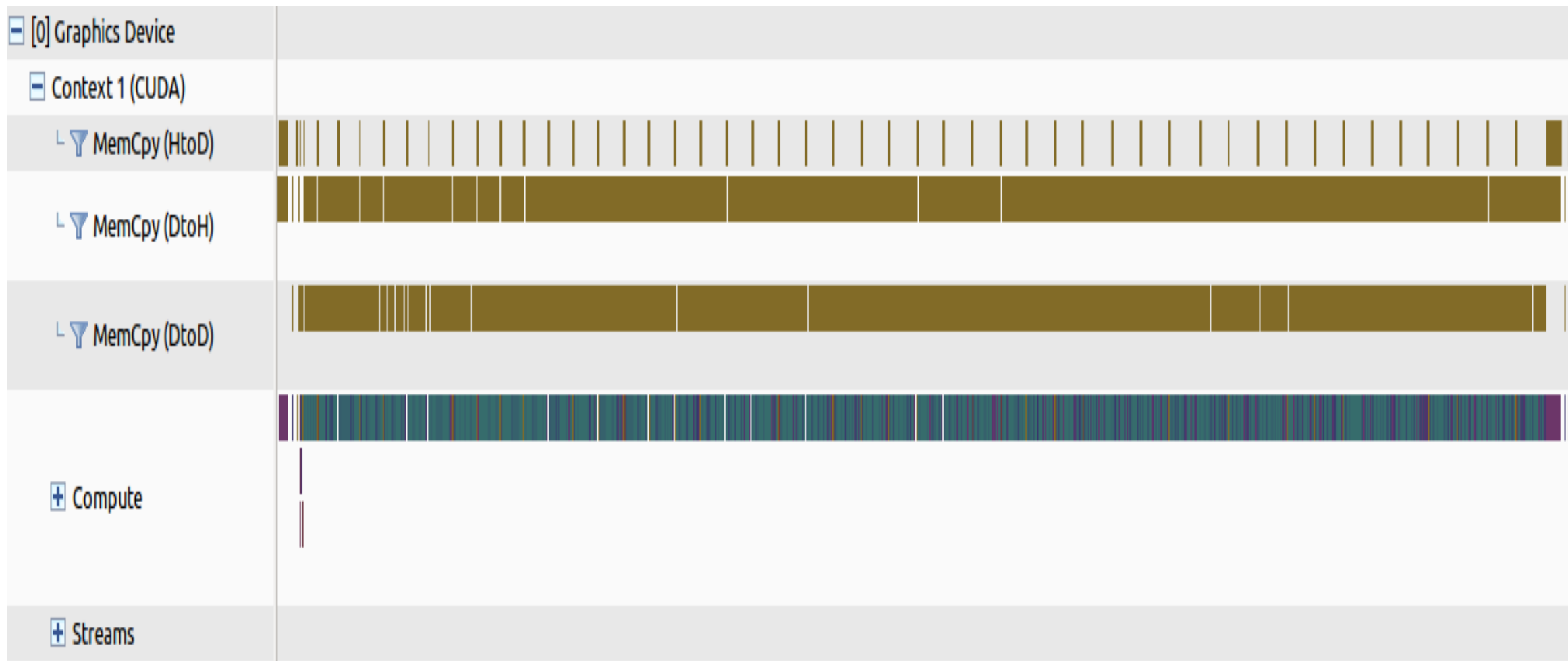
# INFERENCE TIME IS BEAM SEARCH TIME

- Wu, Et. Al. 2016, 'Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation'  
[arXiv:1609.08144](https://arxiv.org/abs/1609.08144)
- Sharan Narang, Jun, 2017, Baidu's DeepBench -  
<https://github.com/baidu-research/DeepBench>
- Rui Zhao, Dec, 2017, 'Why does inference run 20x slower than training.' - <https://github.com/tensorflow/nmt/issues/204>
- David Levinthal, Ph.D., Jan, 2018, 'Evaluating RNN performance across hardware platforms.'

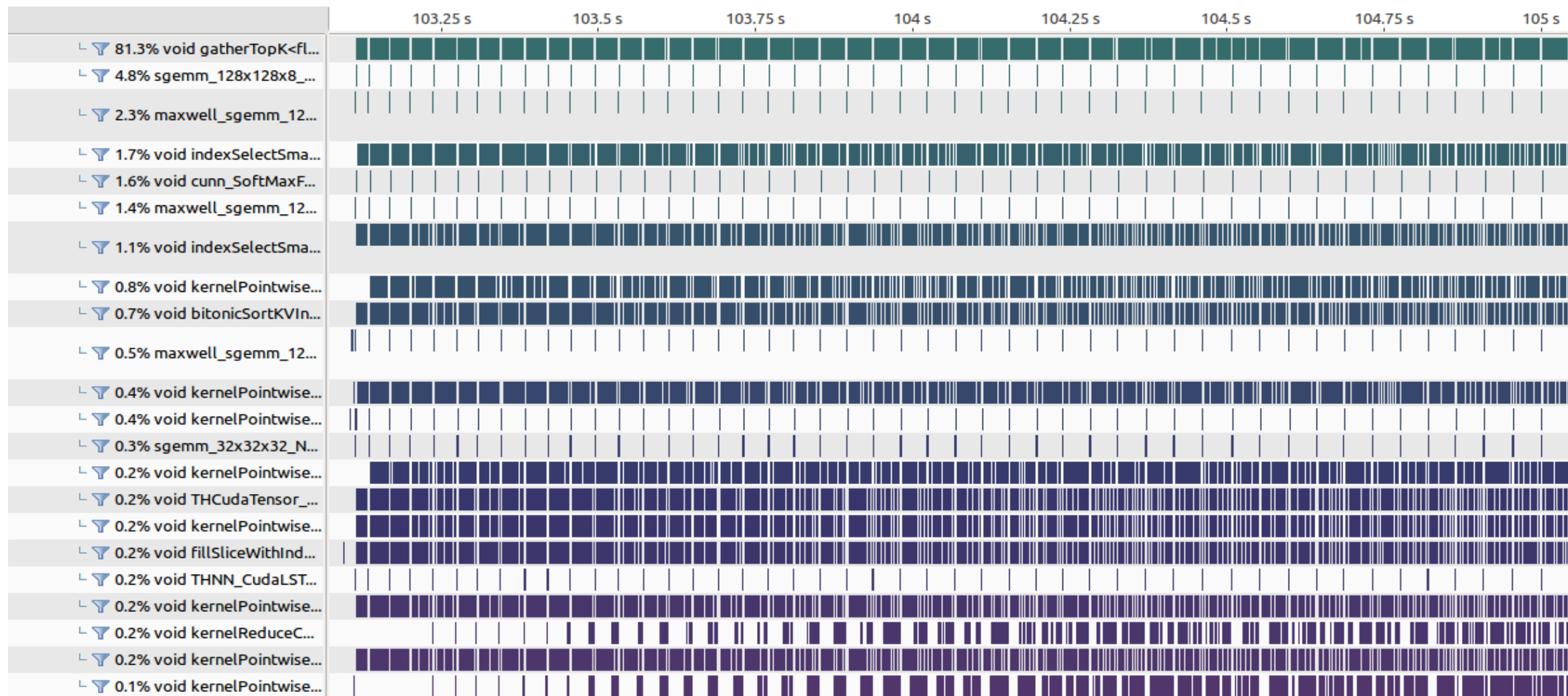
# Agenda

- What is NMT?
- What is current state?
- **What are the problems?**
- How did we solve it?
- What perf is possible?

# PERF ANALYSIS



# KERNEL ANALYSIS



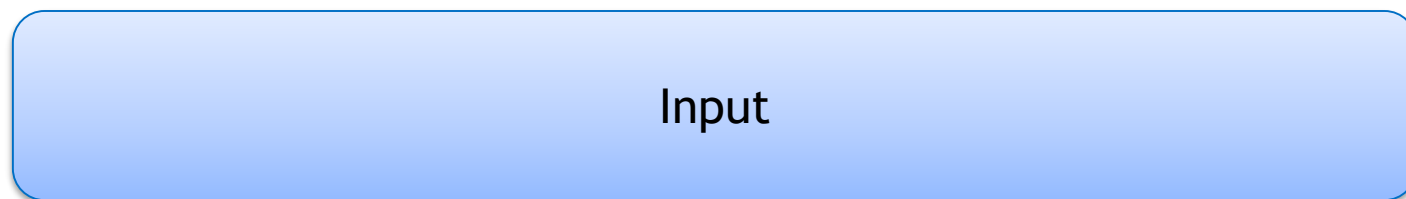
# Agenda

- What is NMT?
- What is current state?
- What are the problems?
- **How did we solve it?**
- What perf is possible?



# ENCODER

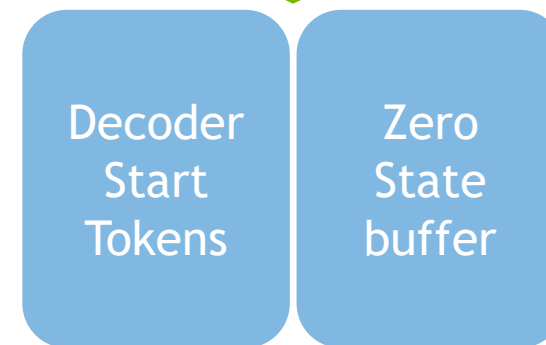
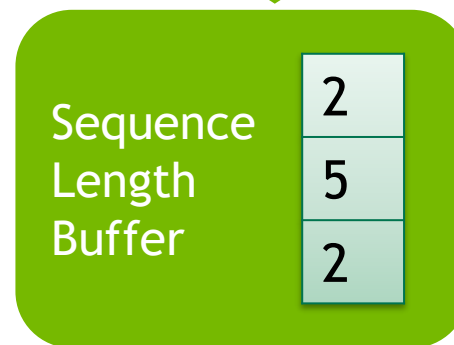
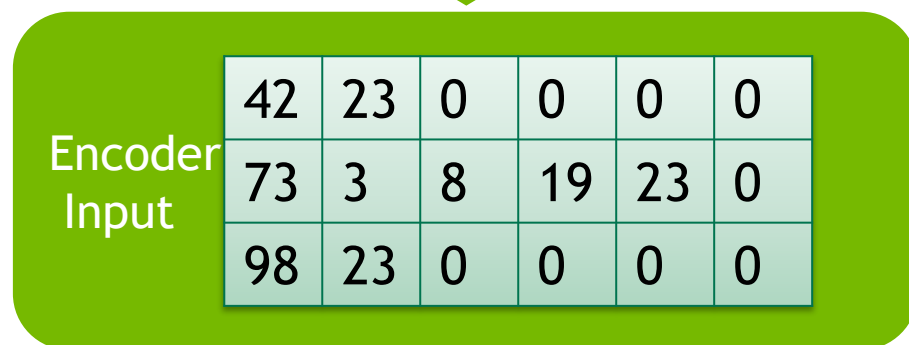


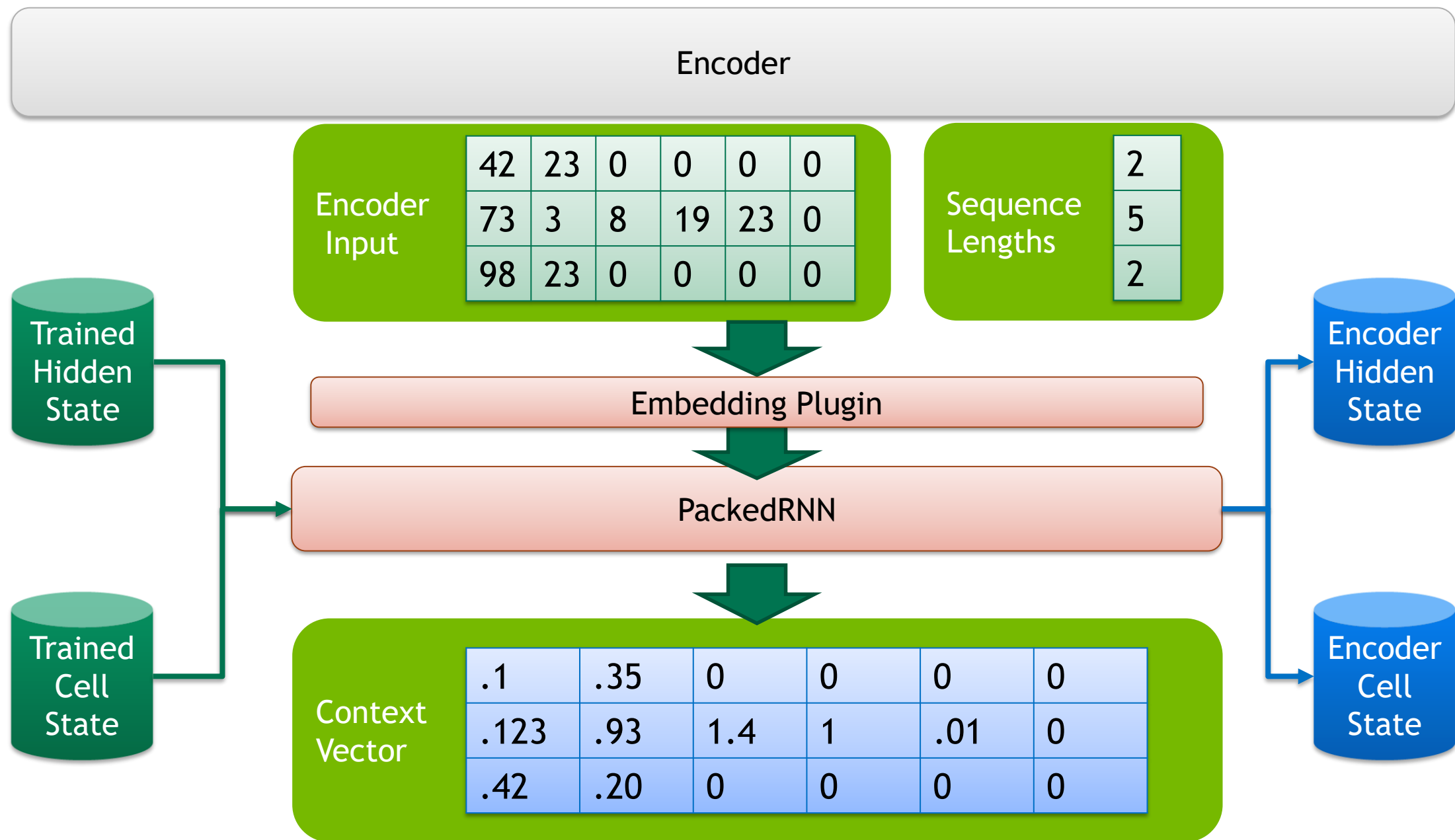


Hello.  
This is a test.  
Bye.

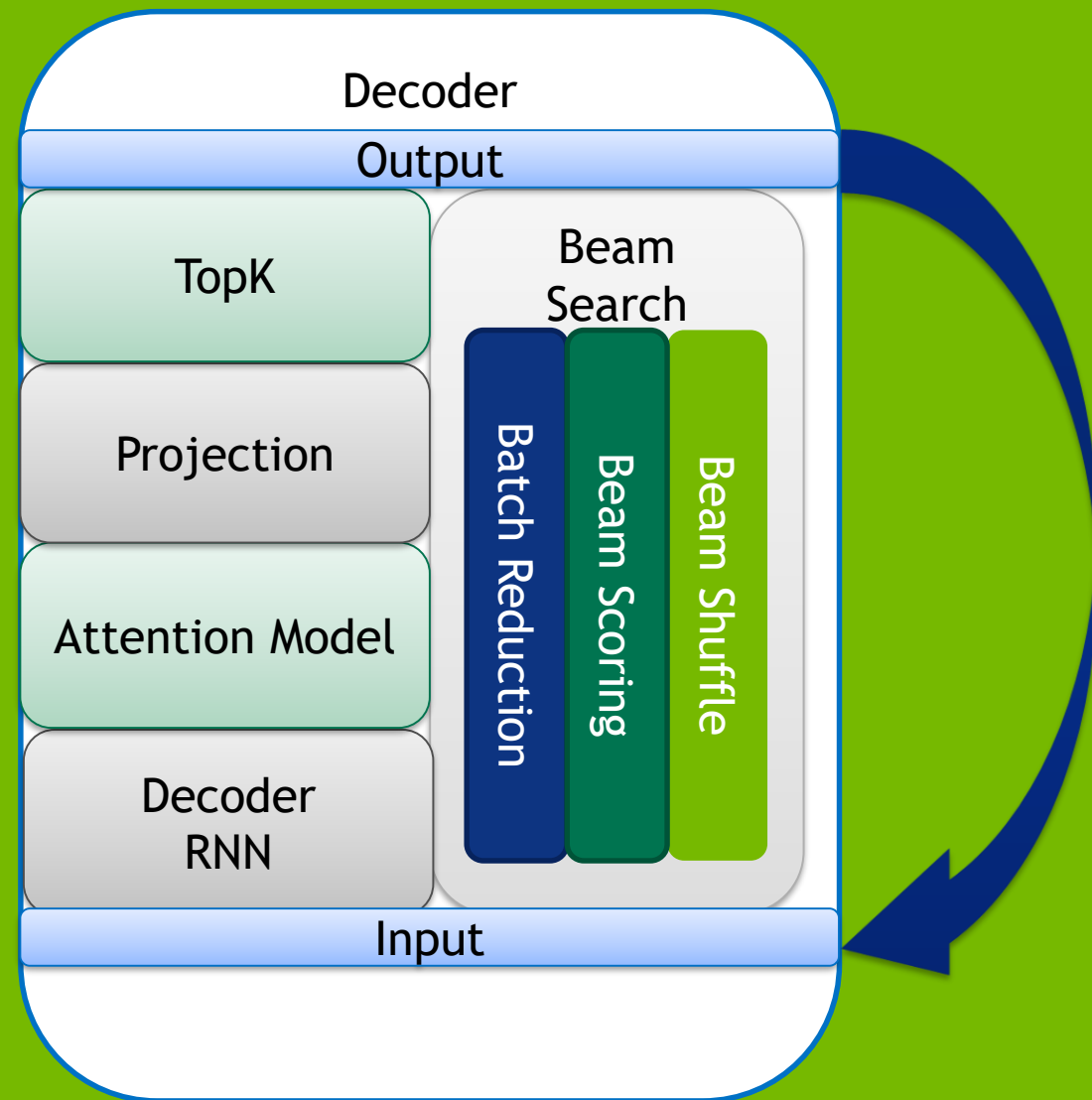
Tokenization

Hello .  
This is a test .  
Bye .





# DECODER



## Decoder, 1<sup>st</sup> Iteration

Decoder  
Input

Batch0	<S>
BatchN	<S>

Start Sentence  
Token

Embedding Plugin

RNN

Decoder  
Output

Batch0	.124
BatchN	.912

Encoder  
Hidden  
State

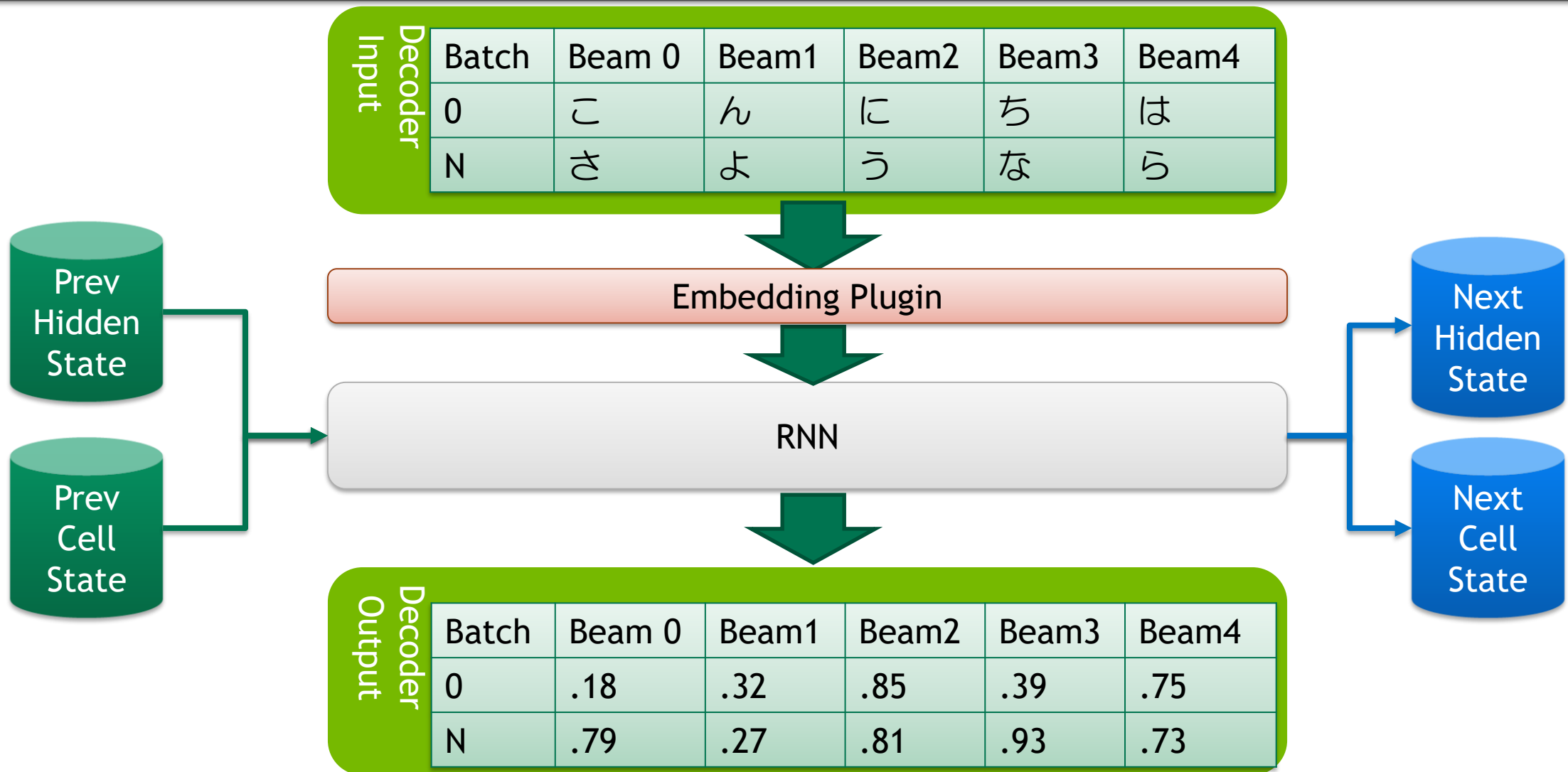
Encoder  
Cell  
State

Decode  
Hidden  
State

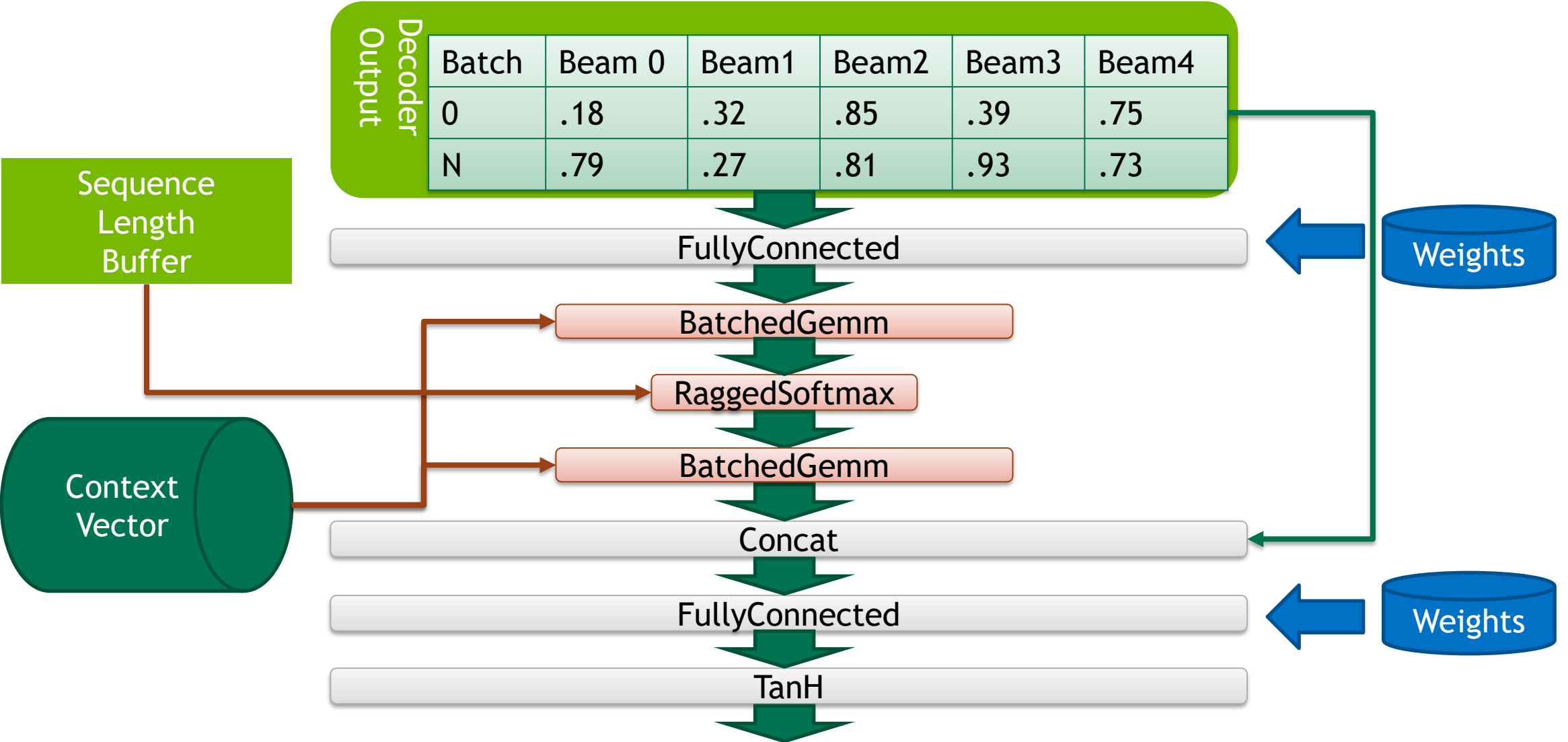
Decode  
Cell  
State



## Decoder, 2<sup>nd</sup>+ Iteration



# Global Attention Model



## Projection

Attention  
Output

Batch	Beam 0	Beam1	Beam2	Beam3	Beam4
0	[.9,...,.1]	[0,...,.3]	[.1,...,0]	[.6,...,.8]	[.3,...,.2]
N	[.4,...,.9]	[.5,...,.2]	[0,...,.7]	[0,...,2]	[.1,...,.9]

FullyConnected

Softmax

Log

Projection  
Output

Batch	Beam 0	Beam1	Beam2	Beam3	Beam4
0					
N					

Weights

## TopK Part 1

Projection  
Output

Batch	Beam 0	Beam1	Beam2	Beam3	Beam4
0					
N					

TopK

Intra-beam  
Output

Batch	Beam 0	Beam1	Beam2	Beam3	Beam4
Index	[1,3]	[2,4]	[9,0]	[5,0]	[7,6]
Prob	[.9,.8]	[.99,.5]	[.3,.8]	[.1,.93]	[.85,.99]

Gather

## TopK Part 2

Gather  
Output

Prob	[.9,.8,.99,.55,.3,.8,.1,.93,.85,.99]
------	--------------------------------------

TopK

Inter-beam  
Output

Indices	[2,9,7,0,8]
Prob	[.99,.99,.93,.9,.85]

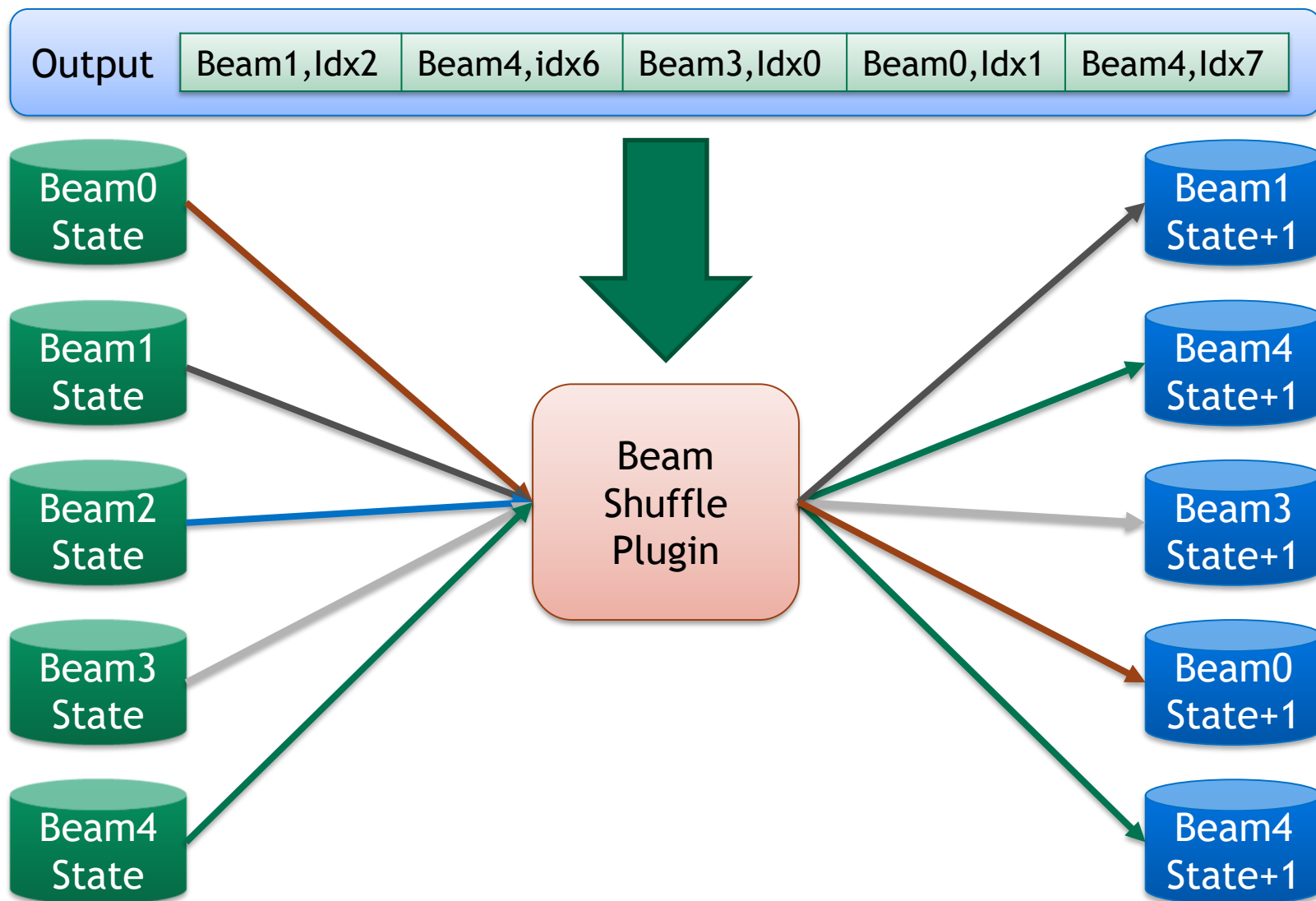
Intra-beam  
Output

Beam Mapping Plugin

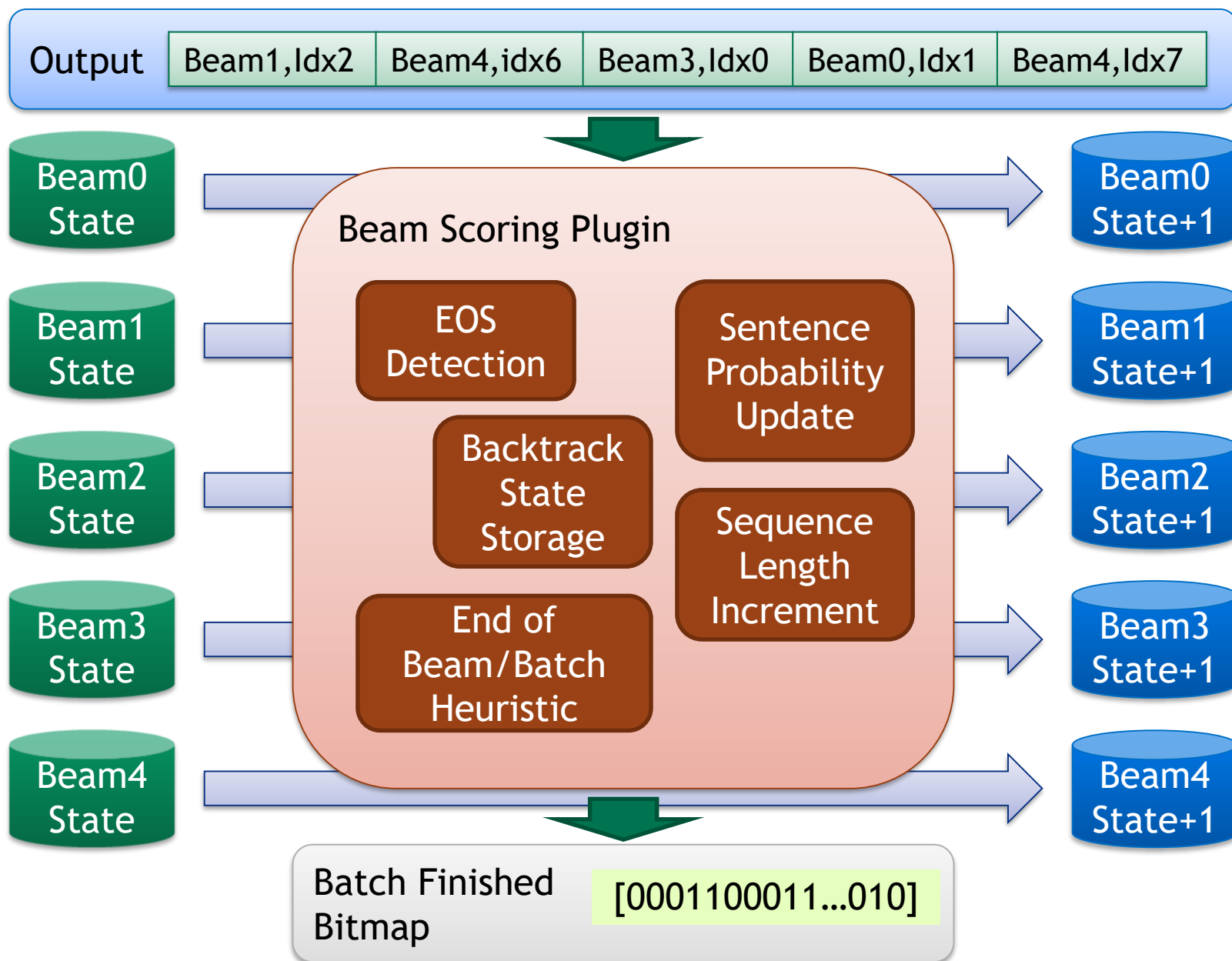
Output	Beam1,Idx2	Beam4,Idx6	Beam3,Idx0	Beam0,Idx1	Beam4,Idx7
--------	------------	------------	------------	------------	------------



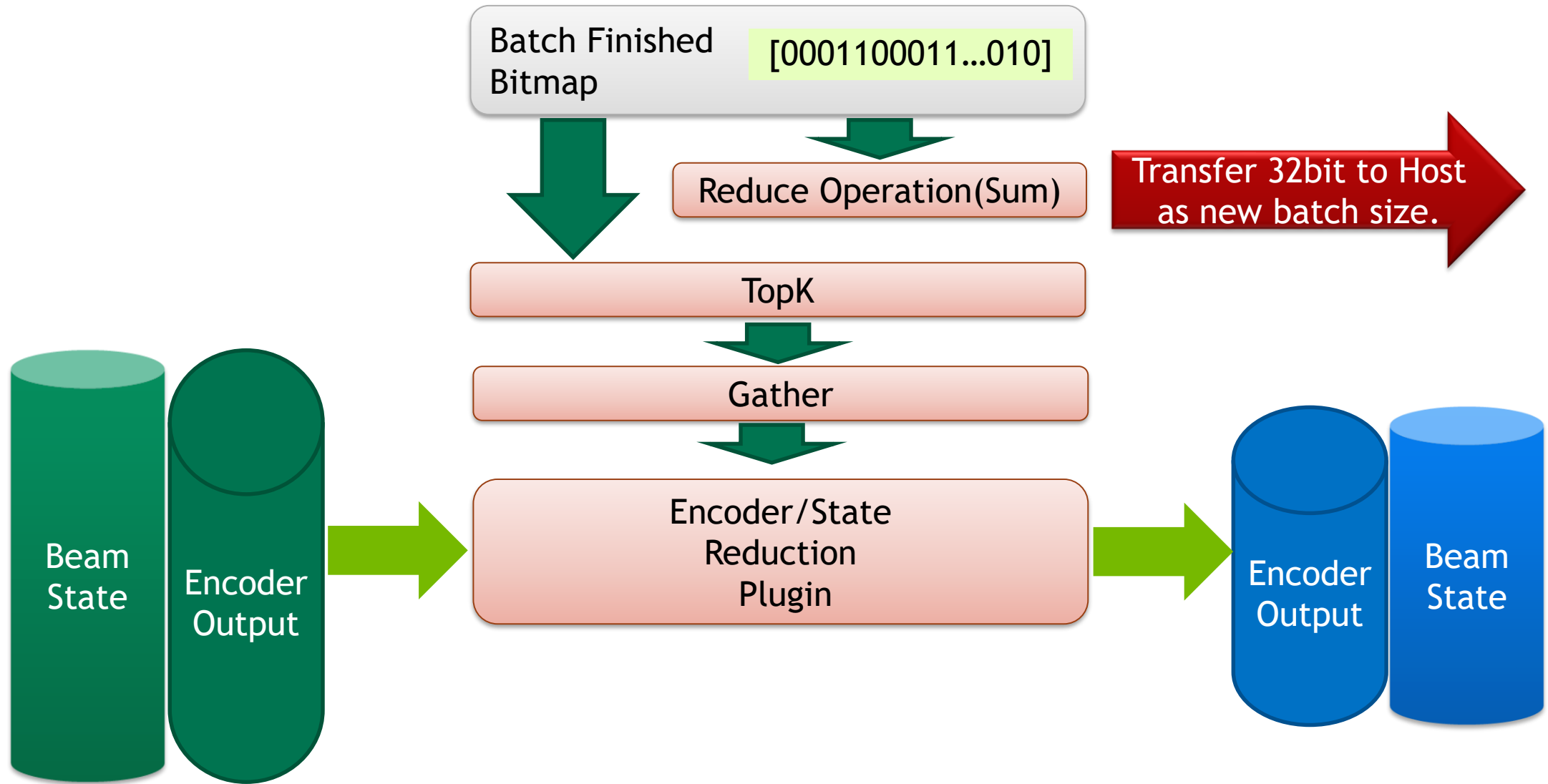
## Beam Search - Beam Shuffle

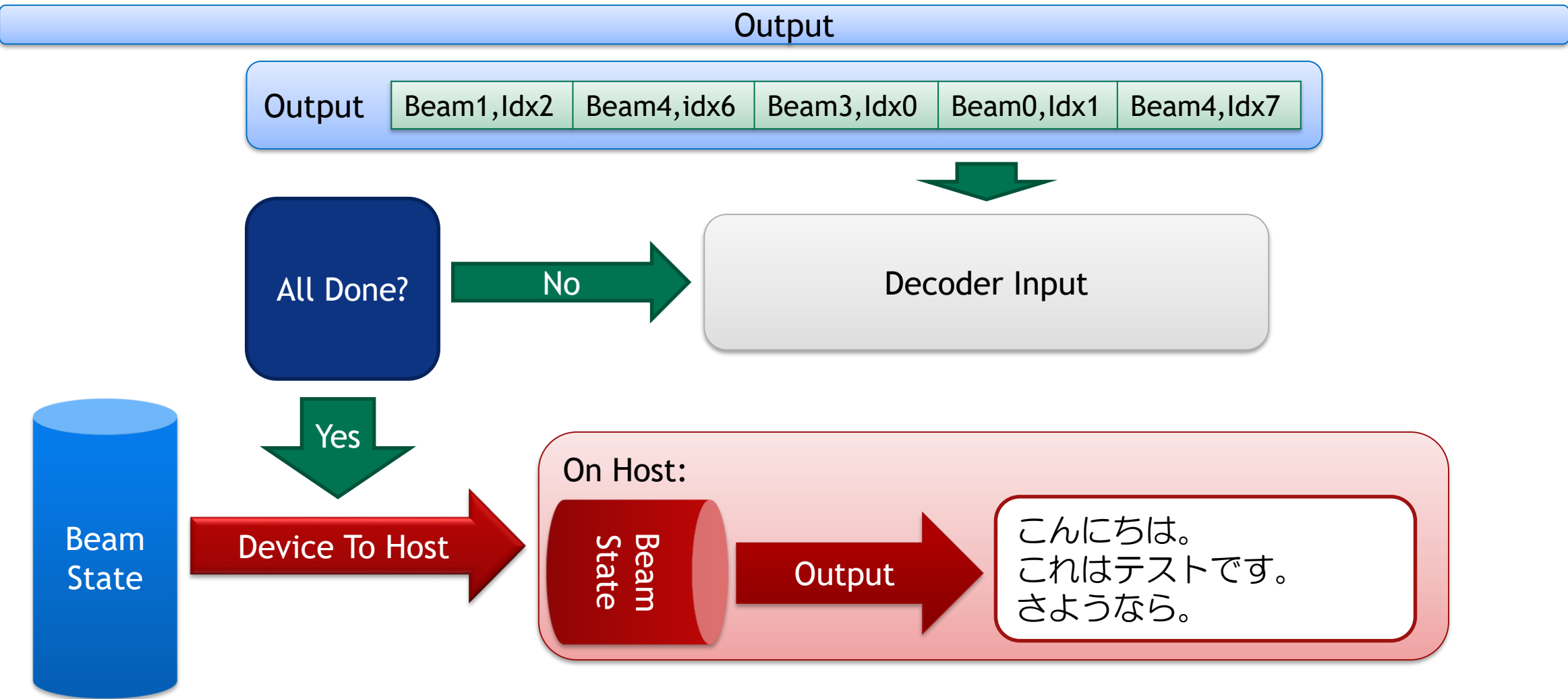


## Beam Search - Beam Scoring

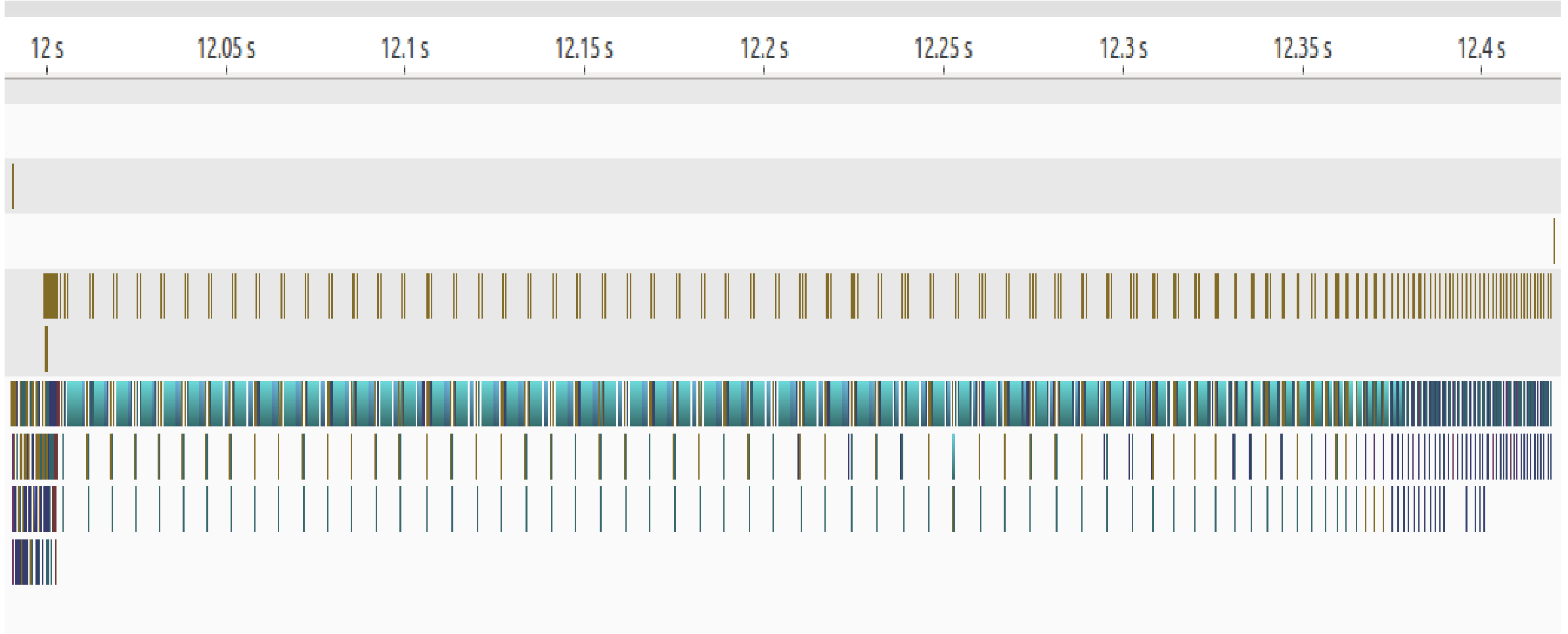


## Beam Search - Batch Reduction

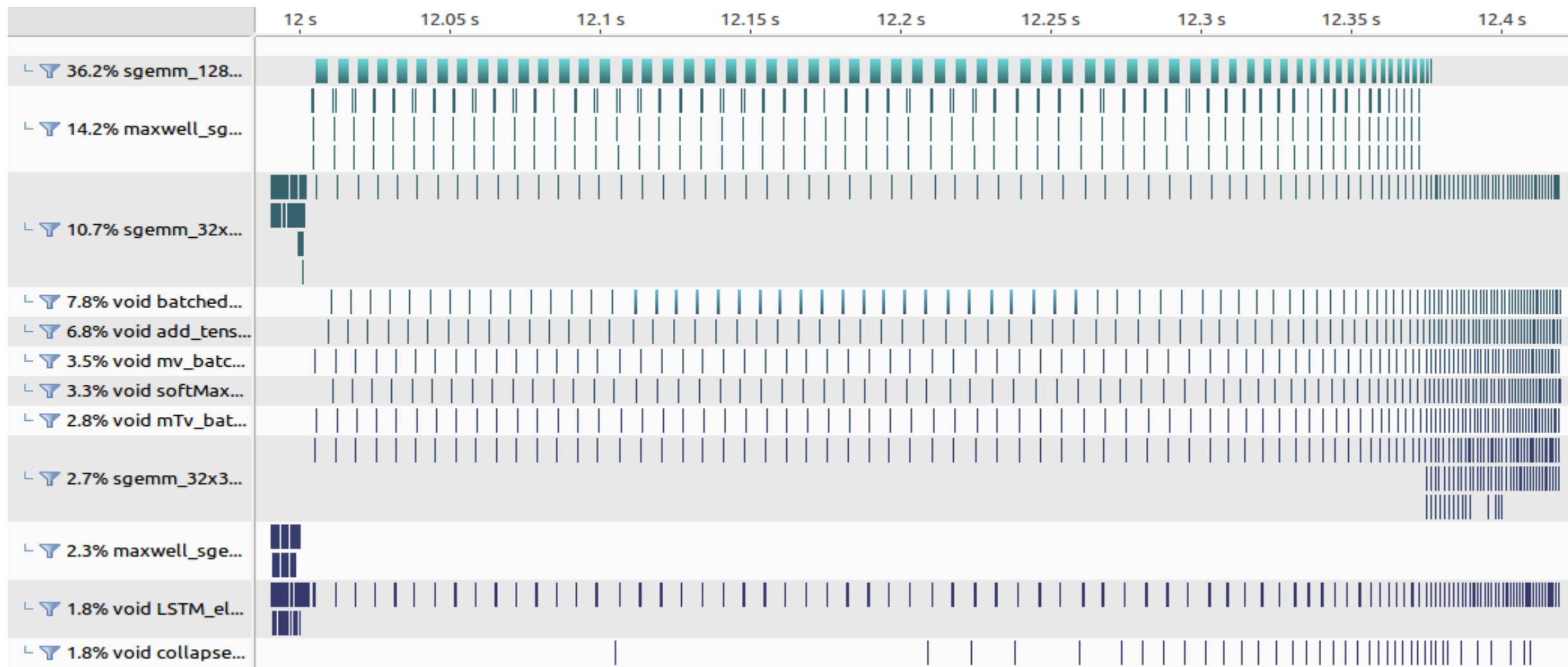




# TENSORRT ANALYSIS



# TENSORRT KERNEL ANALYSIS

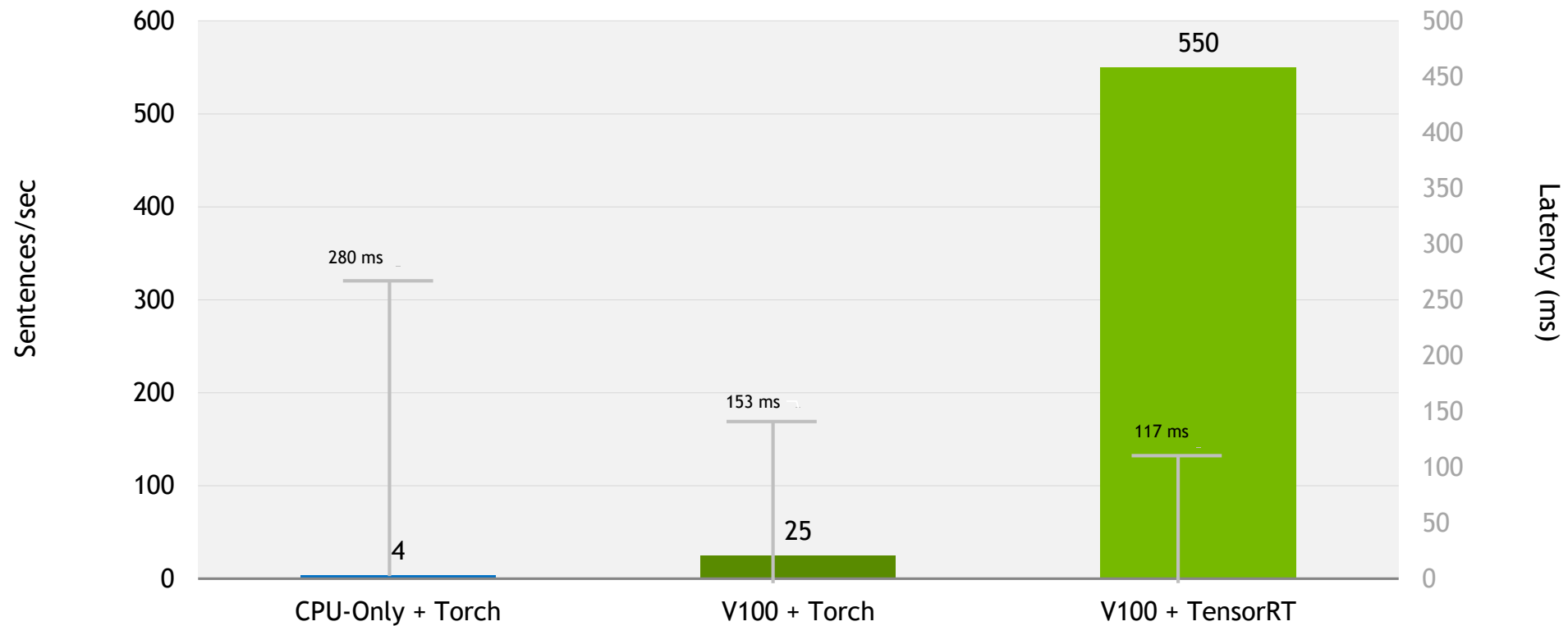


# Agenda

- What is NMT?
- What is current state?
- What are the problems?
- How did we solve it?
- **What perf is possible?**

# RESULTS

140x Faster Language Translation RNNs on V100 vs. CPU-Only Inference  
(OpenNMT)



Inference throughput (sentences/sec) on OpenNMT 692M. **V100 + TensorRT**: NVIDIA TensorRT (FP32), batch size 64, Tesla V100-PCIE-16GB, E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. **V100 + Torch**: Torch (FP32), batch size 4, Tesla V100-PCIE-16GB, E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On. **CPU-Only**: Torch (FP32), batch size 1, Intel E5-2690 v4@2.60GHz 3.5GHz Turbo (Broadwell) HT On



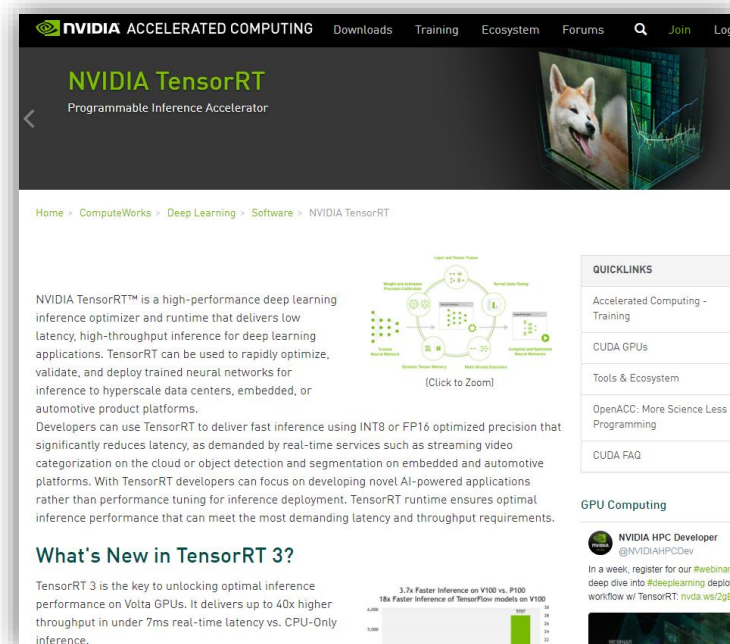
# SUMMARY

- Show that topK no longer dominates sequence inference time.
- Show that RNN Inference is compute bound, not memory bound.
- TensorRT accelerates sequence inferencing.

## PRODUCT PAGE

[developer.nvidia.com/tensorrt](https://developer.nvidia.com/tensorrt)

- Over two orders of magnitude higher throughput over CPU.
- Latency reduction by more than half over CPU.



# LEARN MORE

## PRODUCT PAGE

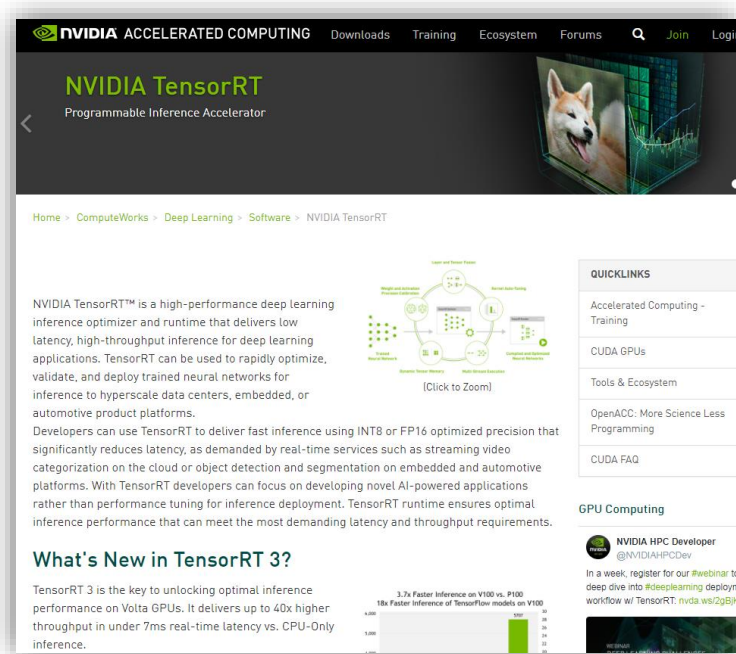
[developer.nvidia.com/tensorrt](https://developer.nvidia.com/tensorrt)

## DOCUMENTATION

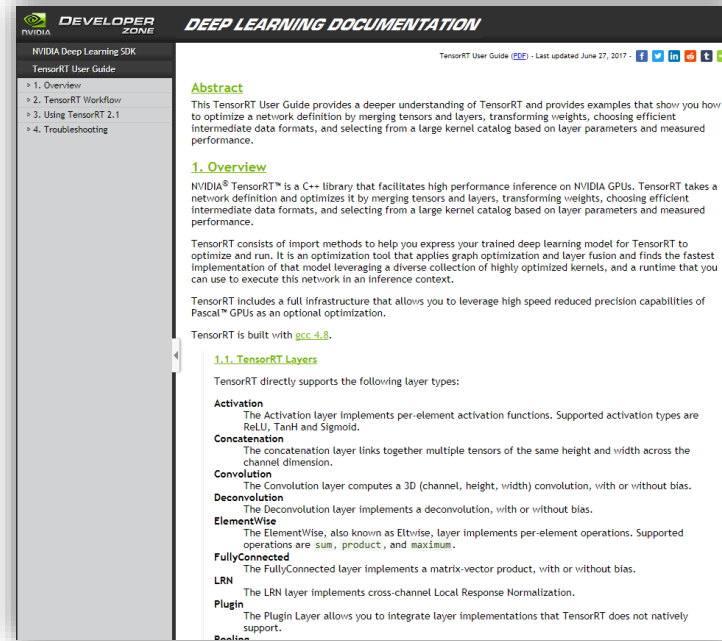
[docs.nvidia.com/deeplearning/sdk](https://docs.nvidia.com/deeplearning/sdk)

## TRAINING

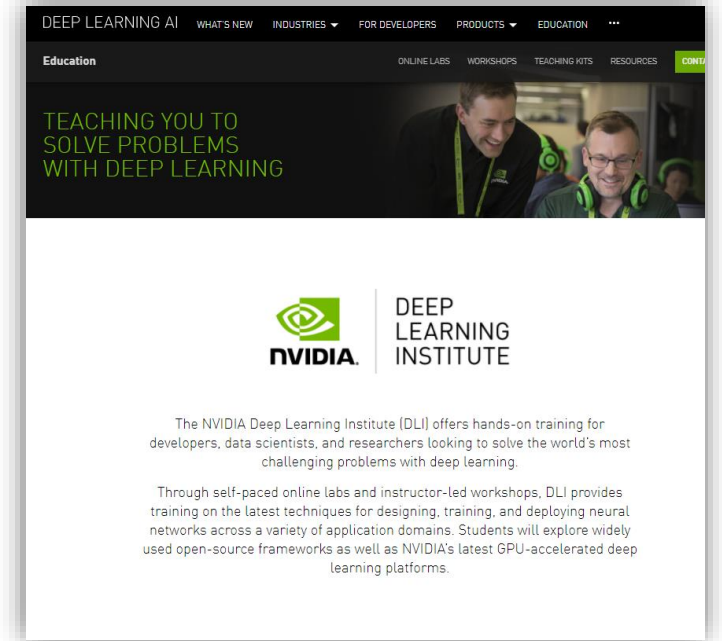
[nvidia.com/dli](https://nvidia.com/dli)



The screenshot shows the NVIDIA TensorRT product page. At the top, there's a navigation bar with links like Downloads, Training, Ecosystem, Forums, and a search icon. The main header features the NVIDIA logo and 'ACCELERATED COMPUTING'. Below this, the 'NVIDIA TensorRT' title is prominent, followed by the subtitle 'Programmable Inference Accelerator'. A hero image shows a dog's face on a screen with a green grid overlay. A breadcrumb trail reads: Home > ComputeWorks > Deep Learning > Software > NVIDIA TensorRT. The main content area describes TensorRT as a high-performance deep learning inference optimizer and runtime. It includes a diagram of the inference workflow and a 'QUICKLINKS' section with links to Accelerated Computing - Training, CUDA GPUs, Tools & Ecosystem, OpenACC: More Science Less Programming, and CUDA FAQ. A 'GPU Computing' section mentions NVIDIA HPC Developer and includes a call to action for a webinar. A bar chart at the bottom shows '3.7x Faster Inference on V100 vs. P100' and '18x Faster Inference of TensorFlow models on V100'.



The screenshot shows the NVIDIA Deep Learning Documentation page. The header includes the 'DEVELOPER ZONE' logo and 'DEEP LEARNING DOCUMENTATION'. A sidebar on the left lists the 'TensorRT User Guide' sections: 1. Overview, 2. TensorRT Workflow, 3. Using TensorRT 2.1, and 4. Troubleshooting. The main content area is titled 'Abstract' and provides a deeper understanding of TensorRT. It includes an 'Overview' section and a '1.1. TensorRT Layers' section. The '1.1. TensorRT Layers' section lists the layers supported by TensorRT: Activation, Concatenation, Convolution, Deconvolution, ElementWise, FullyConnected, LRN, and Plugin. Each layer is described with its function and supported operations.



The screenshot shows the NVIDIA Deep Learning Institute (DLI) training page. The header includes the 'DEEP LEARNING AI' logo and navigation links like WHAT'S NEW, INDUSTRIES, FOR DEVELOPERS, PRODUCTS, EDUCATION, and ONLINE LABS. The main content area features the headline 'TEACHING YOU TO SOLVE PROBLEMS WITH DEEP LEARNING' and a photo of two people working on a computer. Below this, the NVIDIA logo and 'DEEP LEARNING INSTITUTE' are displayed. The text describes the DLI's offerings, including hands-on training for developers, data scientists, and researchers. It mentions that DLI provides training on the latest techniques for designing, training, and deploying neural networks across a variety of application domains. The page also notes that students will explore widely used open-source frameworks as well as NVIDIA's latest GPU-accelerated deep learning platforms.

# Q&A

