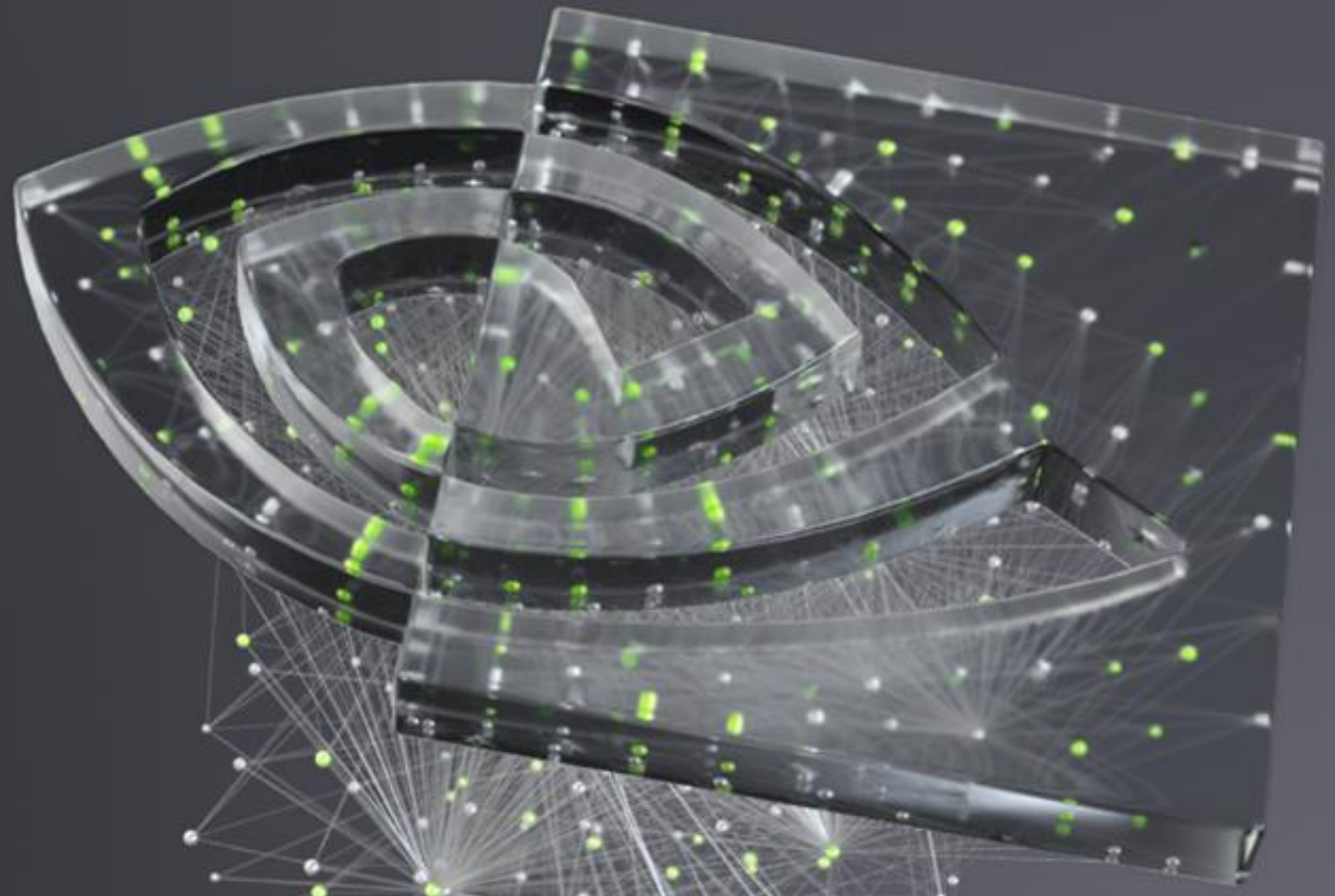


From research to production

Enabling real world applications with Conversational AI

Adam Henryk Grzywaczewski & Zenodia Charpy

About Me

Zenodia Charpy

- Senior Solution Architect @ NVIDIA – Auto & Healthcare focus on Deep Learning.
- My past experience:
  - Azure : Data Scientist & Solution Architect
  - Telia : Data Scientist

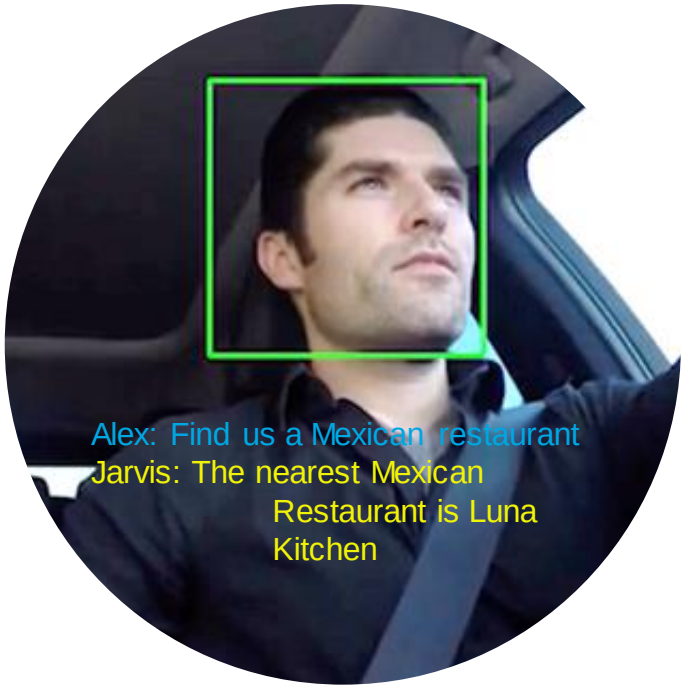
About Me

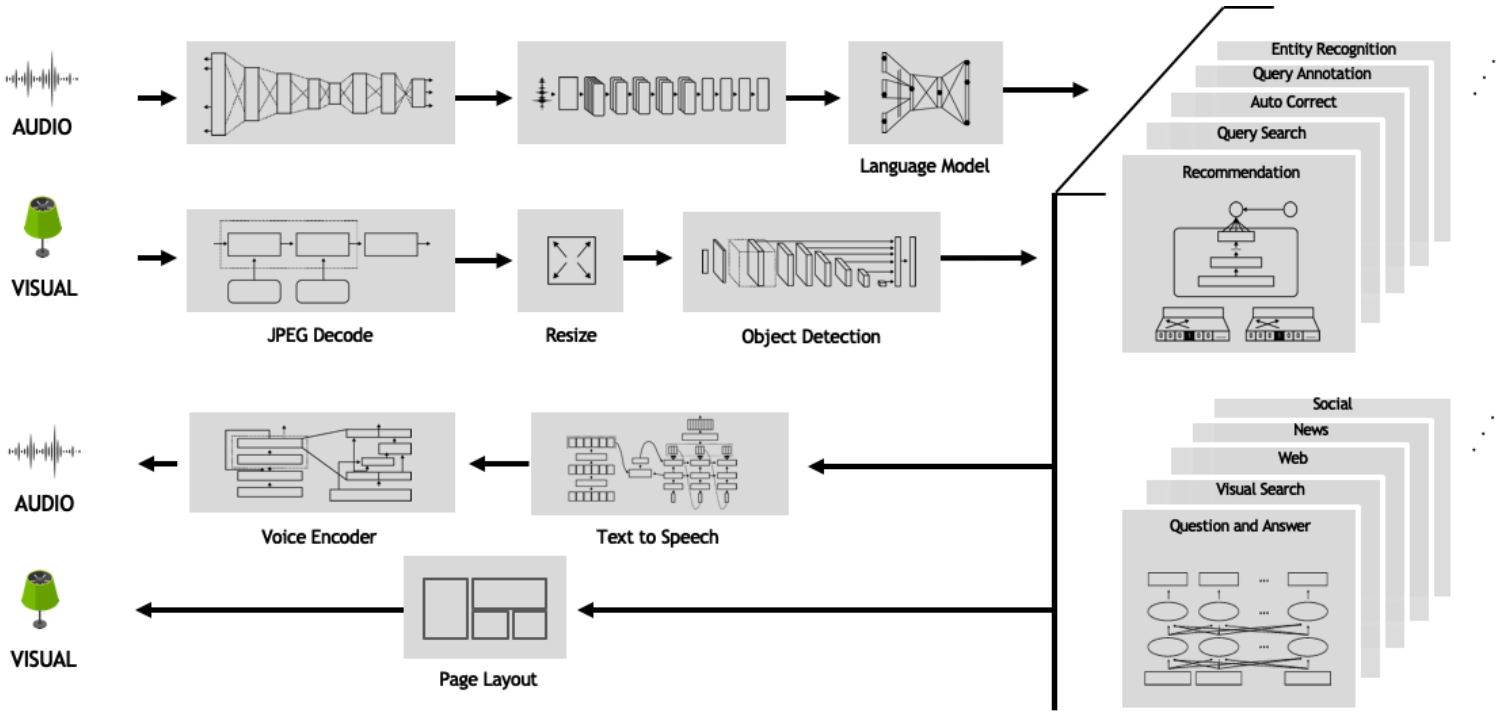
Adam Grzywaczewski

- Senior Deep Learning Data Scientist @ NVIDIA - Supporting delivery of AI / Deep Learning solutions
- Specialising in Deep Learning at scale.
- My past experience:
  - Capgemini: https://goo.gl/MzgGbq
  - Jaguar Land Rover Research: https://goo.gl/ar7LuU

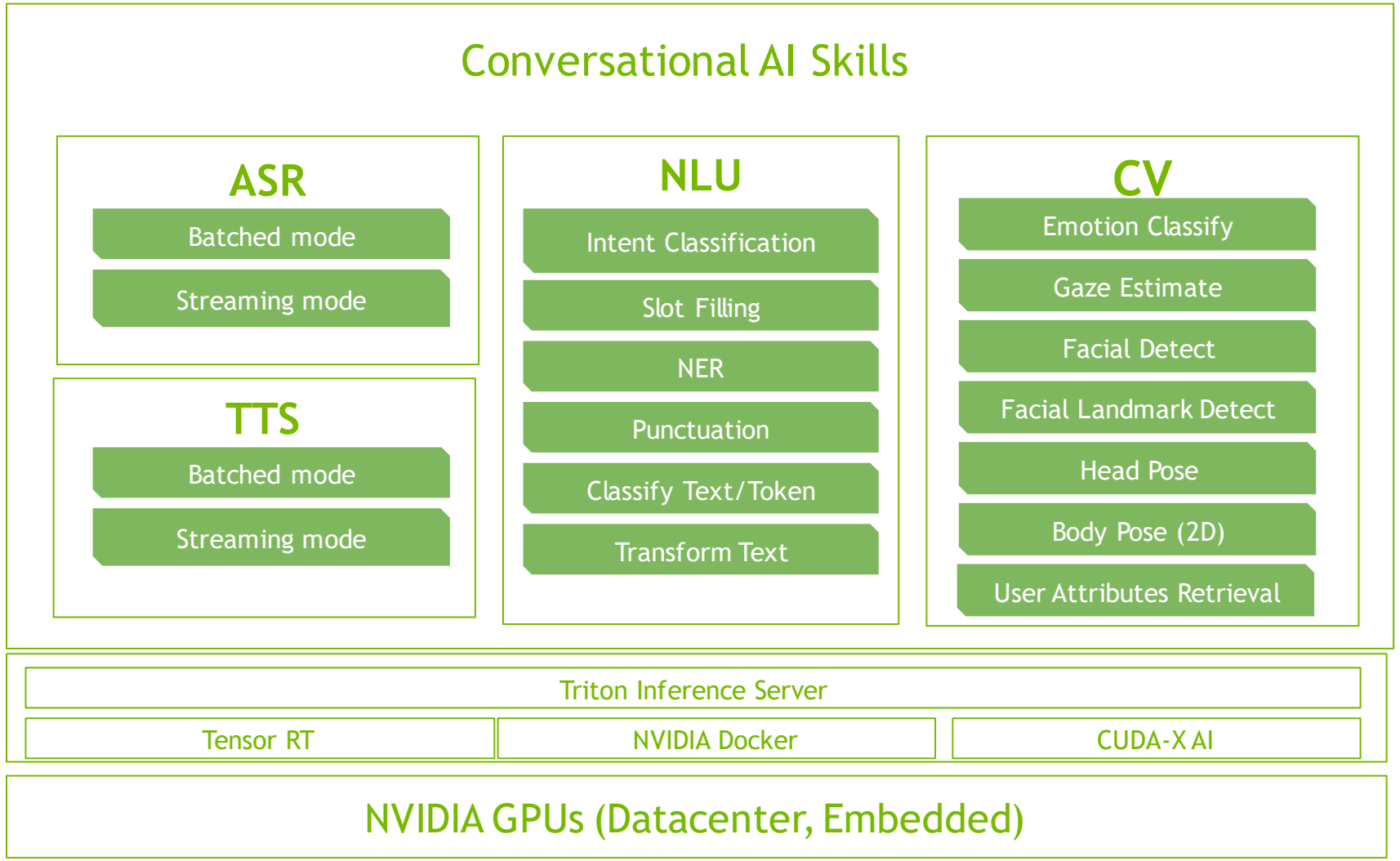
Conversational AI

A Complex System





Conversational AI

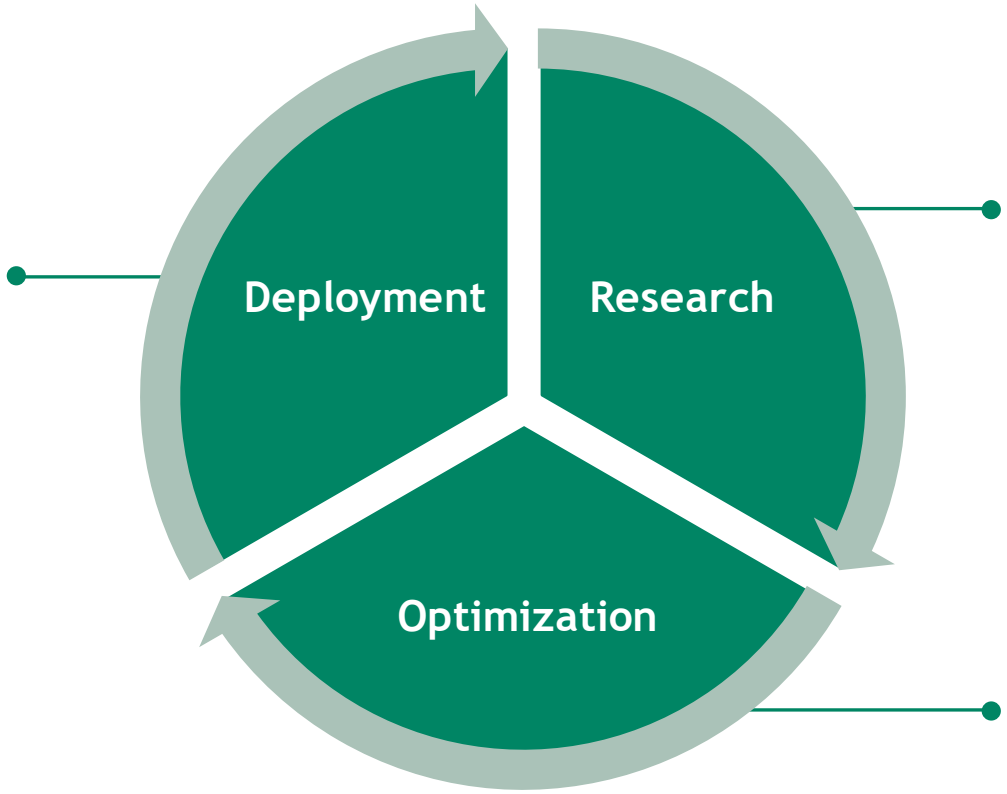


The Workflow

Research, Optimization, Deployment

Jarvis

An integrated & Scalable solution for **serving** Conversational AI models



NeMo / TLT

NeMo: Open source applied research toolkit for **researching, developing and training** conversational AI models

TLT: Transfer Learning Toolkit is a **0-coding** tool for **training** supported Conversational AI models

TensorRT + Triton

TensorRT: an SDK for high-performance deep learning inference.

Triton: Open-source inference serving software.


RESEARCH

NeMo

Open-Source Toolkit for Development of Conversational AI Models

Automatic Speech Recognition

Spoken word to text transcription

Natural Language Processing

Understanding tasks
Named Entity Recognition
Question Answering
Dialog Management
Machine Translation

Text to Speech

Text to spoken language

What is NeMo

Encapsulation of Best Practice

QUARTZNET: DEEP AUTOMATIC SPEECH RECOGNITION WITH 1D TIME-CHANNEL SEPARABLE CONVOLUTIONS

Samuel Kriman[†⋆] *Stanislav Beliaev*[‡⋆] *Boris Ginsburg* *Jocelyn Huang*
Oleksii Kuchaiev *Vitaly Lavrukhin* *Ryan Leary* *Jason Li* *Yang Zhang*

[†]Univ. of Illinois Urbana-Champaign, [‡]Hig

Jasper: An End-to-End Convolutional Neural Acoustic Model

...taly Lavrukhin[1], *Boris Ginsburg*[1], *Ryan Leary*[1], *Oleksii Kuchaiev*[1], *...onathan M. Cohen*[1], *Huyen Nguyen*[1], *Ravi Teja Gadde*[2]

[1]NVIDIA, Santa Clara, USA
[2]New York University, New York, USA
...ginsburg,rleary,okuchaiev,jocohen,chipn}@nvidia.com,rtg267@nyu.edu

...tract

...-the-art results on LibriSpeech ...gnition models without any ex- ...el, Jasper, uses only 1D con- ... ReLU, dropout, and residual ...ng, we further introduce a new ...voGrad. Through experiments, ...d deep architecture performs as ...ex choices. Our deepest Jasper ...yers. With this architecture, we ...eam-search decoder with an ex- ...nd 3.86% WER with a greedy ...an. We also report competitive ...nd the Hub5'00 conversational ...ition, convolutional networks,

It is possible to increase the capacity of the Jasper model by stacking these operations. Our largest version uses 54 convolutional layers (333M parameters), while our smaller model uses 34 (201M parameters). We use residual connections to enable this level of depth. We investigate a number of residual options and propose a new residual connection topology we call *Dense Residual (DR)*.

Integrating our best acoustic model with a Transformer-XL [12] language model allows us to obtain new state-of-the-art (SOTA) results on LibriSpeech [13] test-clean of 2.95% WER and SOTA results among end-to-end models[1] on LibriSpeech test-other. We show competitive results on Wall Street Journal (WSJ), and 2000hr Fisher+Switchboard (F+S). Using only greedy decoding without a language model we achieve 3.86% WER on LibriSpeech test-clean.

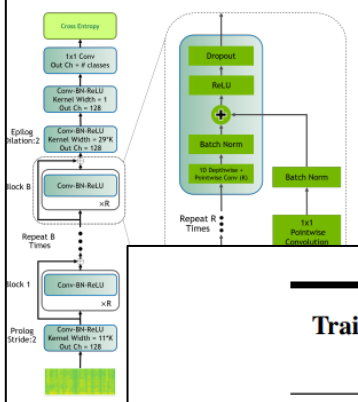
This paper makes the following contributions:

1. We present a computationally efficient end-to-end con-

MatchboxNet: 1D Time-Channel Separable Convolutional Neural Network Architecture for Speech Commands Recognition

Somshubra Majumdar, Boris Ginsburg

...lara, USA
...}@nvidia.com

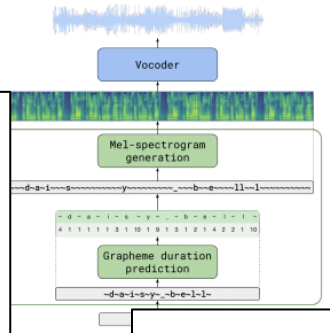


TalkNet: Fully-Convolutional Non-Autoregressive Speech Synthesis Model

...eliaev[1⋆], *Yurii Rebryk*[1], *Boris Ginsburg*

...NVIDIA, Santa Clara, USA
...l of Economics, Saint Petersburg, Russia
...sburg}@nvidia.com, y.a.rebryk@gmail.com

...utoregressive ...el consists of ... first network ... expanded by



Cross-Language Transfer Learning, Continuous Learning, and Domain Adaptation for End-to-End Automatic Speech Recognition

Jocelyn Huang[1], *Oleksii Kuchaiev*[1], *Patrick O'Neill*[2], *Vitaly Lavrukhin*[1], *Jason Li*[1], *Adriana Flores*[1], *Georg Kucsko*[2], *Boris Ginsburg*[1]

[1]NVIDIA, Santa Clara, CA, USA
[2]Kensho, Cambridge, MA, USA
{jocelynh,okuchaiev,vlavrukhin,jasoli,adrianaf,bginsburg}@nvidia.com,

...ess.AS] 8 May 2020

In thi... and c... nition... mode... easily... langu... speci... cases... curac... tune... datas... nifica... large... **Index**... langu...

CORRECTION OF AUTOMATIC SPEECH RECOGNITION WITH TRANSFORMER SEQUENCE-TO-SEQUENCE MODEL

Oleksii Hrinchuk[12⋆] *Mariya Popova*[2⋆] *Boris Ginsburg*[2]

[1]Moscow Institute of Physics and Technology, Moscow, Russia
[2]NVIDIA, Santa Clara, CA, USA
{aleksey.grinchuk@phystech.edu, mariewelt@gmail.com, bginsburg@nvidia.com}

...10697v1 [cs.CL] 23 Oct 2019

ABSTRACT

In this work, we introduce a simple yet efficient post-processing model for automatic speech recognition (ASR). Our model has Transformer-based encoder-decoder architecture which "translates" ASR model output into grammatically and semantically correct text. We investigate different strategies for regularizing and optimizing the model and show that extensive data augmentation and the initialization with pre-trained weights are required to achieve good performance. On the LibriSpeech benchmark, our method demonstrates significant improvement in word error rate over the baseline acoustic model with greedy decoding, especially on much noisier dev-other and test-other portions of the evaluation dataset. Our model also outperforms baseline with 6-gram language model re-scoring and approaches the performance of re-scoring with Transformer-XL neural language model.

Index Terms— speech recognition, spelling correction, pre-trained language models

Language model re-scoring effectively expands the search space of speech recognition candidates; however, it can barely help when the ground truth word was assigned a low score by erroneous ASR model. Traditional left-to-right language models are also prone to error accumulation: if some word at the beginning of the decoded speech is misrecognized, it will affect the scores of all succeeding words by providing them with incorrect context. To address these problems, we propose to train a conditional language model that corrects the errors made by the system operating similar to neural machine translation (NMT) [10, 11] by "translating" corrupted ASR output into the correct language.

There is a plethora of prior work on correcting ASR systems output, and we refer the reader to [12] for a detailed overview. Most closely to our work, [13] propose to train a spelling correction model based on RNN with attention [14] to correct the output of Listen, Attend and Spell (LAS) model. In contrast to this work, our model is based on Transformer

Training Neural Speech Recognition Systems with Synthetic Speech Augmentation

Jason Li, Ravi Gadde*, Boris Ginsburg, Vitaly Lavrukhin

...hin}@nvidia.com

...ASR) system requires a large ...amples produced by a diverse ...ets is one of the main issues ...s this problem, we propose to ...ch. We train very large end-to- ...briSpeech dataset augmented ...ate of the art Word Error Rate ...external language model.

NeMo: a toolkit for building AI applications using Neural Modules

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, Jonathan M. Cohen

NVIDIA
Santa Clara, CA
{okuchaiev,jasoli,chipn,ohrinchuk,rleary,bginsburg,skriman,stanislavv, vlavrukhin,jocook,pcastonguay,mpopova,jocelynh,jocohen}@nvidia.com

...Sep 2019

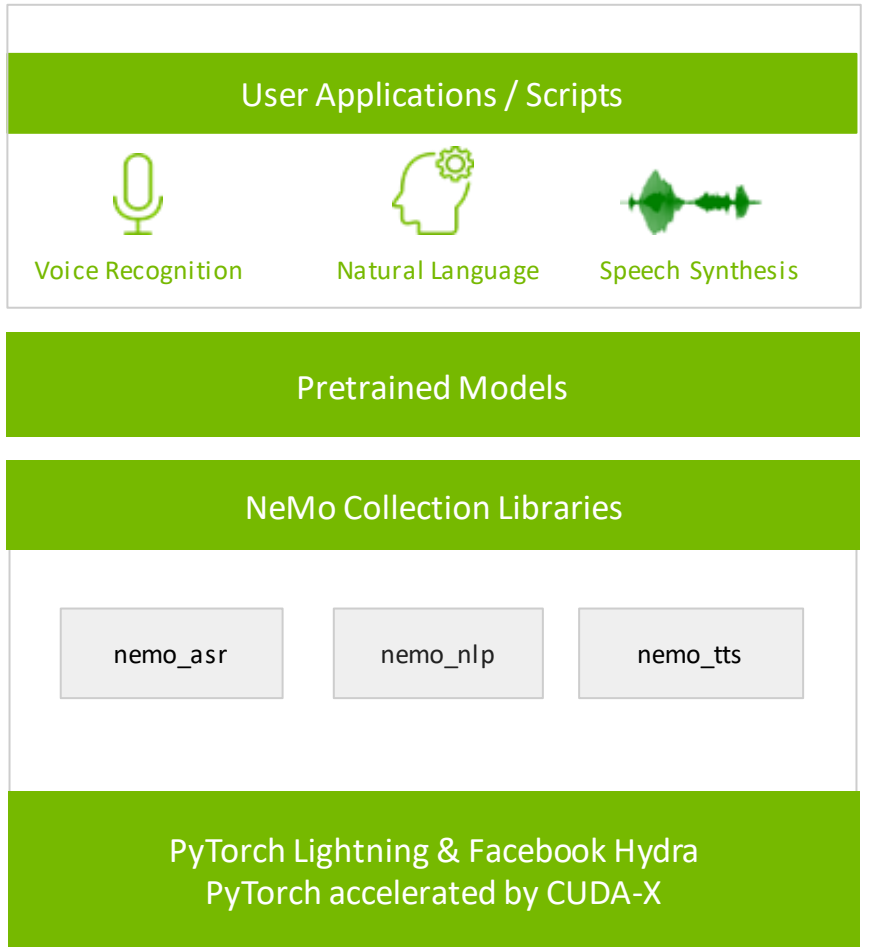
NVIDIA

Broad adoption

Welcome Contribution and Enable Collaboration

Enabling R&D

Build for Flexibility and Ease-of-Use

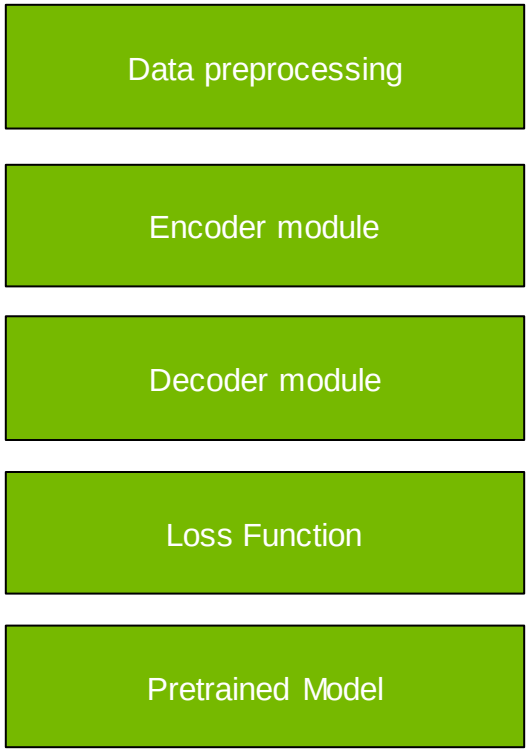


github.com/NVIDIA/NeMo

Neural Modules

Lego-Like Building Blocks Enabling Fast Experimentation

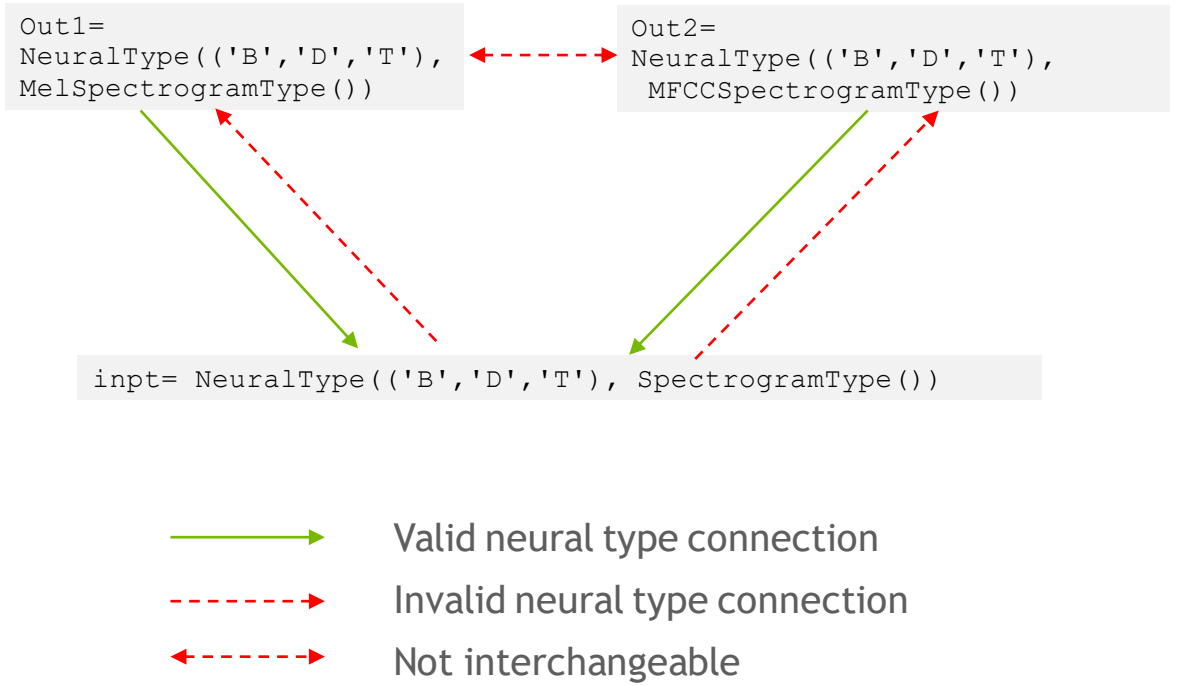
- Neural Modules, building blocks of conversational AI
- Typed inputs and outputs
- Easy management of experiments (Hydra)
- Integration with PyTorch Lightning
- Highly scalable and performant (FP16, Distributed)
- Extensive collection of pretrained models



Hightlight 1

Strong Typed Tensors

- Ensure compatibility between modules
- Catch semantic, rank and dimensionality mismatches
- Simplifing the debug process



Tutorial notebooks available at https://github.com/NVIDIA/NeMo

Hightlight 2

Hydra：Simplify Complex Model Development

Flexible approach for developers to configure and customize the model

Single-stop solution for editing the end-to-end neural network

Configurable with both YAML and Hydra CLI commands

quartznet_15x5_aug.yaml

```
#specify the name of the model you want to use
name: &name "QuartzNet15x5"
model:
 sample_rate: &sample_rate 16000
 repeat: &repeat 5
 dropout: &dropout 0.0
 separable: &separable true
 labels: &labels [" ", "a", "b", "c", "d", "e", "f",
"g", "h", "i", "j", "k", "l", "m", "n", "o", "p",
"q", "r", "s", "t", "u", "v", "w", "x", "y", "z",
"'"]
#manage training data parameters
 train_ds:
   manifest_filepath: ???
   sample_rate: 16000
   ...
#manage validation data parameters
 validation_ds:
   manifest_filepath: ???
   sample_rate: 16000
   ...

 ...
```

QuartzNet Model Customization with .YAML file

Hightlight 3

Up to 4.5x Faster Training on Single GPU, Scale to Multiple GPUs Easily

Tight integration with PyTorch Lightning Trainer to easily invoke training actions.

Scale to multi-GPU and multi-node to speed-up training while retaining the accuracy

Speed-up training up to 4.5X on a single GPU with mixed-precision versus FP32 precision

Ease to use parameters to enable Multi-GPU/node training and mixed-precision

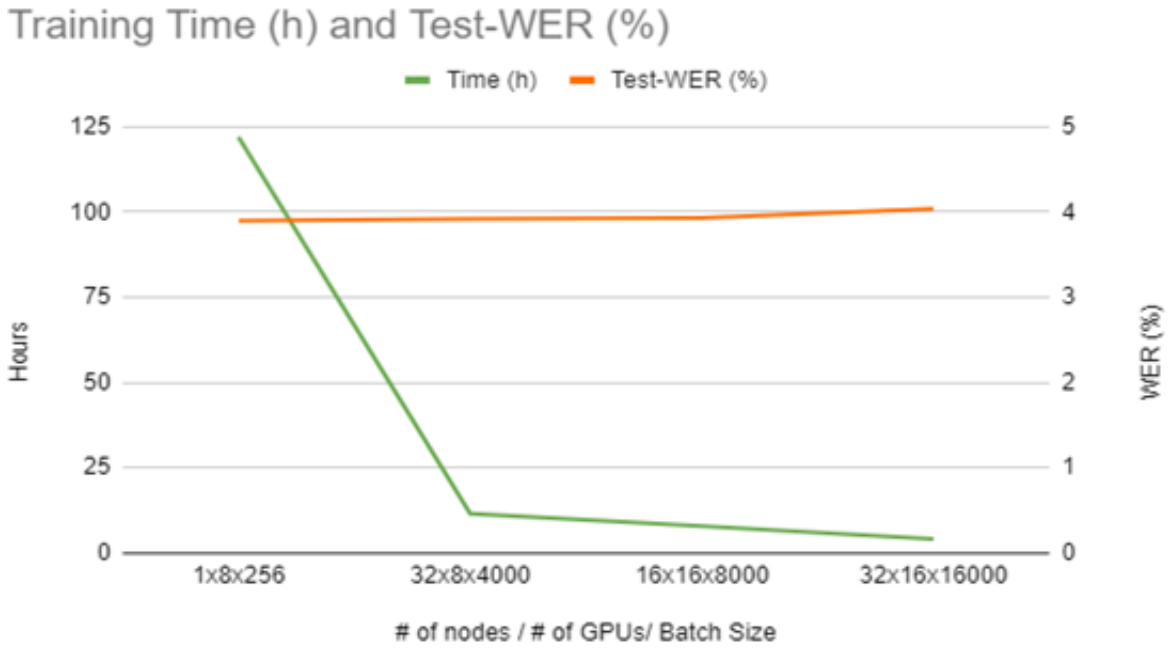
```
trainer = pl.Trainer(**cfg.trainer)
asr_model = EncDecCTCModel(cfg=cfg.model,
trainer=trainer)
trainer.fit(asr_model)
```

Training NeMo model with PyTorch Lightning Trainer API

https://docs.nvidia.com/deeplearning/sdk/mixed-precision-training/index.html

Hightlight 3

Training at Scale - Multi GPU and Multi Node Training



- Reduce total training time
- Distribute workload onto multiple compute instances with a single parameter change
- Maintain very high accuracy (Word Error Rate)

Model Zoo

Key Area of Focus

Speech Processing

- Automatic Speech Recognition
- Speech Classification
- Speaker Recognition
- Speaker Diarization

Natural Language Processing

- Punctuation & Capitalization
- Token Classification (NER)
- Joint Intent and Slot Classification
- Text Classification
- Question Answering
- Dialogue State Tracking
- Information Retrieval
- Machine Translation
- Language Modelling (other tasks)

Text to speech

- Two stage pipelines
- End to end pipelines

Model Zoo

ASR

https://docs.nvidia.com/deeplearning/nemo/

Model Zoo

Other Speech

Speech Command Recognition
The task of classifying an input audio pattern into a discrete set of classes.

Audio Sentiment Classification
Extends the conventional text-based sentiment analysis to depend on the acoustic features extracted from speech.

Speaker Identification

Who is speaking?

Speaker Verification

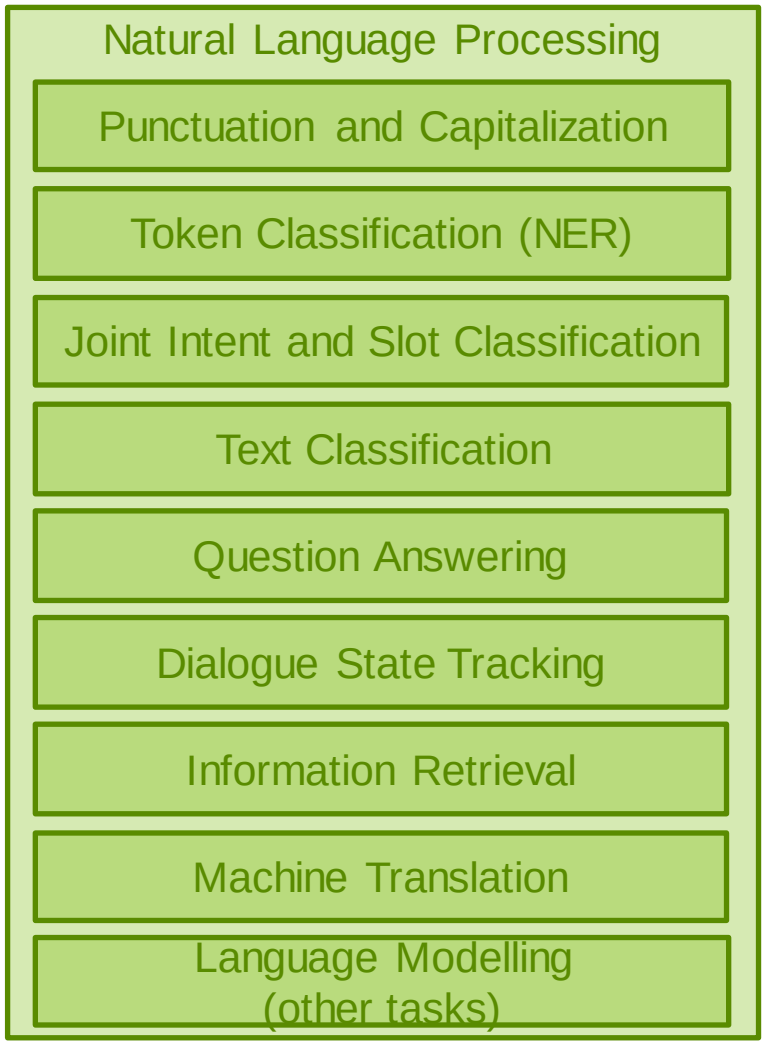
Is the speaker who they claim to be?

Voice Activity Detection (VAD)
The task of predicting which parts of input audio contain speech versus background noise.

Many other
….

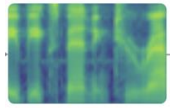
Model Zoo

NLP



Model Zoo

TTS



Two Stage Pipeline

Mel-Spectrogram Generation

Audio Generation (Vocoder)

End to End Pipeline

Audio Generation (TextToWaveform)

Model Zoo

TTS

Two Stage Pipeline

Mel-Spectrogram Generation
- Tacotron2
- GlowTTS
- FastSpeech2
- FastPich
- TalkNet

Audio Generation (Vocoder)
- WaveGlow
- SqueezeWave
- UniGlow
- MelGAN
- HiFiGAN

End to End Pipeline

Audio Generation (TextToWaveform)
- Wavenet
- DeepVoice 3
- 2 Stages in End-2-End
  - FastPitch_HifiGan_E2E
  - FastSpeech2_HifiGan_E2E

What’s Next

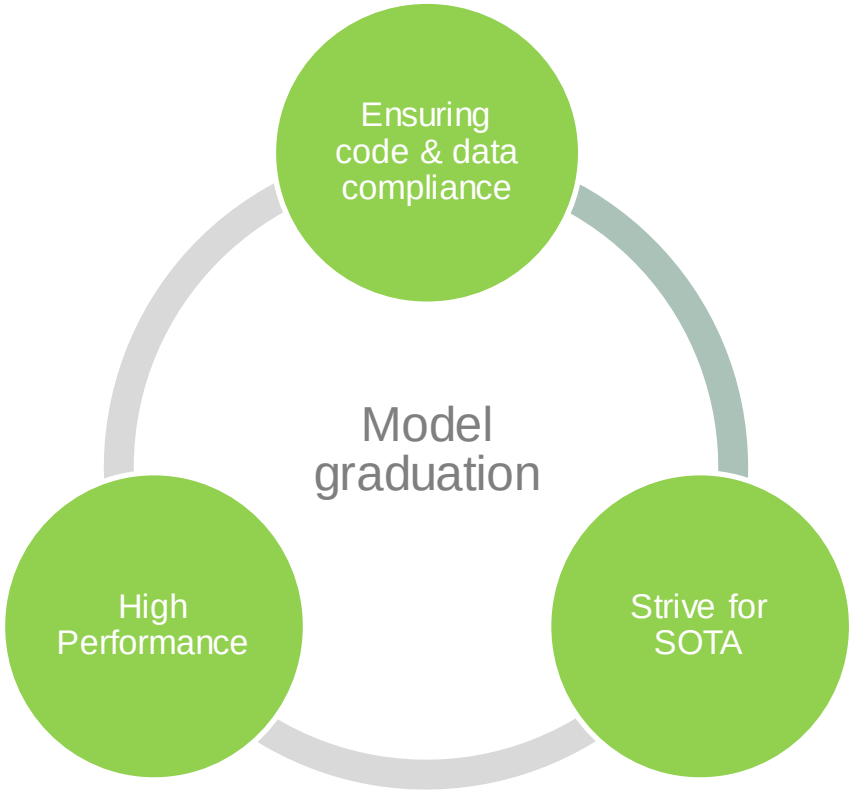
Zenodia Charpy

- Model Graduation
- Optimisation and deployment

GRADUATING THE MODEL

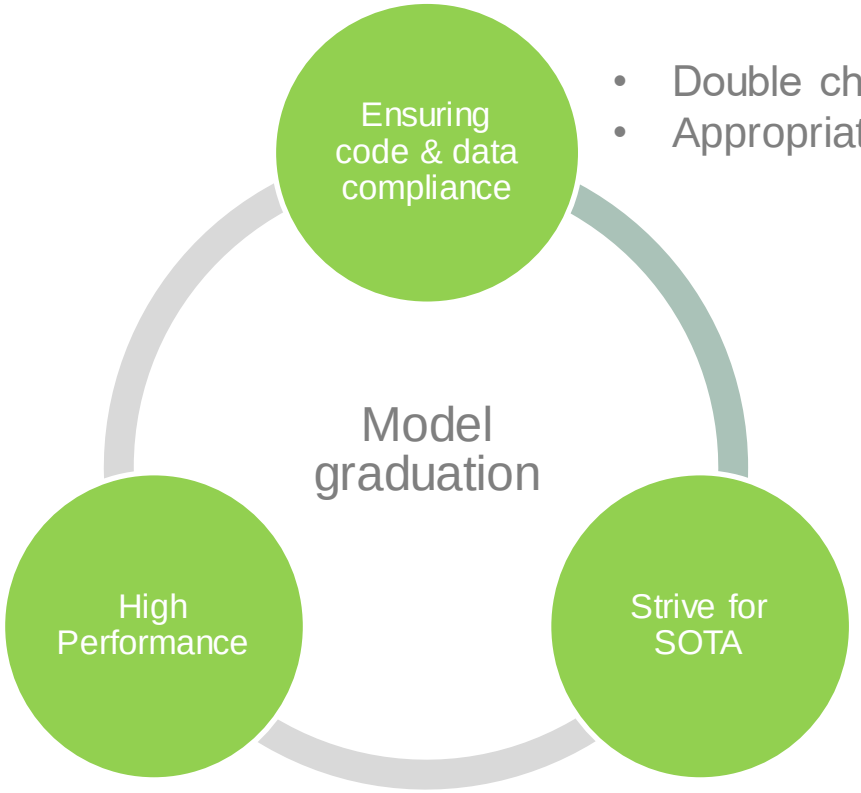
Graduating the model

Three Aspects of The Problem



Ensuring code & data compliance

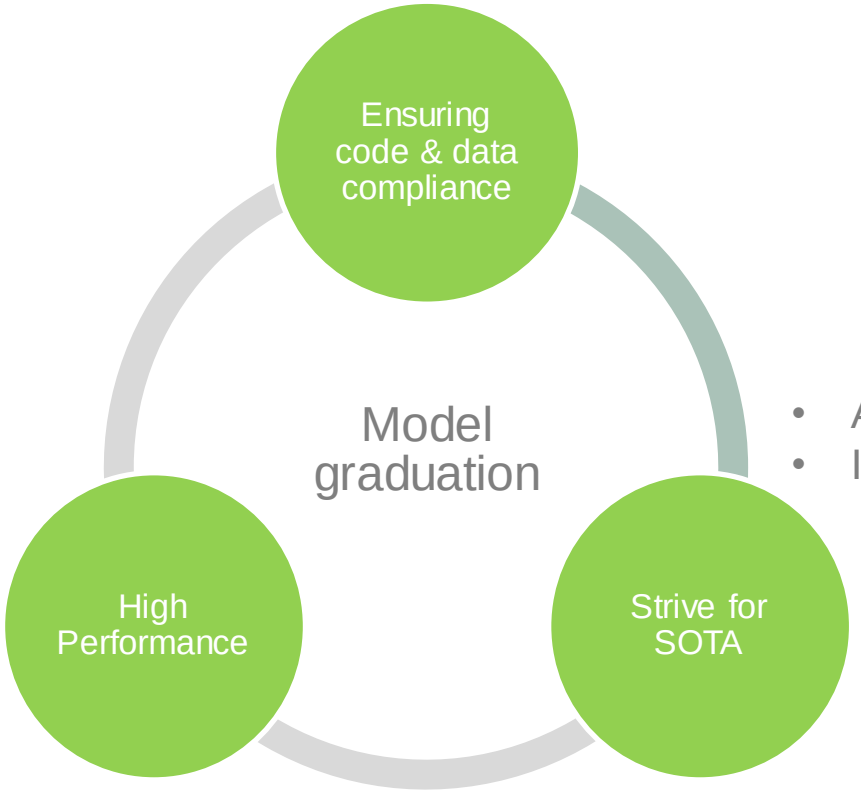
Crediting | Citing | License Compliance



- Double check datasets' legal compliance
- Appropriate crediting & citing in NeMo code

Striving towards State Of The Art

Benchmark | Iterative Improvement

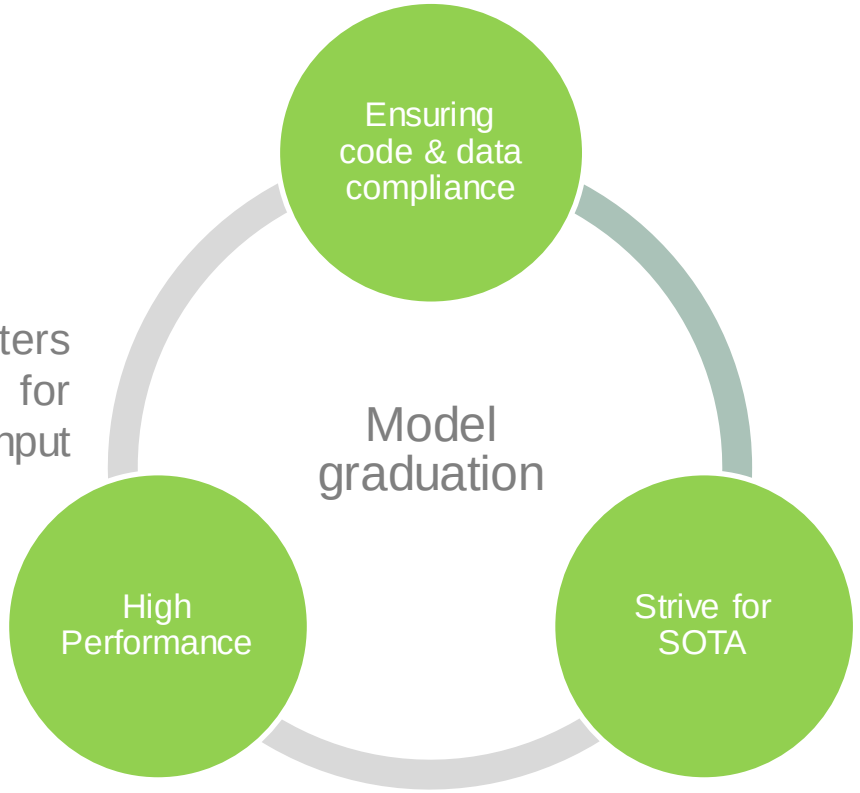


- Authoritive test dataset for benchmarking
- Iterate until SOTA is reached

Ensuring high performance

Code Quality | Model Size | Throughput

- Code quality matters
- Shrink model size & post-quantization for throughput

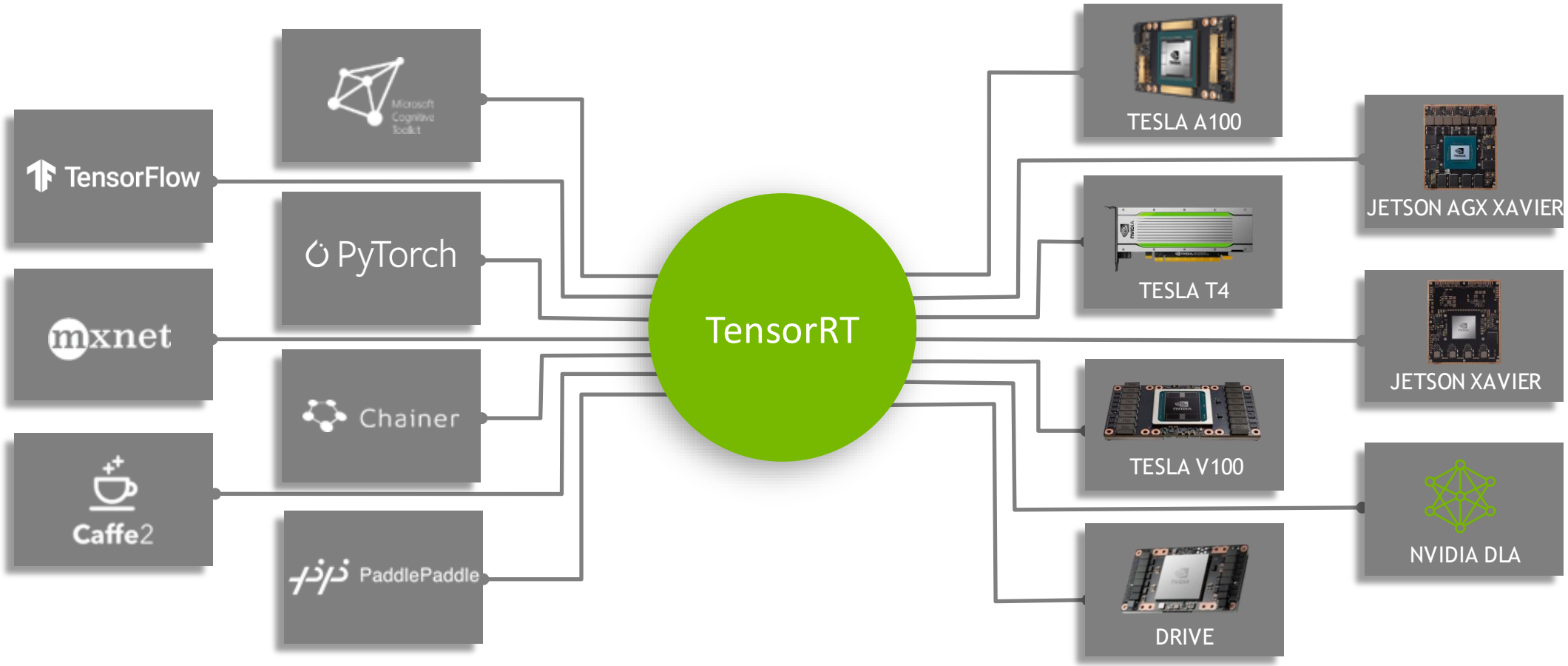


Jasper – 300M https://arxiv.org/pdf/1904.03288.pdf
Quartznet – 20M parameters https://arxiv.org/pdf/1910.10261.pdf

OPTIMIZATION

NVIDIA TensorRT

From Every Framework, Optimized for Each Target Platform



Why should we optimize (with TensorRT)

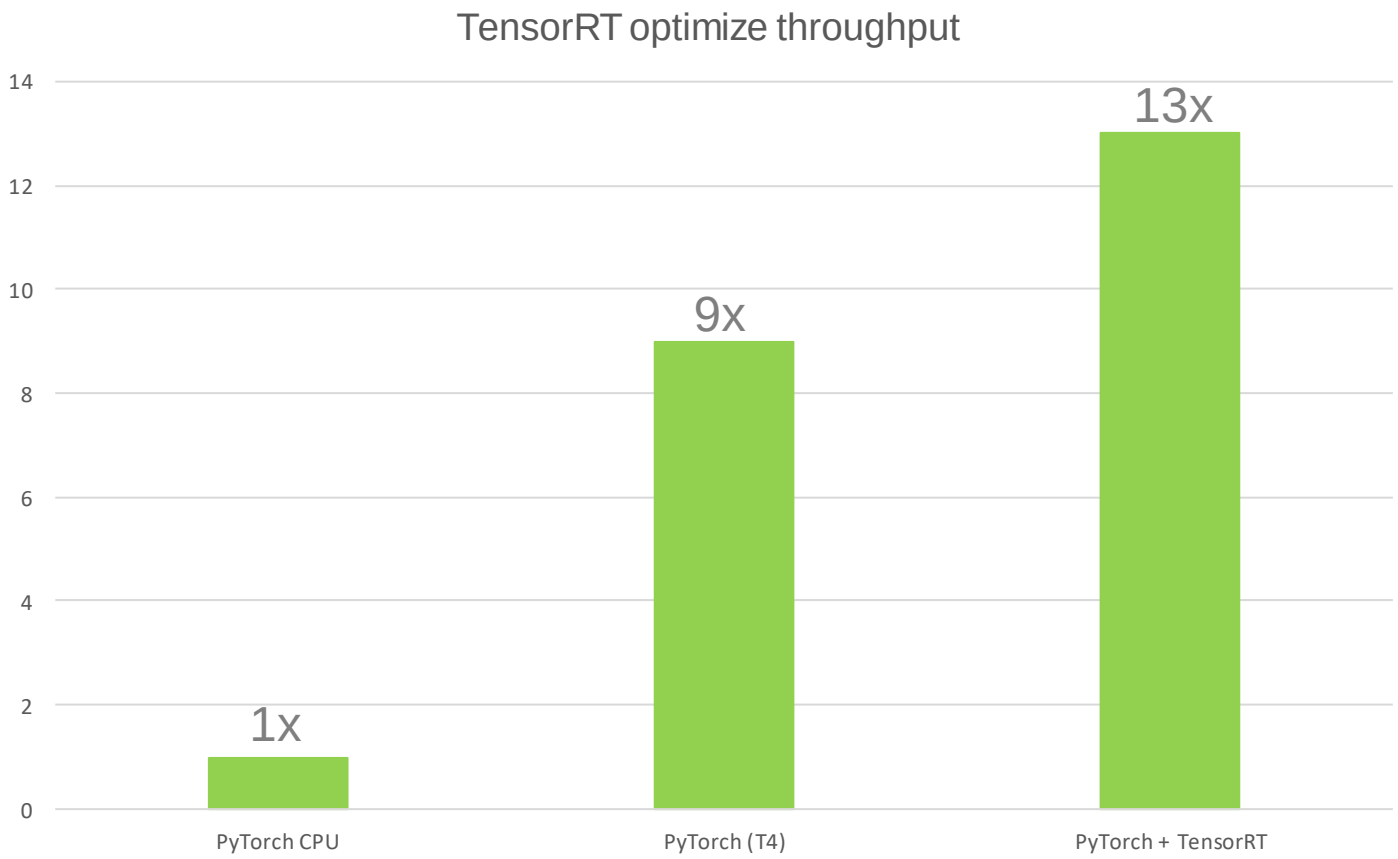
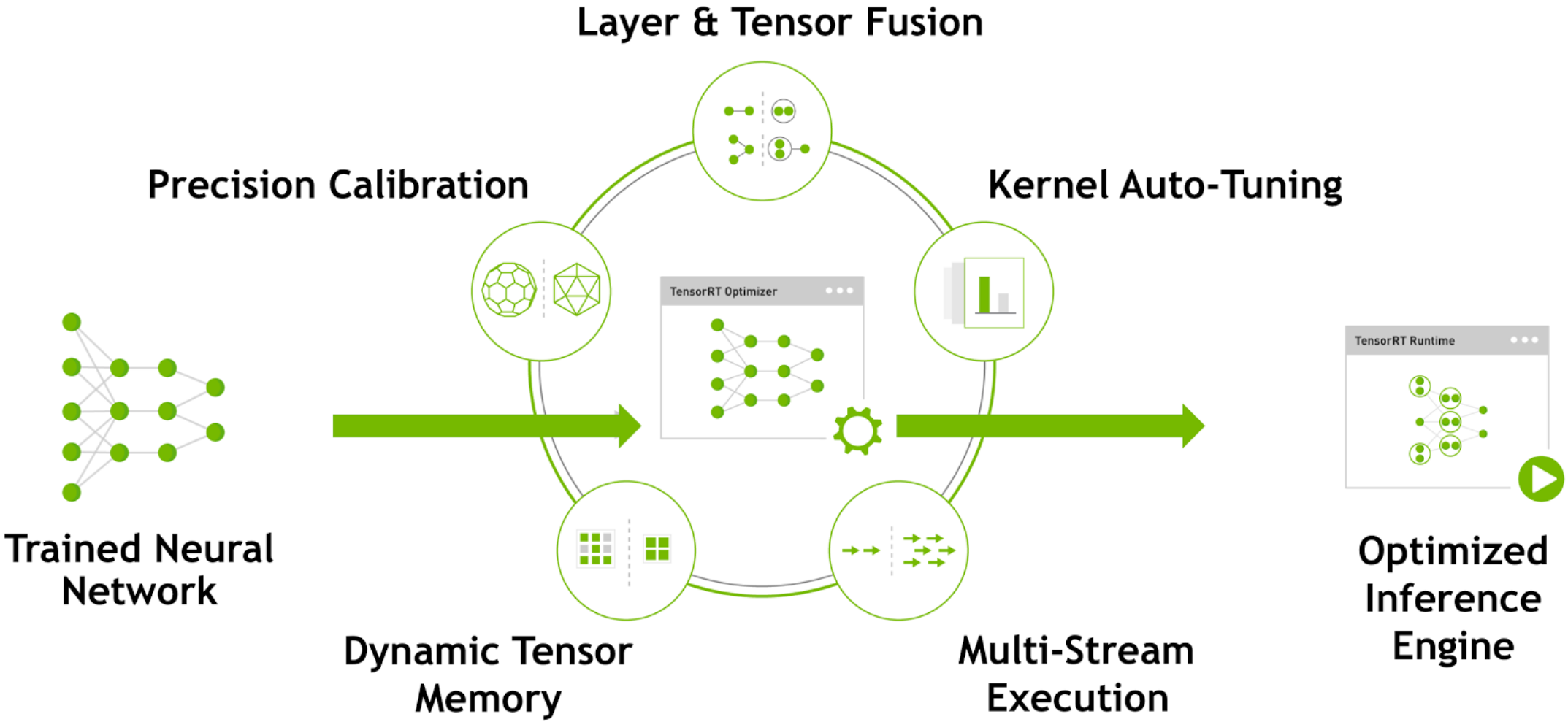


Table 1: Comparison of PyTorch and TensorRT TTS inference latencies on 1xNVIDIA T4 GPU

Source : https://developer.nvidia.com/blog/how-to-deploy-real-time-text-to-speech-applications-on-gpus-using-tensorrt/

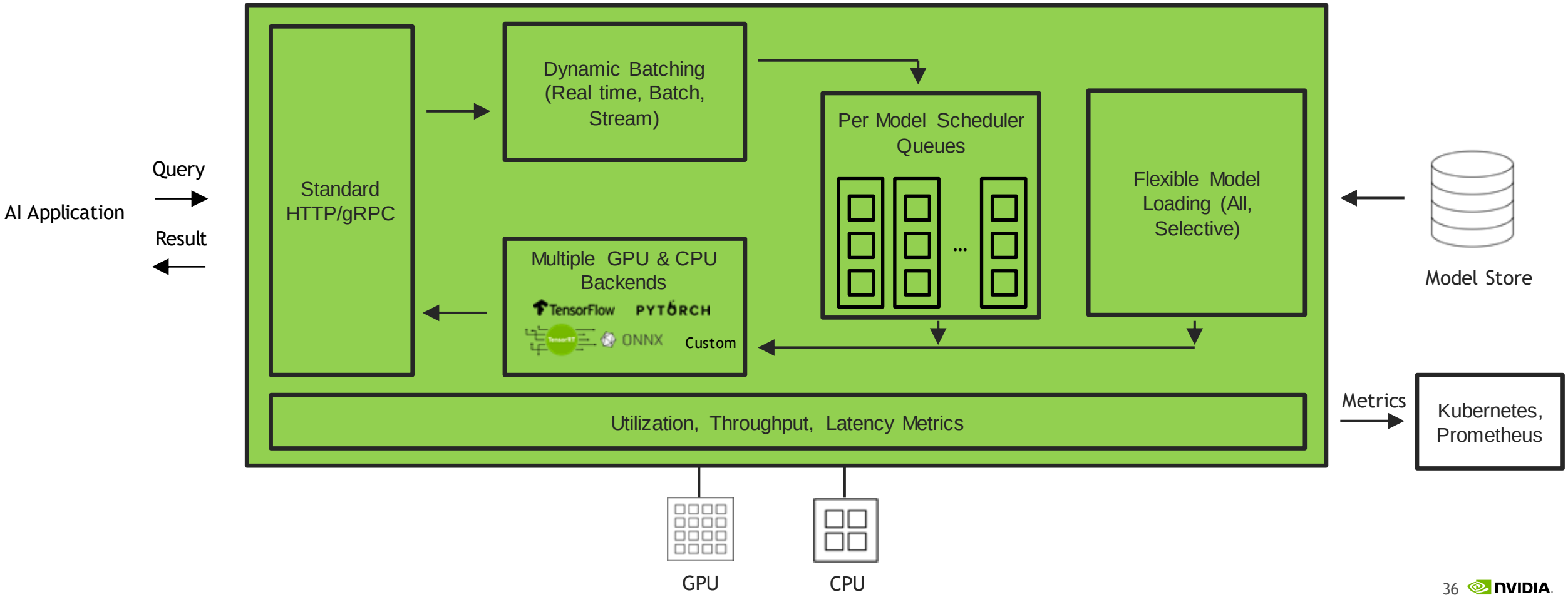
What are we optimizing



DEPLOYMENT

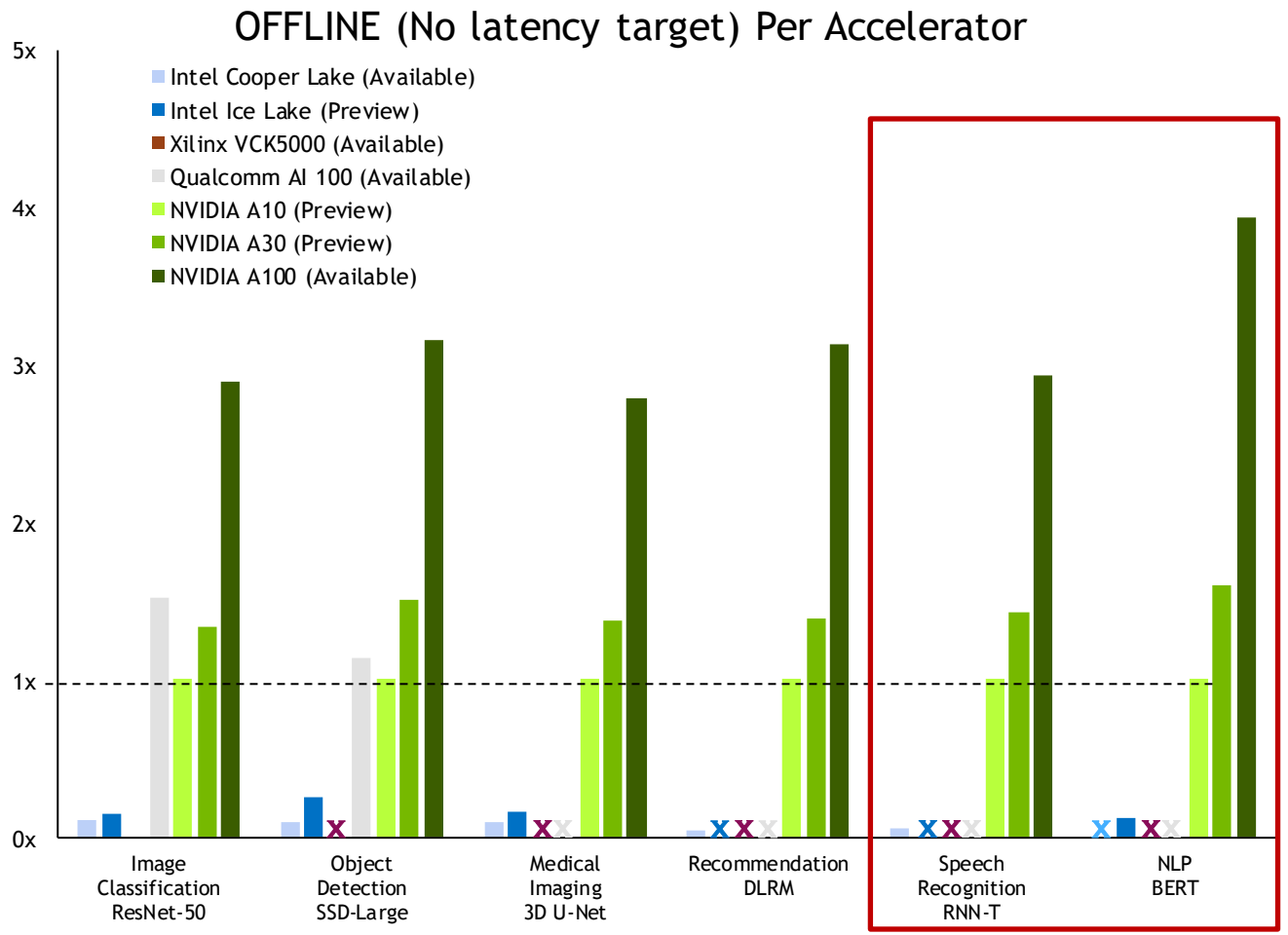
Deploy & Serve with Triton Inference Server

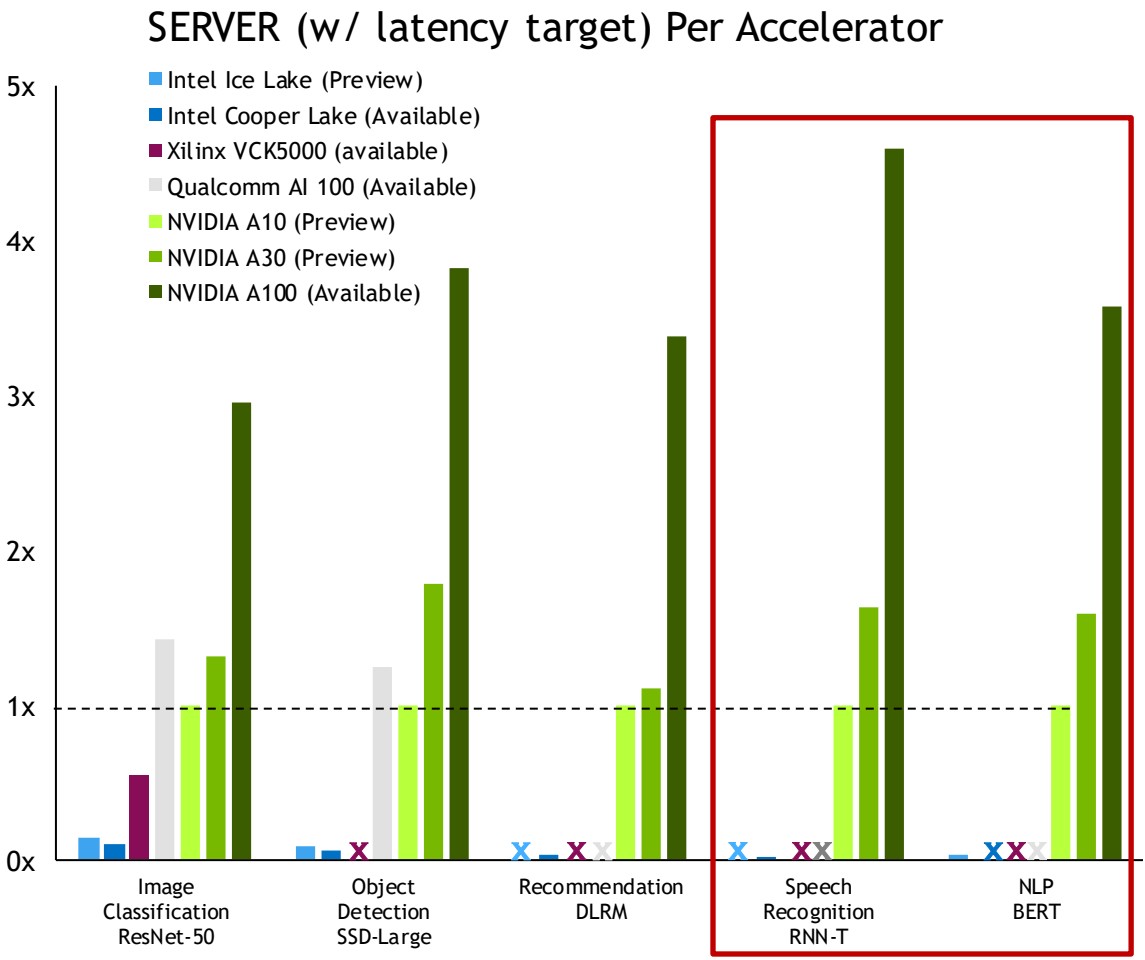
Open-Source Software for Scalable Inference Serving



NVIDIA Top MLPERF Data Center Benchmarks

GPU is Faster Than CPU



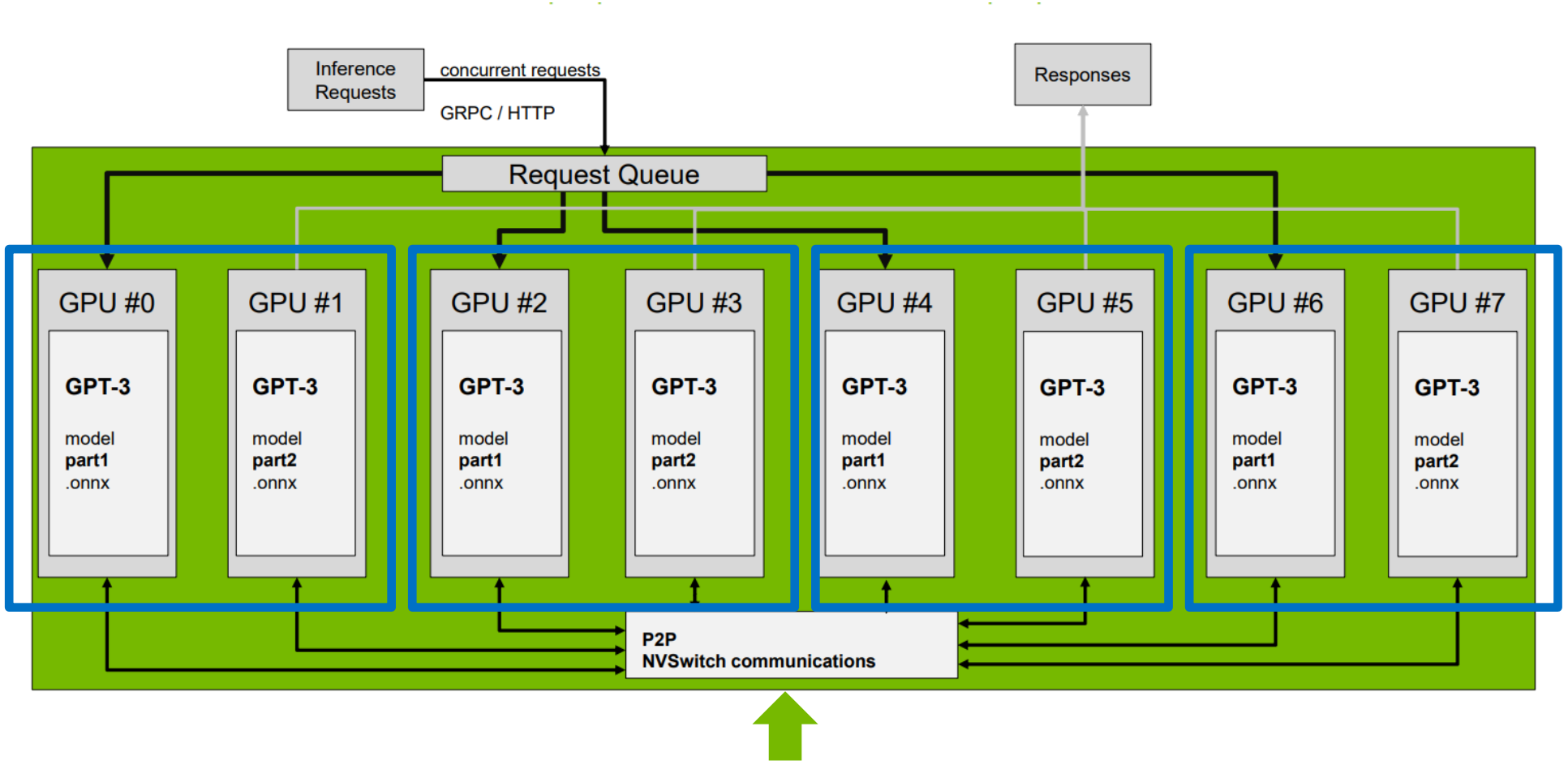


MLPerf v1.0 Inference Closed; Per-accelerator performance derived from the best MLPerf results for respective submissions using reported accelerator count in Data Center Offline and Server. 3D U-Net 99%,: 1.0-19, 1.0-53, 1.0-54, 1.0-56, 1.0-30 ResNet-50: 1.0-17, 1.0-53, 1.0-41, 1.0-35, 1.0-54, 1.0-56, 1.0-30, RNN-T: 1.0-20, 1.0-54, 1.0-56, 1.0-30 SSD-Large: 1.0-17, 1.0-53, 1.0-35, 1.0-54, 1.0-56, 1.0-30 DLRM 99%: 1.0-18, 1.0-54, 1.0-56, 1.0-30, BERT 99%: 1.0-52, 1.0-54, 1.0-56, 1.0-30. MLPerf name and logo are trademarks. See www.mlcommons.org for more information.

X = No result submitted

Deploying large NLP models

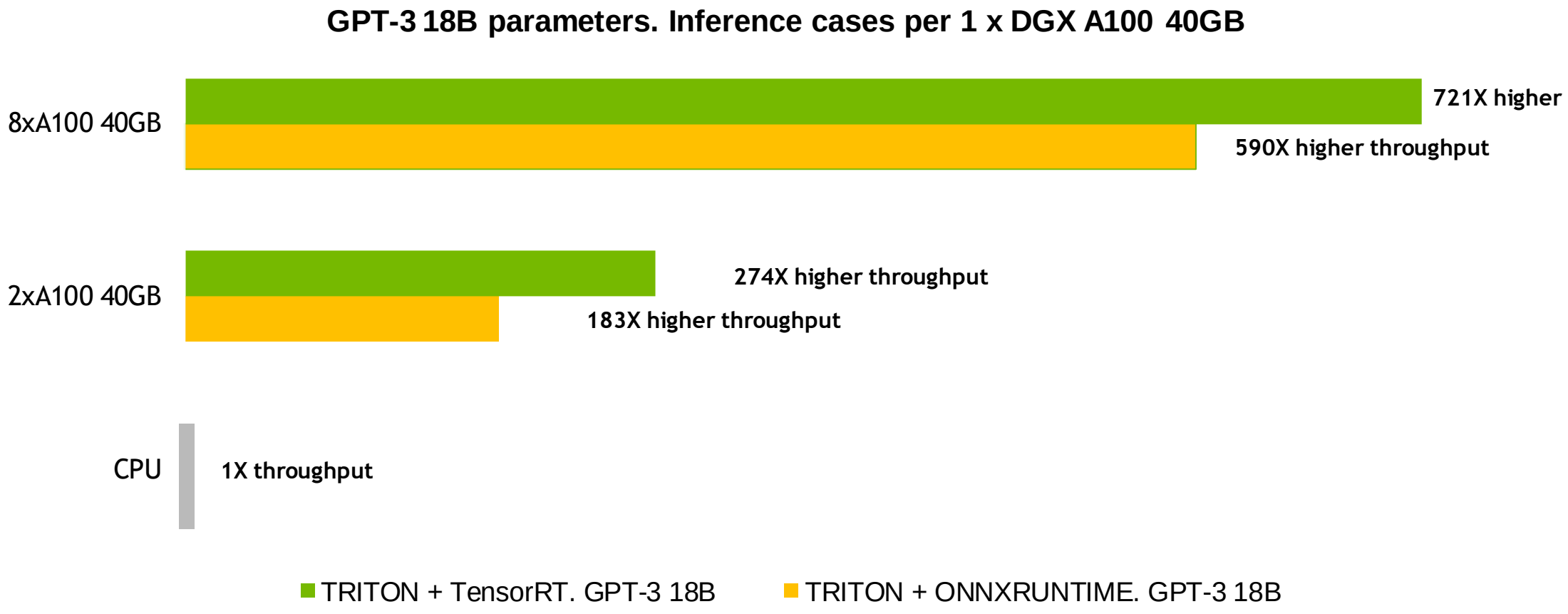
Triton Inference Server Hosting 4 Copies of GPT-3 18B on DGX A100



NVLinks to connect different parts of GPT-3

Deploying large NLP models

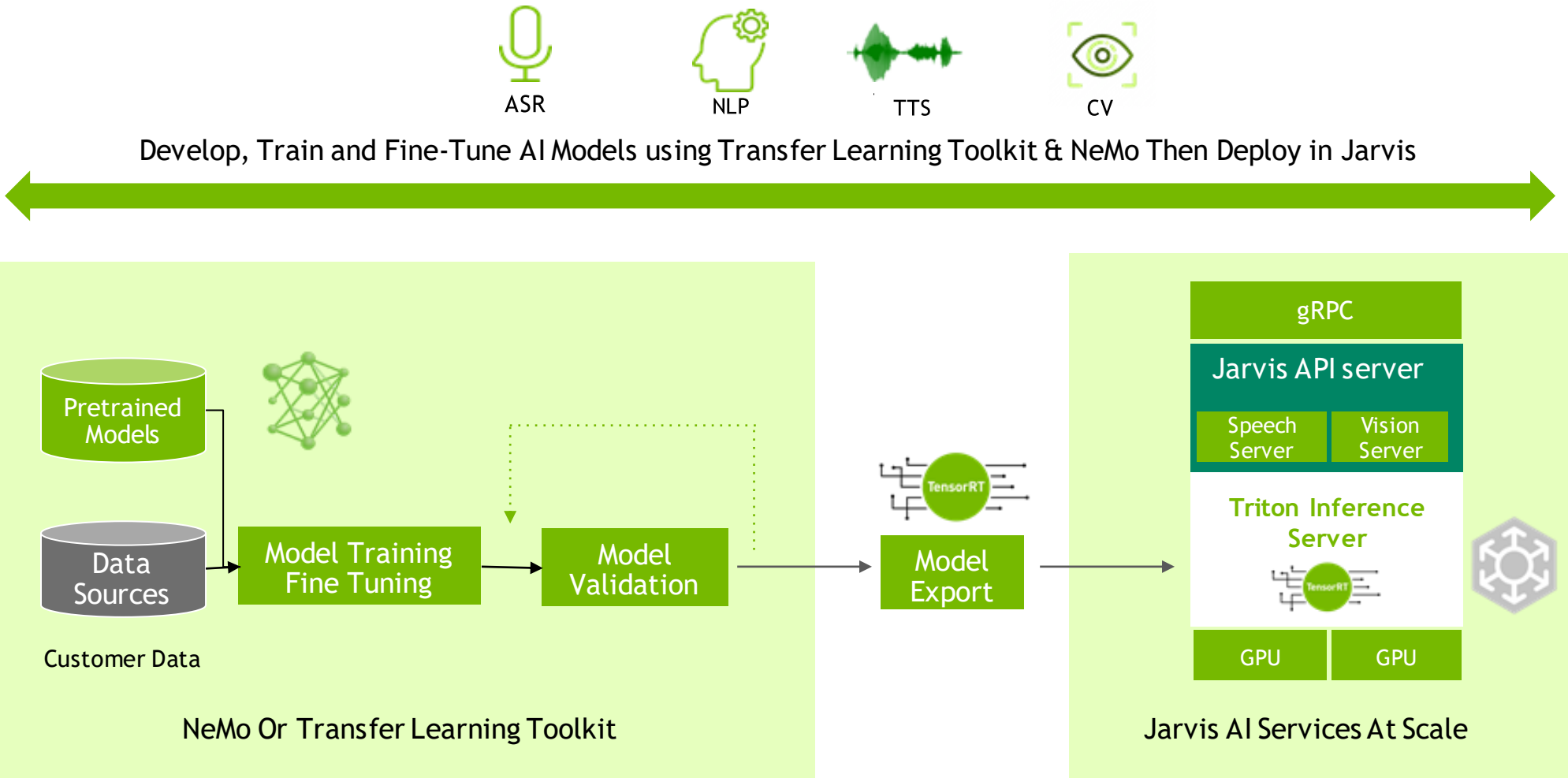
TensorRT + Triton Inference Server Deliver the Highest Throughput



Inference throughput comparisons (requests /per second)
GPT-3 MEGATRON-LM 18B parameters. Seq_length=1024.

**CPU case: Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost), FP32. Container: nvcr.io/nvidia/pytorch:21.03-py3 OnnxRuntime-CPU version.*
Code was not properly optimized for this processor, so with better optimization, the difference in results between GPU and CPU may differ multiple times.
GPU: 8xA100 for 4x models in parallel. FP16. Container: nvidia/tritonserver:21.03-py3

Put It All Together = Multimodal Conversational AI



What's Next

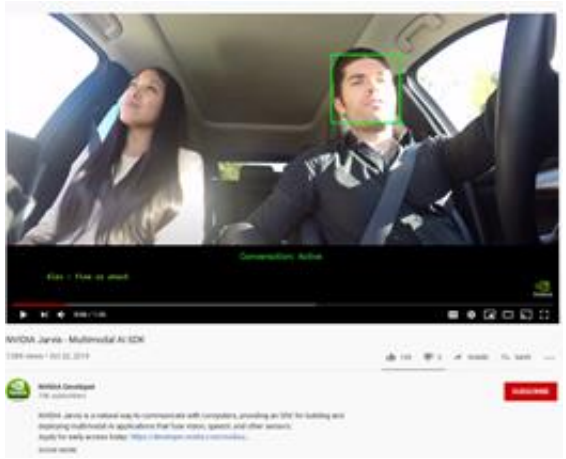
Adam Grzywaczewski

- Packaging
- Final remarks

PACKAGING

NVIDIA Jarvis use cases

A Flexible & Extendible Framework for Conversational AI



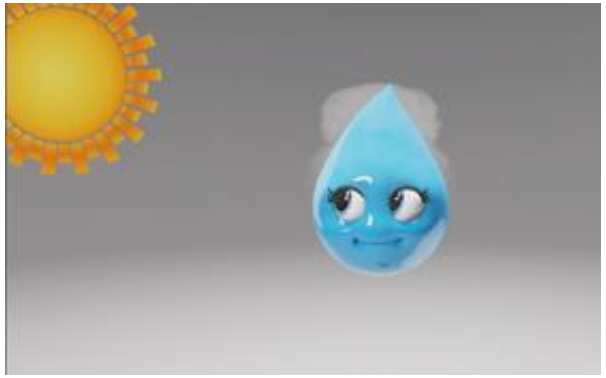
In-car experience (@Youtube)

Chatbot with A2F
(JHH Keynote @Youtube)

Misty: making of 3D AI assistant (@Youtube)

Virtual assistant with multi-domain conversation (recorded)

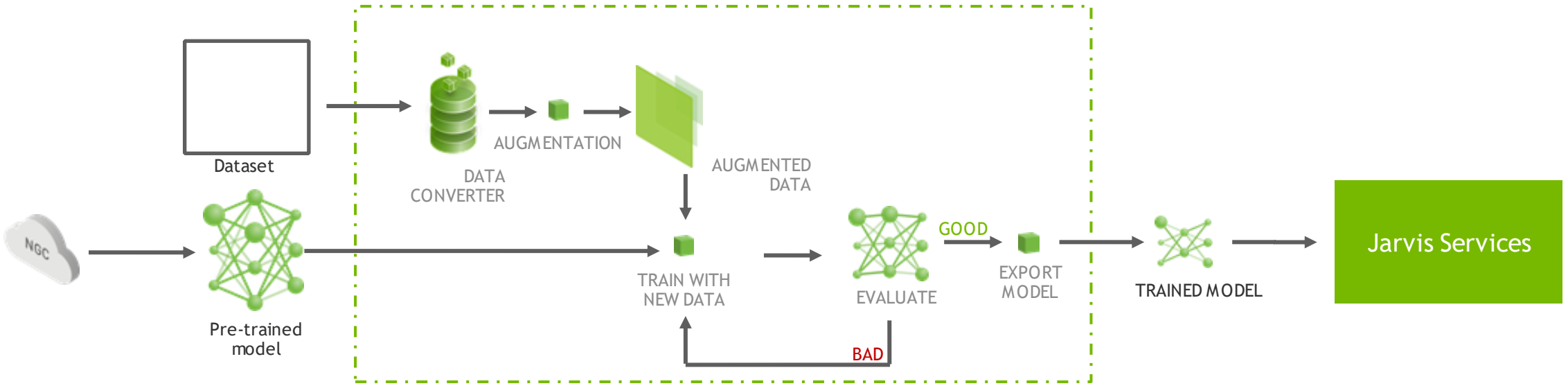
Call center transcription and annotation
(recorded)



Virtual assistant with digital avatar
(recorded)

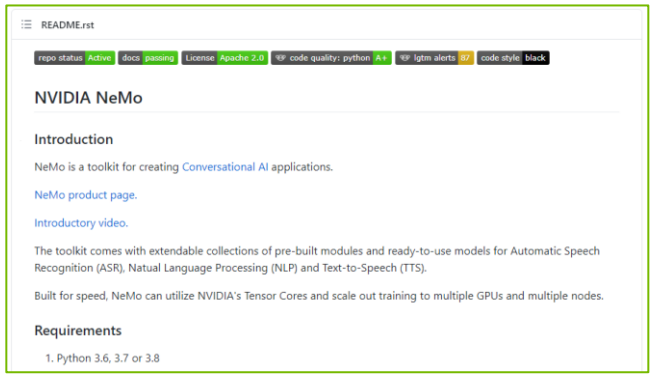
NVIDIA Transfer Learning Toolkit

Bring Your Own Data (BYOD)

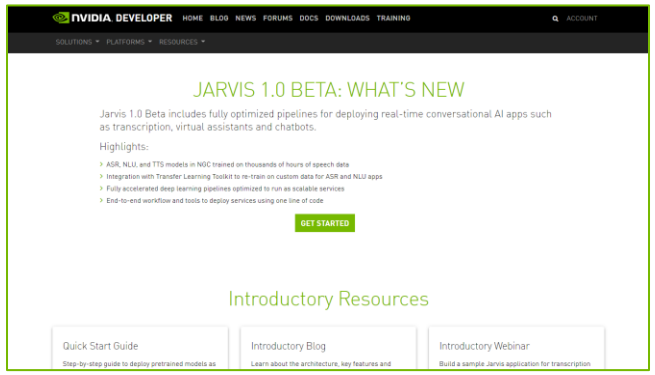


- Increase accuracy by fine-tuning on proprietary data
- Zero Coding approach reduces barrier to entry for enterprises
- Use Tensor Cores to achieve highest training performance
- Integrated with Jarvis to deploy models as real-time services

Try Jarvis and NeMo today



NeMo
github.com/NVIDIA/NeMo



Jarvis
developer.nvidia.com/nvidia-jarvis-getting-started

TLT
developer.nvidia.com/transfer-learning-toolkit

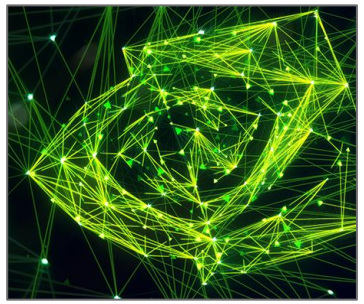
NVIDIA DEEP LEARNING INSTITUTE

Hands-on training in deep learning, accelerated computing, and accelerated data science for developers, data scientists and researchers

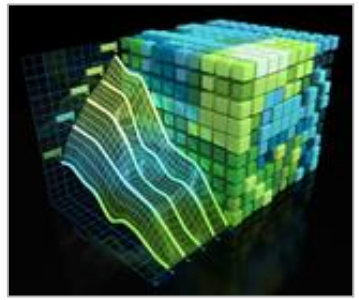
Introductory courses on AI for IT professionals

Take self-paced training online, view instructor-led training catalog and upcoming public workshop schedule, and learn about university resources at www.nvidia.com/dli

For a consultation contact us: nvdli@nvidia.com



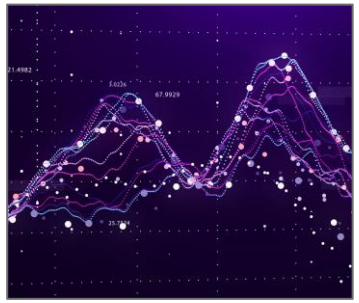
Deep Learning Fundamentals



Accelerated Computing Fundamentals

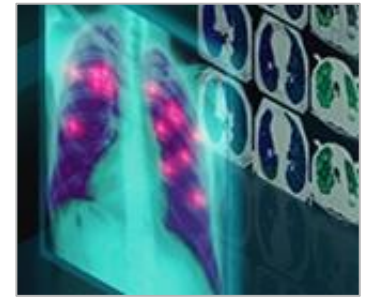
Accelerated Data Science Fundamentals

Intro to AI in the Data Center



AI for Anomaly Detection

AI for Autonomous Vehicles



AI for Healthcare

Conversational AI and NLP

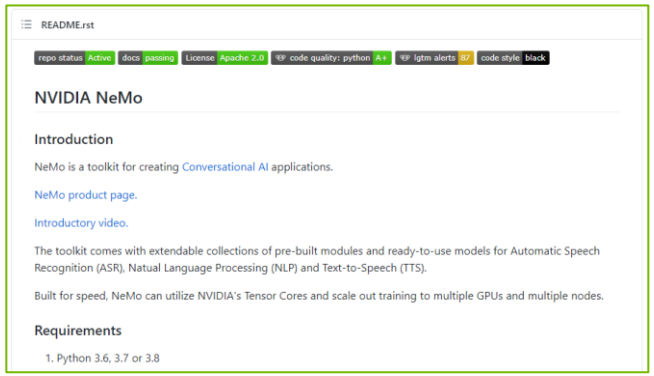
AI for Industrial Inspection

AI for Intelligent Video Analytics

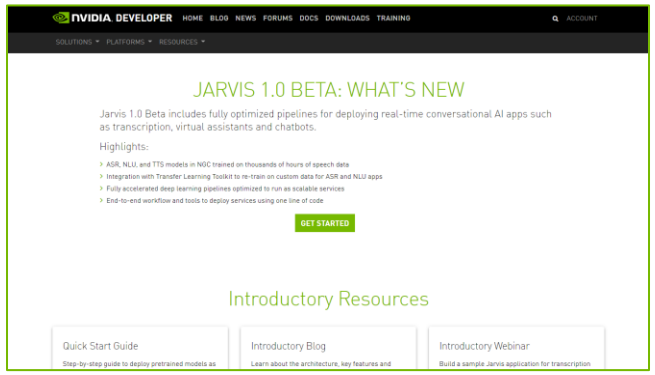
AI for Predictive Maintenance

Networking

Try Jarvis and NeMo today



NeMo
github.com/NVIDIA/NeMo



Jarvis
developer.nvidia.com/nvidia-jarvis-getting-started

TLT
developer.nvidia.com/transfer-learning-toolkit