Ministry of Electronics &
Information Technology

सुस्वागतम्
நல்வரவு
ಸುಸ್ವಾಗತಂ
സുസ്വാഗതം
సుస్వాగతం
સુસ્વાગતમ
ସୁସ୍ୱାଗତମ
সুস্বাগতম
ਸੁਆਗਤਮ
سوسواگتم
خوش آمدید

NPSF User Workshop 2023

24 July – 24 Aug

Presentation
On

Inference of Custom FRS Model

Muthukumaravel – Quantiphi

31st July

National PARAM Supercomputing Facility

Agenda

- Current Problems of Face Recognition System(FRS)
- Overview of FRS
- Training Pipeline for FR Model
- FRS Using NVIDIA Metropolis
- Intro to DeepStream
- Intro to Triton Inference Server
- DeepStream Pipeline for FRS
- End to End demo Training and Inferencing

Current Problems of Face Recognition System(FRS)

Inferencing Challenges

Model Latency

Pipeline Scalability

Higher memory consumption

Model optimization

Current Problems Faced by Face Recognition System (FRS)

Limited Scalability

Face recognition systems often struggle to handle large-scale datasets and real-time processing

Performance Constraints

Traditional face recognition systems face challenges in achieving high accuracy and speed simultaneously

Resource-intensive

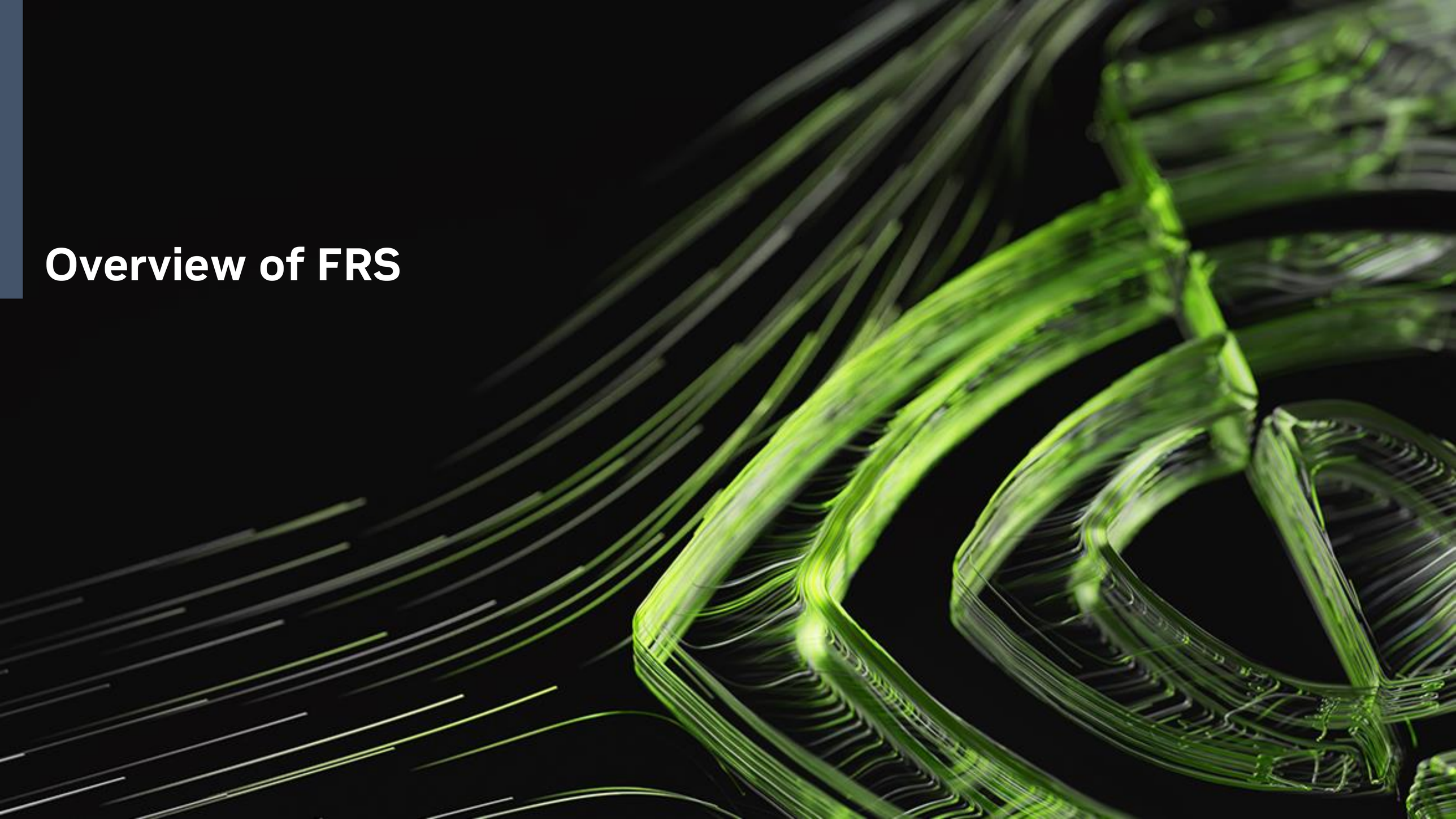
Training and inference processes require significant computational resources and time

Privacy Concerns

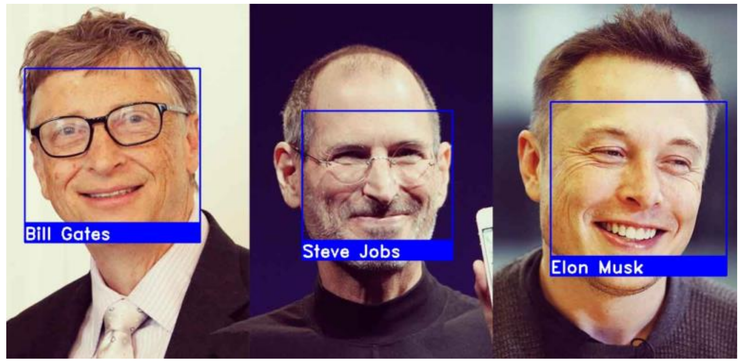
Ensuring privacy and security in face recognition systems is a crucial concern

DeepStream, Multi-GPU, and DGX SuperPod can help solve these problems by enabling scalable and efficient face recognition solutions that leverage powerful hardware for training and real-time inferencing

Overview of FRS



Overview of FRS



Face Recognition System (FRS)

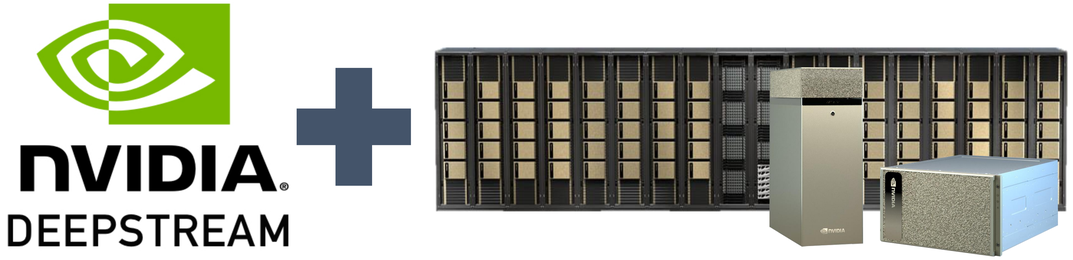
It refers to the technology that identifies and verifies individuals based on their facial features.

DeepStream

It is an NVIDIA platform that provides a scalable and efficient solution for video analytics, including face recognition.

NVIDIA DGX SuperPod

It is a powerful AI infrastructure that delivers exceptional performance for training and inferencing workloads.



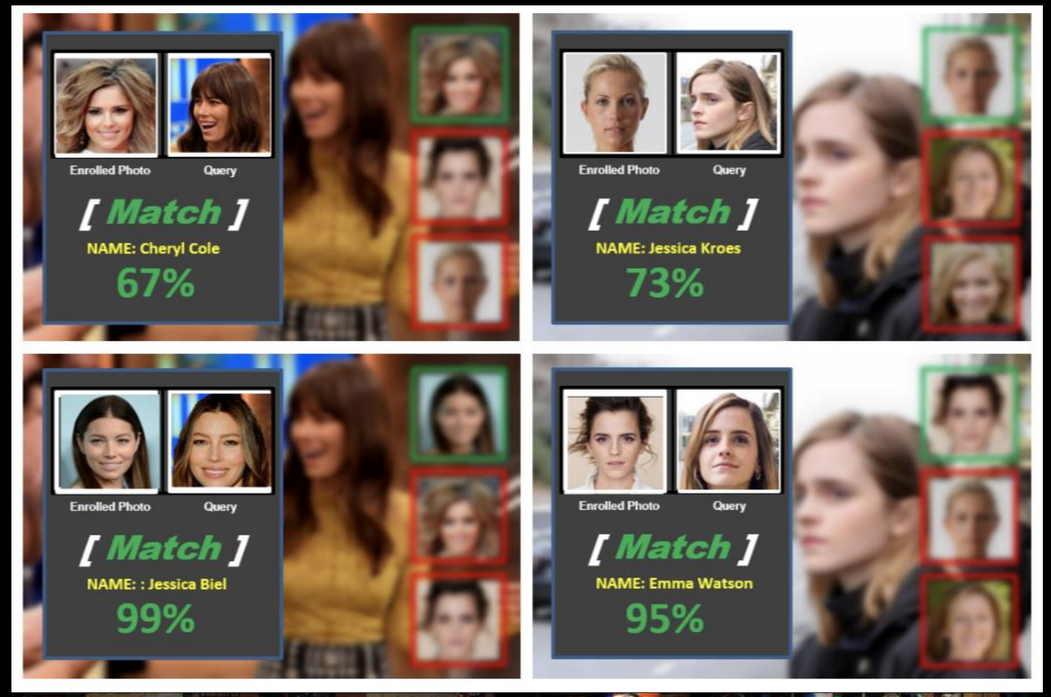
Combining DeepStream & DGX SuperPod

We can build a robust and scalable face recognition system that addresses the challenges mentioned earlier.

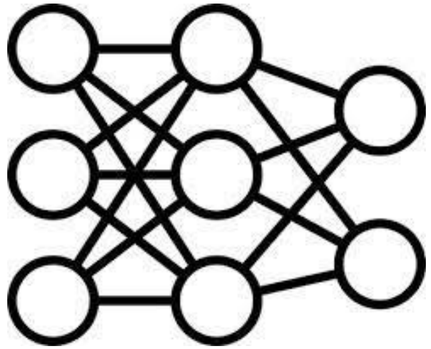
FR Model Training

FR Model Training

- For Face Recognition[FR] Model Training, we are using the face.evoLVe repository on GitHub which is a comprehensive library that collects and implements a wide range of popular deep-learning based methods for face recognition.
- The face.evoLVe is a popular deep learning framework focused on face recognition tasks. It offers an all-in-one solution for face recognition and verification.
- The IR_SE_50 model in face.evoLVe utilizes the ResNet architecture with SE-ResNet modules, achieving state-of-the-art performance.
- The Inception-ResNet architecture combines Inception and ResNet, capturing multi-scale features and utilizing residual connections.
- It captures multi-scale features using inception modules and adapts feature maps with channel-wise attention for informative representation.
- The Squeeze-and-Excitation mechanism enhances network representation by modeling channel-wise dependencies.
- The FR model consists of the IR_SE_50 backbone for feature extraction and the Focal loss function for training and better representation learning.
- The IR_SE_50 model backbone achieves state-of-the-art performance on benchmarks like LFW and MegaFace.
- Refer https://github.com/ZhaoJ9014/face.evoLVe for more info



FR Model Training



Model

- The IR_SE_50 model serves as a backbone architecture for our system, forming the foundation for subsequent layers and components of the face recognition model.
- It combines two key components: the Inception-ResNet (IR) architecture and the Squeeze-and-Excitation (SE) mechanism.
- This configuration effectively captures and utilizes discriminative features, allowing for enhanced face recognition performance.
- The IR_SE_50 model is designed to extract discriminative features from face images, enabling accurate face recognition and verification.

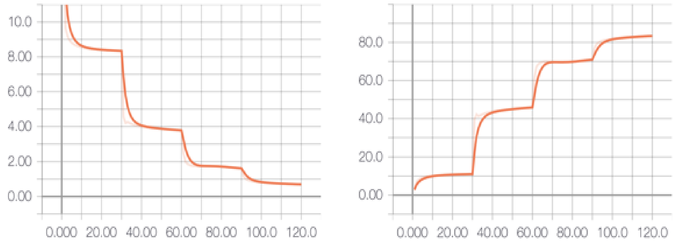
Dataset

- The dataset used to train our IR_SE_50 Model is Clean version of the CASIA-WebFace dataset. The CASIA-WebFace dataset serves as our reliable source for face verification and identification tasks.
- It has an extensive collection of 494,414 face images belonging to 10,575 genuine individuals, this dataset was carefully gathered from various sources on the internet
- This dataset includes images captured under different lighting conditions, facial expressions, and poses, enabling robustness in face recognition algorithms.
- CASIA-WebFace emphasizes the diversity of facial appearances, making it suitable for training models that can generalize

FR Model Training: Benchmarking

- To assess the performance of our Face Recognition Model, we conducted benchmarking on the emore dataset, which includes LFW, CFP-FF, CFP-FP, AgeDB-30, calfw, cplfw, and vgg2_fp.
- Our model incorporates the IR-SE-50 backbone, an ArcFace head, and the Focal loss function, all of which contribute to its face recognition capabilities.
- During training, we utilized the CASIA-WebFace-Clean dataset to train the model and enhance its accuracy and performance.

Training Statistics



Training Loss | Training Acc

Validation Statistics:

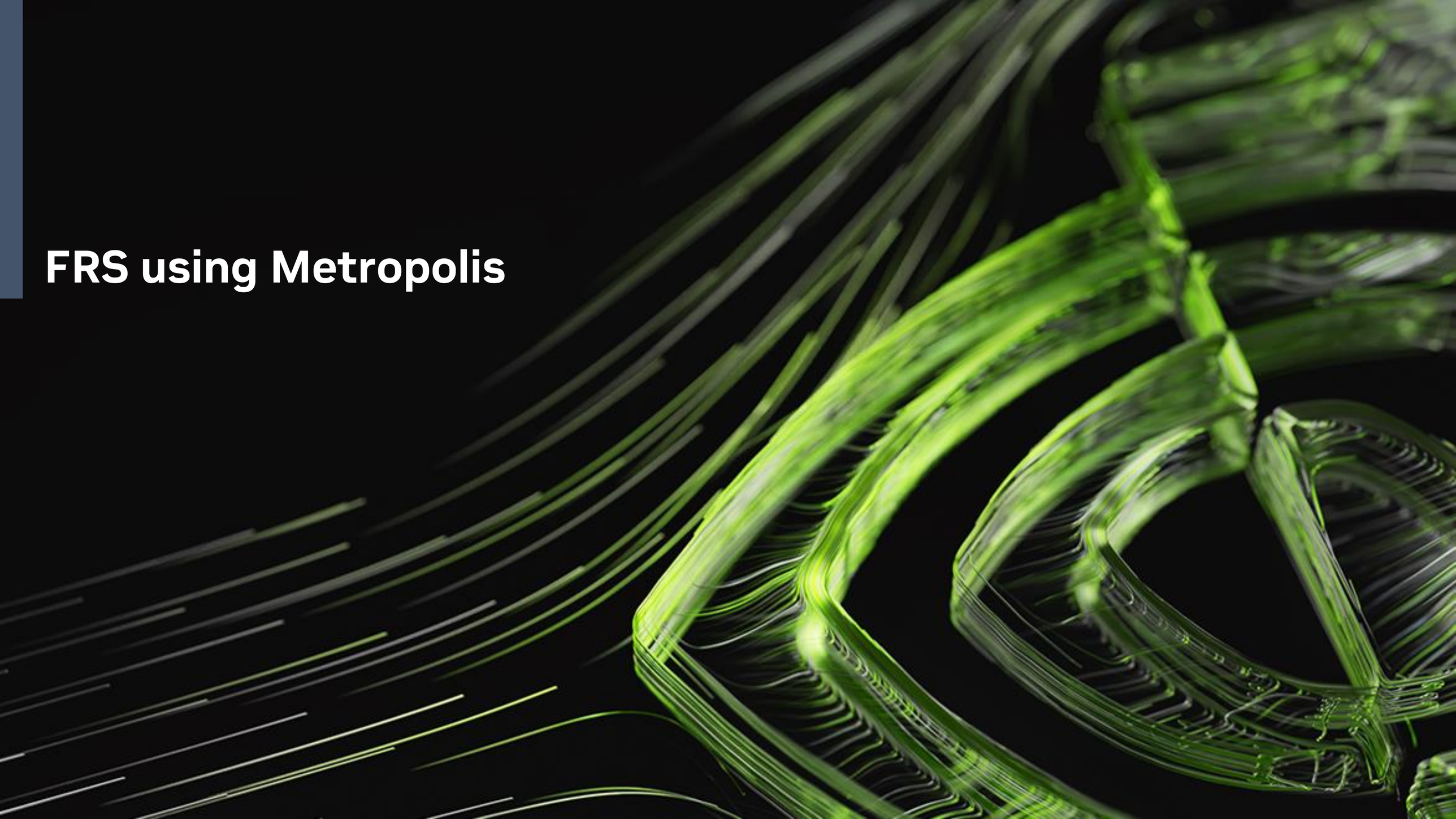
LFW | CFP_FF | CFP_FP | AgeDB | CALFW | CPLFW | VGGface2_FP

Model:

Backbone	Head	Loss	Training Data
IR-SE-50	ArcFace	Focal	CASIA-WebFace-Clean

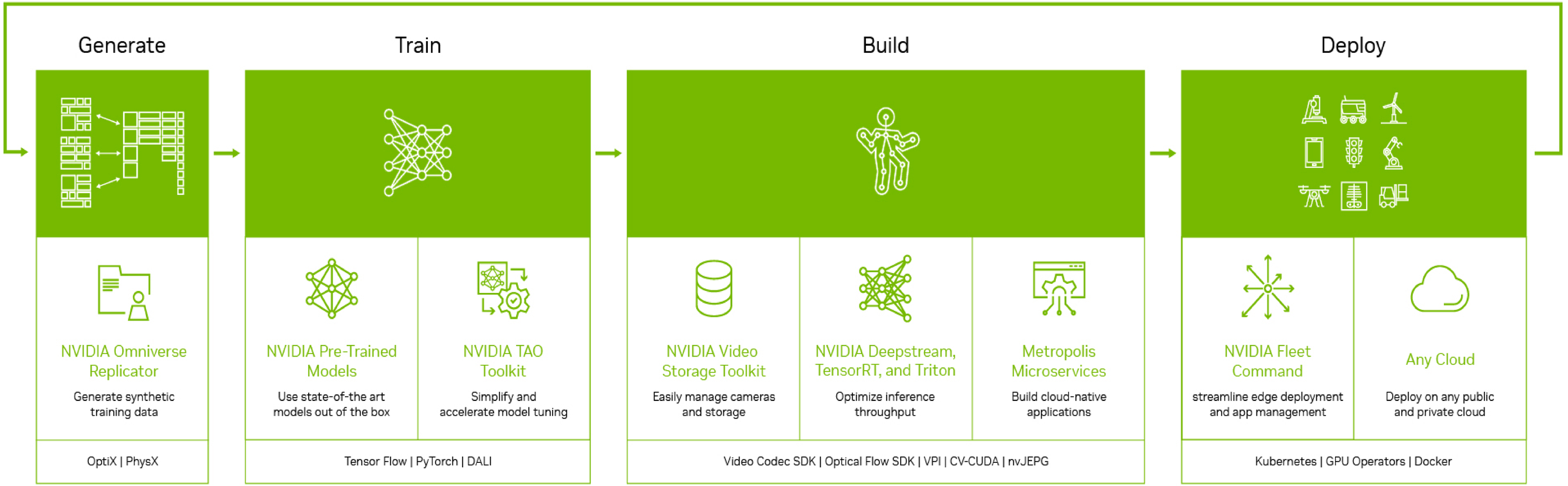
LFW	CFP_FF	CFP_FP	AgeDB	CALFW	CPLFW	Vggface2_FP
99.78	99.69	98.14	97.53	95.87	92.45	95.22

FRS using Metropolis



METROPOLIS STACK

Fast-Track Data Generation, AI Model Creation, App Development, Inference, and Scalability



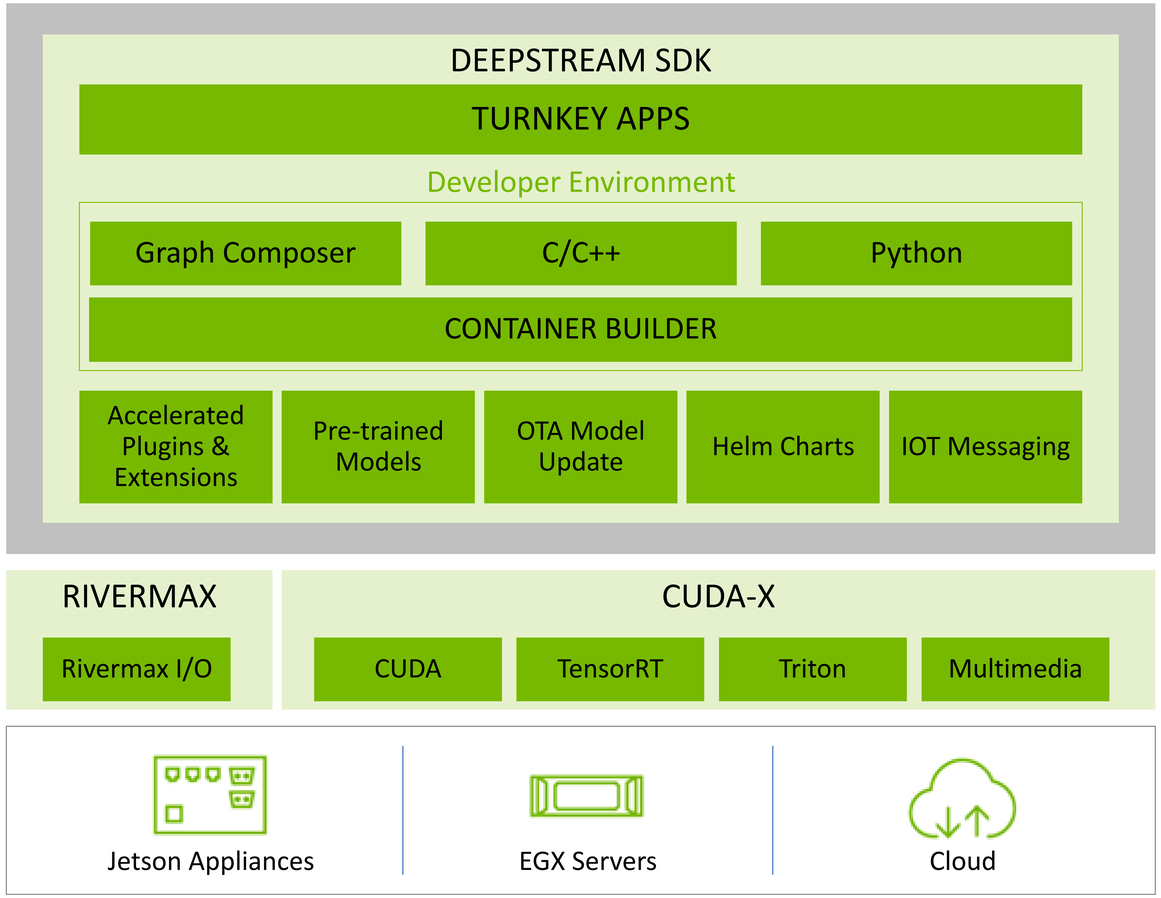
- Leading AI revolution, to meet every challenge with smarter, faster applications.
- Benefits:
  - Faster builds
  - Lower cost
  - Flexible deployments

Intro to DeepStream

DEEPSTREAM

Introduction

- Platform for building end-to-end services and solutions
- Enable real-time analytics
- Flexible and scalable framework that can be customized to meet specific use cases and requirements
- Support for a wide range of hardware platforms and protocols, making integration easier
- Creates powerful vision AI applications.
- Rich and multi-modal real-time sensor data.
- Deploy AI services in cloud native containers and orchestrate them using Kubernetes.
- Increase stream density.



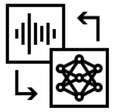
DeepStream SDK

Triton Inference Server

TRITON INFERENCE SERVER

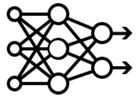
SDK for high-performance deep learning inference

NVIDIA Triton™ Inference Server, is an open-source inference serving software that helps standardize model deployment and execution and delivers fast and scalable AI in production.



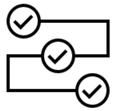
Support for Multiple Frameworks

Triton supports all major training and inference frameworks, such as TensorFlow, NVIDIA® TensorRT™, PyTorch,, Python, ONNX, XGBoost, scikit-learn RandomForest, OpenVINO, custom C++, and more.



High-Performance Inference

Triton supports all NVIDIA GPU-, x86-, Arm® CPU-, and AWS Inferentia-based inferencing. It offers dynamic batching, concurrent execution, optimal model configuration, model ensemble, and streaming audio/video inputs to maximize throughput and utilization.

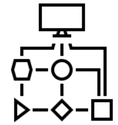


Designed for DevOps and MLOps

Triton integrates with Kubernetes for orchestration and scaling, exports Prometheus metrics for monitoring, supports live model updates, and can be used in all major public cloud AI and Kubernetes platforms. It's also integrated in many MLOps software solutions.

An Integral Part of NVIDIA AI

The **NVIDIA AI platform**, which includes Triton, gives enterprises the compute power, tools, and algorithms they need to succeed in AI, accelerating workloads from speech recognition and recommender systems to medical imaging and improved logistics.



Supports Model Ensembles

Because most modern inference requires multiple models with preprocessing and post-processing to be executed for a single query, Triton supports model ensembles and pipelines. Triton can execute the parts of the ensemble on CPU or GPU and allows multiple frameworks inside the ensemble.

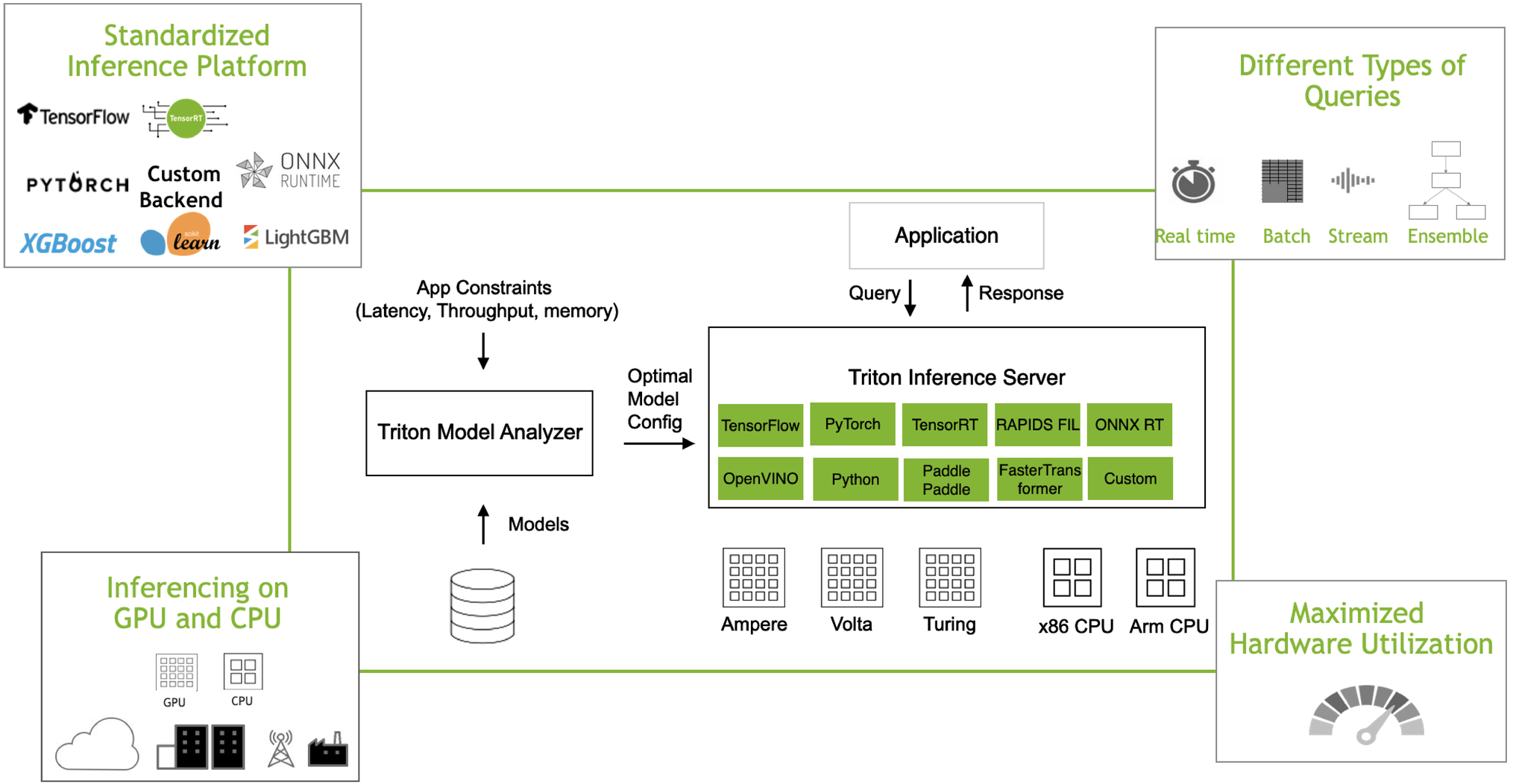
- Fast and Scalable AI in Every Application
- Achieve higher throughput
- Scale Inference With Ease

- Native Support in Python
- Model Orchestration With Management Service

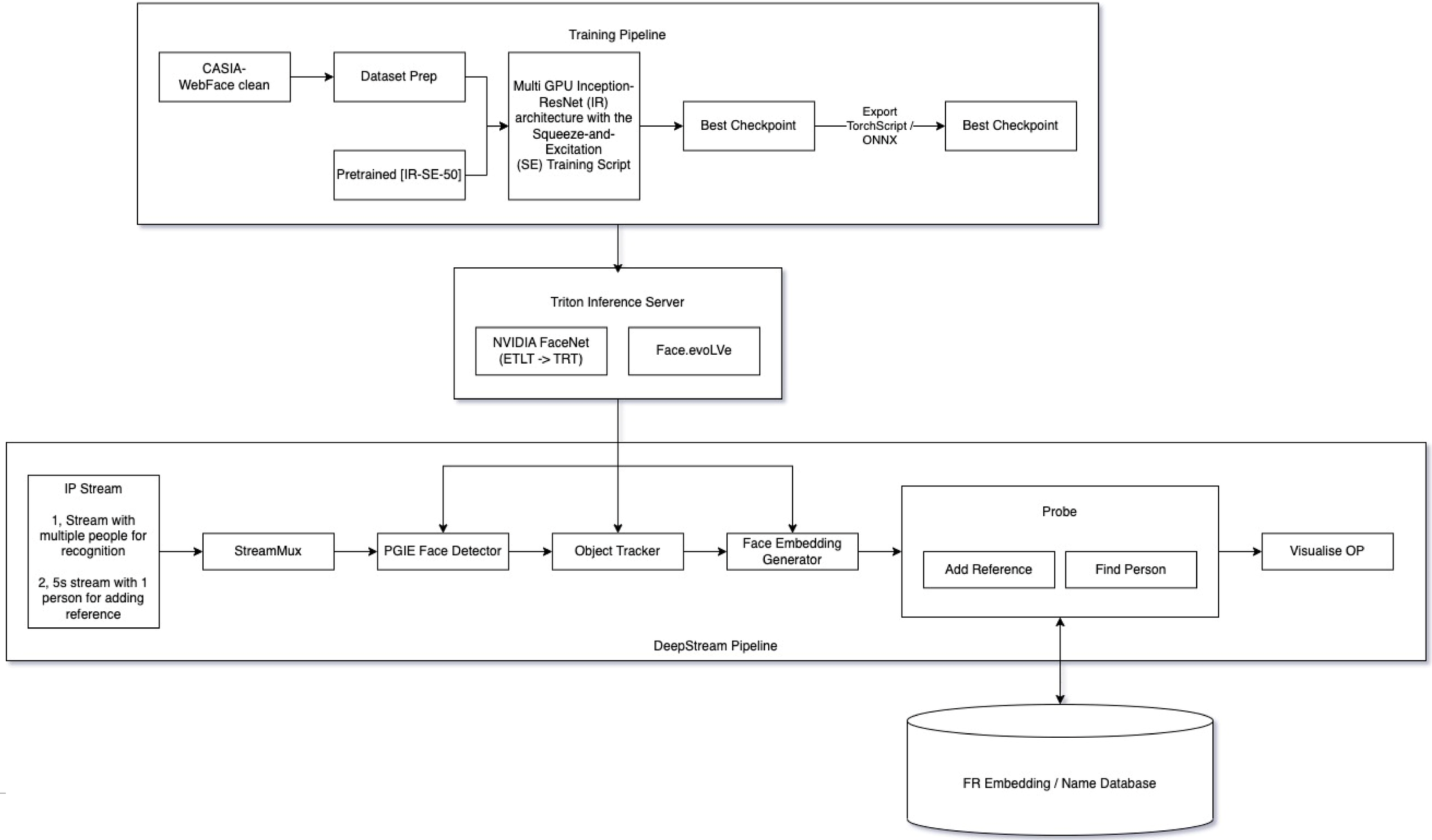
TRITON INFERENCE SERVER

SDK for high-performance deep learning inference

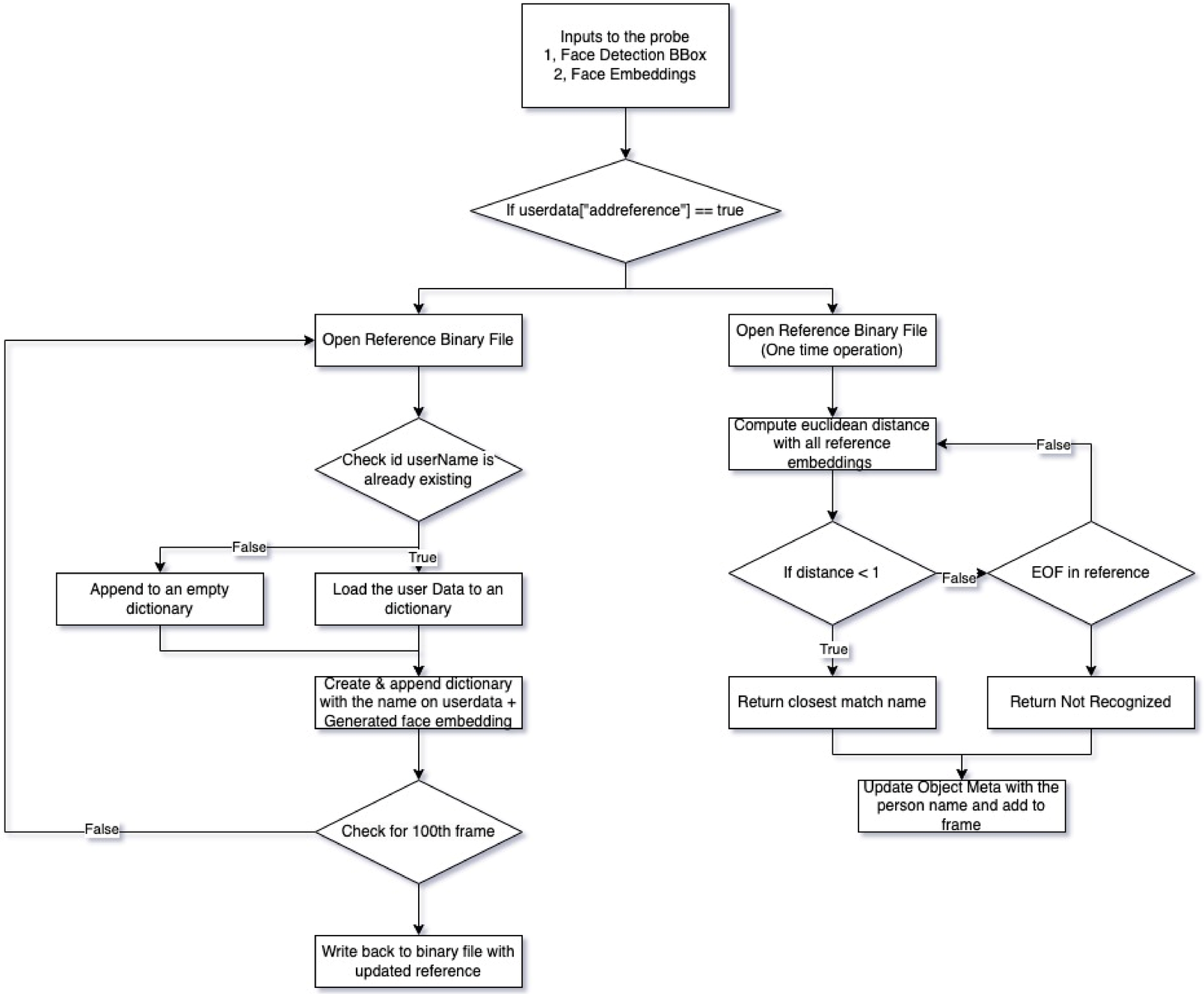
NVIDIA Triton™ Inference Server, is an open-source inference serving software that helps standardize model deployment and execution and delivers fast and scalable AI in production.



Basic FRS Architecture using DeepStream



Custom Probe flow



Exascale Computing
Microprocessor and Quantum Computing
AI and Language Computing
IoE, Dependable and Secure Computing
GenNext Applied Computing
10010111001100
Advanced Computing For Human Advancement
Thank You !