InferX ™X1 Software

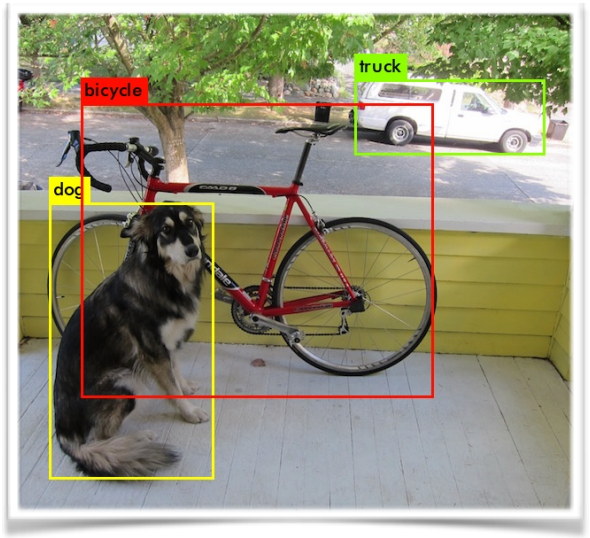
Why Software is Critical for AI Inference Accelerators

Jeremy Roberson Linley Fall Processor Conference, April 22nd 2021

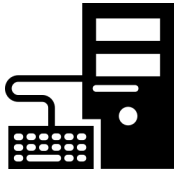
Customers want a solution that maximizes throughput while minimizing cost, power

PROBLEM	SOLUTION
1. Maximizing performance on hardware is difficult	InferX SDK hides the complexity from the end user, and ensures optimal tuning out of the box
2. Integration into existing workflow is also a cost	Design an API that prioritizes simplicity

Embedded Applications at the Edge



Embedded Application



Desktop tower

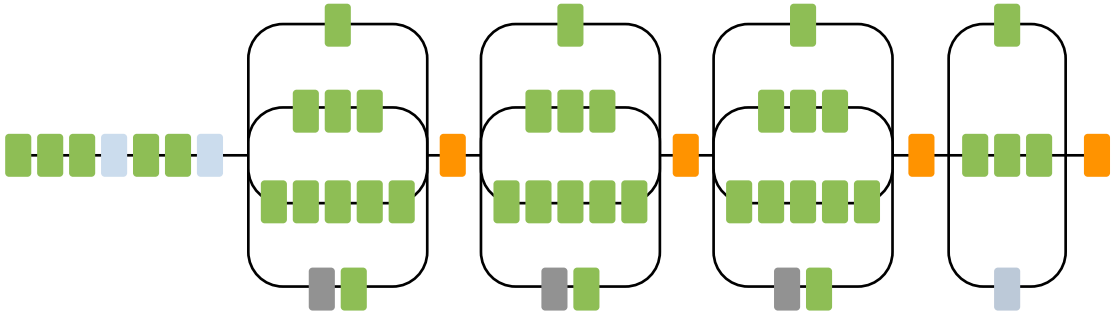
NN Model Frameworks and Infrastructure

Conv

AvgPool

MaxPool

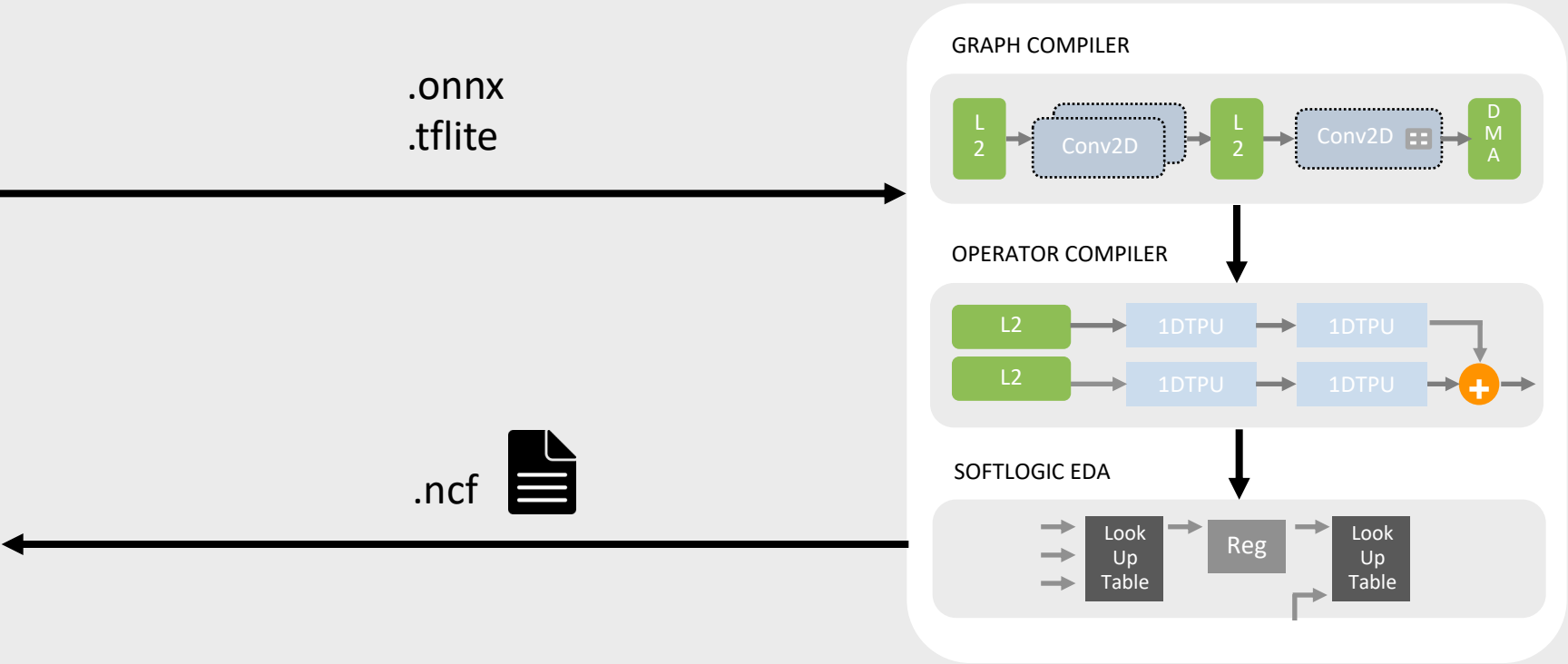
Concat



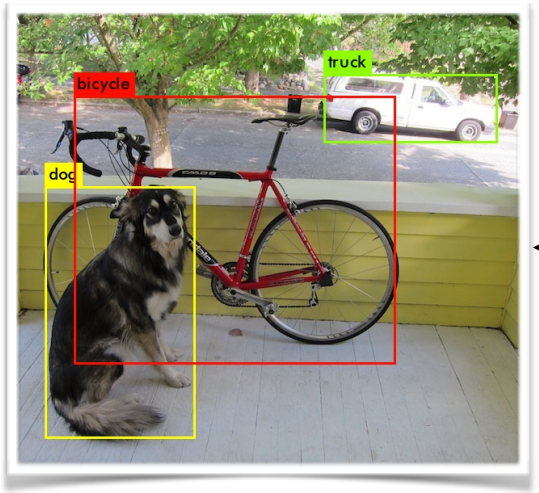
TensorFlow PyTorch mxnet Caffe

.onnx
.tflite

InferX X1 Compiler Suite



Embedded Applications at the Edge



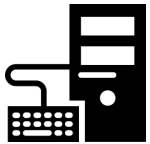
Embedded Application

Runtime API

X1 drivers

ONNX Runtime

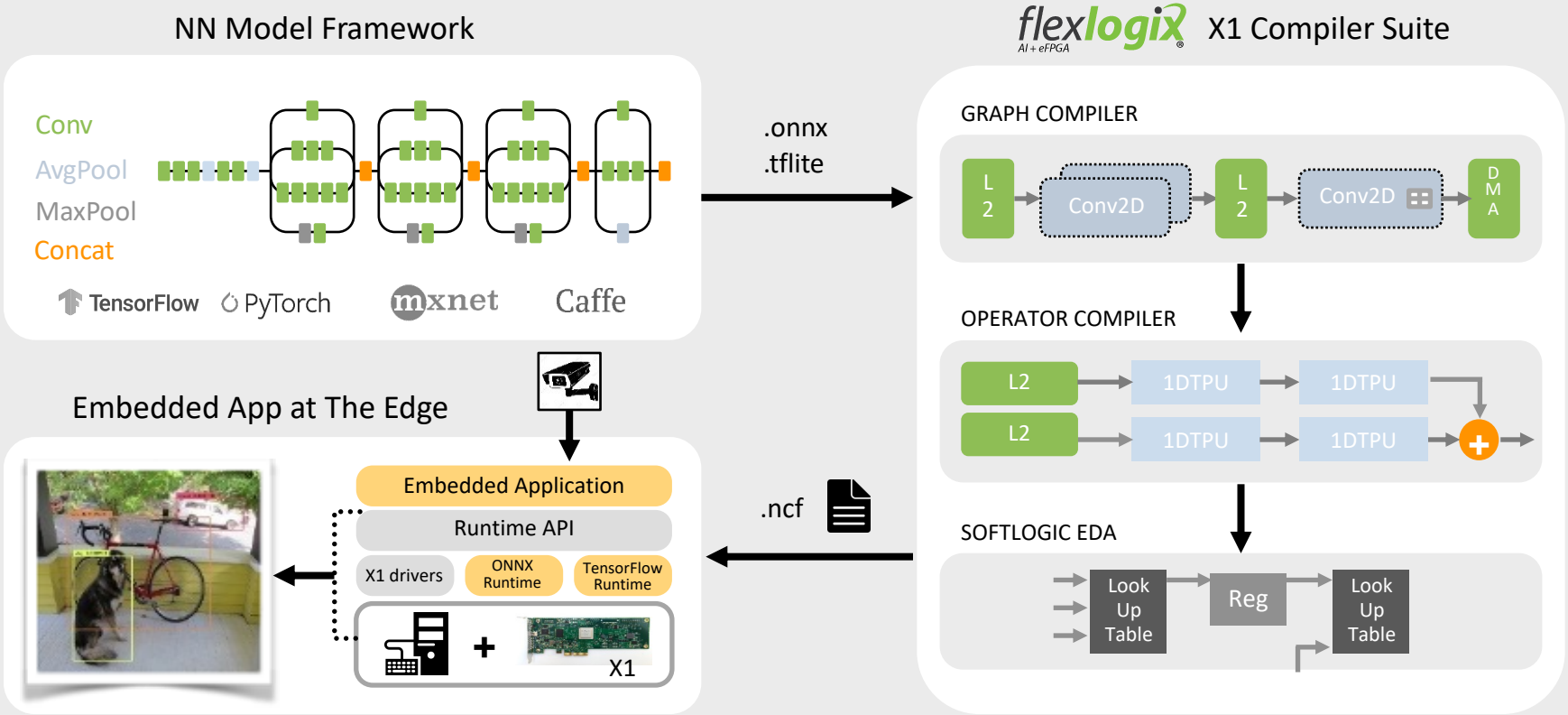
TensorFlow Runtime



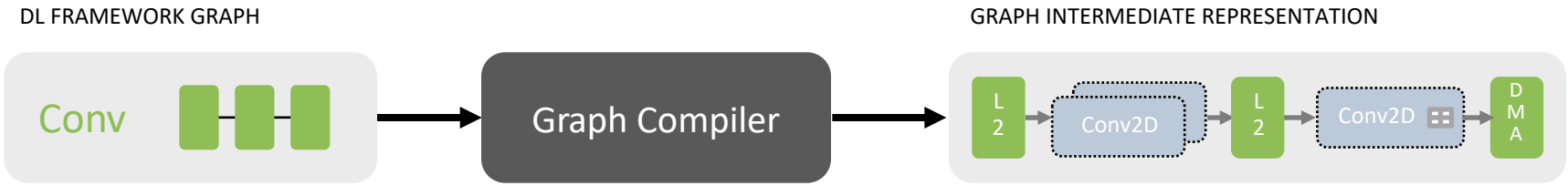
+

.ncf

Putting it All Together and Going to the Next Level



Graph Compiler: Adds Value on Multiple Fronts

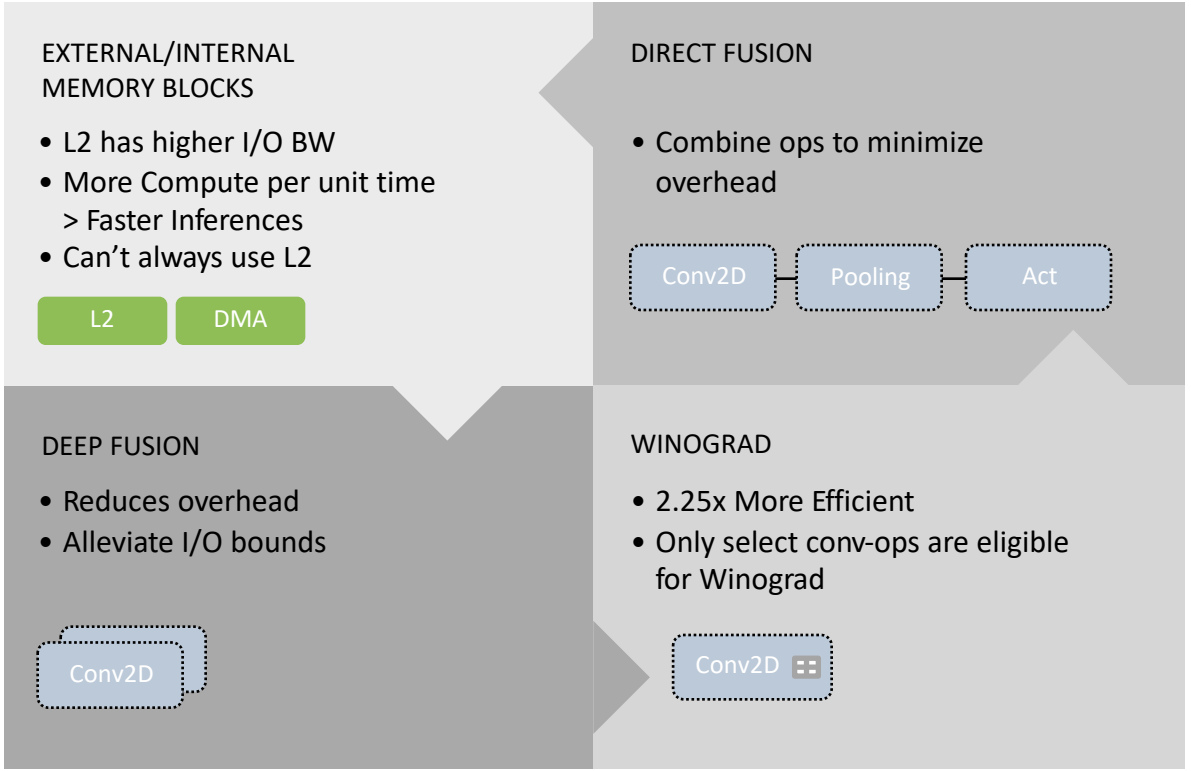


BENEFITS:

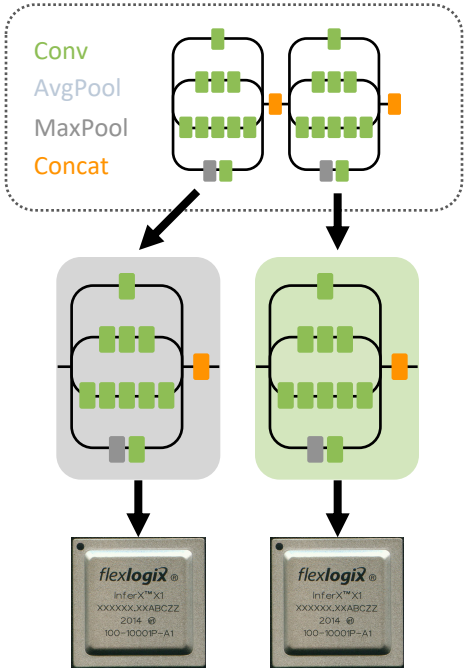
1. Maximizes Inference Speed
2. Accelerates Bigger Models
3. Broadens Coverage of Neural Networks

1. Maximizes Inference Speed

- X1 HW blocks were architected with Deep Learning models in mind
- Model Compiler's job is to best utilize those HW blocks



2. Accelerates Bigger Models



GRAPH PARTITIONING

4 INFERENCES ON 1 INFERX X1

Inference Time

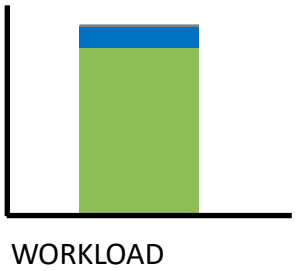
4 INFERENCES ON 2 INFERX X1

Inference Time

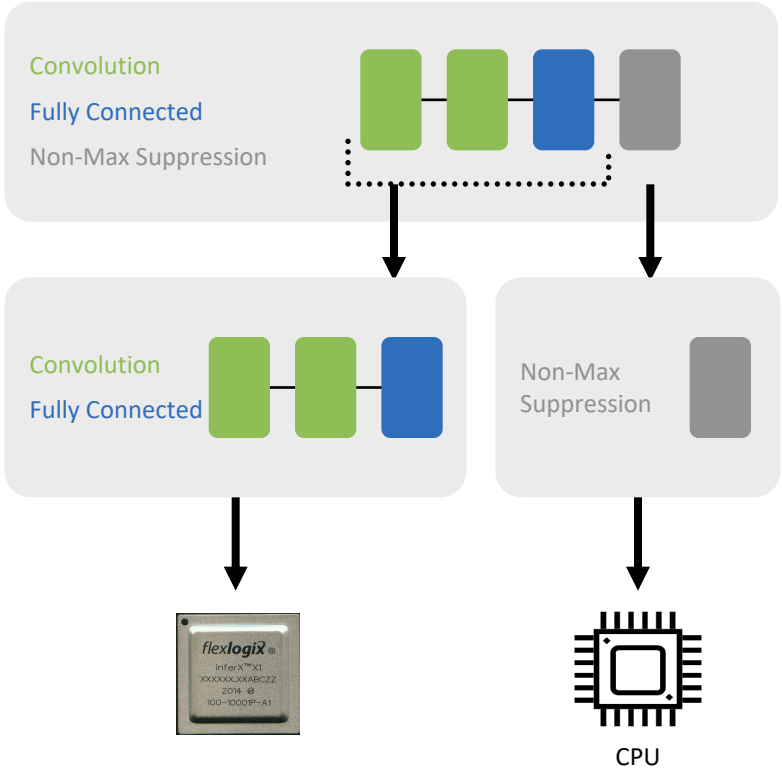
Significant Speedup !

3. Broadens Coverage of Neural Networks

- Certain ops can be offloaded to other HW with minimal impact to end inference time
- Better model coverage
- Done in automatically so users don't have to worry about manual model partitioning



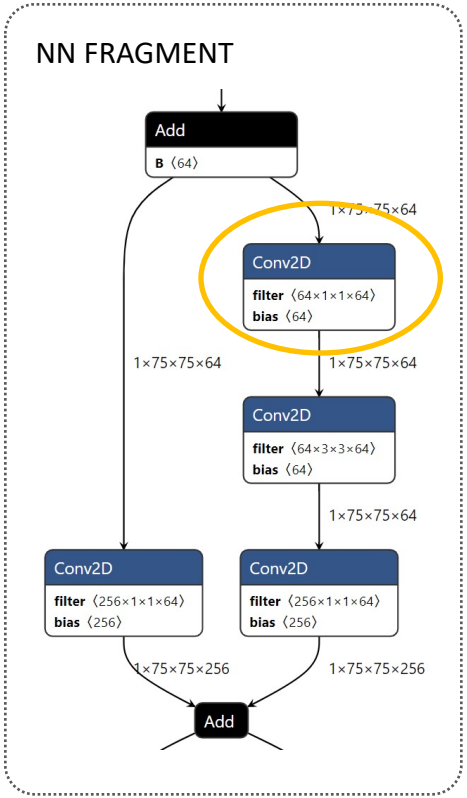
WORKLOAD

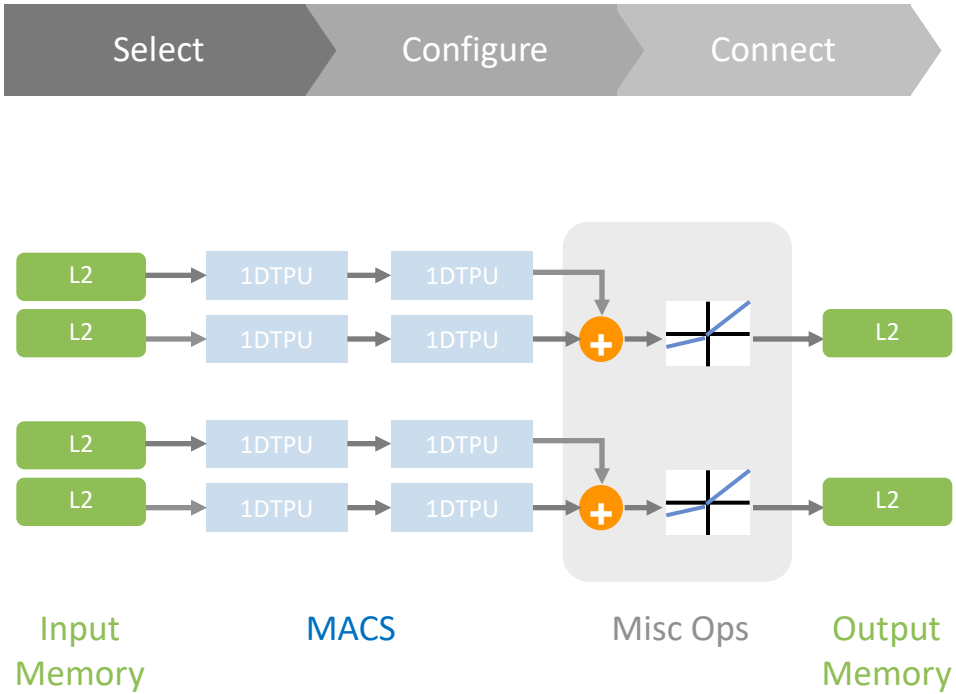


Operator Compiler: Unleashes The Power of eFPGA

SELECTION OF OPERATORS

Add
Convolution
AvgPool
MaxPool
Affine
Concatenate
And more





Runtime Software: Simplifies Deployment at The Edge

INFERX RUNTIME C++ API

FUNCTIONS	DESCRIPTION	ARGS
StartInference	Executes single inference	Data_handle
InferMulti	Executes multiple inferences	times
WaitForInferenceReady	Waits for inference to finish	{ }
GetInferOutputData	Retrieves inference results	data_out
SetModel	Loads model into memory	filename
SetData	Established Data	data
Initialize	Initialize X1 HW	{ }

SW Specialists are NOT required to master InferX Runtime!

Putting it All Together

Design

The InferX X1 HW is designed around existing/future DL models for future proofing

Software

The InferX Compiler Software is a 3-in-1 product that maximizes performance in a fully automated way for easy intergration

Runtime API

The InferX Runtime API is streamlined for ease-of-use

Customer Impact

Makes optimizing performance easy and lowers integration cost

Availability and Release Schedule

	DESCRIPTION	DATE
InferX SDK		
	Compiler with INT8 support	Q3 2021
	BF16 support	Q4 2021
Chips		
	X1-800	Sampling Q3 2021 Production Q4 2021
	X1-533	Sampling June 2021 Production August 2021
Boards		
	X1P1	Sampling June 2021 Production August 2021
	X1M	Sampling Q3 2021 Production Q4 2021

Let Us Benchmark Your Model

Do you have high resolution models that need to run in realtime?
We will show you how fast X1 can run them.

Dana McCarty
VP Sales and Marketing
for Inference Products
dana@flex-logix.com

TensorFlow Lite

ONNX

Thank you

Jeremy Roberson jeremy@flex-logix.com