InferX™ DSP Solution
#1 PPA DSP IP & Software

flexlogix reconfigurable®

DSP Performance of the Fastest FPGAs at a fraction of $/W

✔ Dozens of TeraMACs/sec. in 5nm (INT16 inputs, INT40 accumulation for accuracy)

✔ Real and Complex data types

✔ InferX uses 1-dimensional tensor processors (vector * matrix) controlled by eFPGA

✔ Tensor processor (TPU) MACs are configurable: 128 int16 MACs or 512 int8 MACs

✔ An InferX TPU has ~10x the DSP performance of an EFLX DSP tile in ~¼ the area

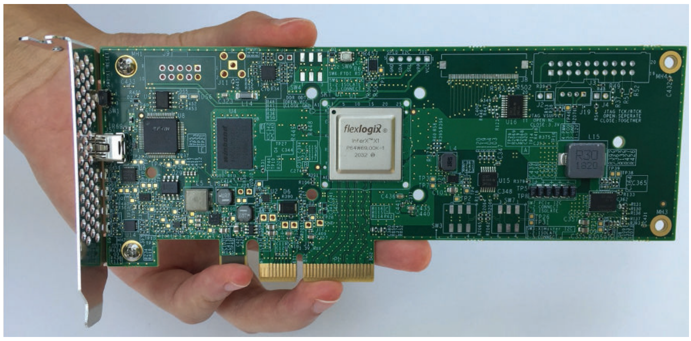
✔ EFLX eFPGA is a hard macro; InferX is soft IP delivered as RTL with constraints

✔ Scalable architecture from 1 TPU to 128+ with the same software flow

✔ High level software flow from Simulink to InferX Compiler to Verilog generation

✔ Streaming mode or packet mode

✔ Proven in silicon



Dozens of Use Cases

InferX DSP Benchmarks in N5 (128-16K MACs) & GF12 (128-2K MACs)

N5 1GHz	1 TPU 128 MACs	16 TPUs 2K MACs	128 TPUs 16K MACs
Complex INT16 1K/2K/4K FFT	500 MS/s (MegaSamples/sec)	8.5 GS/s (GigaSamples/sec)	68 GS/s (GigaSamples/sec)
Real INT16x16 FIR 256 taps	0.25 GS/s	4 GS/s	32 GS/s
Real INT16x16 FIR 4096 taps	16 MS/s	0.25 GS/s	2 GS/s
32x32 Complex INT16 Matrix Inversion	10K-Inv/sec	0.35M-Inv/sec	2.8M-Inv/sec
Area (est.)	0.8 mm^{2}	3.6 mm^{2}	28.8 mm^{2}

GF12 250MHz	1 TPU 128 MACs	4 TPUs 512 MACs	16 TPUs 2K MACs
Complex INT16 1K/2K/4K FFT	125 MS/s (MegaSamples/sec)	500 MS/s (MegaSamples/sec)	2 GS/s (GigaSamples/sec)
Real INT16x16 FIR 256 taps	60 MS/s	250 MS/s	960 MS/s
Real INT16x16 FIR 4096 taps	4 MS/s	16 MS/s	64 MS/s
Area (est.)	2 mm^{2}	4 mm^{2}	14 mm^{2}

We support InferX from 40nm to 18A: ask us for the node you need.

This is a subset of operators: ask about our full DSP library

InferX™ DSP Software
Simulink Streaming or Packetized

flexlogix® reconfigurable

A fully featured software package for all program phases

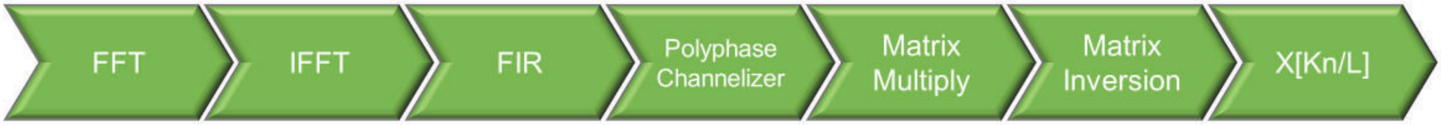
DSP Systems Engineers: provides a fully integrated Simulink model library. Each block has a bit accurate model allowing System Designers to seamlessly launch time-domain simulations of the entire DSP pipelines.

System Integration and Verification: provides a cycle accurate simulation/verification environment. Also automatically does RTL generation, synthesis, and place+route to seamlessly allow full validation at the IP level.

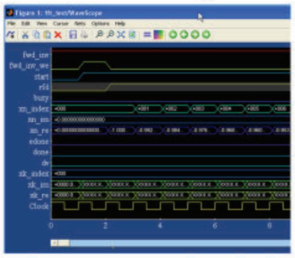
SoC Architects: provides IP configuration options as well as full PPA metrics to support early analysis and optimization of SoC architectures.

InferX DSP Software Overview

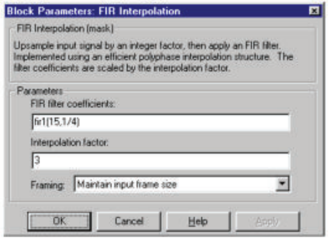
Standard Simulink Blockset


FFT → IFFT → FIR → Polyphase Channelizer → Matrix Multiply → Matrix Inversion → X[Kn/L]

Bit Accurate Modeling



Simplified Configuration


Block Parameters: FIR Interpolation

FIR Interpolation (mask)

Upsample input signal by an integer factor, then apply an FIR filter. Implemented using an efficient polyphase interpolation structure. The filter coefficients are scaled by the interpolation factor.

Parameters

FIR filter coefficients: fir1(15,1/4)

Interpolation factor: 3

Framing: Maintain input frame size

Flexible Precision

Input Precision	Output Precision
Real int8	Real int16
Complex int8	Complex int16
Real int16	Real int32
Complex int16	Complex int32

Int40 accumulation

Automatic CodeGen For flexible HW targets

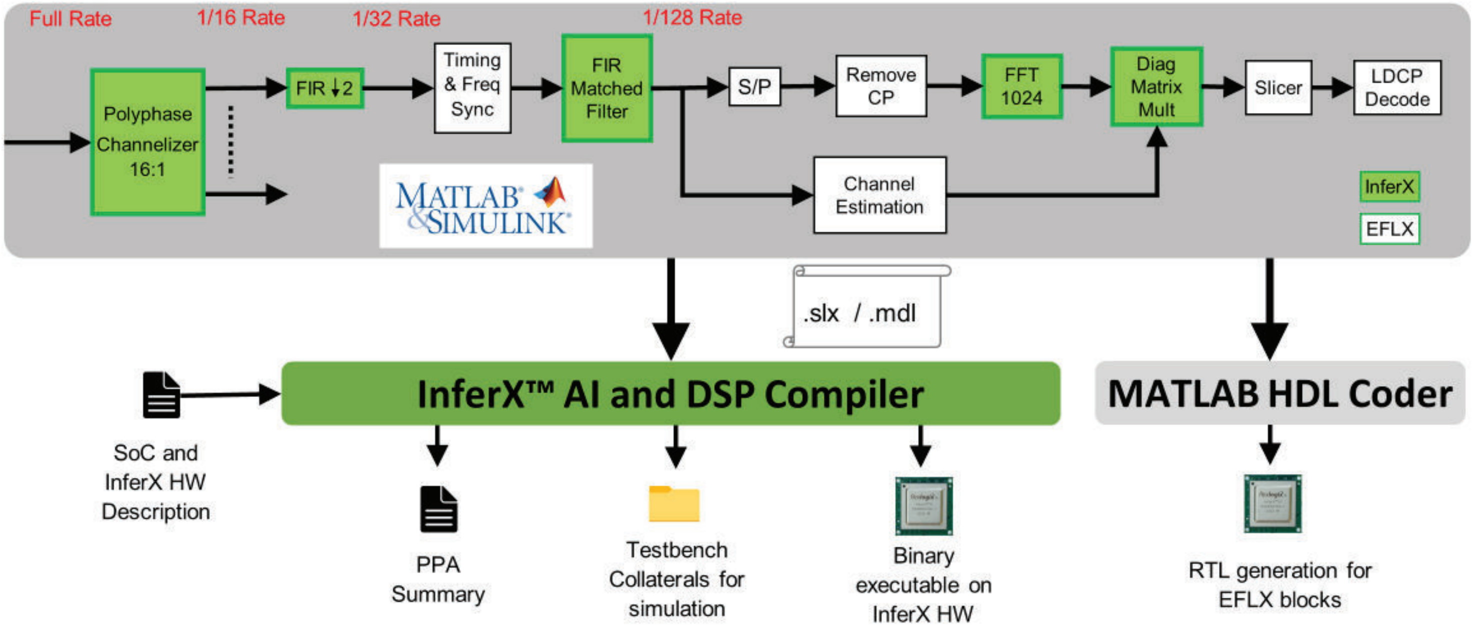
Leverage Simulink HDL Coder

EFLX

Simulink to Binary via InferX Compiler

InferX

InferX DSP Software Flow



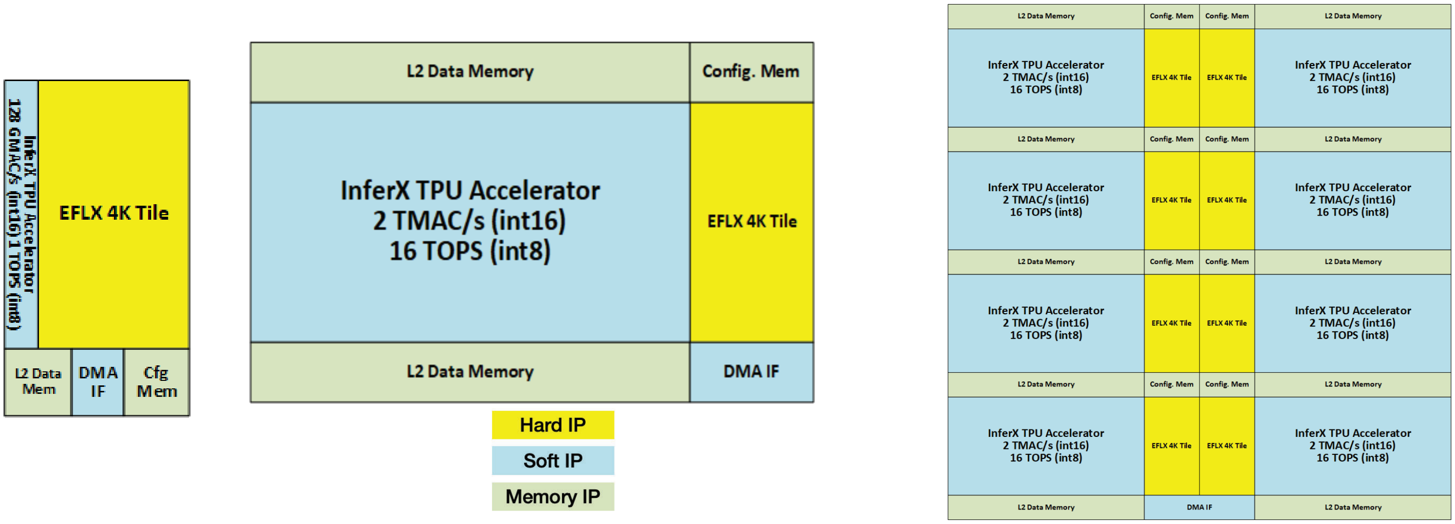
InferX™ Hardware
#1 PPA DSP IP in your SoC

flexlogix reconfigurable®

Fast streaming DSP at low cost and low power for your SoC

- ✔ INT16 inputs, INT40 accumulation for high accuracy; Real and Complex data types
- ✔ An InferX TPU has ~10x the DSP performance of an EFLX DSP tile in ~¼ the area
- ✔ GF22/12 & TSMC 40/28/16/12/7/6: 1-16 TPUs with existing EFLX eFPGA
- ✔ TSMC N5/4/3 & Intel18A: 1-128 TPUs or more with EFLX3.0 eFPGA with extra I/O
- ✔ Rapidly reconfigurable in a few microseconds (N5/N3/Intel18A)
- ✔ NOC/AXI bus interface to your SoC
- ✔ High DFT test coverage both DC and AC for high quality test; -40C to +125C Tj

Scalable DSP Performance from 1 to 16 to 128 TPUs (N5 shown)



TSMC IP Alliance Member

Flex Logix® is a TSMC IP Alliance Member based on the work it has done with TSMC over many years to develop IP meeting TSMC9000 compliance for design methodology, validation in silicon & documentation. Flex Logix has implemented IP in TSMC 40, 28, 16, 12 and 7nm; has started on 5nm; and has design files for 3nm. Dozens of customer chips are working in silicon with our IP; dozens more are in design.

tsmc

GlobalFoundries Ecosystem Member

Flex Logix® is an GlobalFoundries Ecosystem Member. Flex Logix has EFLX eFPGA and InferX IP available for GF22FDX and GF12. Dozens of customer chips are working in silicon with our IP and dozens more are in design.

GlobalFoundries™

Intel Foundry Services IP Alliance Member

Flex Logix® is an Intel Foundry Services IP Alliance Member. Flex Logix has early access to Intel 18A design databases to implement EFLX eFPGA and InferX DSP/AI for a major mutual customer.

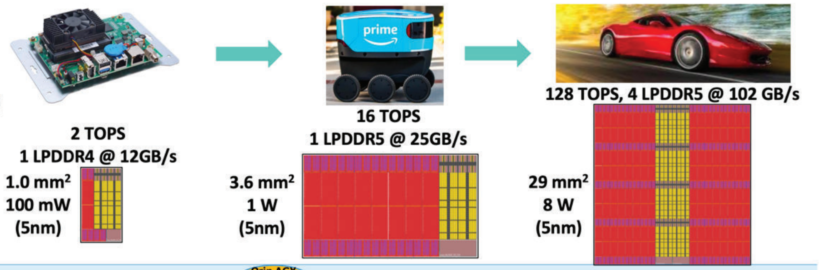
intel®

InferX AI Solution
World Class AI in your SoC

flexlogix®
reconfigurable

InferX IP is linearly scalable (computed based on 5nm, 1 GHz)

Over **4x TOPS efficiency** (IPS/TOPS) compared to Orin AGX

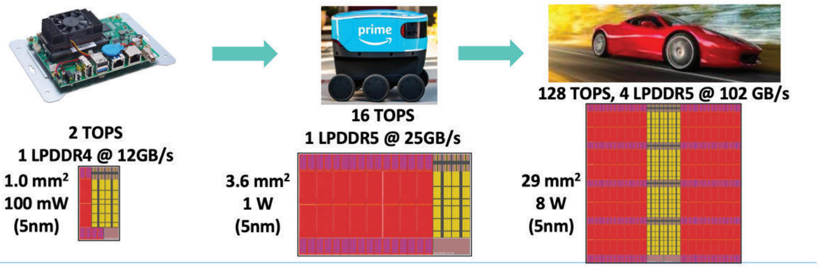


	2 TOPS		16 TOPS		128 TOPS
YOLOv5s (640×640)	35 IPS	Orin AGX 135 TOPs 125 IPS	234 IPS		1250 IPS
YOLOv5l6 (1280×1280)	1.6 IPS		14 IPS	Orin AGX 135 TOPs 31 IPS	115 IPS
ResNet50 (512x512)	16 IPS		144 IPS		1060 IPS
ResNet50 (1024x1024)	4 IPS		34 IPS		230 IPS
DETR 2020 (Transformer) (512x512)	17 IPS		211 IPS		603 IPS
DETR 2020 (Transformer) (1024×1024)	2.5 IPS		33 IPS		178 IPS

InferX IP uses bandwidth efficiently in a shared-memory system

InferX does **not** require dedicated DRAM: friendly to shared-memory system

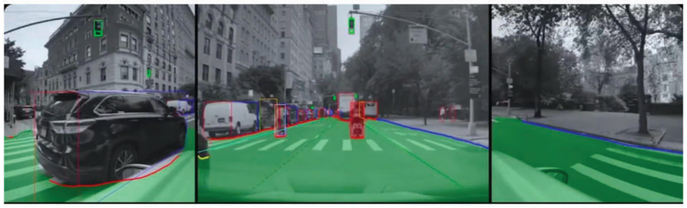
- 4K transfers to/from DRAM
- Performance not very sensitive to latency
- Use a fraction of available BW
- Over **4x DDR efficiency** (IPS/GB) compared to Orin AGX

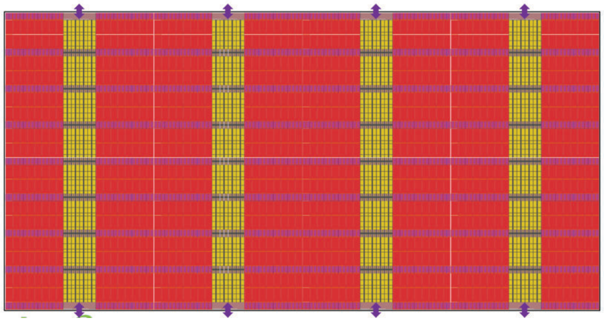


	DRAM BW & Capacity		DRAM BW & Capacity		DRAM BW & Capacity
YOLOv5s (640×640)	3 GB/s, 256MB (35 IPS)	Orin AGX 204.8 GB/s 125 IPS	14 GB/s, 256MB (234 IPS)		65 GB/s, 256MB (1250 IPS)
YOLOv5l6 (1280×1280)	2 GB/s, 512MB (1.6 IPS)		15 GB/s, 512MB (14 IPS)	Orin AGX 204.8 GB/s 31 IPS	60 GB/s, 512MB (115 IPS)
ResNet50 (512x512)	2 GB/s, 128MB (16 IPS)		18 GB/s, 128MB (144 IPS)		52 GB/s, 256MB (1060 IPS)
ResNet50 (1024x1024)	2 GB/s, 256MB (4 IPS)		16 GB/s, 256MB (34 IPS)		67 GB/s, 256MB (230 IPS)
DETR 2020 (Transformer) (512x512)	2 GB/s, 128MB (17 IPS)		14 GB/s, 128MB (211 IPS)		85 GB/s, 256MB (603 IPS)
DETR 2020 (Transformer) (1024×1024)	2 GB/s, 128MB (2.5 IPS)		6 GB/s, 128MB (33 IPS)		28 GB/s, 256MB (178 IPS)

InferX IP scales to over 1000 TOPS for next-generation ultra-HD AI models

1024 TOPS
HBM @ 820 GB/s
240 mm²
64 W
(5nm)





YOLOv5l6 – HD (1280×1280)	576 IPS
YOLOv5l6 – Ultra HD (5120×2560)	120 IPS
ResNet50 – HD (1024x1024)	1950 IPS
ResNet50 – Ultra HD (4096x2048)	232 IPS
DETR 2020 (Transformer) – HD (1024x1024)	1950 IPS
DETR 2020 (Transformer) – Ultra HD (2048x2048)	120 IPS

Other Nodes

Ask about AI benchmarks on other nodes such as N3 and 18A.

InferX #1 AI PPA

TOPS is a measure of PEAK AI throughput (2 times the number of Multiplier-Accumulator operations per second).

There are Dense TOPS and Sparse TOPS - Sparsity sacrifices accuracy. InferX TOPS are Dense TOPS.

What matters in your SoC is getting the inferences/second your application needs for the NN model and image size you want at the smallest silicon area and power.

As the data to the left shows, InferX outperforms Orin AGX using far fewer TOPS. InferX is more efficient. InferX is 4 to 10 times more inferences/second than Orin AGX for the same number of TOPS.

Adapt to New Models in Field

When you are designing your AI SoC you may be focused on beating Nvidia, but your IP options are DSP-derived VLIW architectures.

Unlike these other architectures, InferX is very programmable so it is possible to upgrade post-silicon, in the field to run any new operator and model that is created during the operating life of your chip.

InferX incorporates eFPGA as well as Tensor Processors. The eFPGA can be programmed to run anything.

4/2024 TGI