High-Efficiency Edge Vision Processing Based on Dynamically Reconfigurable TPU Technology

Cheng C. Wang, Co-Founder and CTO
Flex Logix Technologies

Flex Logix: A technology leader

Geoff Tate, CEO

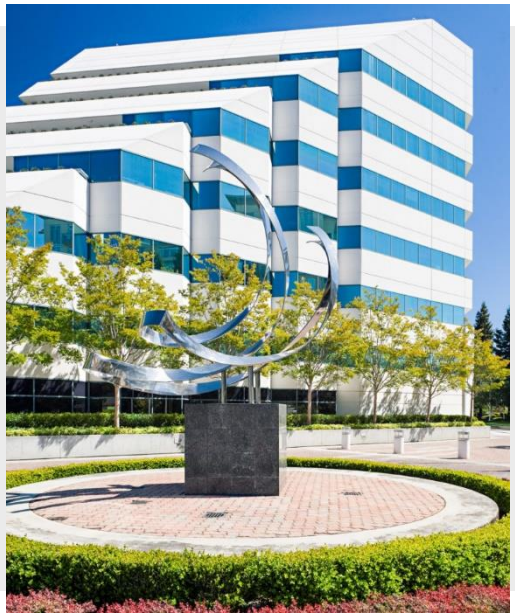
- Experienced executive taking company public
- Rambus: 4 people to IPO to $2B

Cheng Wang, CTO, Co-founder

- Industry expert with track record in tech innovation
- Winner: ISSCC Outstanding Paper Award, the premier chip design award. (Recent winners include IBM, Toshiba, Nvidia and Sandisk)

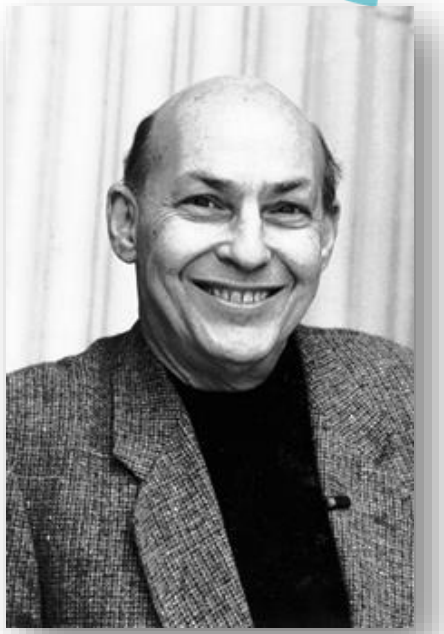
Flex Logix:

- Founded in 2014
- Profitable embedded FPGA IP business
- Edge AI Inference accelerator based on tensor array + programmable logic
- logic
- Backed by top technology and innovation investors
- Growing rapidly – We're hiring!
- Based in Mountain View, CA, with offices in Austin, TX



Short history of vision processing

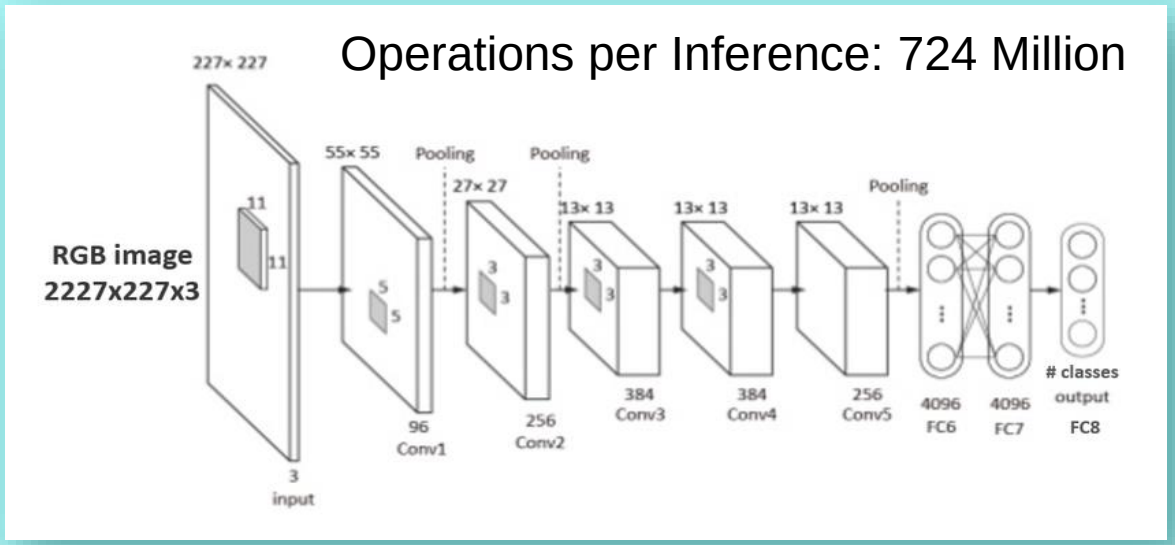
In 1966, American computer scientist and co-founder of the MIT AI Lab Marvin Minsky hired a first-year undergraduate student, Gerald Sussman, to spend the summer linking a camera to a computer and getting the computer to describe what it saw.



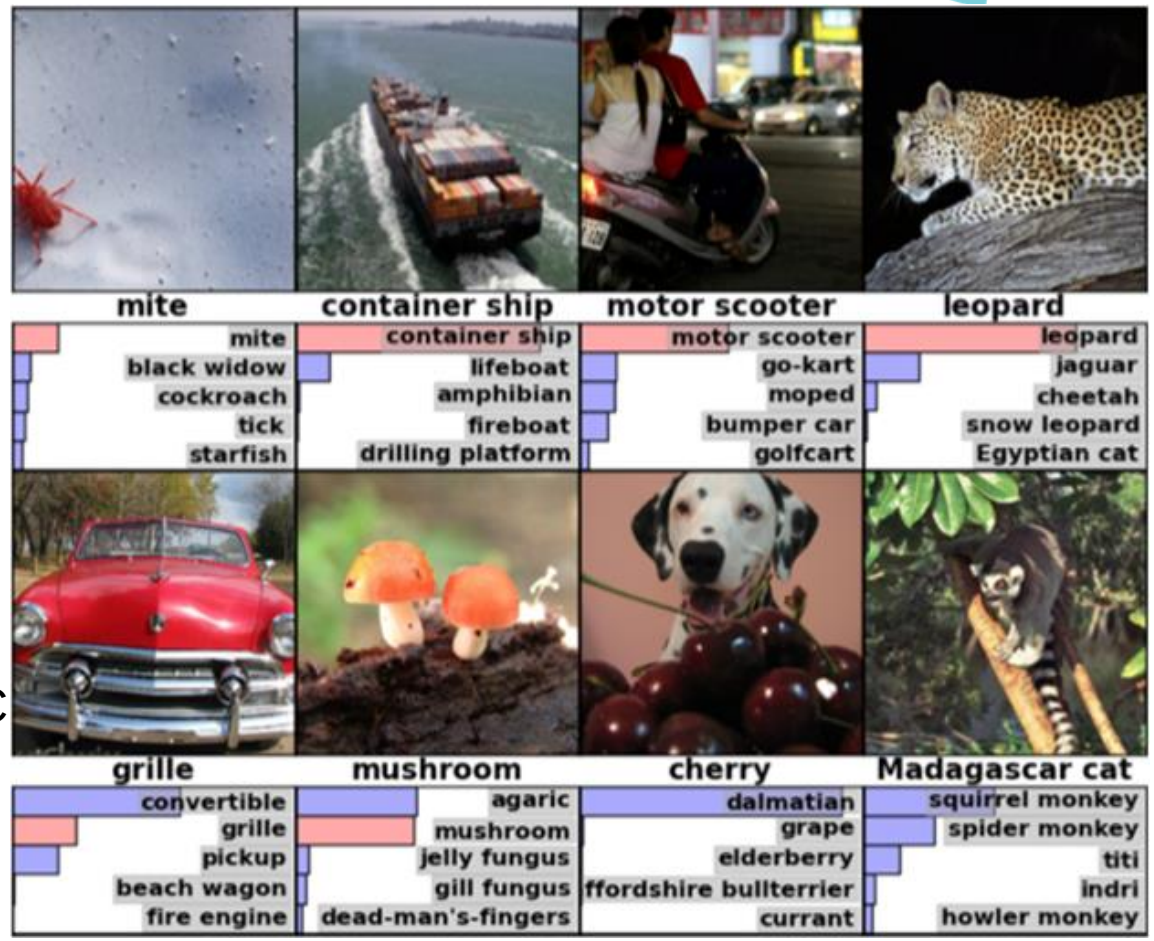
Needless to say, Sussman didn't make the deadline.

Vision turned out to be one of the most difficult and frustrating challenges in AI over the next four decades.

AlexNet 2012 (ten years ago)
ImageNet competition winner



- Operator Types: 11x11, 5x5, 3x3, MaxPool 3x3s2, FC
- Total Layers: 8
- Output is classification to 1000 classes
- Operations per Inference: 724 Million



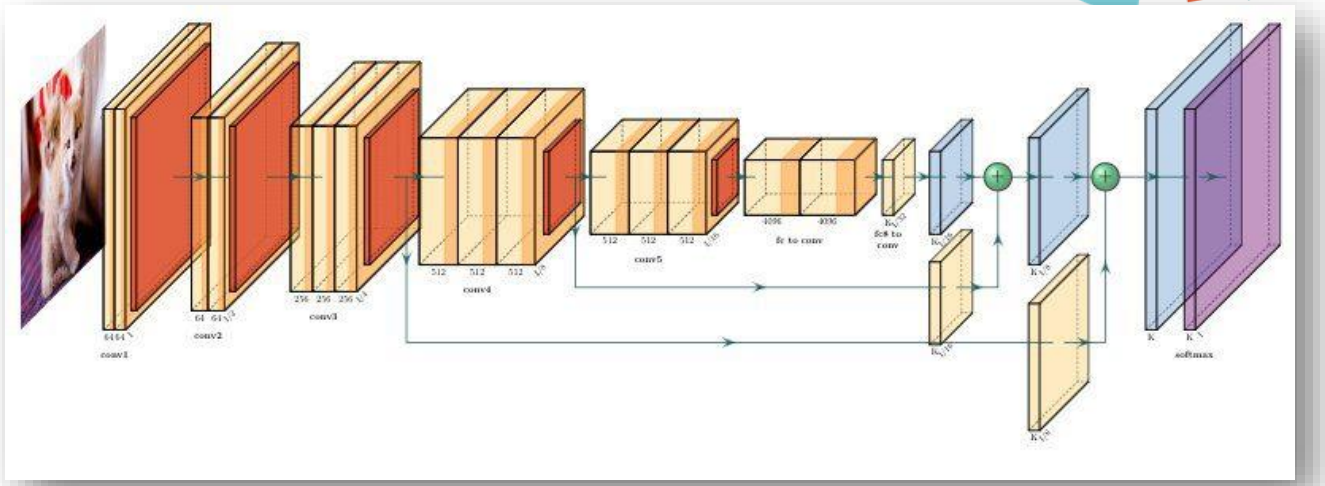
What are the tough new problems?

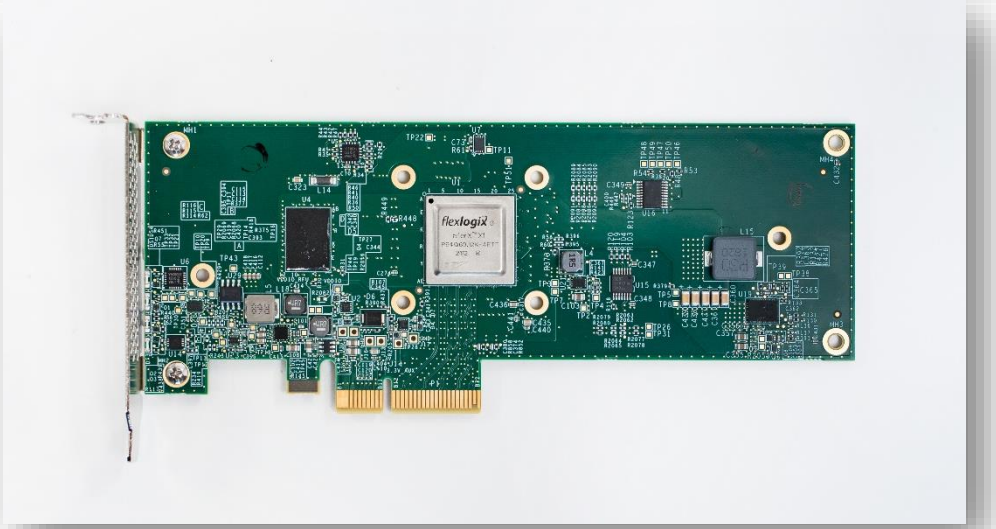
- General Purpose Processing?
- Operating Systems?
- Rugged/Industrial Computers?
- Network Connectivity?
- Software Paradigm?
- Imaging Sensor?

All Solved Problems

What are the tough new problems?

- Software Complexity
  - How do you program a trillion operations?
- TeraOp Processing Efficiency
  - How do you fit TeraOps in a factory?





Miniaturizing AI vision systems

Providing TeraOps of vision processing into smaller form-factors is a real challenge

Industrial Vision Computers

High Power – 250 W Card
System Power – 750 W

$$$$

$

Compact Vision Box

Low Power – 8 W Card
System Power – 30 W

flexlogix
AI + eFPGA

GPUs offer flexibility but at a price

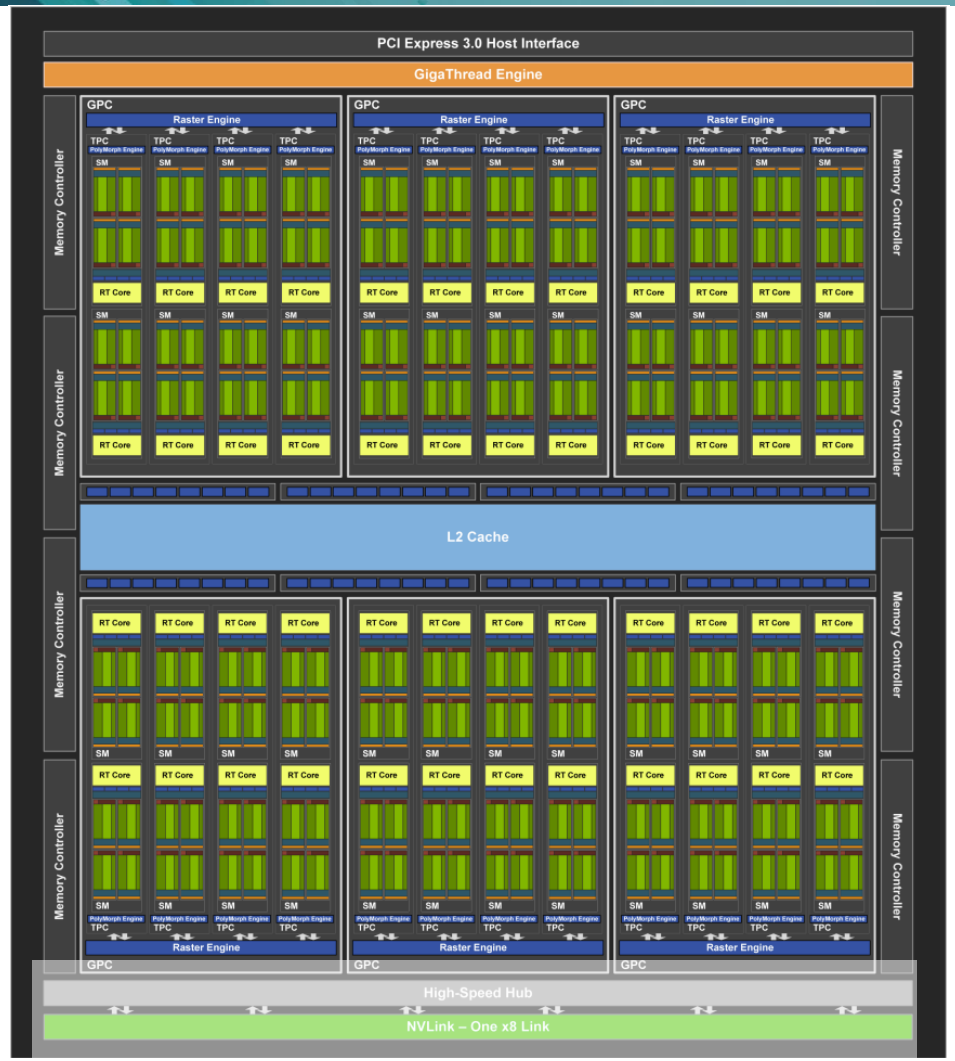
- Good for inference and training
- Tons of memory bandwidth
- Numerous precision choices
- Good SW ecosystem to get started

- But ..
- Large, expensive, power hungry
  - Further amplified at the system level
- Easy to get started, but hard to optimize for
- Support difficulties
- Supply-chain difficulties

Flexible but…

- Large
- Expensive
- Power hungry

It’s all about the memory



Turing TU104 Full Chip Diagram

	TU104
CUDA Cores	3072
SMs	48
Texture Units	192
RT Cores	48
Tensor Cores	384
ROPs	64
Memory Bus Width	256-bit
L2 Cache	4MB
Register File (Total)	12MB
Architecture	Turing
Manufacturing Process	TSMC 12nm "FFN"
Die Size	545mm2

GPUs are architected for DDR, not local memories

- Computation is designed to access GDDR memory
- NVLink used for further expansion
- Only 4 + 12 MB of L2 + RF

Highly parallelized version of Von Neumann architecture

- Still inefficient, but highly flexible

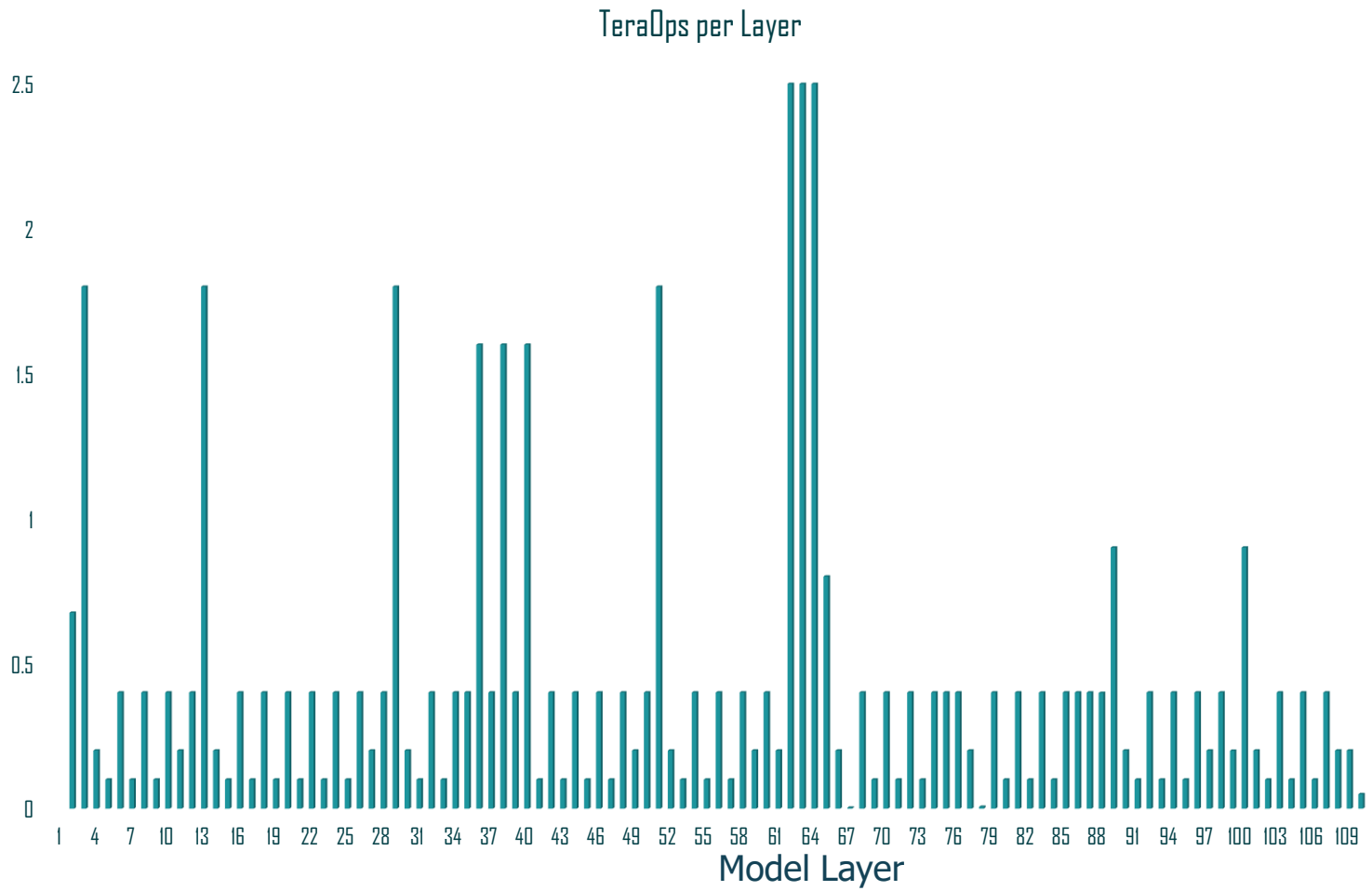
Fast model evolution – Flexibility is key

- Models are evolving rapidly, even the **same-model** is going through incremental changes
- Different versions, flavors (s/m/l/xl), image sizes, are added
- New operators are introduced regularly
- Flexible architecture is a MUST
    - Dedicated ASICs cannot chase a moving target

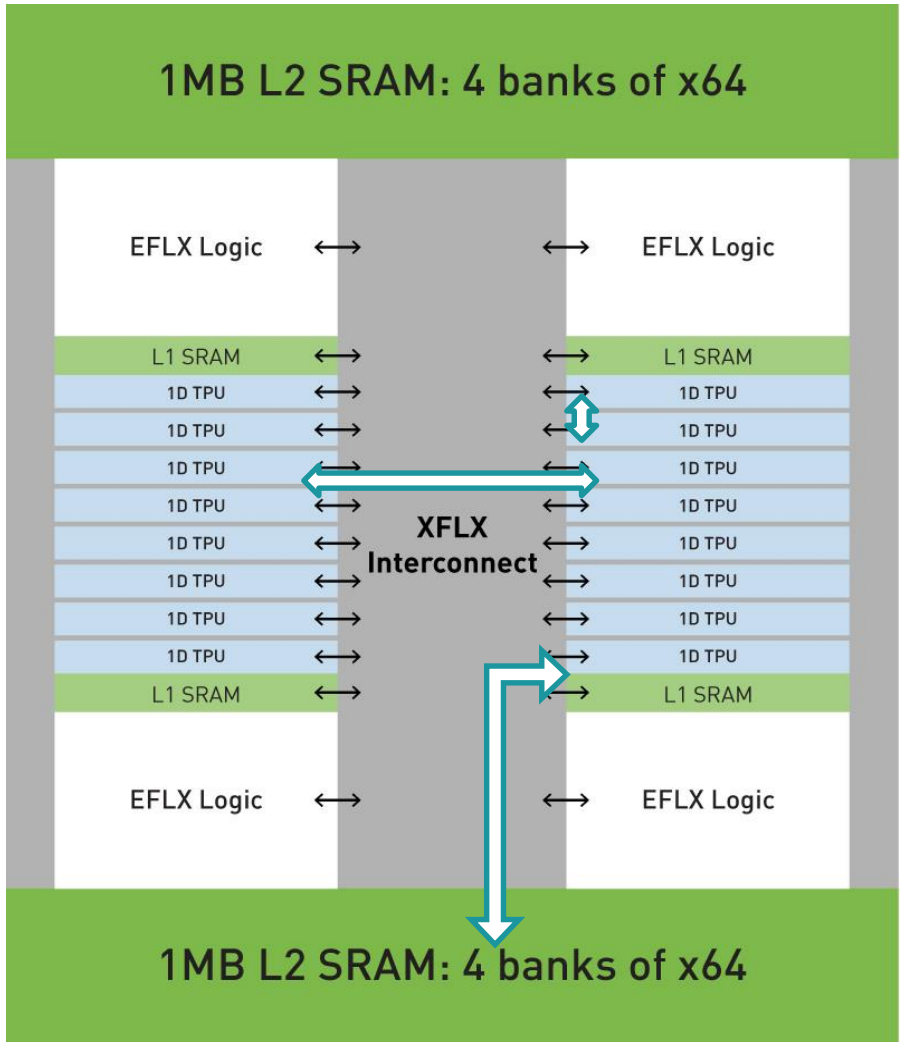
YOLOv5	Changes
V1 May, 2020	Initial Release
V2 July, 2020	LeakyReLU(0.1)
V3 August 2020	HardSwish activations added
V4 January 2021	siLU() replaces leakyReLU and hardswish
V5 April 2021	Added P6 (1280x1280) models
V6 October 2021	SPP -> SPPF, C3 changes

Load-balancing difficult for streaming architectures

- Graph streaming reduces DRAM requirements, but BW matching is difficult
- Each operator requires different amount of compute
- How to maintain efficiency with load imbalance?



Dynamic TPU:
Flexible, balanced & memory-efficient



Efficient data access:

Each 1D TPU core can stream data from:

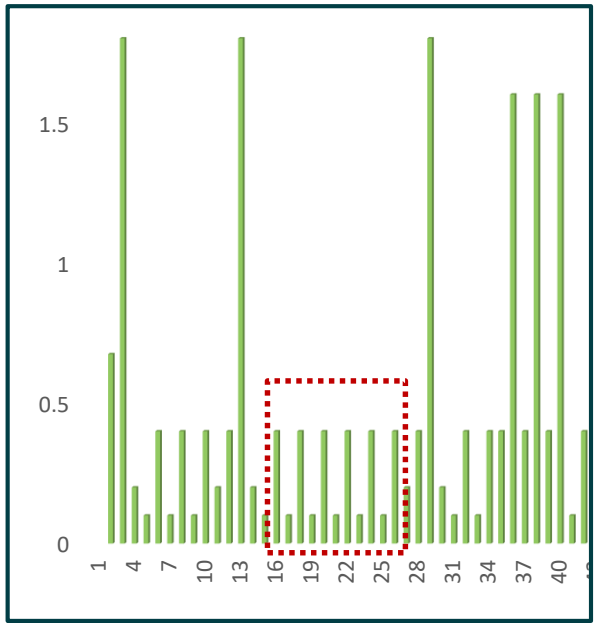
- Neighboring 1D TPU (dedicated)
- Any 1D TPU (via XFLX)
- L2 SRAM (via XFLX)
- DDR (via NoC)

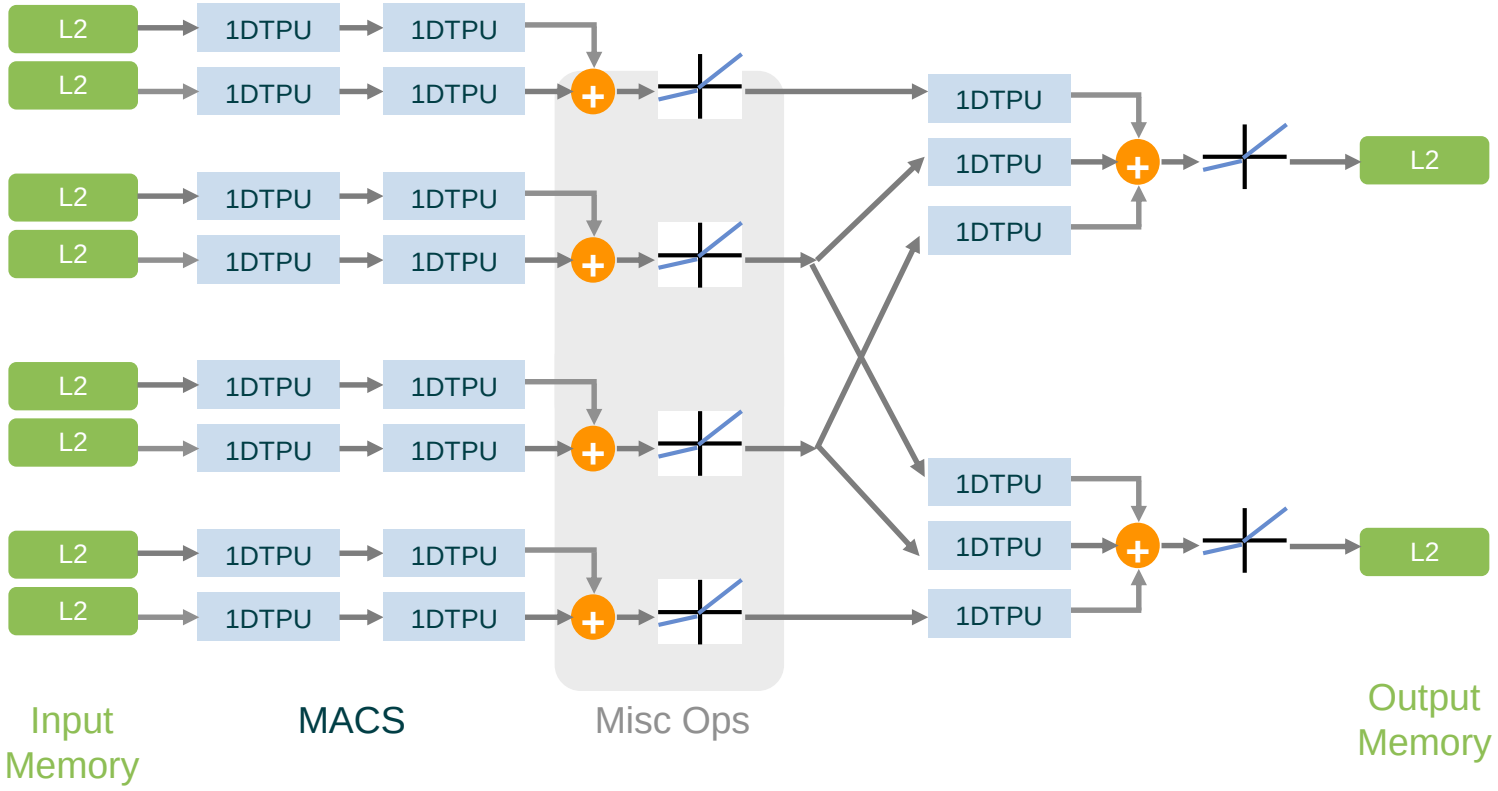
Flexible control and data path:

- "Future proof" compute, activations, and generic operators via EFLX

Layer fusion – Match workloads at sublayer level

- Stream data at a sub-graph level with efficient BW matching





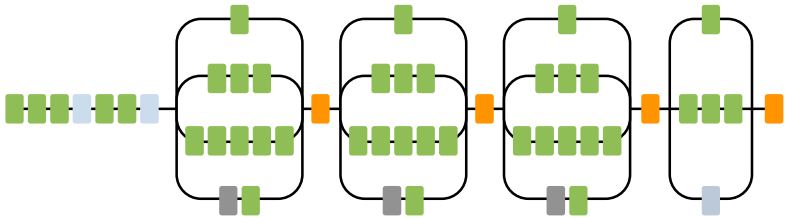
Comparing GPU to our dynamic TPU

GPU	Dynamic TPU
Lots of MACs	Fewer MACs, but more efficiently used
Computes predominantly via GDDR	Compute via local connections, XFLX connections, flexible L2s, and DDR
Lots of SW but hard to achieve efficiency	Easy to achieve efficiency with X1 XDK
Many GDDR Memory (256-bit)	Few LPDDR (32-bit)
75 – 300 W (typ.)	6 – 10 W (typ.)
Flexible	Equally flexible via EFLX & XFLX
Large, expensive & brute-force	Small, low-cost & efficient

System level view of InferXDK software

NN Model Framework

Conv
AvgPool
MaxPool
Concat



TensorFlow

Neural Net

Embedded App at The Edge

Embedded Application
Runtime API
X1 drivers
X1 Runtime
+
X1

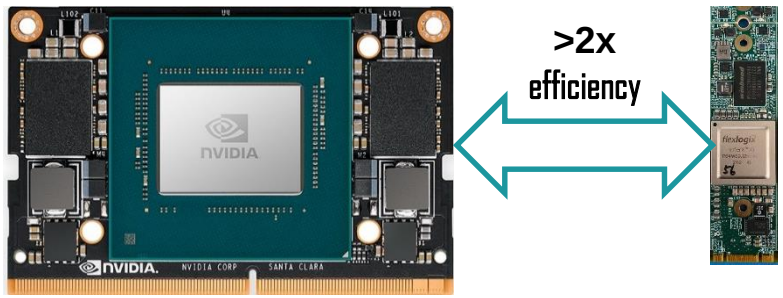
.ncf

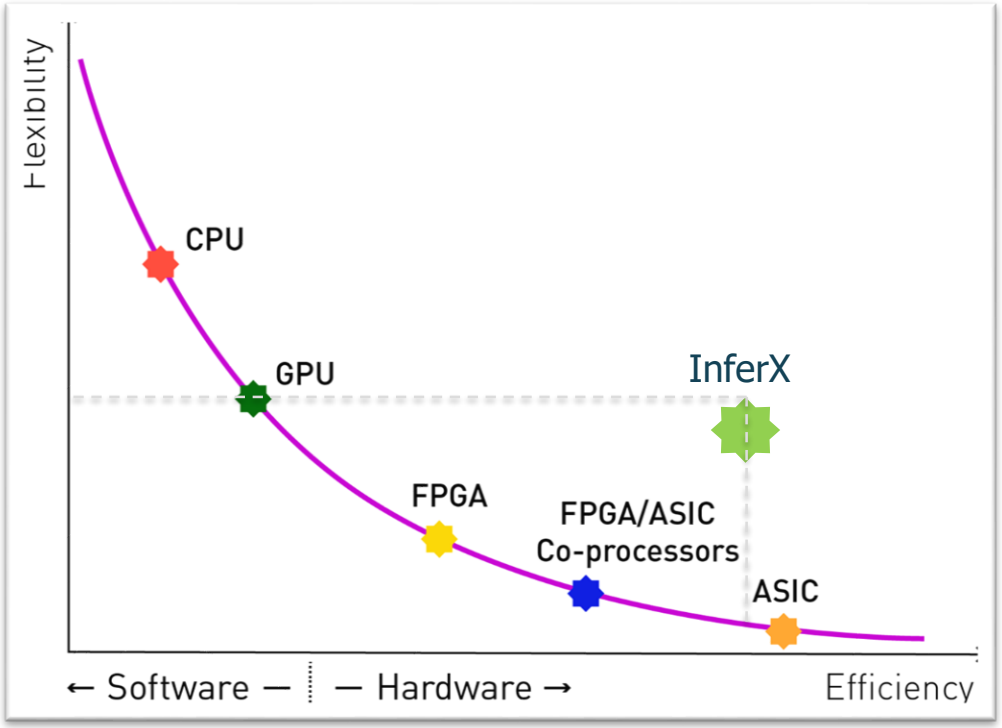
flexlogix AI + eFPGA

InferXDK Software Development Toolkit

Superior performance with SW flexibility

- X1 provides **ASIC performance**/efficiency with flexibility of software
- InferX SDK directly converts neural network graph model to **dynamic InferX hardware instance**
- Much more flexible & future proof vs ASIC solutions
- Much higher efficiency (**Inf/W** & **Inf/$**) vs CPU and GPU based solution
  - Thus enabling compact form factors such as M.2 2280 B+M





2019: Xin Feng, Computer vision algorithms and hardware implementations: A survey

Conclusions

- Putting TeraOps of performance in low power edge devices is today's challenge
- InferX was designed from scratch to solve this problem
- > 10x improvement in efficiency versus comparable GPUs
- Smart architecture reduces memory bandwidth and capacity requirements
- While supporting complex models at high throughput
- Designed to fit in small and low power system form factors

Come visit us in the expo hall for demonstrations for InferX technology

Thank you!

flexlogix
AI + eFPGA