AI On a Battery

Kris Ardis – Executive Director of Micros, Security & Software BU

An AI revolution is underway…

…allowing machines to see, hear and sense the world around them.

maxim integrated™

Gap Between Big Machines and Little Machines

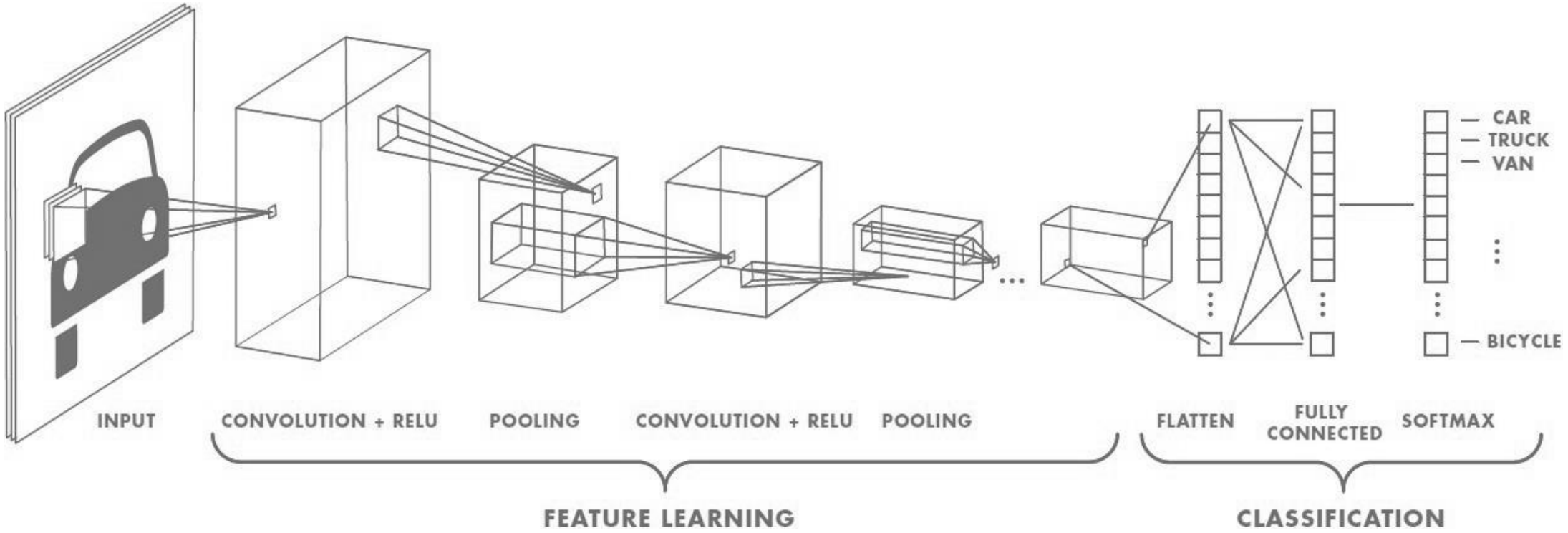
“Unconstrained”
Today: self-driving

Cost, size, and power constraints
Today: simple wakewords

maxim integrated™

Convolutional Neural Networks are the workhorse

…but are computationally expensive!



Millions/Billions of Multiplications!

What if We Could Close the Gap?

Cameras and other battery-powered devices could trigger on smarter warnings, recognize authorized entry

Robots and tools could listen to our commands, avoid obstacles better, change their behavior based on complex sensor data

Handheld or smaller devices could react to complex spoken commands or a wide array of specific sounds

…and so many things we can't even imagine.

Energy

Need to run on batteries and other constrained power sources

Latency

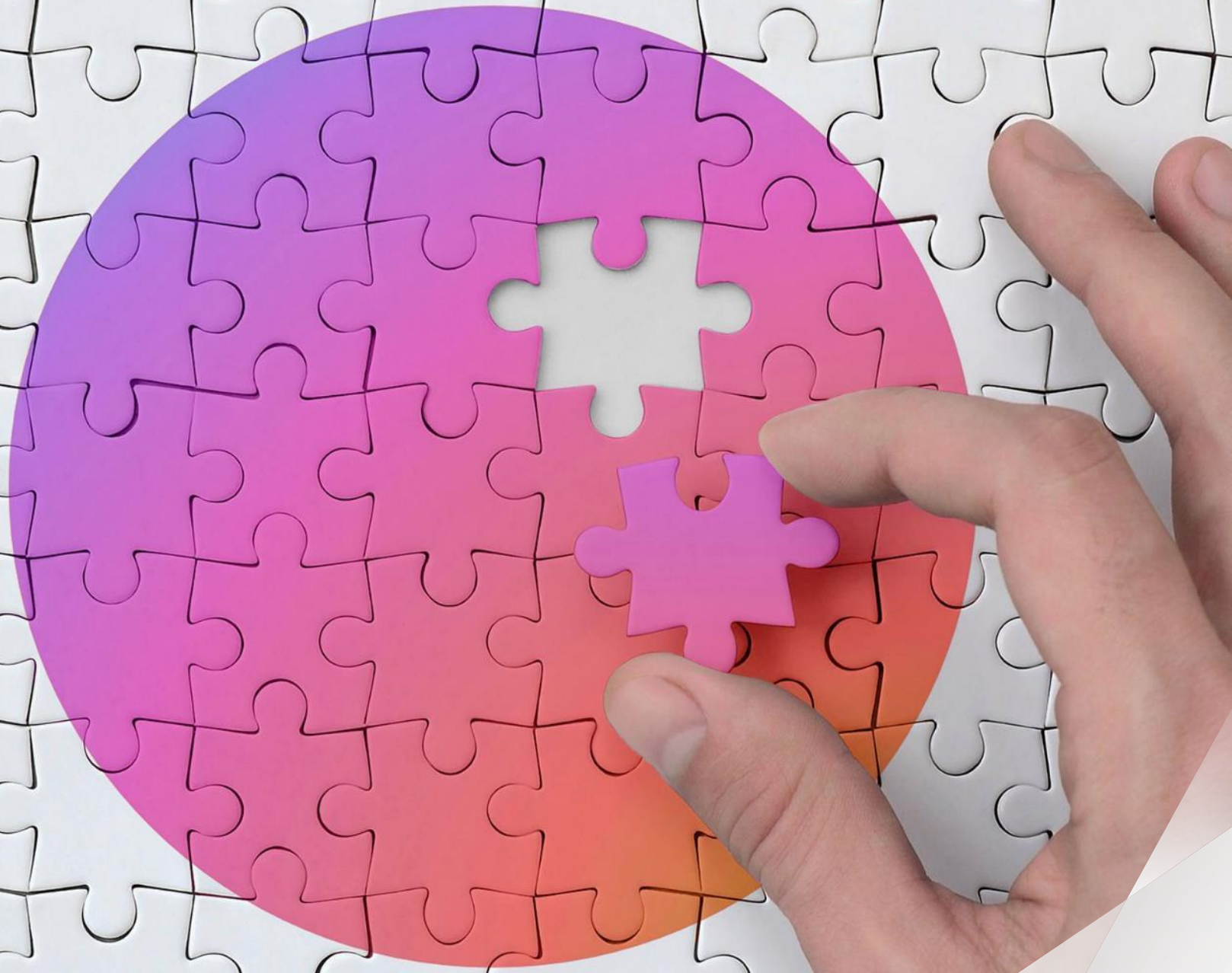
Needs to run fast enough for real-time insights and safety applications

Size

Needs to add intelligence without adding bulk to enable wide adoption

Cost

GPU cost points are ok for cars, but will never enable large deployments



Why Maxim?

Energy
Focus in wearables and low-power IoT applications

Latency
Experience making specialized hardware to solve tricky problems

Size
Strategy of integration to build the smallest end products

Cost
New approach enables mass deployment of embedded AI

Introducing the MAX78000

Ultra low power micro

- ARM Cortex-M4F 100 MHz
- 16 KB cache
- RISC-V Smart DMA 60 MHz, cache
- 4-ch DMA
- 512 KB Flash
- 128 KB SRAM

SIMO/DVS

- AES 128/192/256
- TRNG
- Unique ID

Clocking

100 MHz RO	32 kHz XO (RTC)
60 MHz RO	7.3728 MHz RO
8-30 kHz RO	Ext. Square Wave (up to 80 MHz)

External interfaces

I^2S M/S	3 × UART 1 × LP
3 × I^2C (Hi-speed)	
2 × SPI M/S	9 × Timers 6 × 32-bit, 2 × LP, 2 × Watchdog, 1 × Wakeup
4 × Pulse train	
1-Wire master	
Parallel camera	8-ch 10-bit ADC
SWD	4 µPower Comparator

CNN accelerator

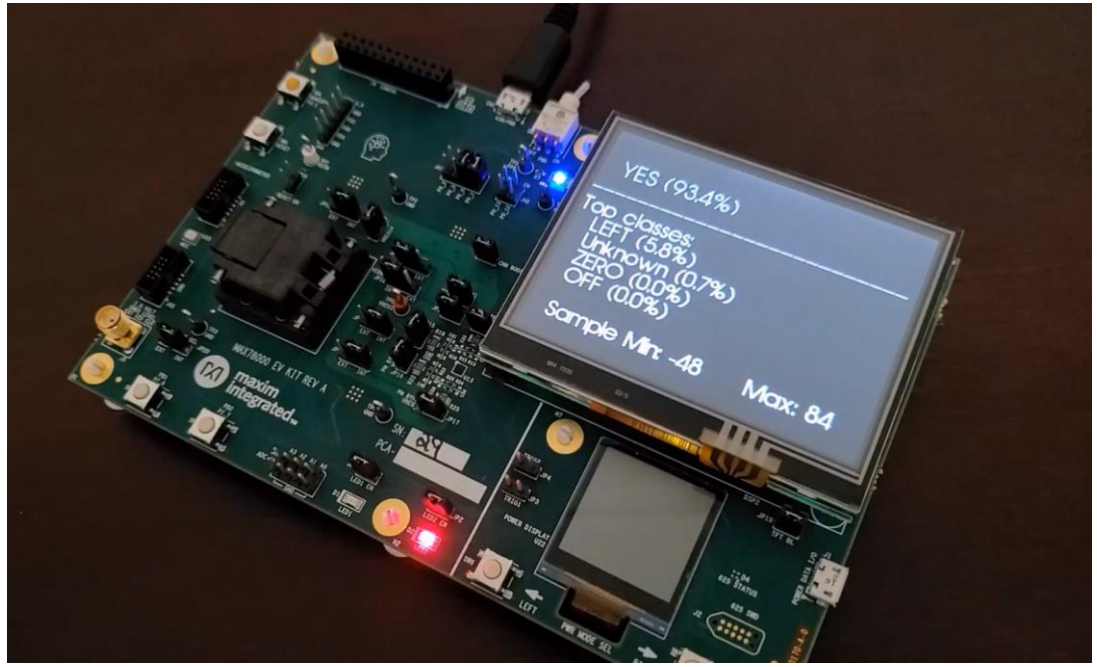
Parallel processors	64
Max layers	32…64[1]
Max input/output channels in any layer	1024
Max weights	432 KB[2] (up to 3.5 M weights)
Data memory	512 KB + 384 KB
Max. input dimensions	1023 × 1023 (per channel, streaming) 181 × 181 (per channel, preloaded)

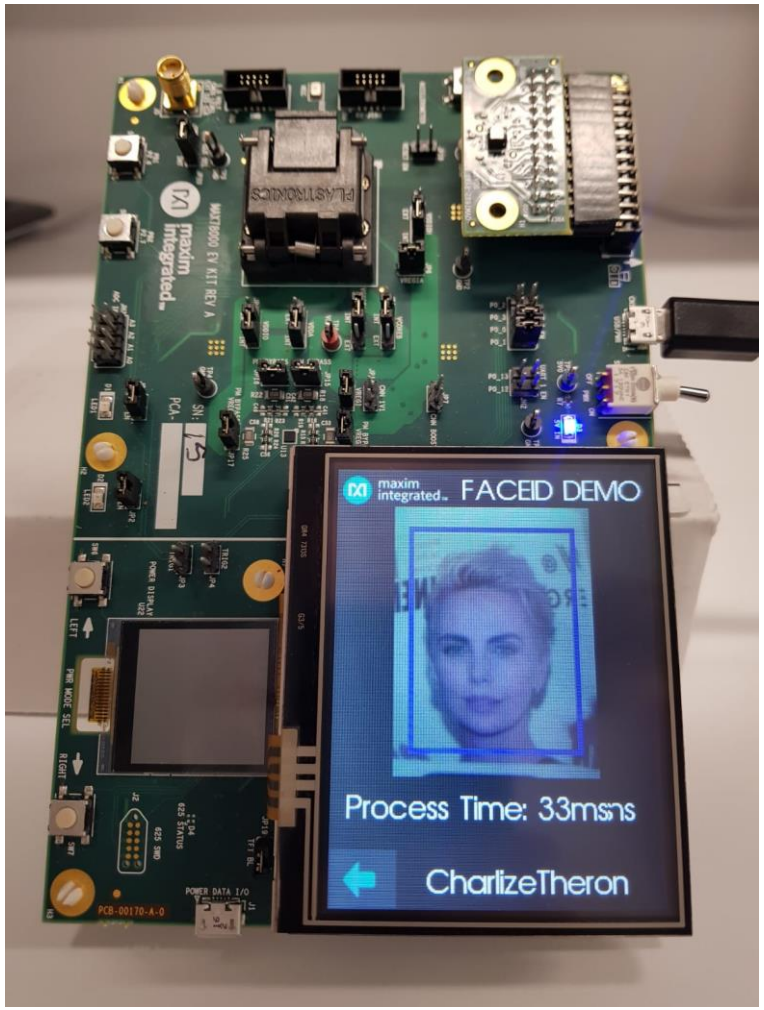
1 – Up to 64 with pooling every other layer, up to 32 with no pooling
2 – Weights can be 1-bit, 2-bit, 4-bit, or 8-bit, selectable per layer

MAX78000's Neural Network Accelerator

- New, novel architecture designed to minimize data movement, maximize parallelism and optimize energy spend
- No μC involvement except to load and start
- No external memory required
- Highly optimized for Convolutional Neural Networks
- Flexible clock control to run fast at higher current or run slow at lower current

What can the MAX78000 do?





Audio and image analysis demos included in the box
(<40% of network used)

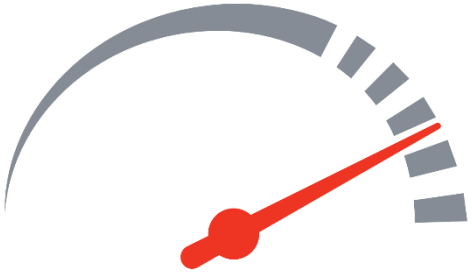
MAX78000: Meeting the challenge of AI at the edge

Energy

Running same network, vs. Low Power Cortex M4F:

- 1,100x lower energy on MNIST
- 600x lower energy on keyword spotting

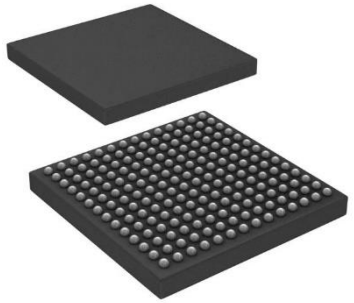
Latency



Running same network, vs. 96MHz Cortex M4F:

- 400x faster on MNIST
- 200x faster on keyword spotting

Size



- 8mm x 8mm BGA package
- Few external components needed
- Less than ½ size of any GPU or big processor AI solution

Cost

- Significantly less than GPU solutions
- No external memory
- Few external components, simpler placement further reduce system cost

Developing with the MAX78000

IAR SYSTEMS

eclipse

KEIL™ Tools by ARM

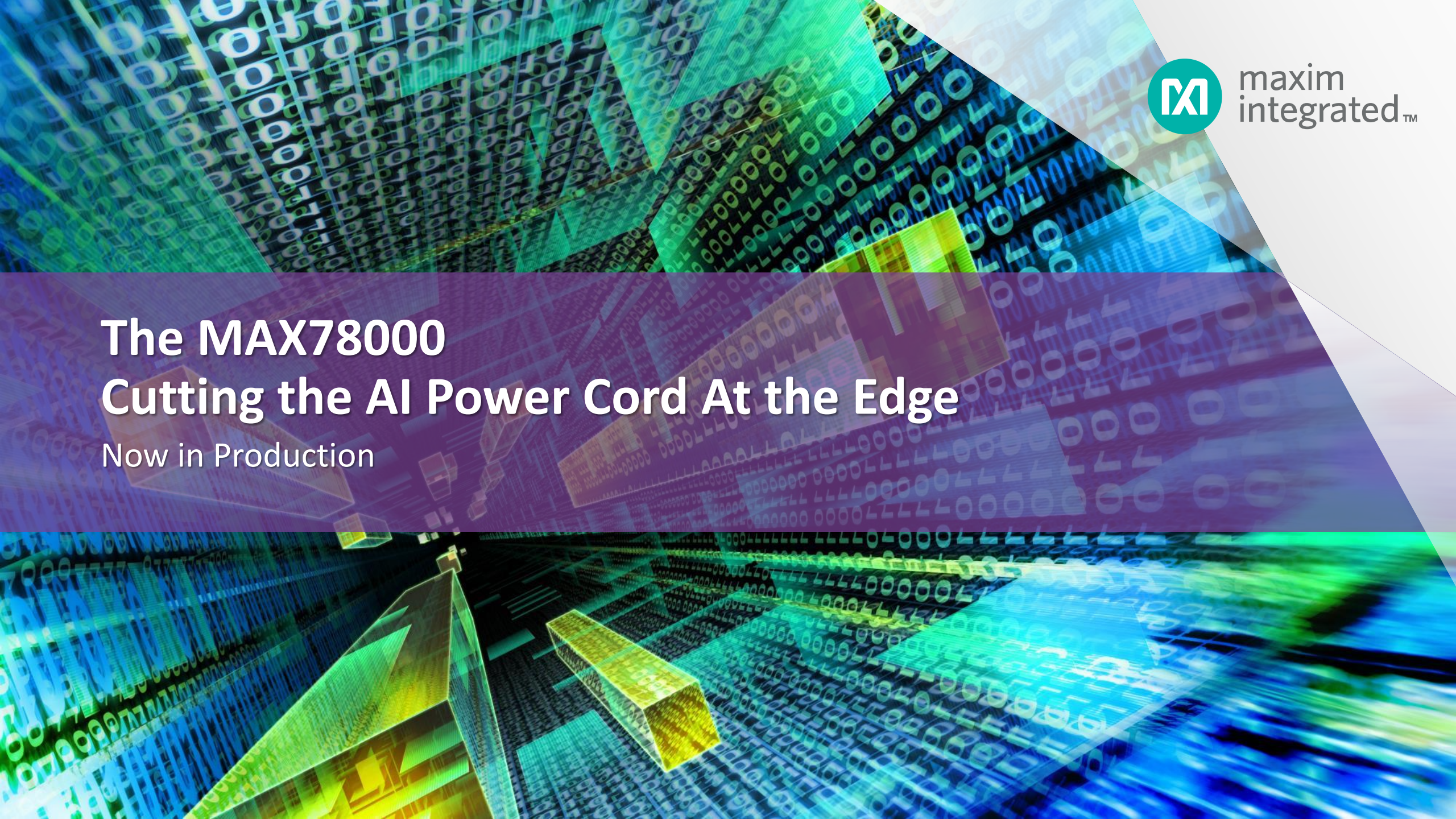
TensorFlow

PyTorch

GitHub

https://github.com/MaximIntegratedAI
Documentation, tools, examples…

maxim integrated™



The MAX78000
Cutting the AI Power Cord At the Edge

Now in Production

maxim integrated™