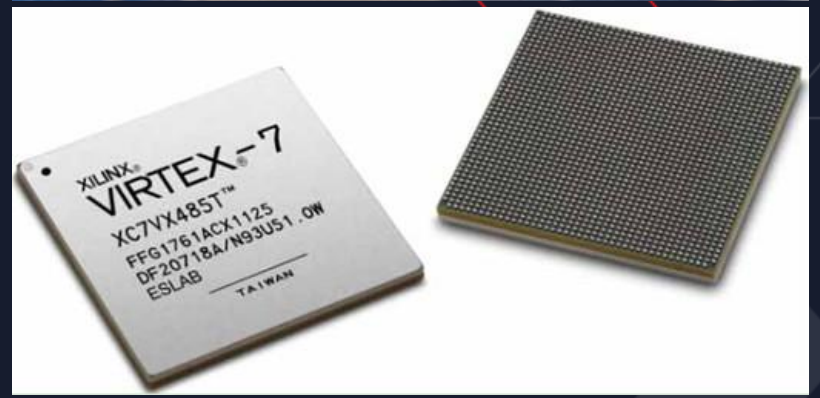
详解AI芯片技术与产业发展路径

姚颂
赛灵思人工智能高级总监
前深鉴科技创始人&CEO
songyao@xilinx.com
2020年7月4日

XILINX.

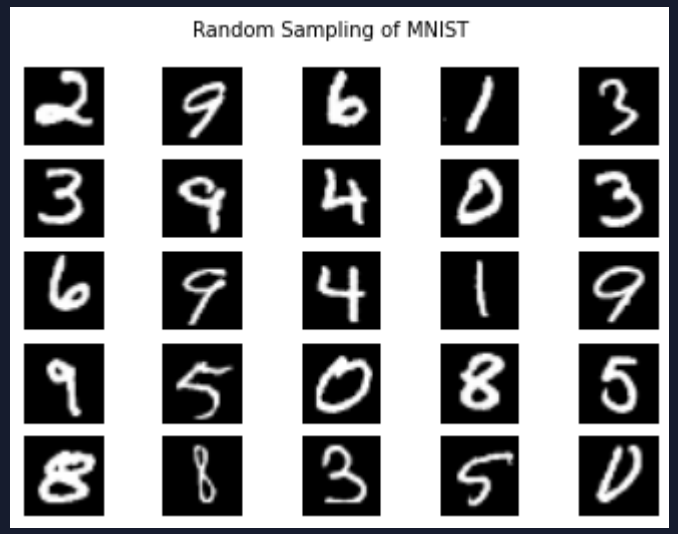
从深鉴科技到Xilinx（赛灵思）

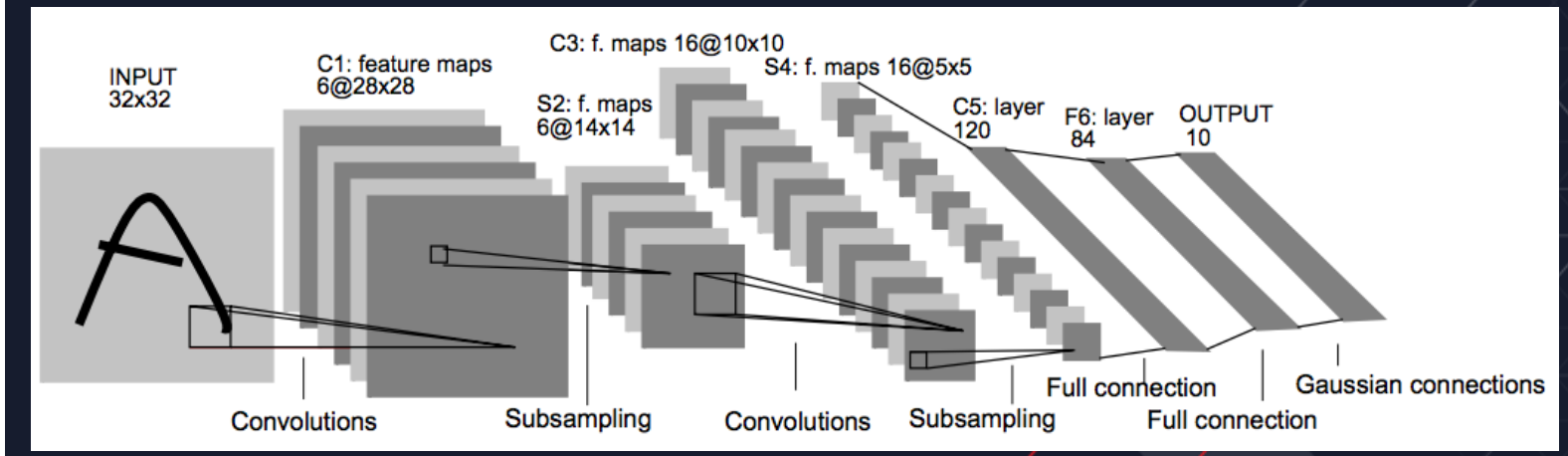


- Xilinx：成立于1984年
- 世界第一家Fabless芯片设计公司
- 全球4,000+员工，年营收超过30亿美金
- FPGA芯片发明人
- FPGA领域>56%全球市场份额

AI需要芯片
芯片需要AI

深度学习算法在上世纪80-90年代已经被认真研究



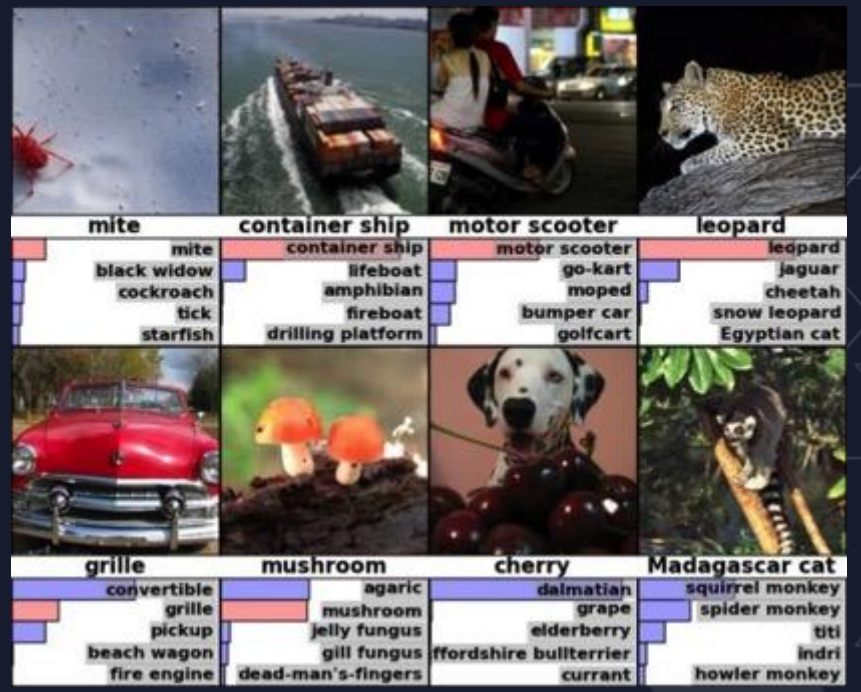


MNIST, Prof. Yann Lecun, 1998
手写数字识别

直到GPU与CUDA的出现才让深度学习复兴

猫脸识别
1000台服务器，16000个CPU

Andrew Ng & Jeff Dean, Google, 2012

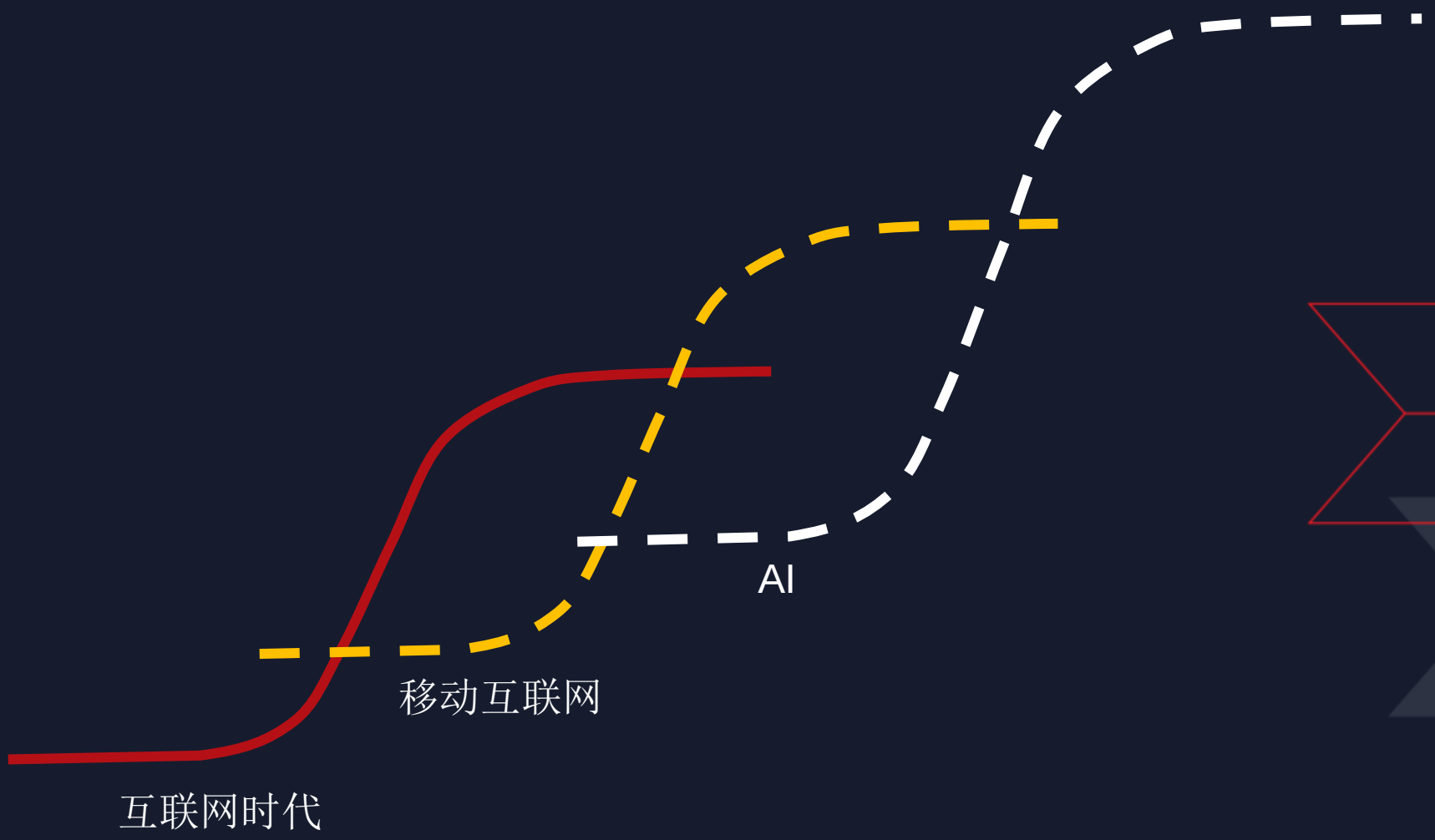


AlexNet for ImageNet
1台服务器，2个GPU

Geoffrey Hinton, 2012

三大因素共同驱动了人工智能的兴起

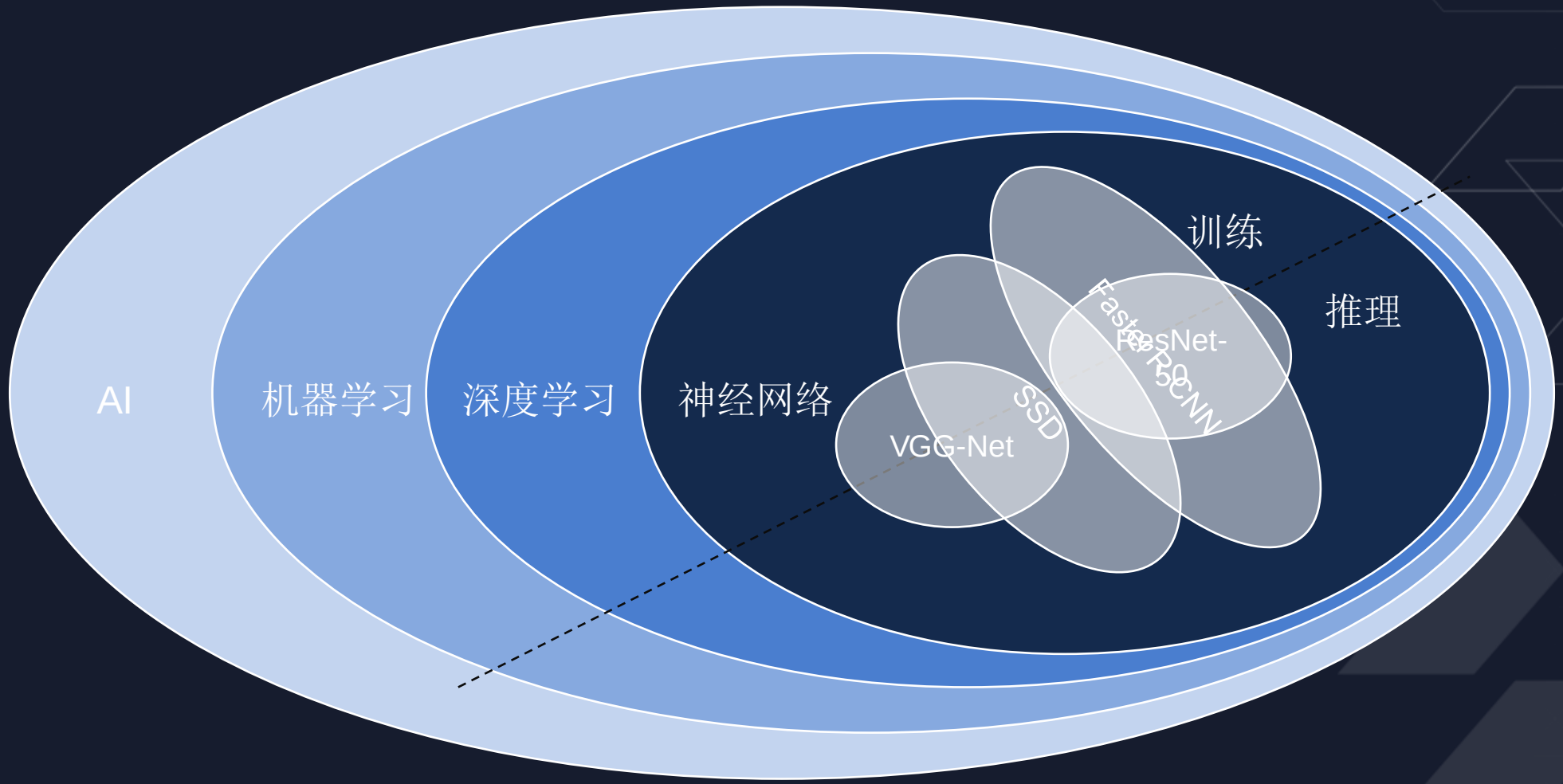
关键应用又引领了芯片的一次又一次升级



XILINX.

什么是AI芯片？

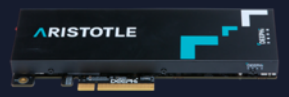
只要为AI优化就是AI芯片



XILINX.

AI芯片分类：根据AI的阶段与应用场景

Horizon Journey



DeePhi DPU

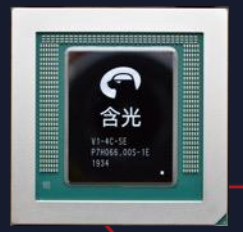
Movidius DSP

Mobileye Eye Q5

推理

Xilinx FPGA

Google TPU



平头哥 含光NPU

寒武纪 MLU

终端

云端

Intel Habana

燧原 云燧T10

nVIDIA GPU

训练

XILINX.

AI芯片最重要解决的是带宽不足的问题

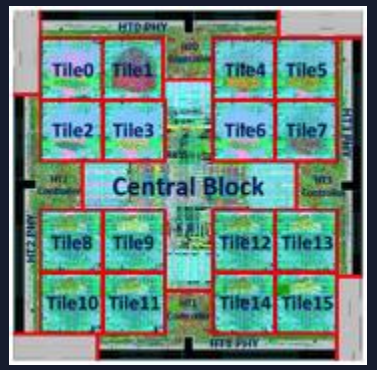
Application	*MLP0*	*MLP1*	*LSTM0*	*LSTM1*	*CNN0*	*CNN1*	*Mean*	*Row*
Array active cycles	12.7%	10.6%	8.2%	10.5%	78.2%	46.2%	28%	1
Useful MACs in 64K matrix (% peak)	12.5%	9.4%	8.2%	6.3%	78.2%	22.5%	23%	2
Unused MACs	0.3%	1.2%	0.0%	4.2%	0.0%	23.7%	5%	3
Weight stall cycles	53.9%	44.2%	58.1%	62.1%	0.0%	28.1%	43%	4
Weight shift cycles	15.9%	13.4%	15.8%	17.1%	0.0%	7.0%	12%	5
Non-matrix cycles	17.5%	31.9%	17.9%	10.3%	21.8%	18.7%	20%	6
RAW stalls	3.3%	8.4%	14.6%	10.6%	3.5%	22.8%	11%	7
Input data stalls	6.1%	8.8%	5.1%	2.4%	3.4%	0.6%	4%	8
TeraOps/sec (92 Peak)	12.3	9.7	3.7	2.8	86.0	14.1	21.4	9

无法完全存储于片上SRAM 性能很低

完全存储于片上SRAM 性能很高

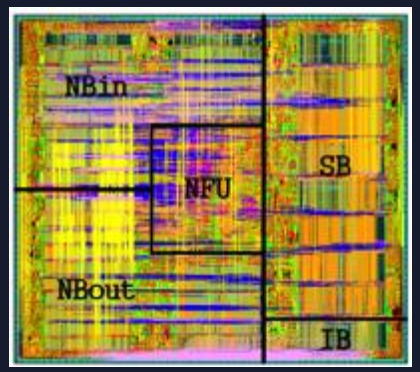
Google TPU的利用率数据

超大的片上存储是高效但奢侈的



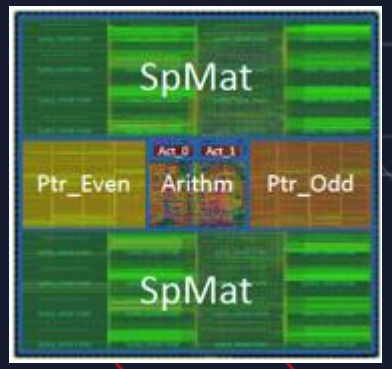
DaDianNao: 36MB eDRAM

Source: Yunji Chen et al., "DaDianNao…", Micro 2014



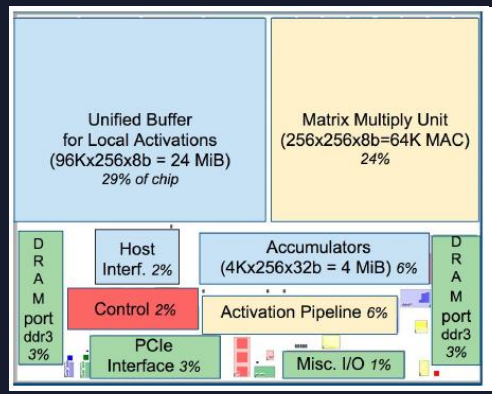
ShiDianNao: 256KB SRAM

Source: Zidong Du et al., "ShiDianNao …", ISCA 2015



EIE:10.13MB SRAM

Source: Song Han et al., "EIE: …", ISCA 2016

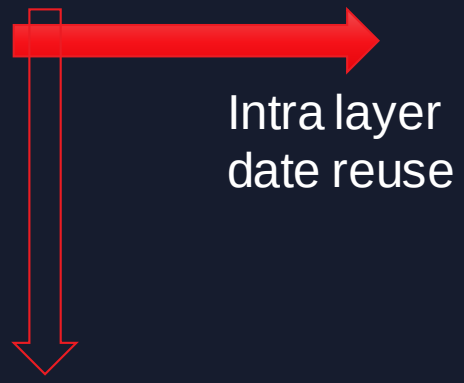


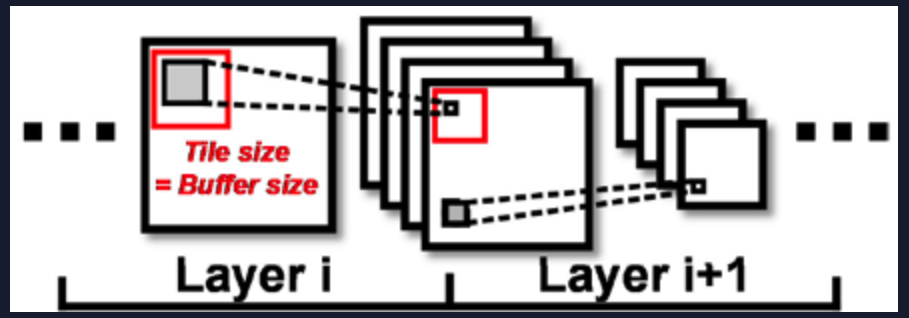
TPU:28MB SRAM

Source: Norman et al., "In-Datacenter: …", ISCA 2017

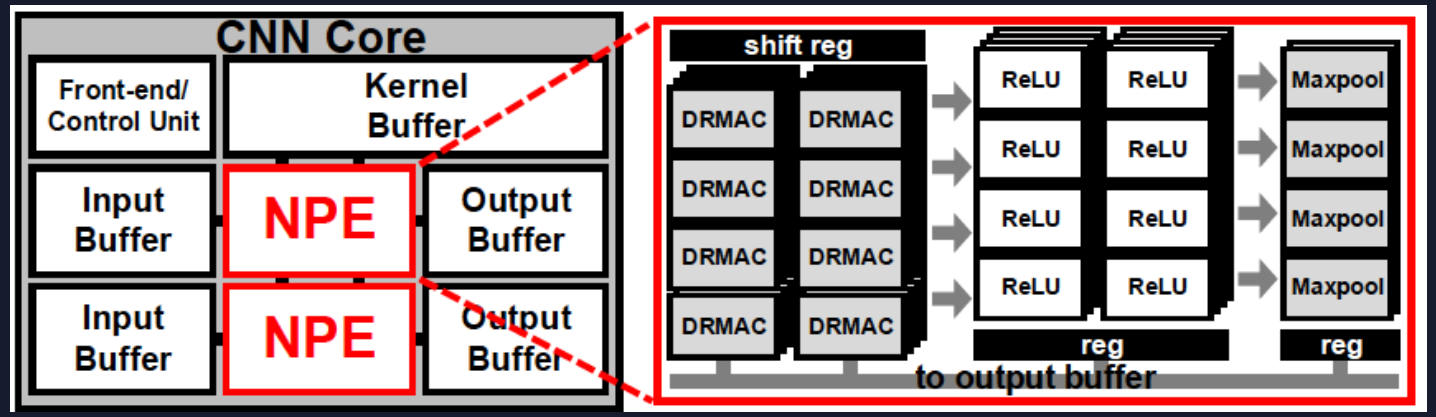
XILINX.

AI芯片的微架构设计核心要减少访存次数

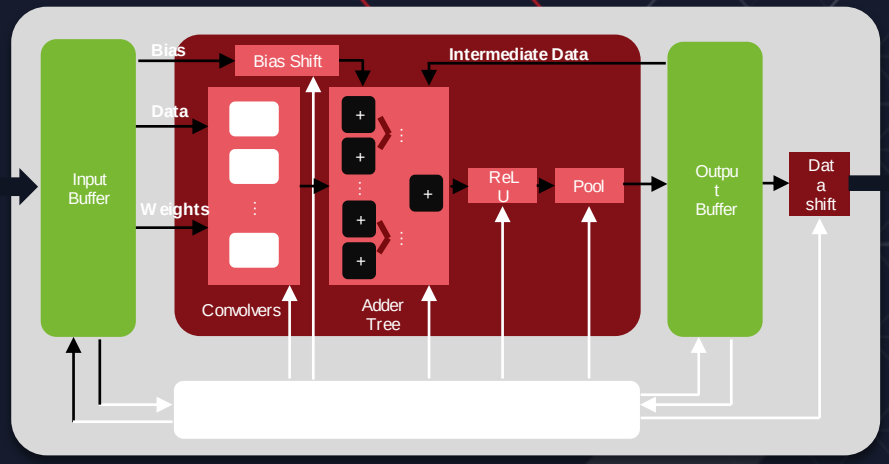




Jaehyeong Sim et al., "A 1.4TOPS/W…", ISSCC 2016

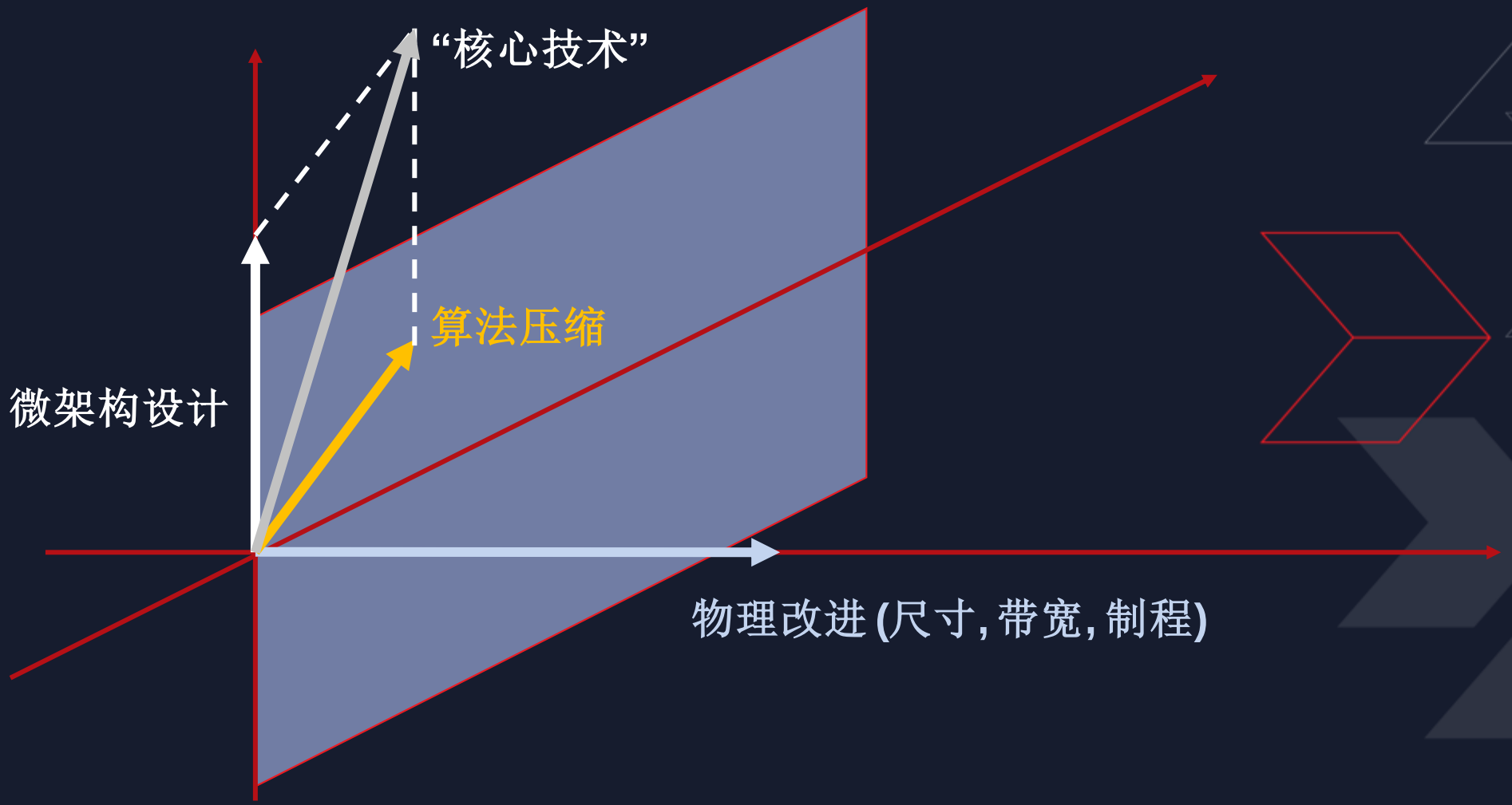


Jaehyeong Sim et al., "A 1.4TOPS/W…", ISSCC 2016

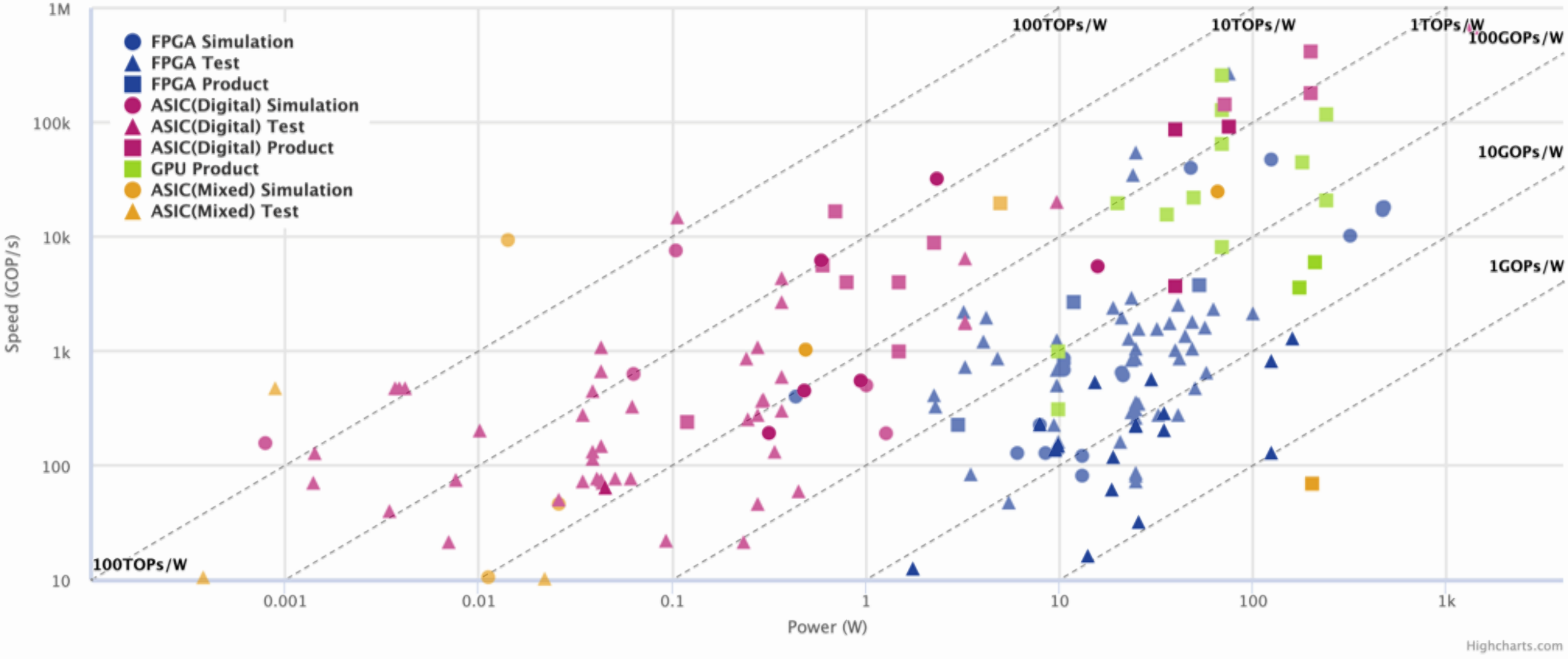


Jiantao Qiu et al., "Going deeper…", FPGA2016

AI计算优化的三个维度



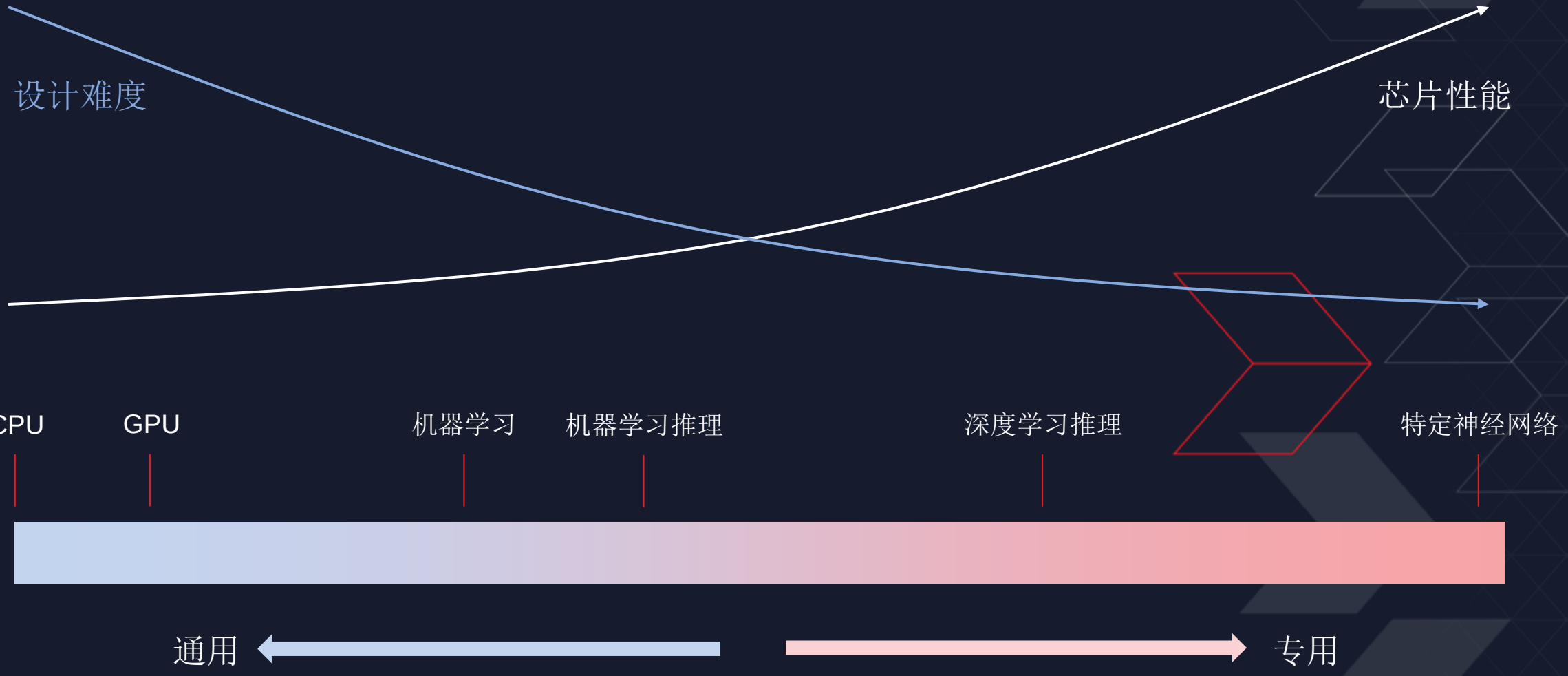
AI芯片近些年进步非常大



从2016年到2019年的AI芯片能效指标变化（越左上角约好）

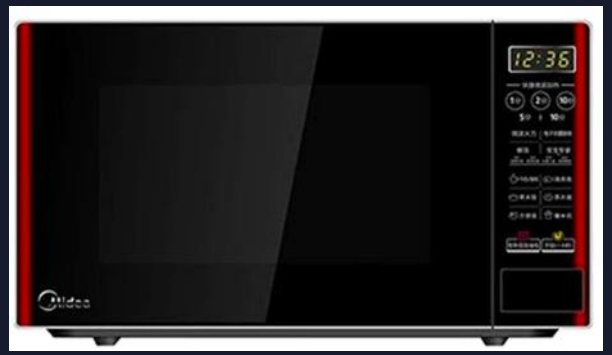
https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

AI芯片并没有那么难



XILINX.

同是AI芯片，需求千差万别



智能微波炉的语音控制
极低成本

无线耳机的语音唤醒与识别
极低功耗，较低成本

自动/辅助驾驶
低延迟，高可靠

无线智能摄像机
极低功耗，极低成本

“AI芯片”理清概念

> AI芯片 ≠ 新技术 ≠ 产品 ≠ 能赚钱的商品

> 全新微架构设计 = 新技术

> AI芯片 = 新技术（有的话）的载体

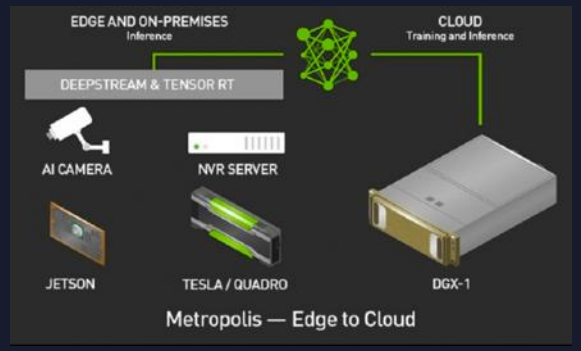
> AI芯片 + 软件 = 产品

> AI芯片 + 软件 + 生态 + 行业优化 = 有竞争力的产品

> 生态与软件的重要性远大于芯片本身

追根溯源：
英伟达为什么这么牛？

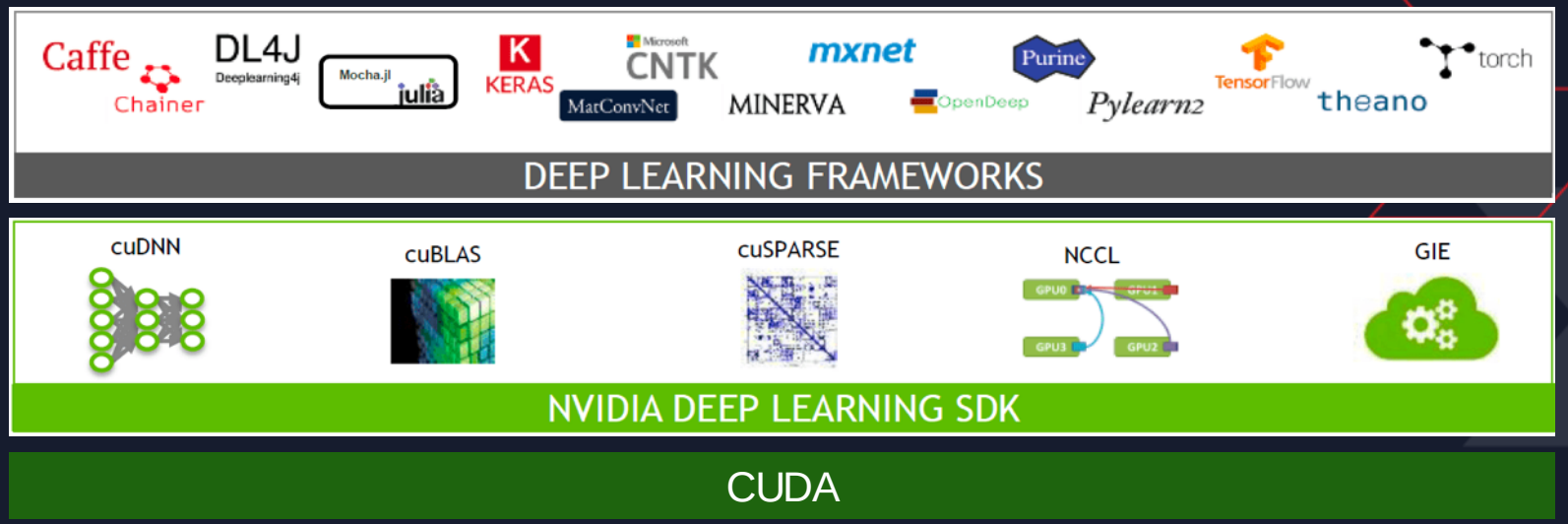
英伟达的软件堆栈是其核心竞争力



DeepStream in Security Industry

DRIVEWORKS SDK	DETECTION	LOCALIZATION	DRIVING	VISUALIZATION
	Detection/Classification	Map Localization	Vehicle Control	Streaming to cluster
	Sensor Fusion	HD-Map Interfacing	Scene understanding	ADAS rendering
	Segmentation	Egomotion (SFM, Visual Odometry)	Path Planning solvers	Debug Rendering

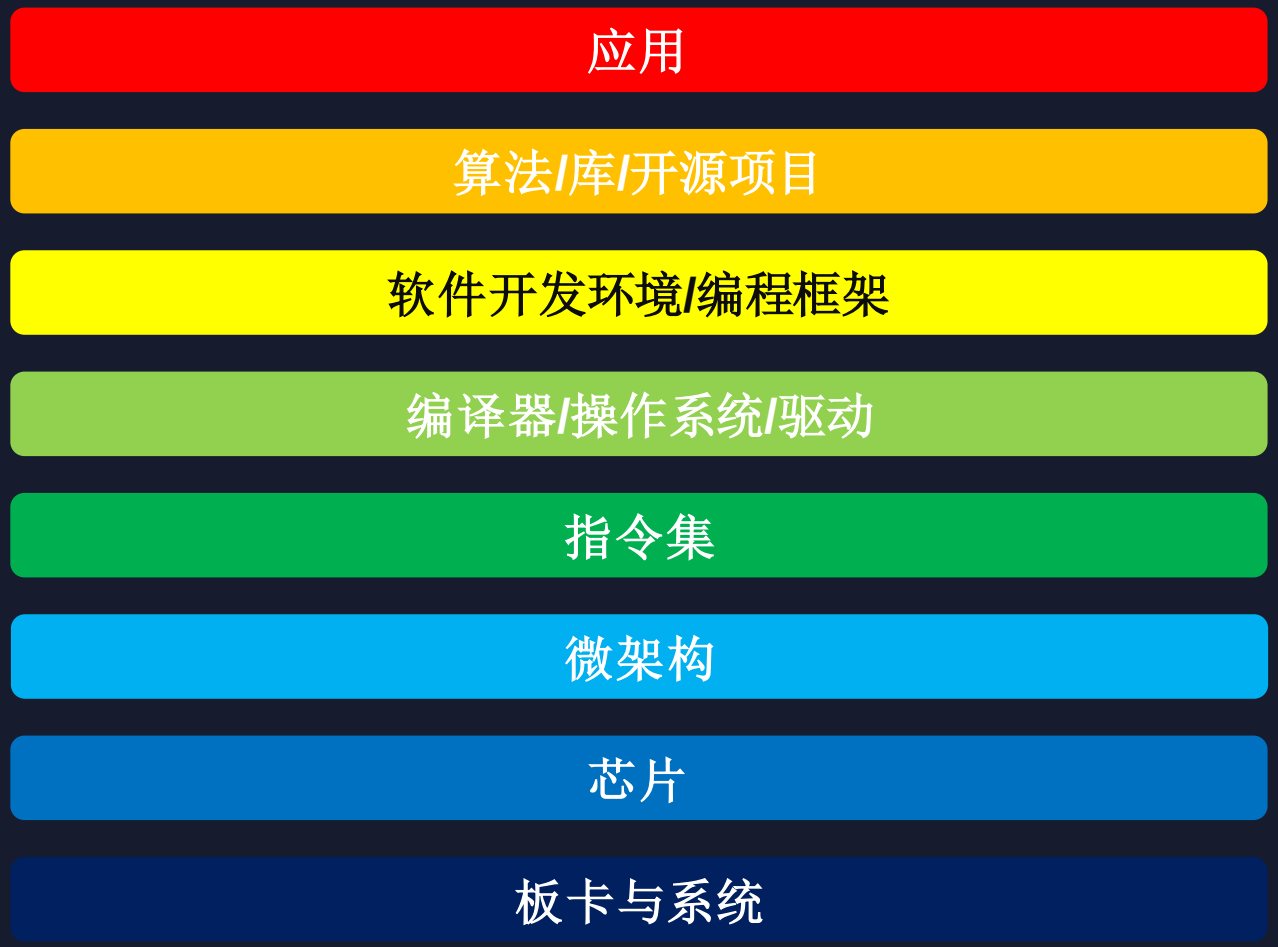
DriveWorks in Automotive Industry



英伟达已经拥有180万AI开发者，99.99%的初学者都会从英伟达GPU开始

XILINX.

英伟达让AI开发者只要专注在应用和算法



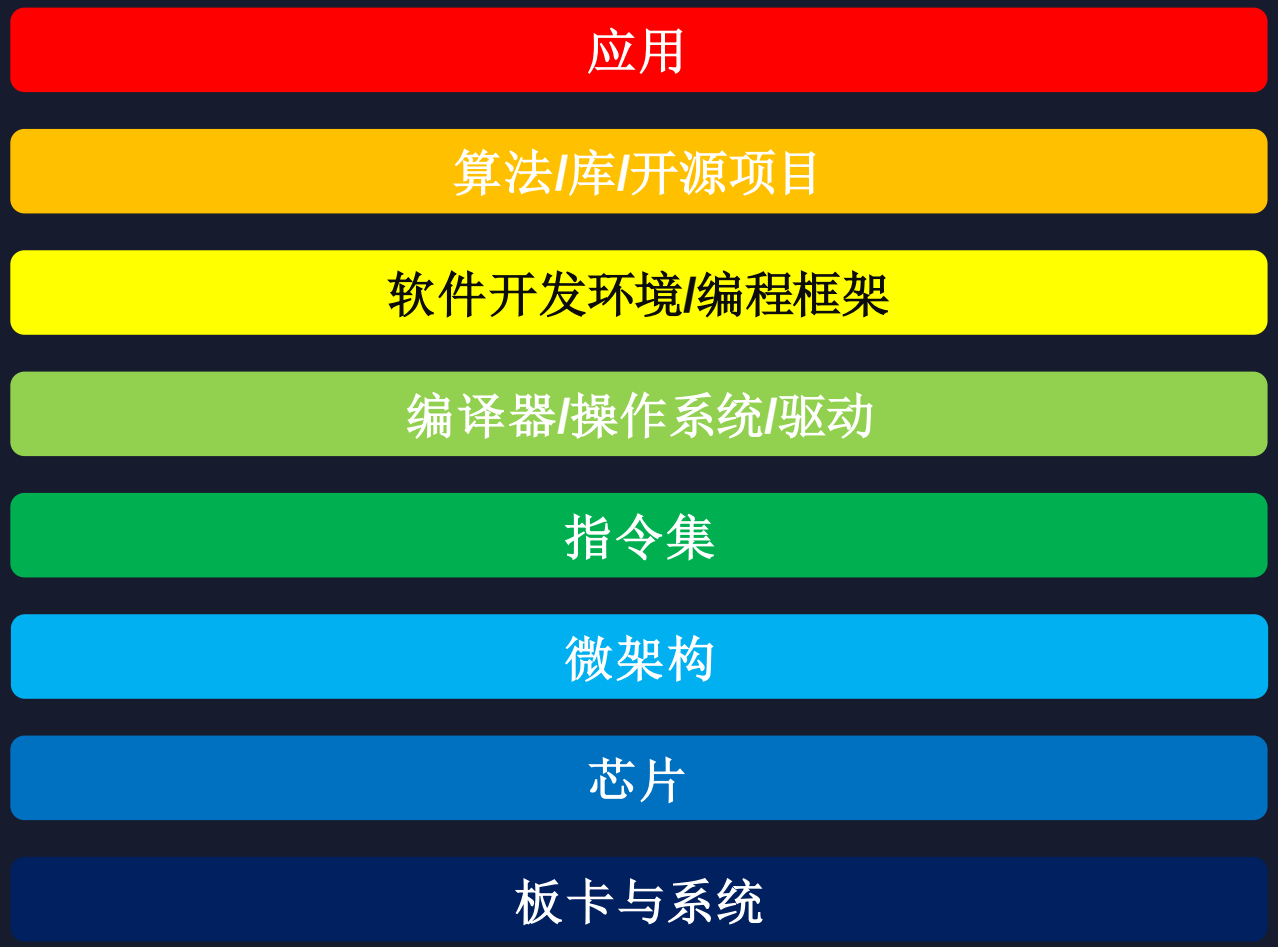
开发者专注部分

英伟达与开源社区提供

英伟达提供

XILINX.

AI芯片没有软件和生态是没有人用的



开发者专注部分

没有生态，谁来提供？

没有软件，怎么使用？

AI芯片的芯片部分

产品的硬件结构怎么设计？

从深鉴时期我们就非常注重软件与应用

算法与应用实例

编译器与软件环境

DEEPHi DNNDK

- AI Model Zoo
- AI Optimizer
- AI Quantizer
- AI Compiler
- Runtime

IP核

DPU IPs

订制板卡

Custom Board | Zedboard | Ultra96 | ZCU104 | ZCU102 | U2x0 | U50

支持各类FPGA: Z7020/ZU2/ZU3/ZU5/ZU7/ZU9/ZU11/U50/U200/U250/U280 ...

Xilinx正在努力构筑软件与应用生态

应用案例

+

解决方案

+

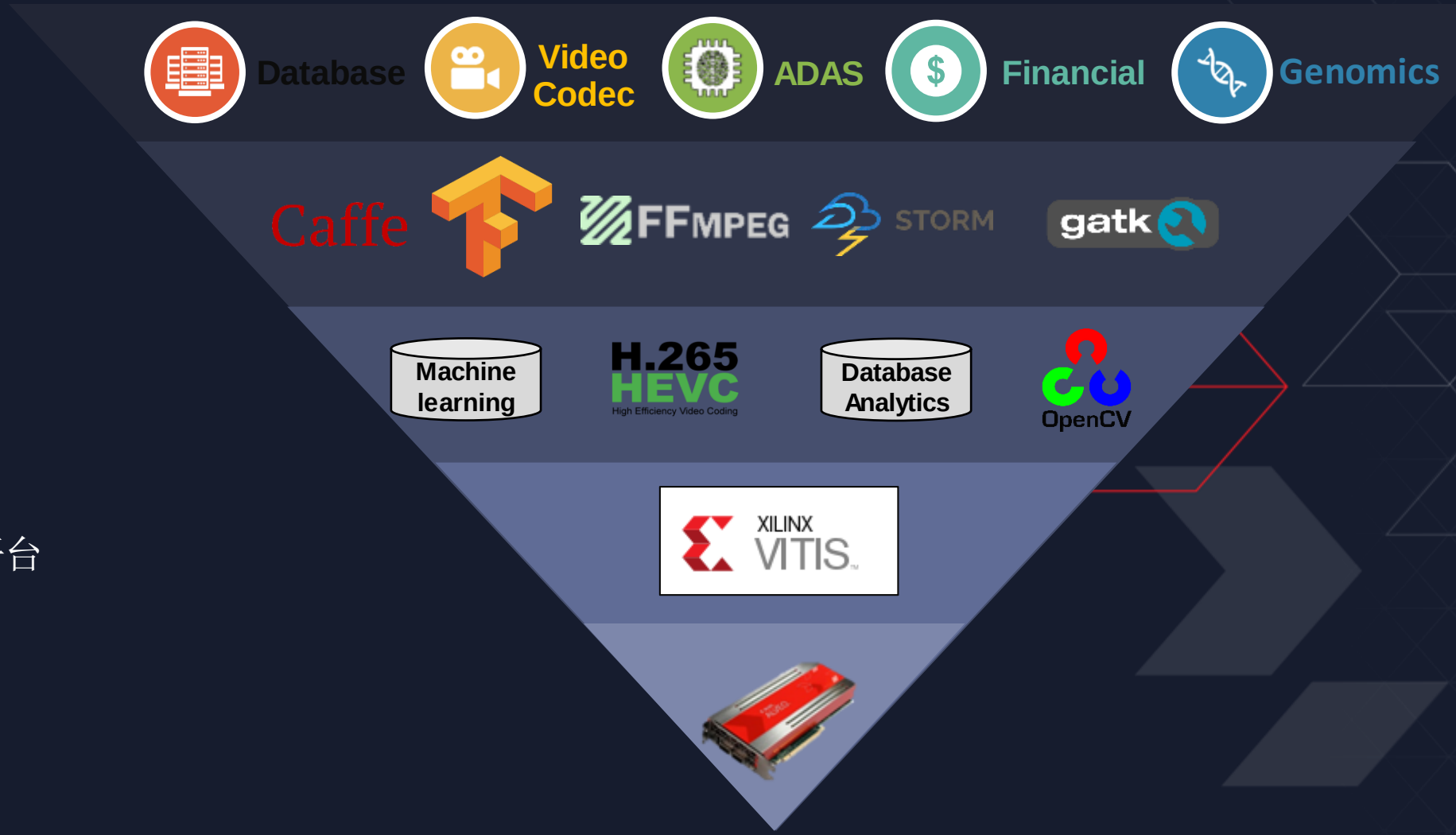
IP核

+

统一软件开发平台

+

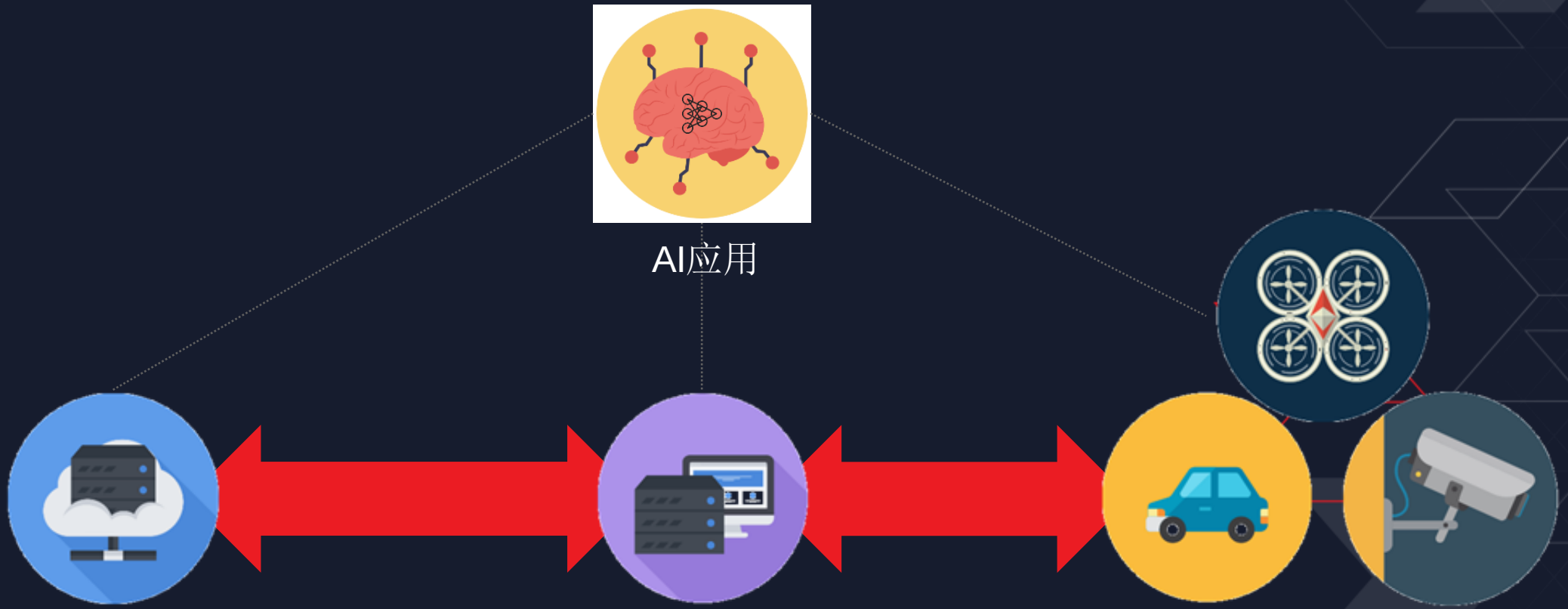
芯片



XILINX

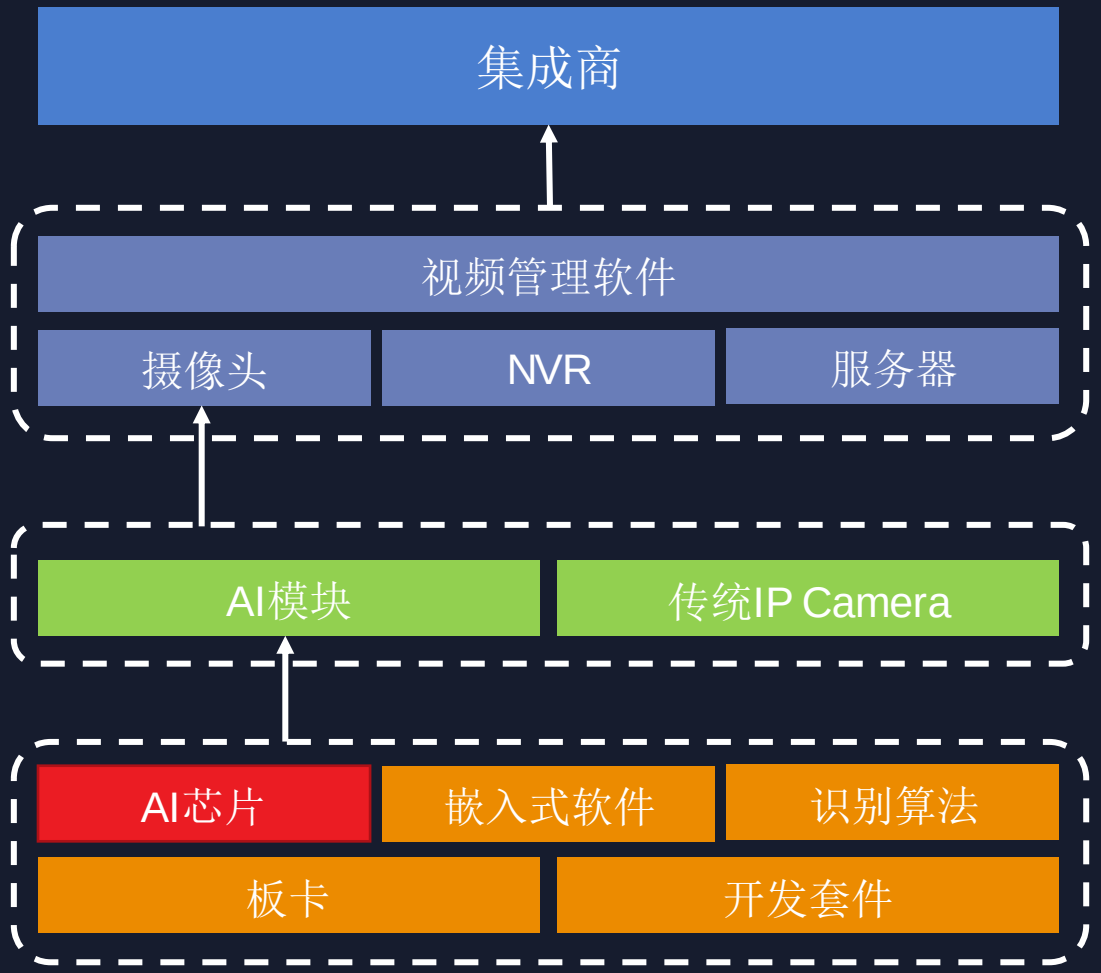
AI芯片产业如何发展？

根本问题：AI应用的范围非常之广

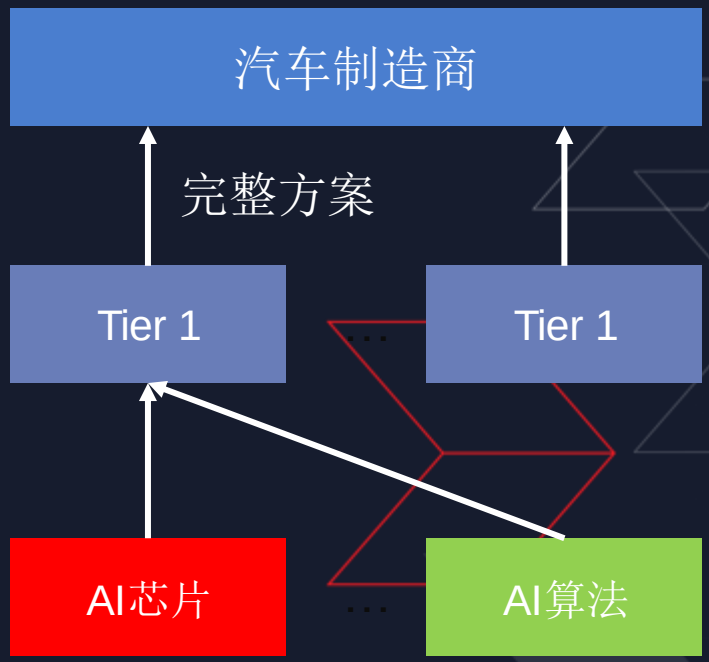


没有任何一款AI芯片可以覆盖多个应用领域

AI芯片只是是系统中的一个组成成分

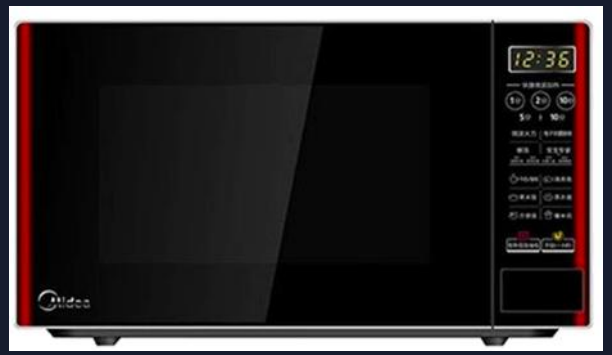


智能安防行业简单拆解



辅助驾驶行业简单拆解

同是AI芯片，需求千差万别，必须专注行业



智能微波炉的语音控制
极低成本

无线耳机的语音唤醒与识别
极低功耗，较低成本

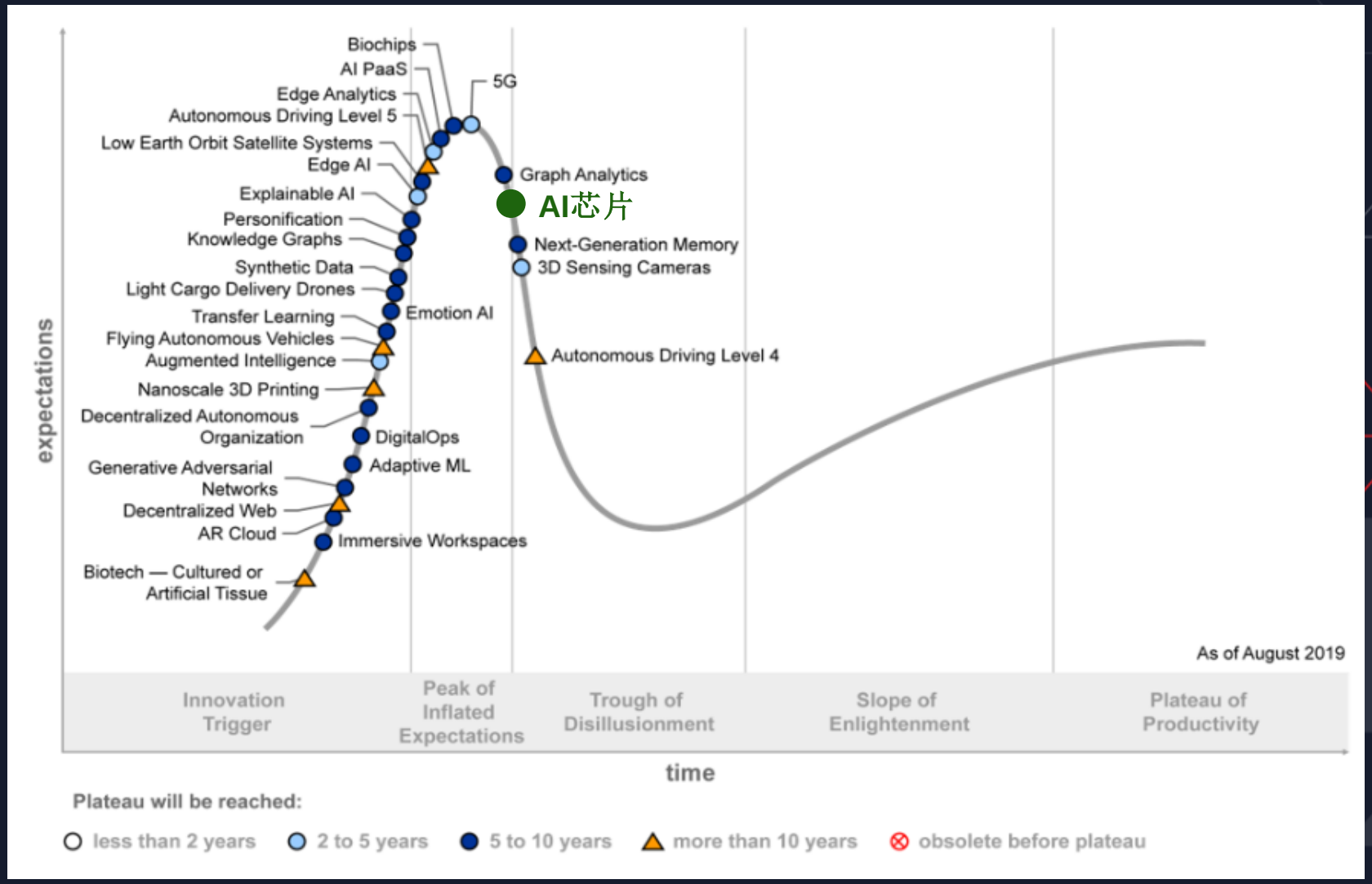
自动/辅助驾驶
低延迟，高可靠

无线智能摄像机
极低功耗，极低成本

市场格局：按垂直行业划分

1. 云端基本没有创业企业机会（投入极高，竞争极其激烈，需求没有想象的高）
2. 终端必须打透垂直行业

AI芯片洗去浮华，价值会逐渐浮现



XILINX.

Adaptable.
Intelligent.

XILINX