XDF XILINX DEVELOPER FORUM

FPGA在讯飞的研究与应用

FPGA Research and Application at Iflytek

Presented By

科大讯飞 iFLYTEK

江宏武 (Hongwu Jiang)
高级架构师 (Senior Architecture)
2018.10.16

XILINX

分享提纲

- 科大讯飞介绍
- FPGA在讯飞的研究成果与经验
- 下一步安排

简介

- 科大讯飞 iFLYTEK 成立于1999年12月；
- 是国内以语音及人工智能技术为产业化方向的龙头骨干企业；
- 是中国声谷整体建设布局的核心区；
- 近几年销售收入持续保持增长态势，每年增长30%以上；
- 连续10年增长超过10%的仅有的10家上市公司之一；
- 2016、2017连续两年荣获CCTV中国十佳上市公司；

中国声谷核心区

技术创新能力

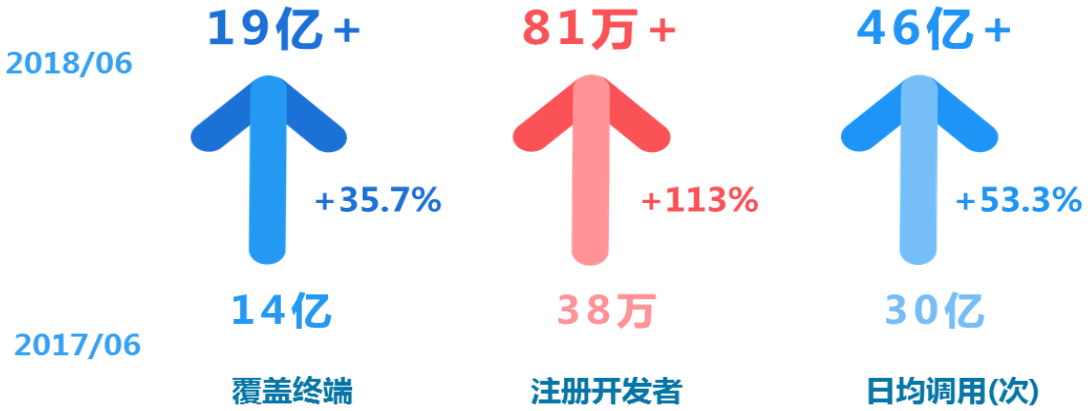
语音识别、声纹识别、语音合成、机器翻译、常识推理、阅读理解等技术全面处于国际领先水平

- 语音合成技术(Blizzard Challenge 2006-2018 十三连冠）
- 语音识别技术（2016 CHiME Challenge国际比赛冠军）
- 机器翻译技术（IWSLT 2014，NIST 2015均荣获第一）
- 常识推理技术（Winograd Schema Challenge 2016第一名）
- 知识发现技术（NIST TAC Knowledge Base Population -KBP2016 第一名）
- 机器阅读理解技术（Stanford Question Answering Dataset 2017 第一名）
- 机器阅读理解技术（SemEval 2018 第一名）
- 声音模仿技术（ Voice Conversion Challenge 2018 第一名）

MIT TR：2017全球50大最具技术创造力企业

生态拓展能力

讯飞开放平台高速增长的用户与使用量



智能硬件 117.9%
2018 1263.3W
2017 579.6W

智能家居 171.1%
2018 979.1W
2017 361.1W

运动健康 101.6%
2018 25.8W
2017 12.8W

金融理财 109.5%
2018 211.8W
2017 101.1W

机器人 236.6%
2018 927.8W
2017 275.6W

智能穿戴 289.8%
2018 2.1W
2017 5.4K

游戏 280.7%
2018 2357.3W
2017 619.2W

教育学习 27.1%
2018 2539.0W
2017 1998.3W

产业化应用

AI+医疗

AI+教育

AI+政务

AI+安全

AI+金融

AI+家居

AI+...

A.I.+产业 典型行业应用

行业大数据　核心技术　行业专家

分享提纲

- 科大讯飞介绍
- FPGA在讯飞的研究成果与经验
- 下一步安排

FPGA在讯飞的历史

☞ 由关注到投入再到应用落地

- 2014年，开始关注FPGA在AI领域的应用。
- 2015年，发布了基于Open CL的AI加速示例。
- 2016年，组建FPGA团队，同年完成DNN/RNN语音识别加速方案。
- 2017年，在云和端同时开展FPGA应用研究，研发第一代加速卡。
- 2018年，第一代FPGA加速卡正式量产发货；

  语音识别实例性能进一步提升，实时率突破1000，优于P4近25%；

  研发集成HBM2存储的第二代加速卡。

猎鹰IM1000加速卡

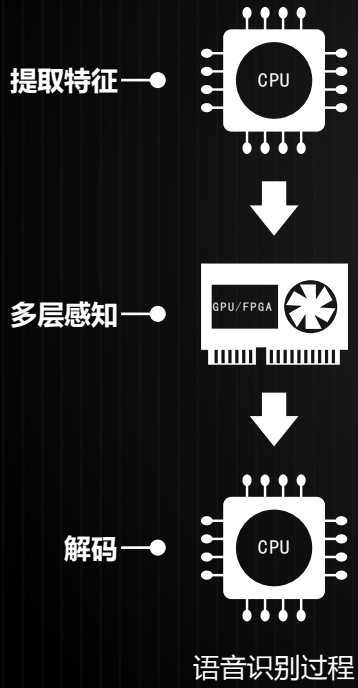
☞ 已批量投产

猎鹰 IM1000

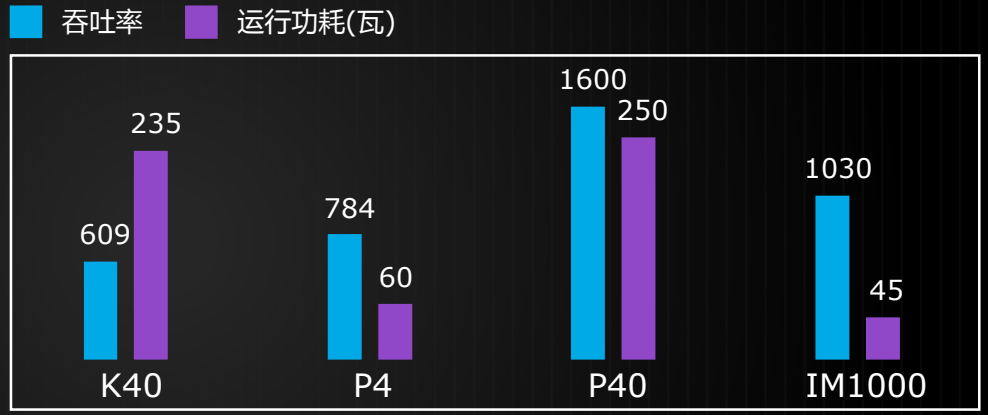
尺寸规格	Half width and half height
主芯片	KU115
主机接口	PCI Express 3.0 ×8
存储器	2 Channel DDR4 with ECC @ 2400MT/s，4GB/Channel
状态指示	4 LEDs
典型功耗	60 w
监测系统	Voltage, Temperature monitor
供电接口	PCIE Slot

IM1000在语音识别任务中的效果

☞ 实现最优性价比



语音识别过程

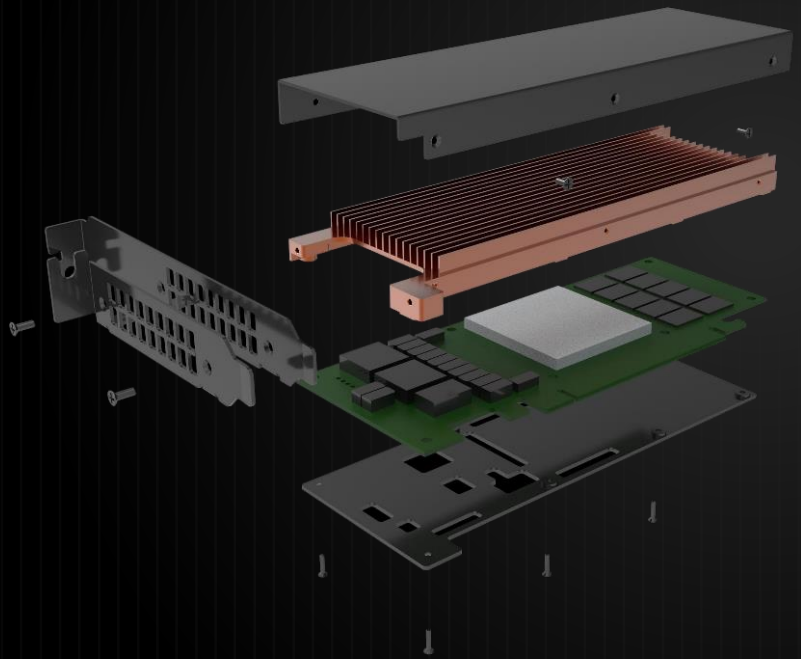


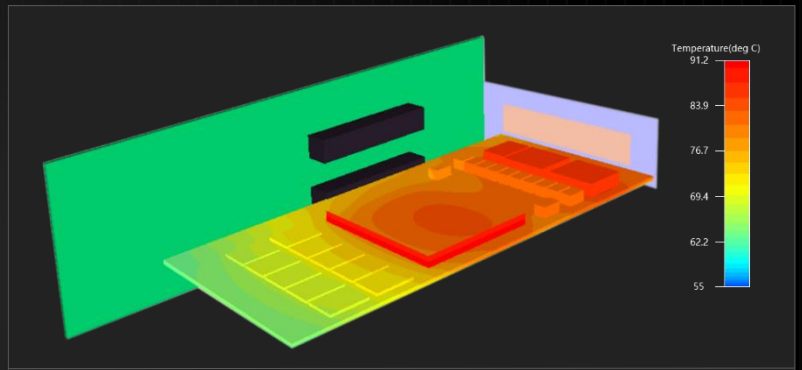
语音识别率

语种	GPU 32bit	FPGA 16bit
维语	80.14%	80.12%
汉语	81.77%	81.95%

IM1000结构

☞ 实际运行不过70℃

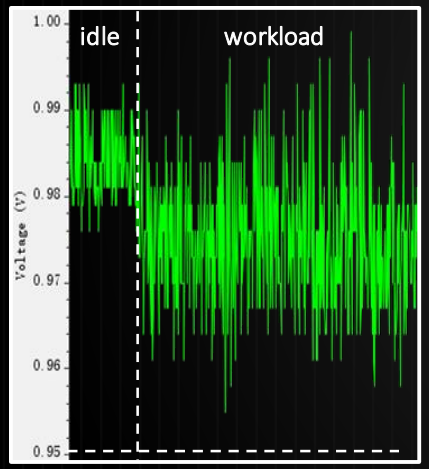




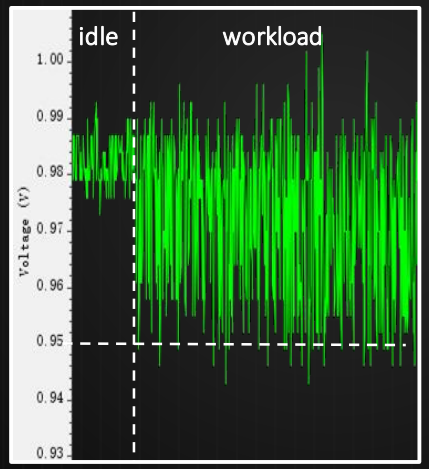
实际运行温度

IM1000电源

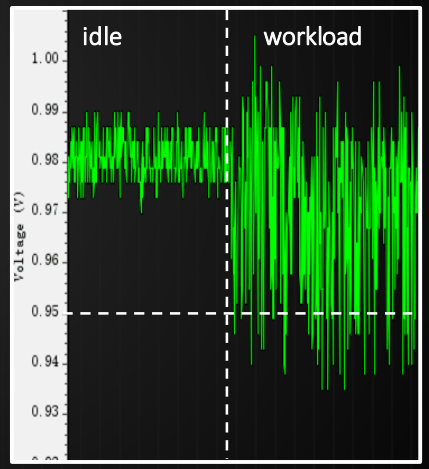
☞ 建议使用数字可编程电源，模拟电源建议预加重输出



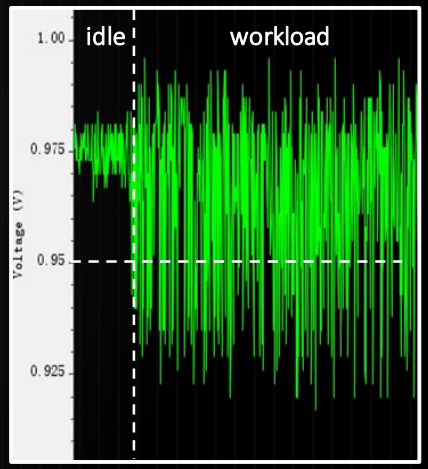
200 Lane @ 400MHz



300 Lane @ 400MHz



400 Lane @ 400MHz



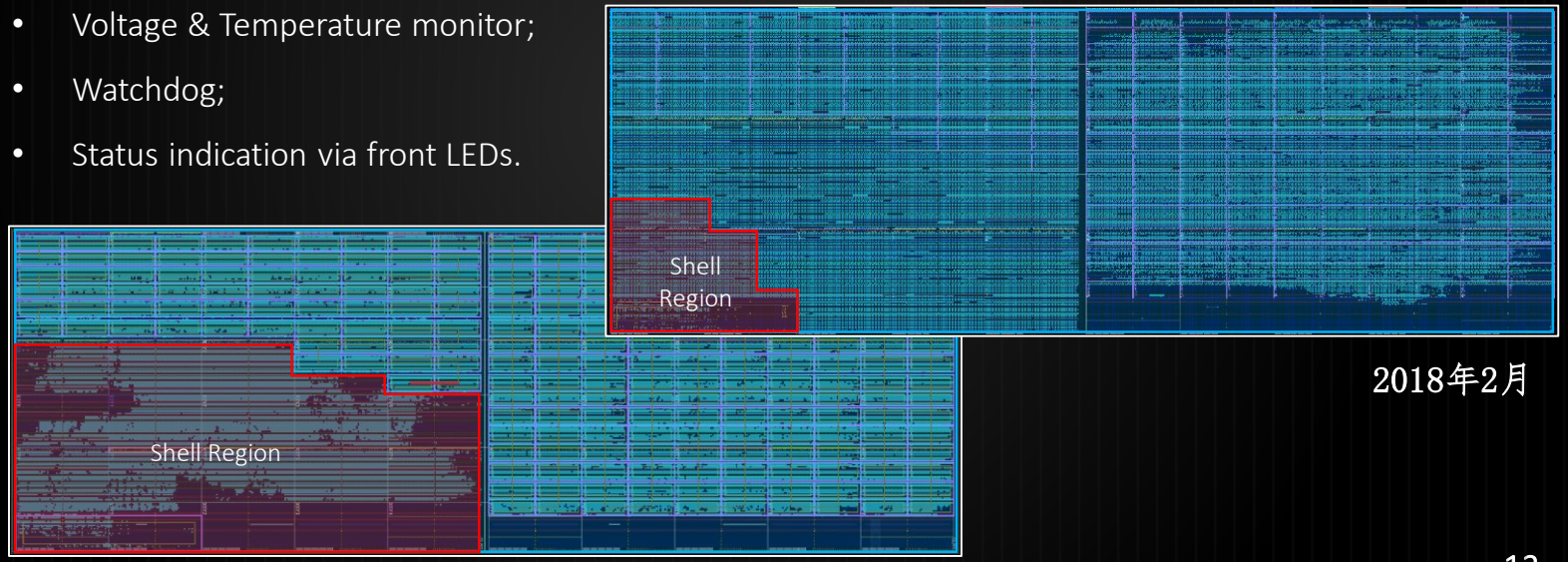
500 Lane @ 400MHz

精简Shell

☞ Shell Region面积高度优化

Shell Region include：

- PCIE interface with DMA;
- Voltage & Temperature monitor;
- Watchdog;
- Status indication via front LEDs.



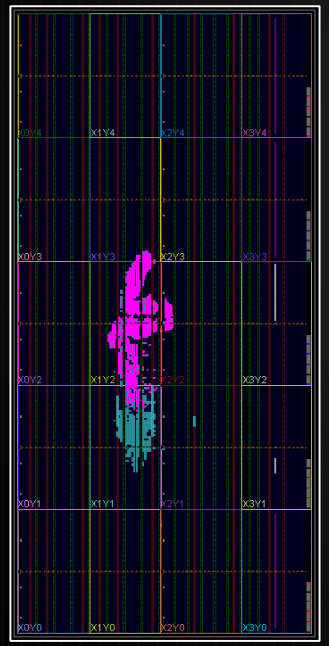
2018年2月

2017年1月

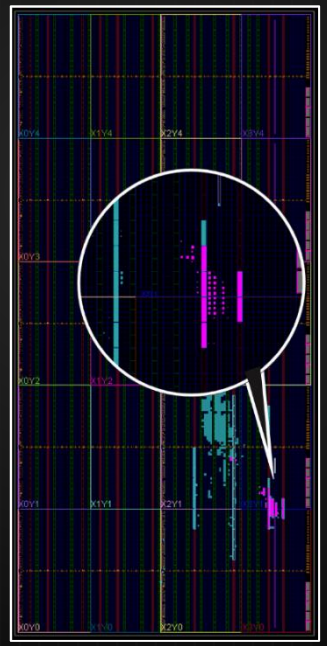
特色IP

☞ 以超低芯片面积实现高精度激活函数电路，支持Int16和fp32

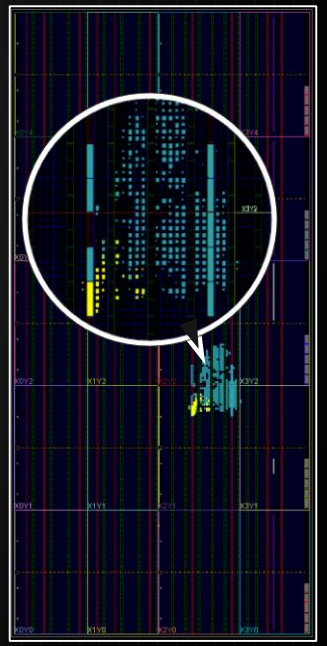
注：从左至右图中，分别以红色，红色，黄色表示实现同等精度的激活函数需要占用的面积（越小越好）



2016年中



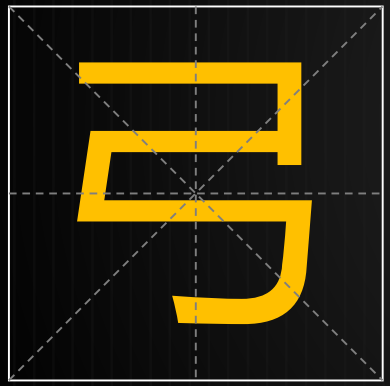
2017年底

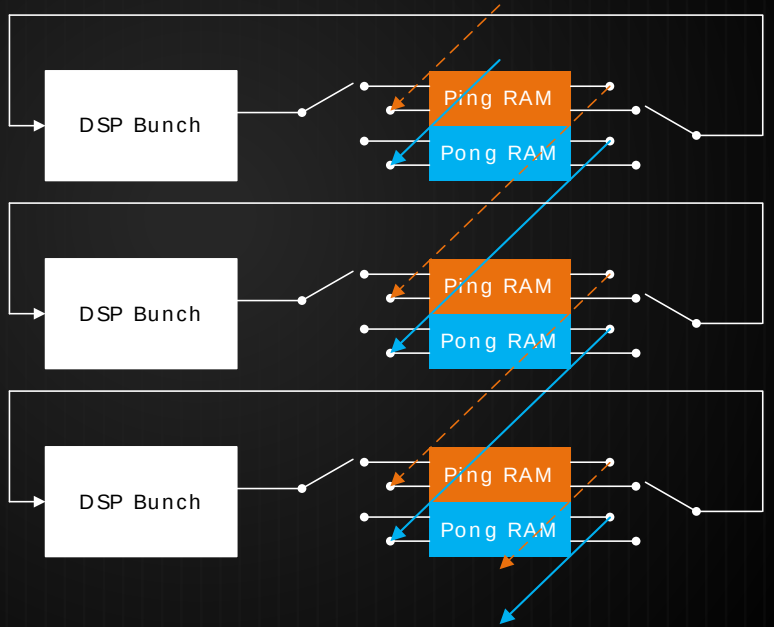


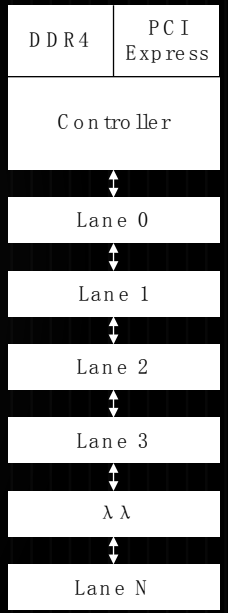
2018年初

高效GEMM（1/2）

☞ 简洁的链状结构

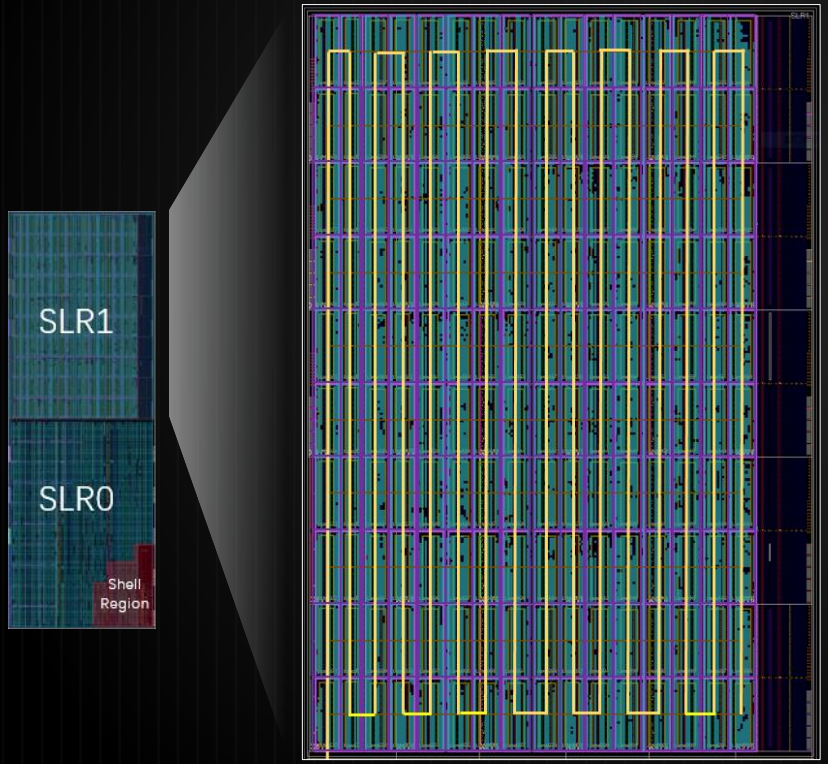






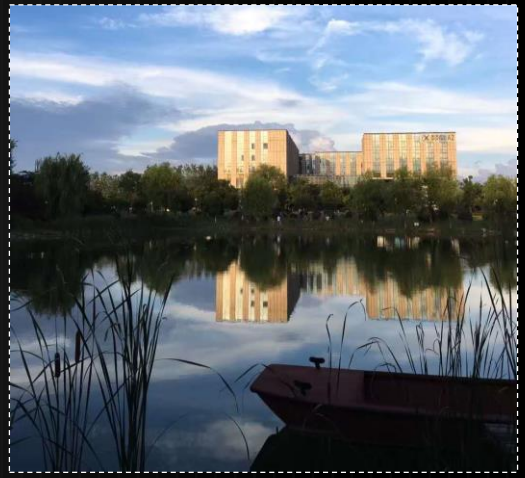
高效GEMM (2/2)

☞ 实现友好的布局布线



Utilization	
LUT	40%
LUTRAM	10%
FF	61%
BRAM	70%
DSP	90%
IO	35%
GT	13%
BUFG	1%
MMCM	13%
PLL	13%
PCIe	17%

讯飞欢迎你



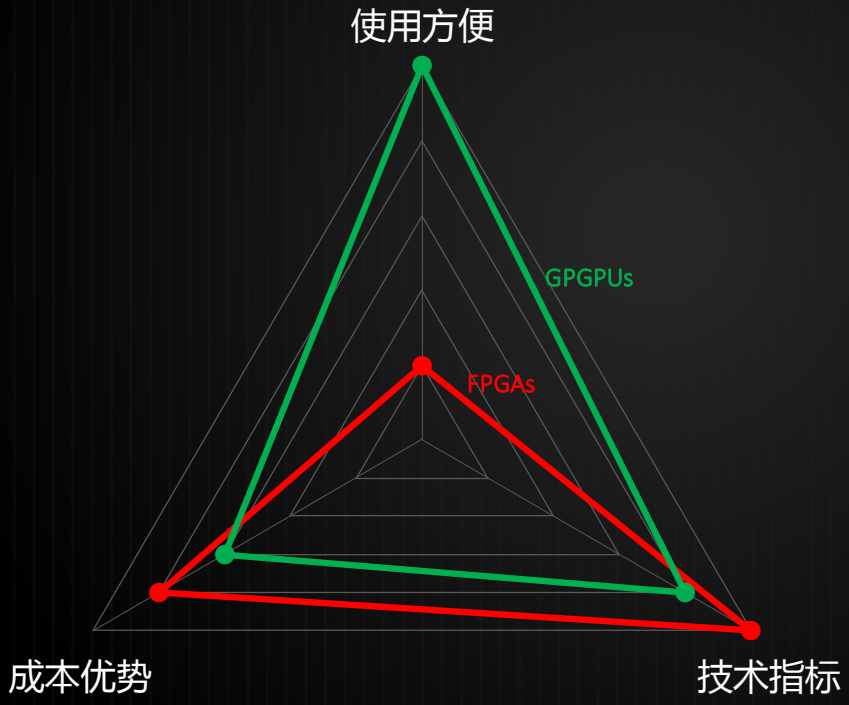
mailto：hr@iflytek.com，hwjiang@iflytek.com

分享提纲

- 科大讯飞介绍
- FPGA在讯飞的研究成果与经验
- 下一步安排

降低FPGA在业务端的使用门槛（1/2）

☞ 今天FPGA在AI领域应用仍然不是很方便



评估是否方便使用的几个维度：

1）获取硬件的难易度；

2）技术开发难易度；

3）从业者数量和经验；

降低FPGA在业务端的使用门槛（2/2）

☞ 构建讯飞FPGA应用技术框架

在中国，用人工智能改变世界！

Adaptable.
Intelligent.