

UISEE 驭势

2018 THINK IN CLOUD BEIJING

嵌入式设备上的实时深度学习方法实践

潘争

— 目录

01

视觉感知特点

02

效率精度平衡的卷积网络

03

卷积网络的压缩

04

嵌入式GPU+CPU的加速

05

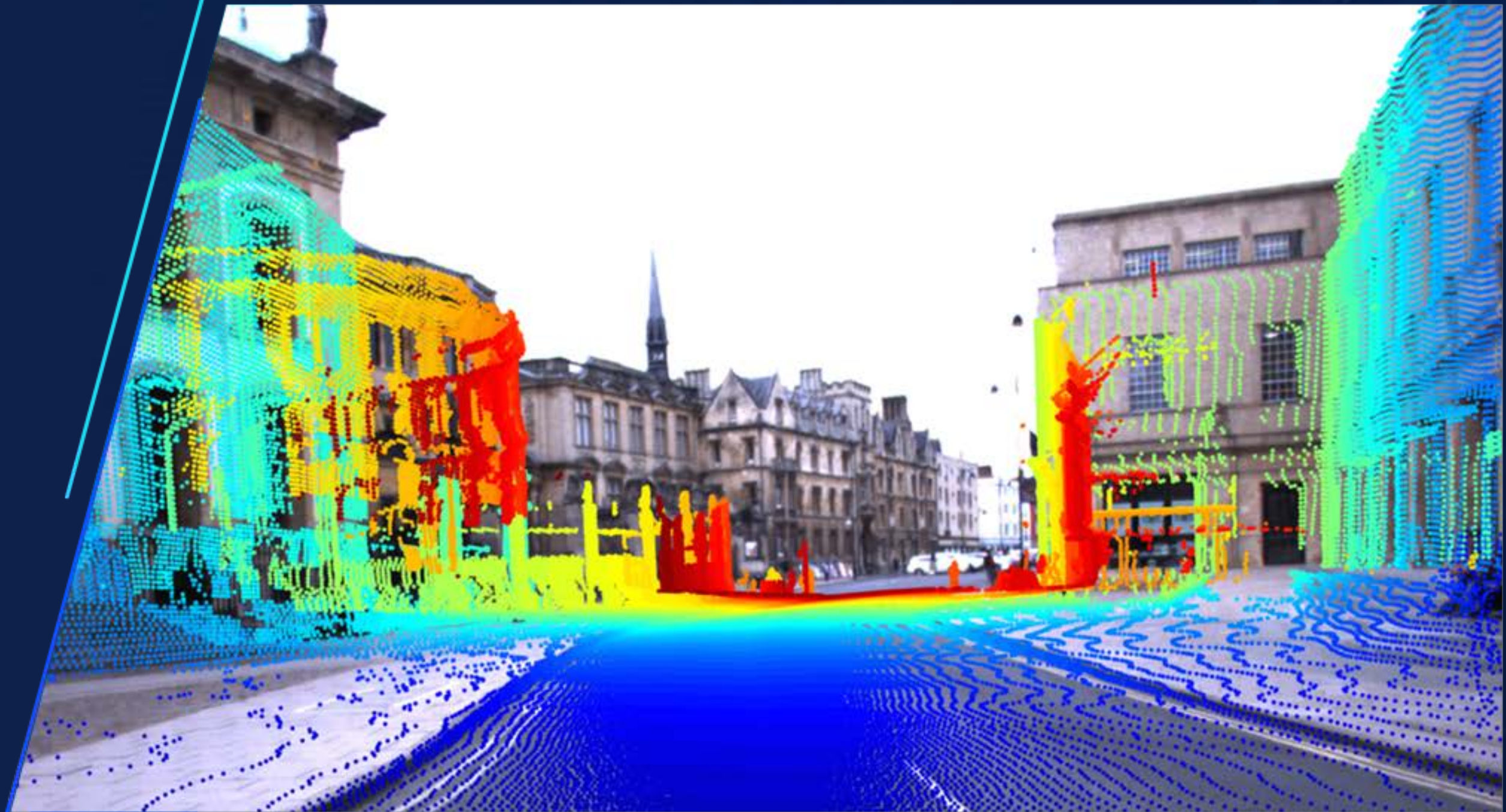
低成本车规级FPGA的加速

01

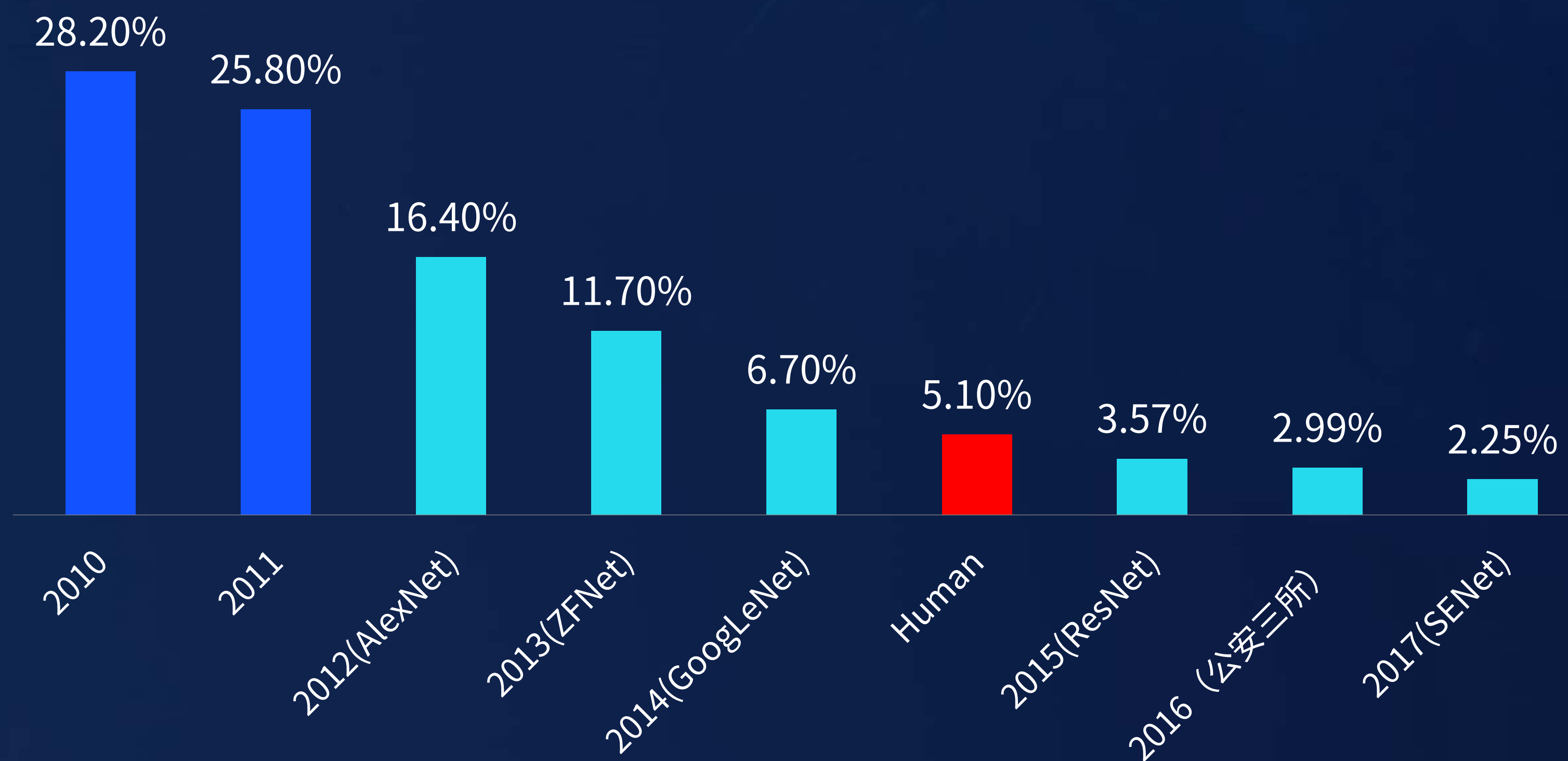
视觉感知的特点

— 视觉感知的特点

- 信息更丰富
- 视野更宽阔
- 基建更配合
- 硬件更便宜



— 视觉识别算法飞速进步ImageNet Top-5错误率



— 视觉感知从demo到deploy

■ Demo



精度要高



不计成本



不管标准



不算功耗

■ Deploy



低成本



低功耗



合车规



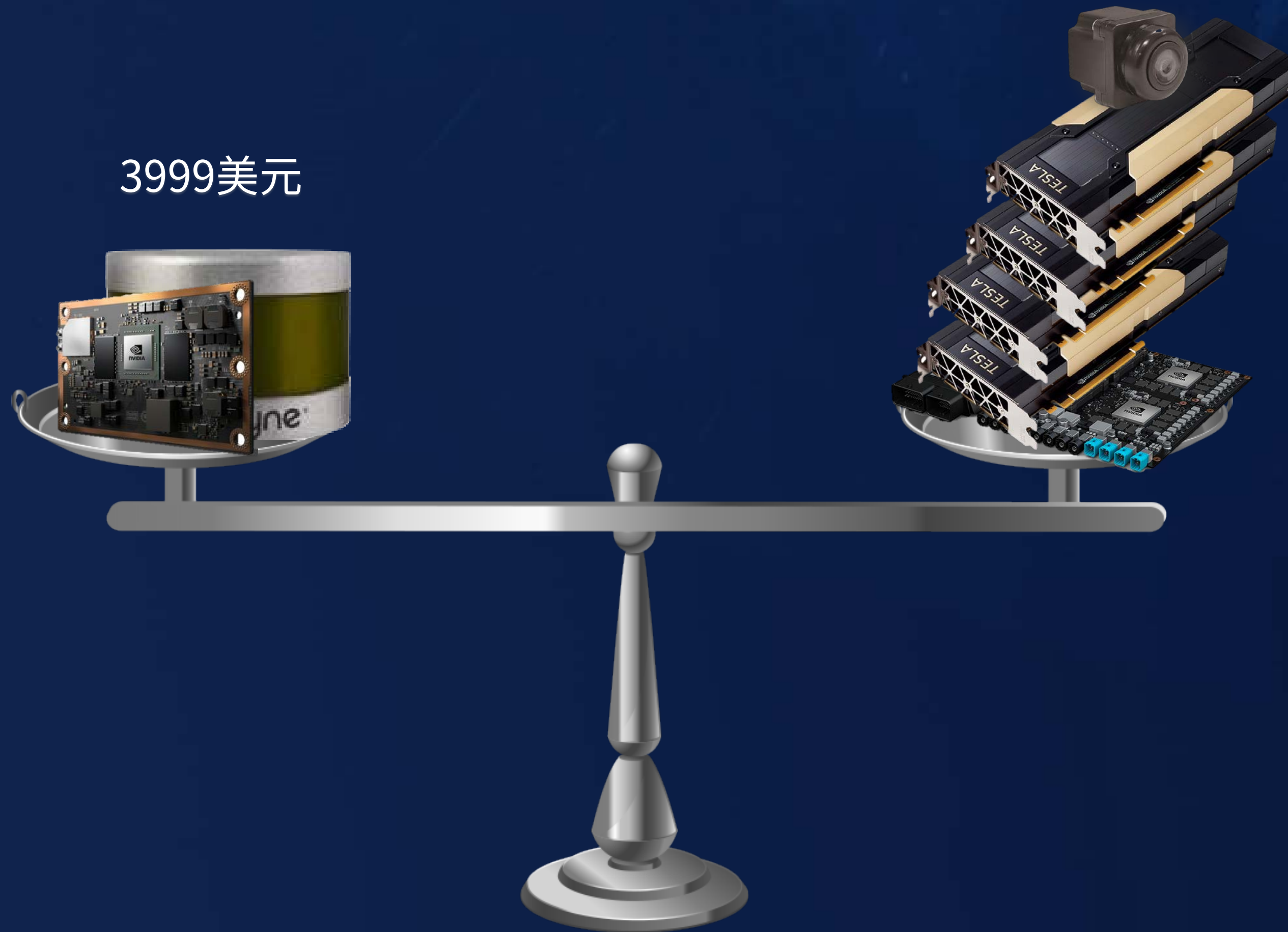
实时性



精度用户满意

— 低成本？

3999美元



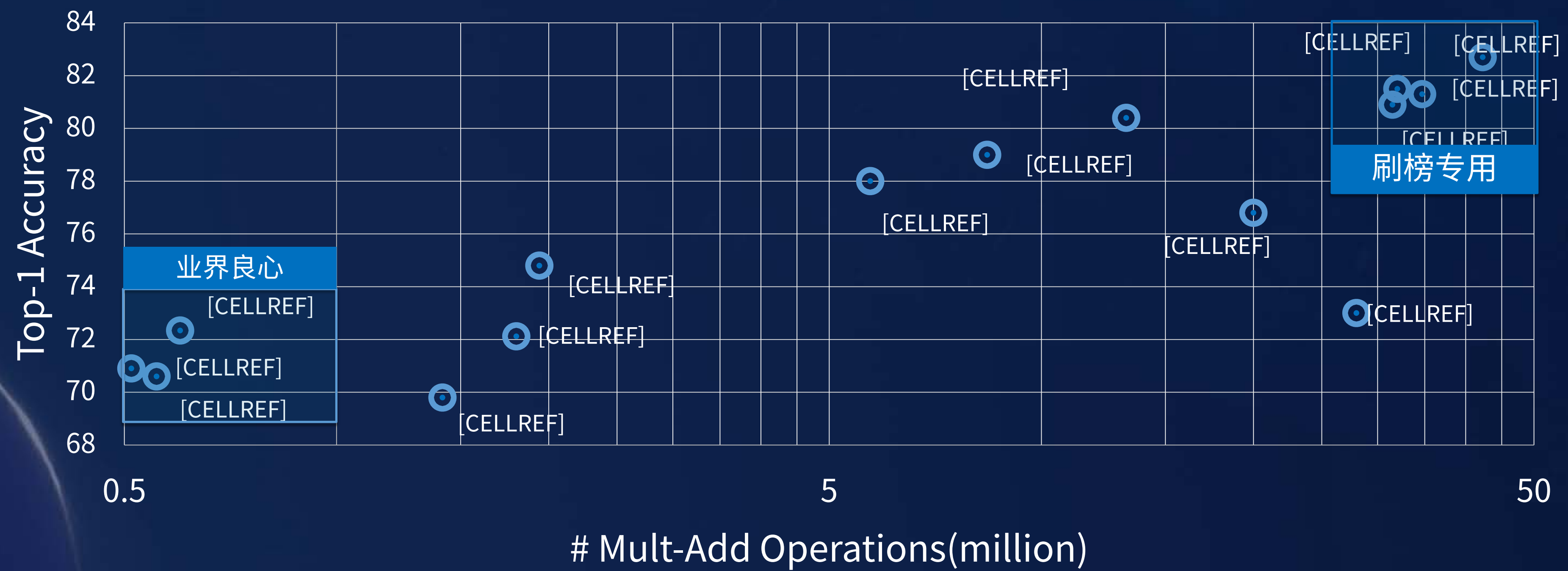
— 硬件更便宜



02

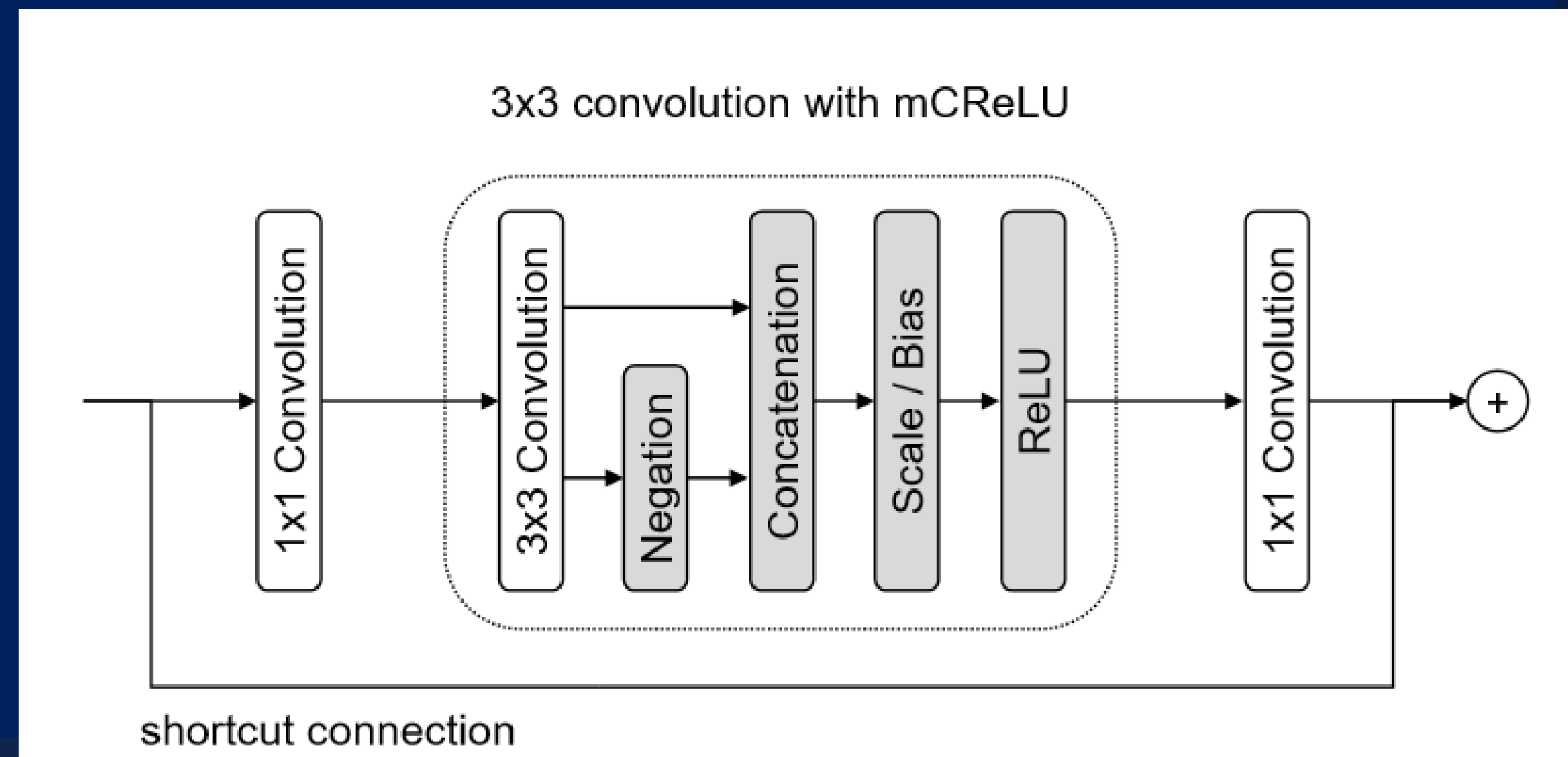
效率精度平衡的卷积网络

效率与精度的取舍



PVANet

Hong S, Roh B, Kim K H, et al. PVANet: Lightweight Deep Neural Networks for Real-time Object Detection. arXiv, 2016.

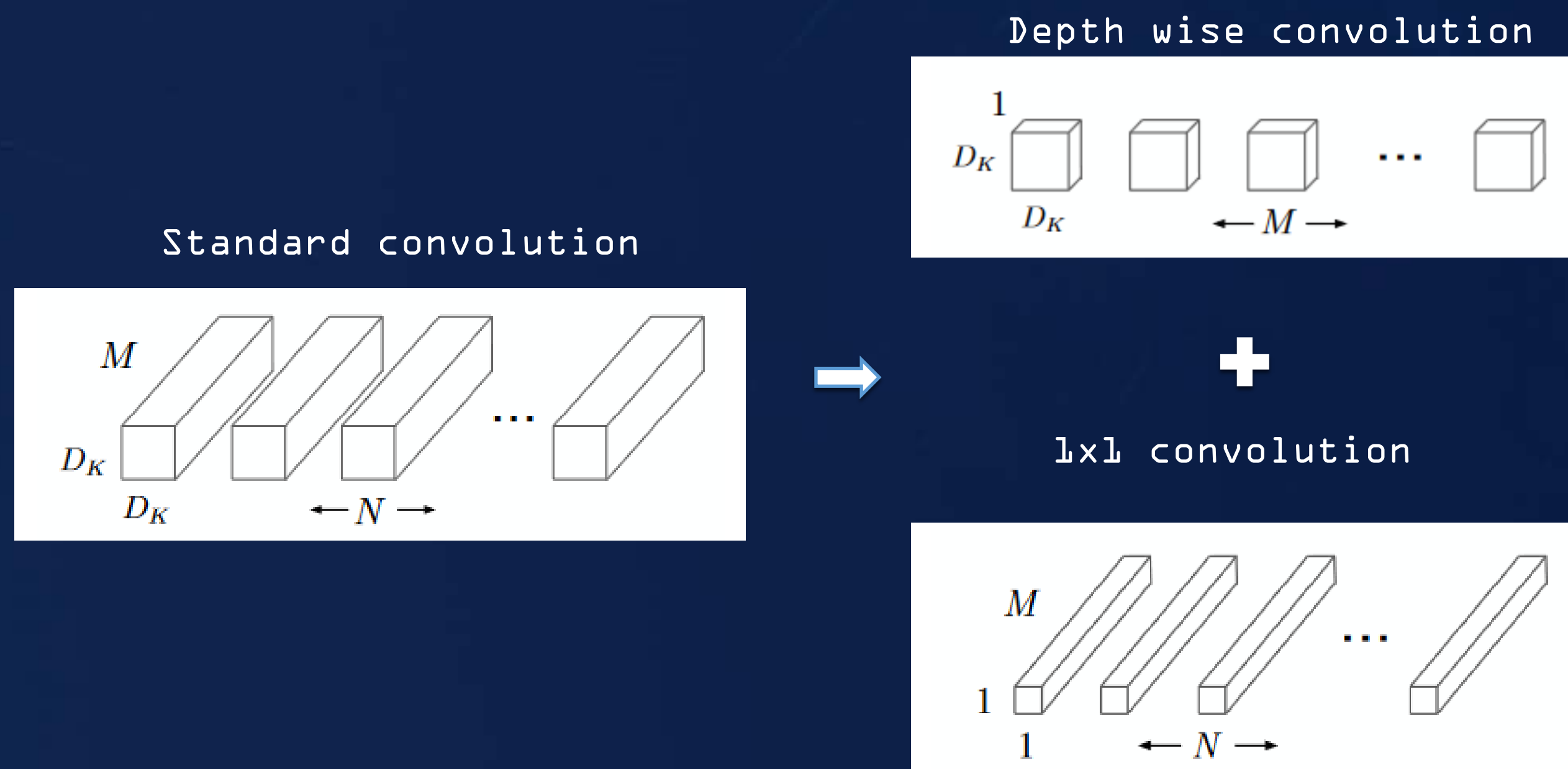


— PVANet+fasterrcnn 物体检测



— MobileNet

Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv, 2017.



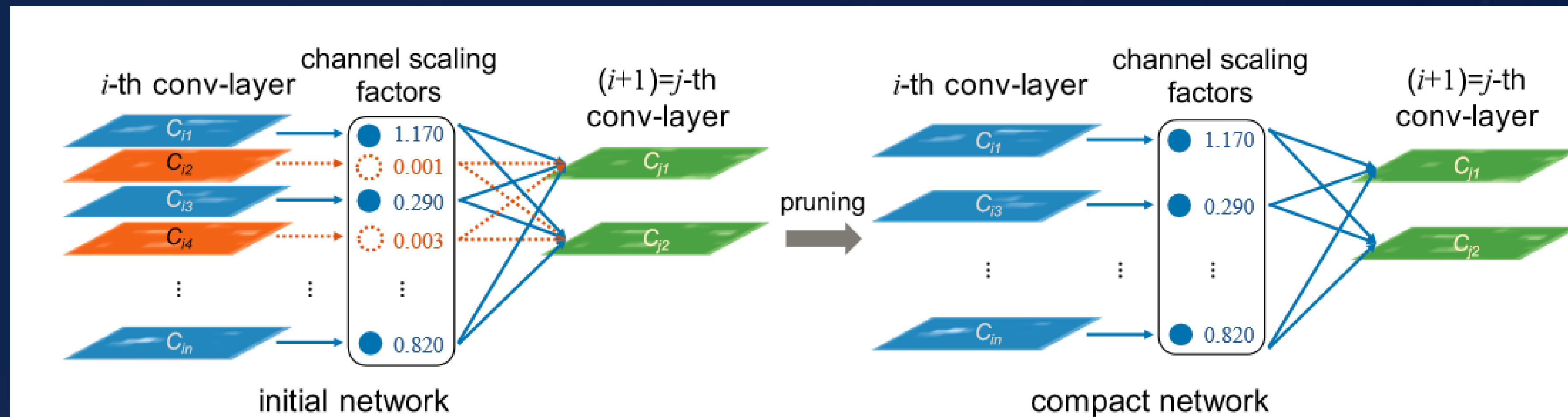
$$\frac{\text{dw conv} + 1 \times 1 \text{ conv}}{\text{std. conv}} = \frac{1}{D_k^2} + \frac{1}{N}$$

03

卷积网络的压缩

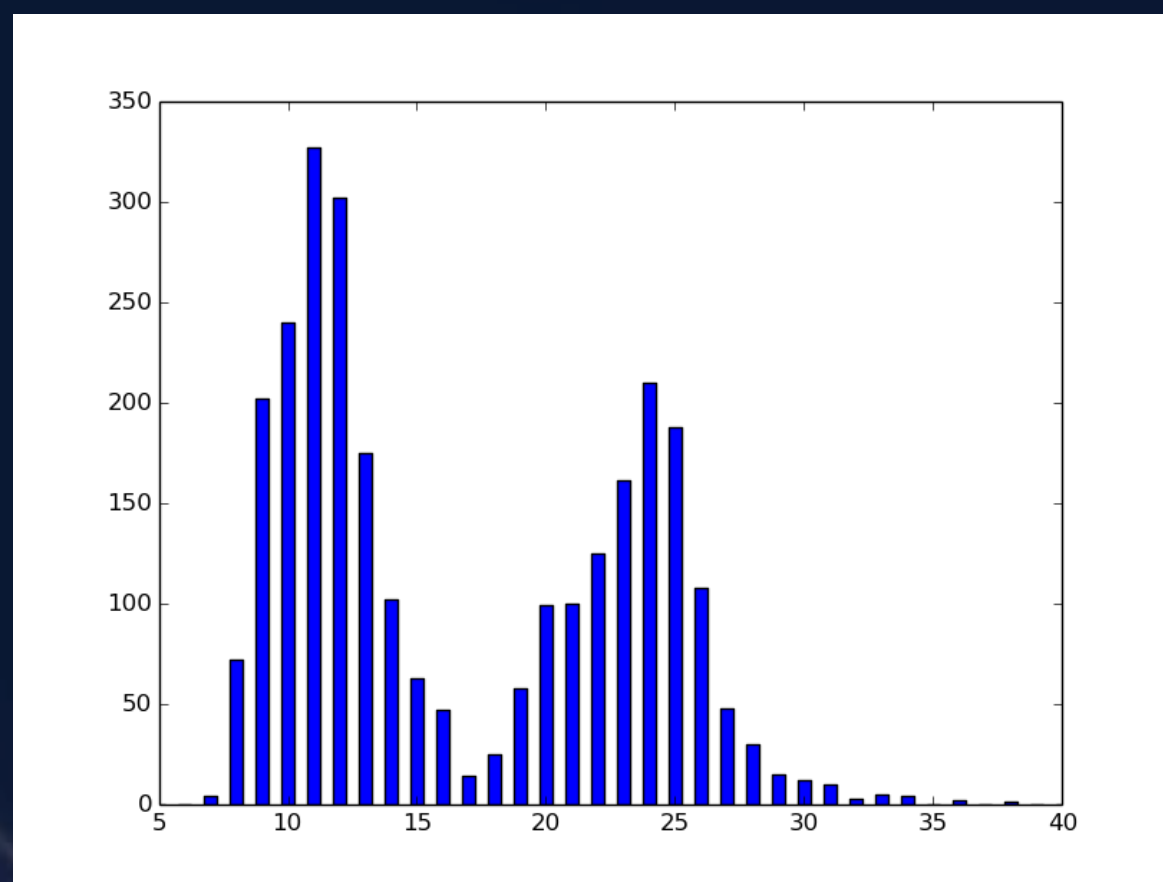
— Network slimming

Liu Z, Li J, Shen Z, et al. Learning Efficient Convolutional Networks through Network Slimming. arXiv, 2017.

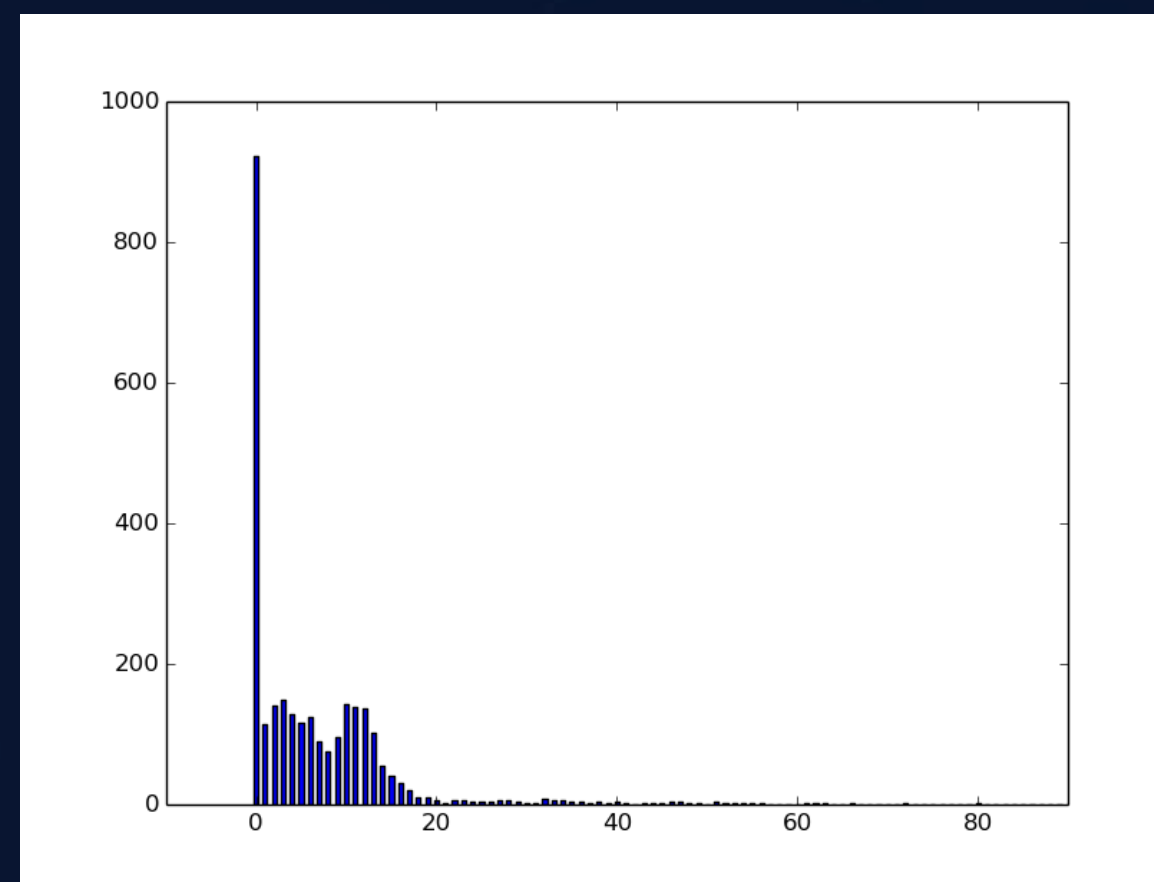


$$L(w) = \sum_i l(f(x_i, w), y_i) + \lambda \sum_{\gamma \in \Gamma} ||\gamma||_1$$

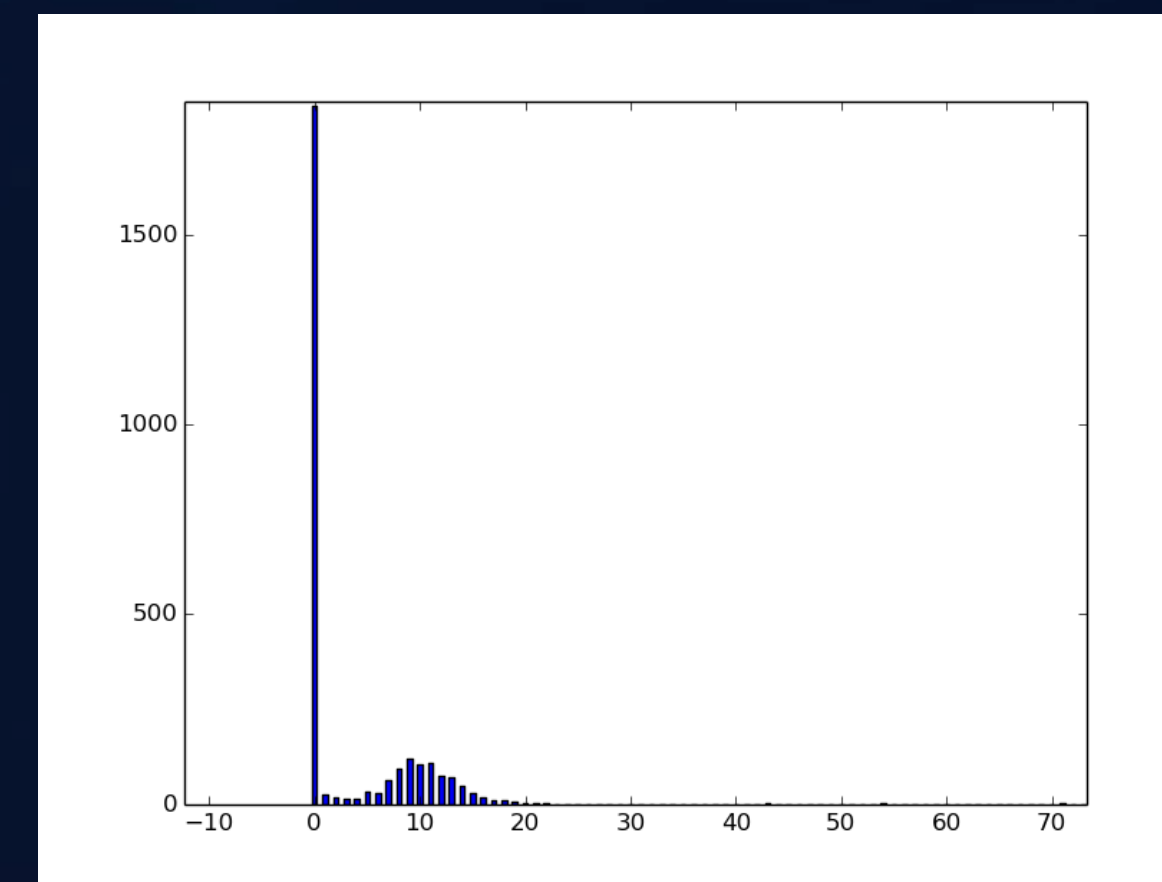
— 网络参数稀疏化效果



$\lambda = 0$

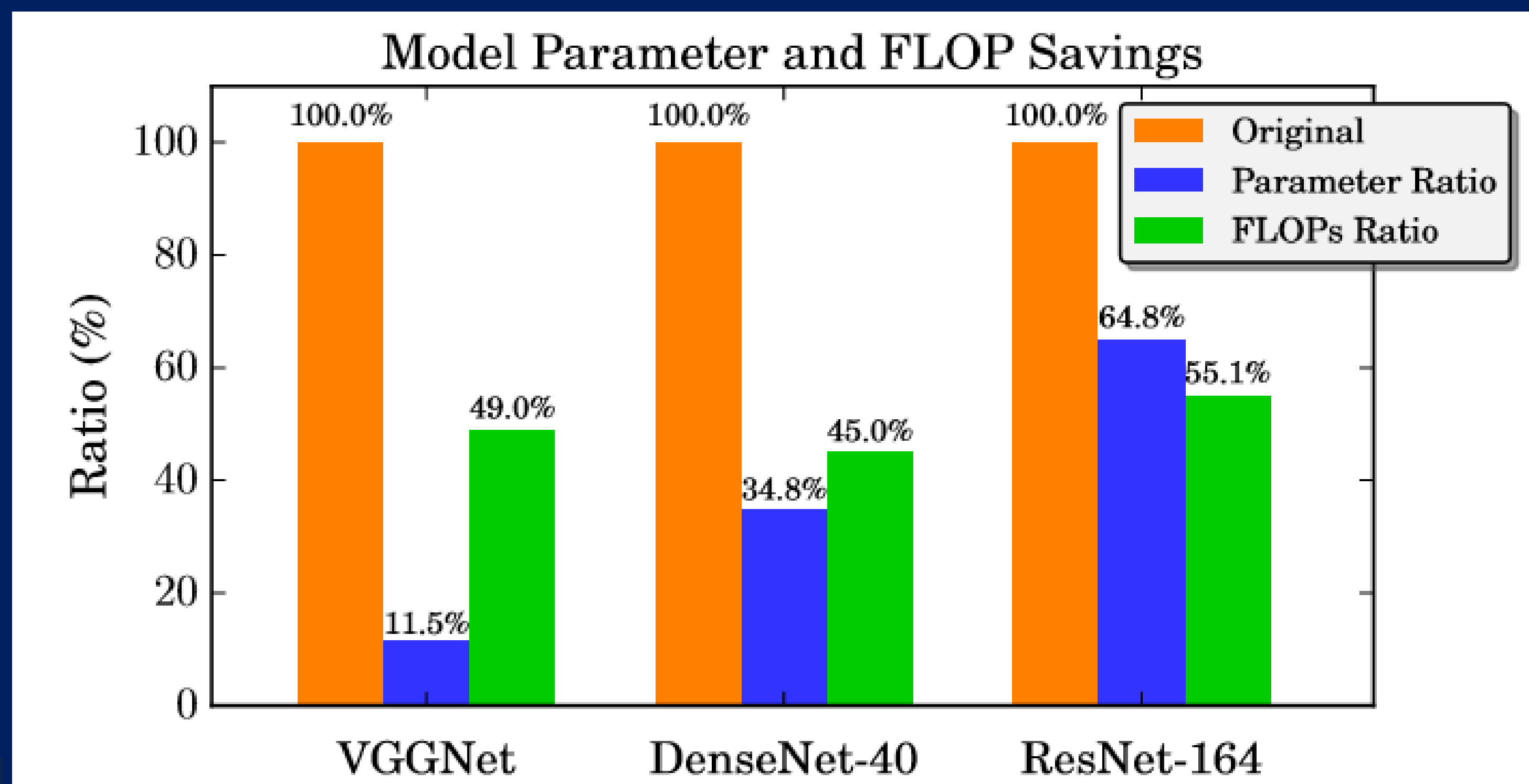


$\lambda = 0.0001$



$\lambda = 0.001$

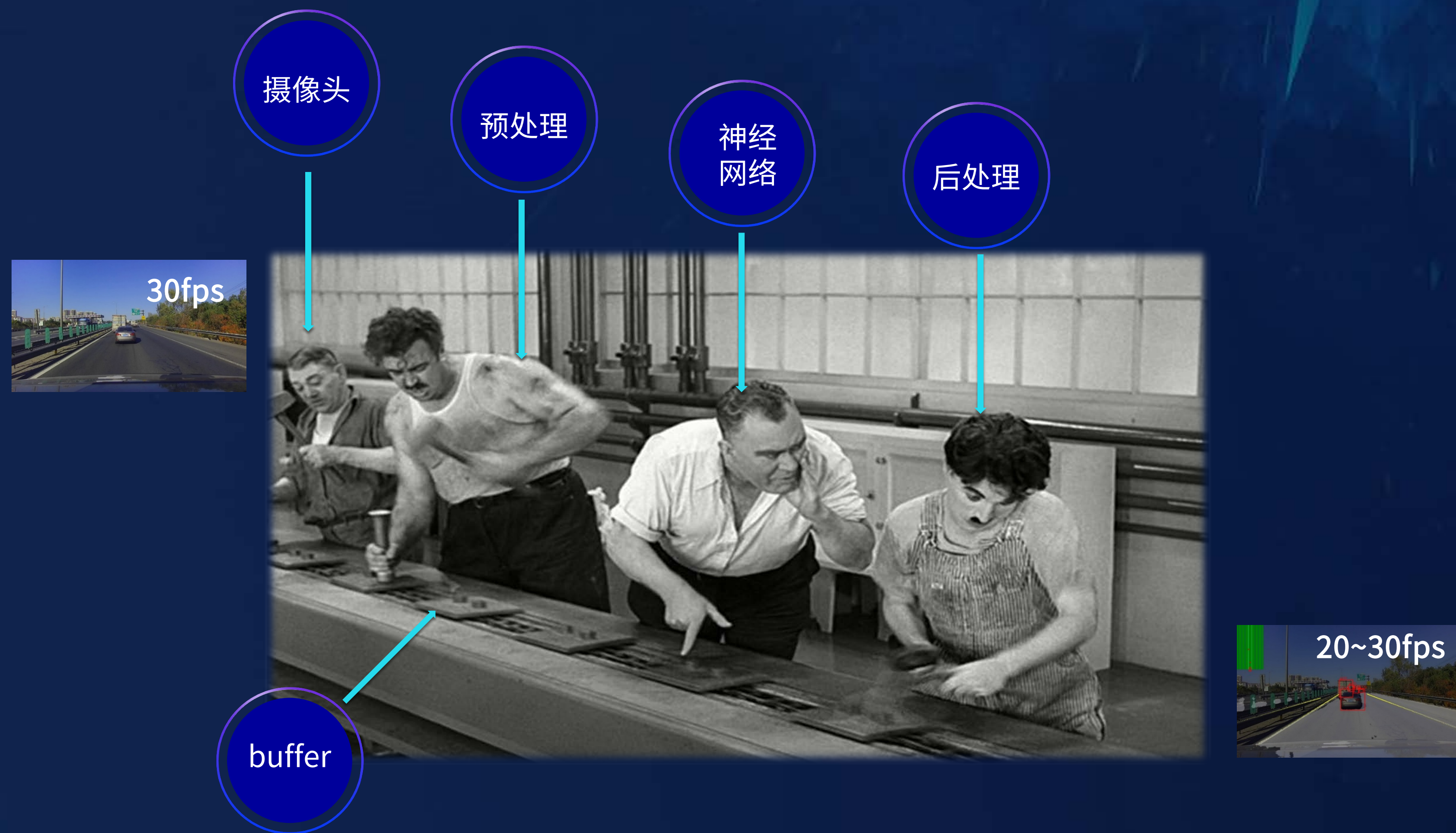
— 网络压缩结果



04

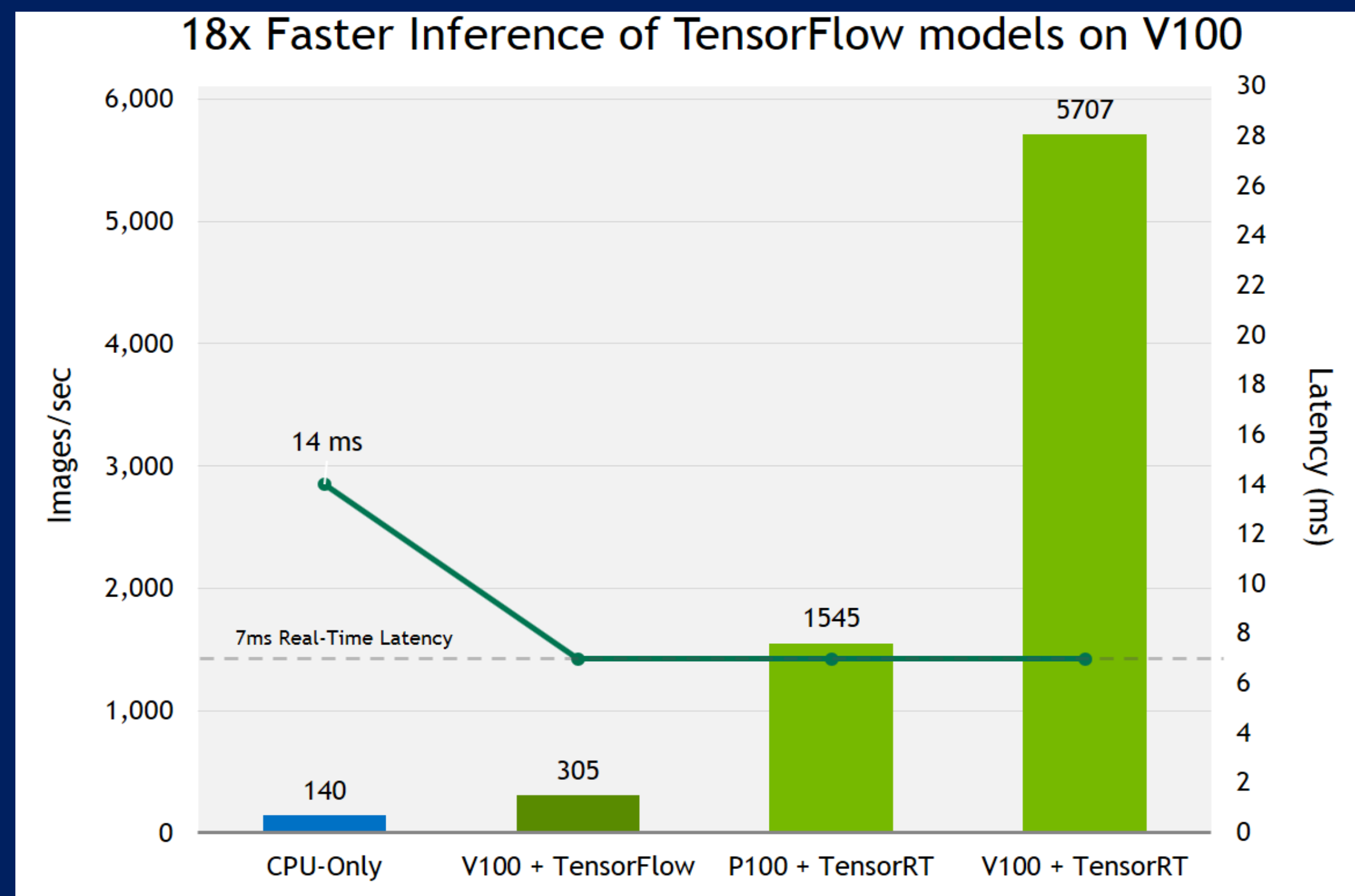
嵌入式GPU+CPU的加速

— Pipeline

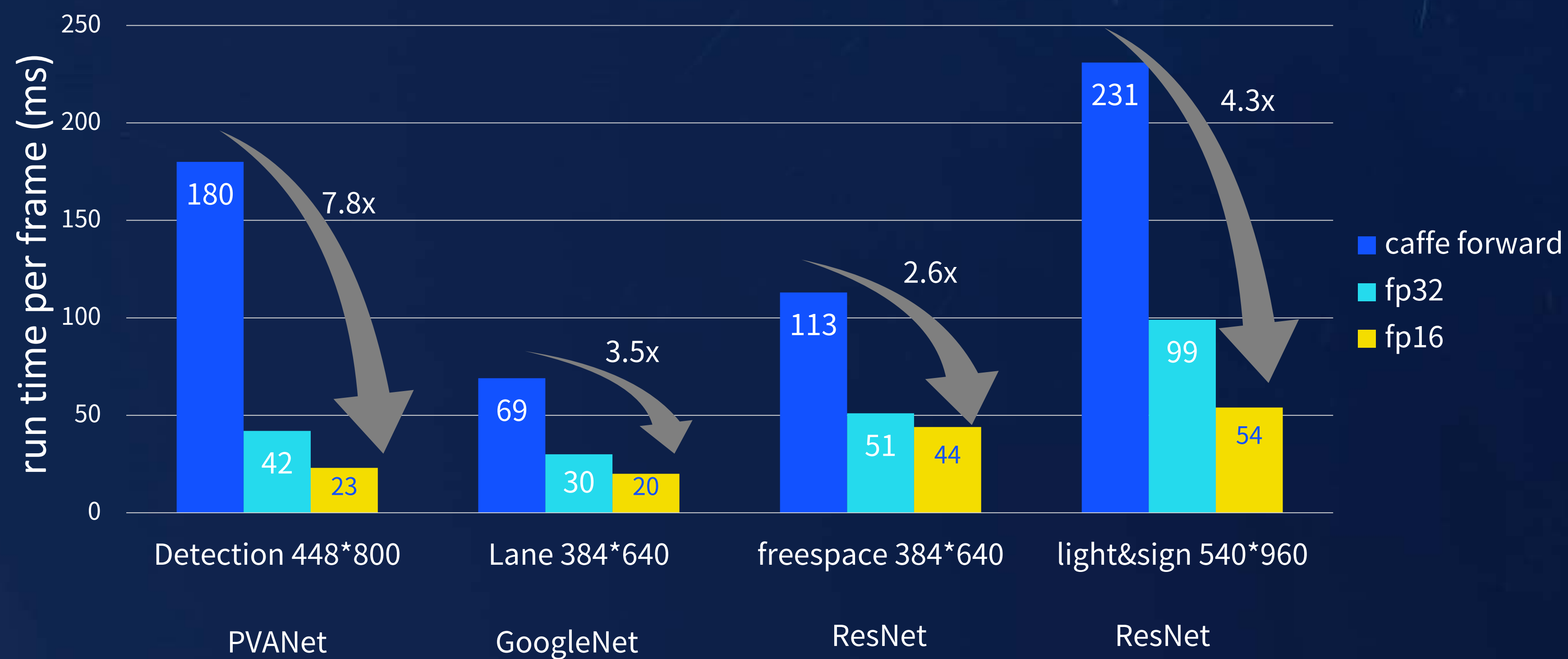


— TensorRT

- FP16和INT8自动量化
- 多层合并
- 自动选择并行算法
- 显存动态优化
- 多任务并发



— TX2上TensorRT 加速效果



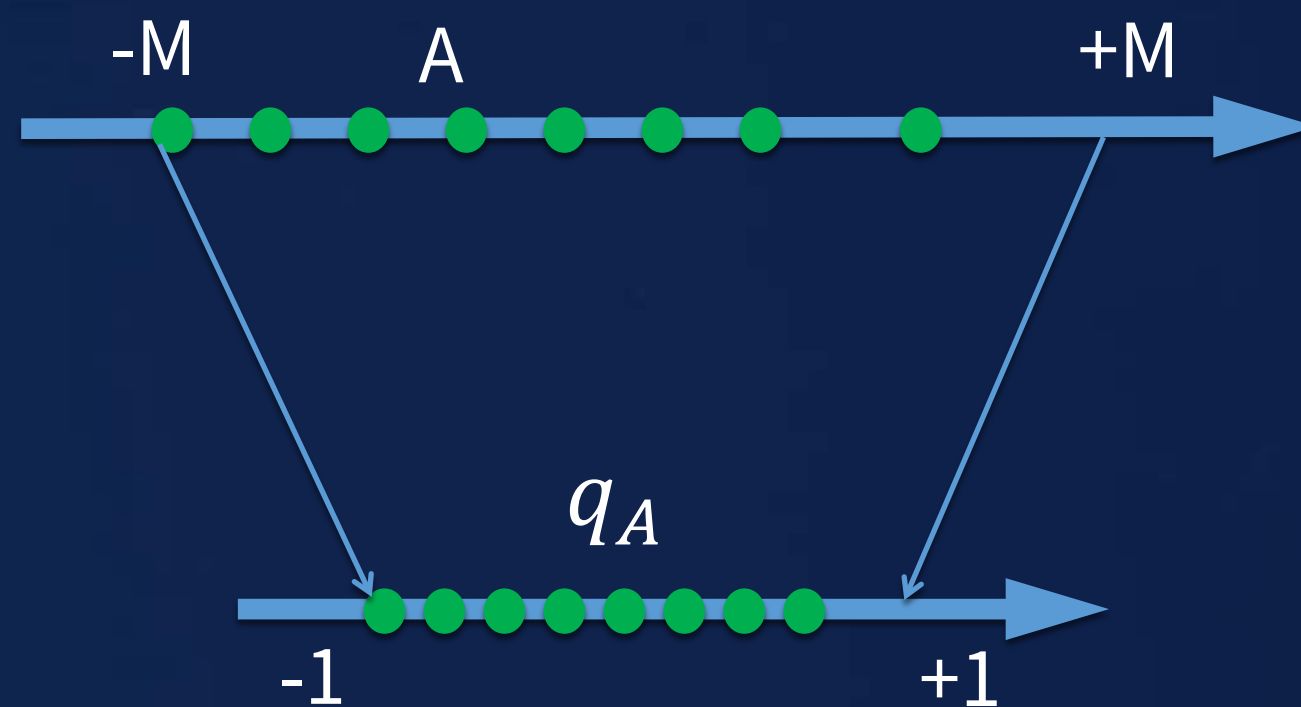
05

低成本FPGA的加速

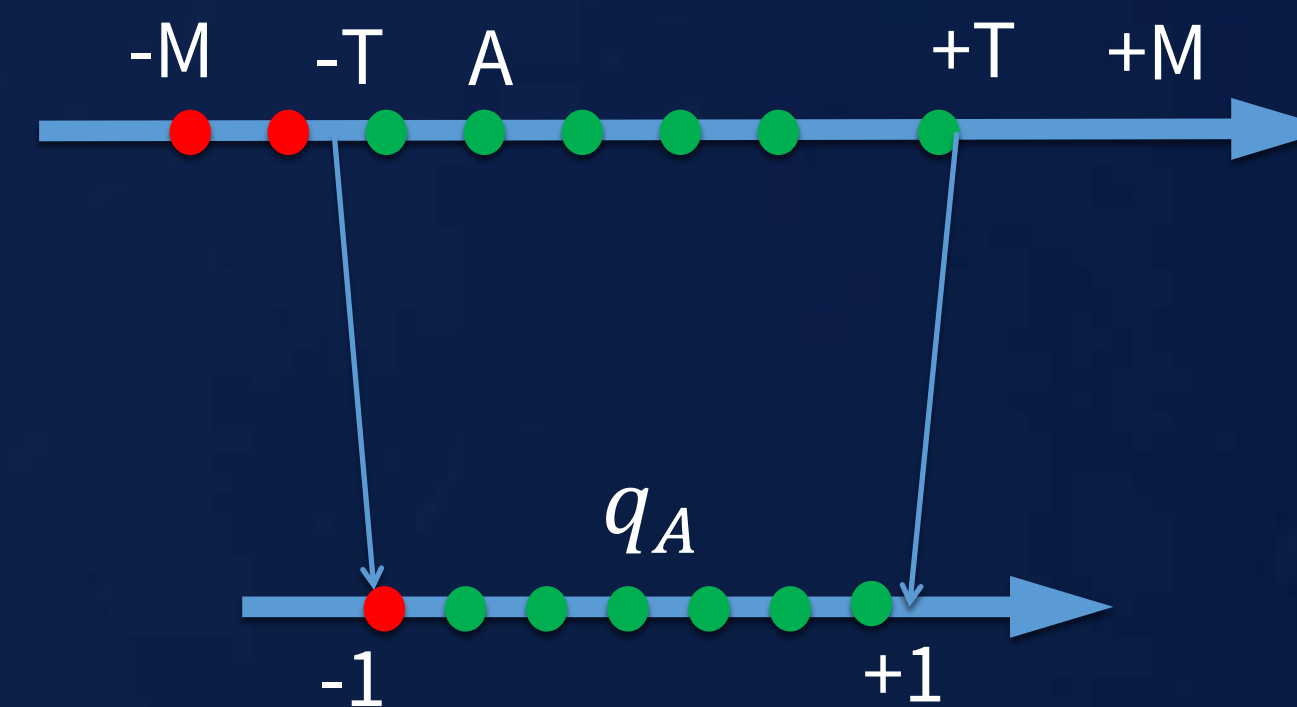
— FPGA定点化

- 表示范围FP32: $-3.4 \times 10^{38} \sim 3.4 \times 10^{38}$, INT8: $-128 \sim 127$
- 定点小数 $A = (A_0)A_1 \cdots A_k \cdots A_n$ (A_0 为符号位, A_i 为0/1)代表的小数为 $(-1)^{A_0} [A_1 * 2^{k-1} + A_2 * 2^{k-2} + \cdots + A_k * 2^0 + \cdots + A_n * 2^{k-n}]$
- 定点小数表示范围在 $\pm 2^k (1 - 0.5^n)$ 之间, 精度 (最小单位) 为 2^{k-n}
- 用INT8定点表示FP32: $A = 2^k * (-1)^{A_0} * 0.A_1A_2 \cdots A_n$ FP32 Value $A =$ FP32 scale factor $s_A * \text{INT8 Value } q_A$
- $$\sum_j A_j B_j = s_A s_B \sum_j q_{Aj} q_{Bj}$$

— 表示范围与精度的取舍



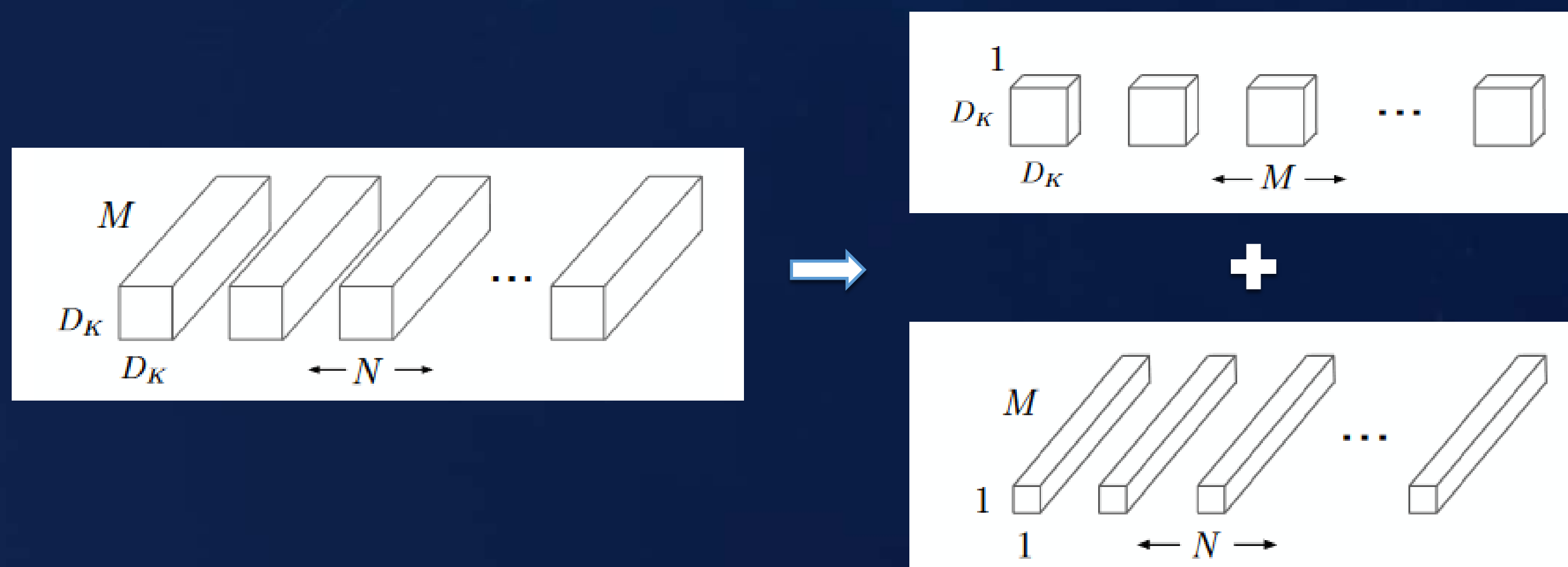
$M = \max |A| \leq 2^k$
表示范围大，精度差



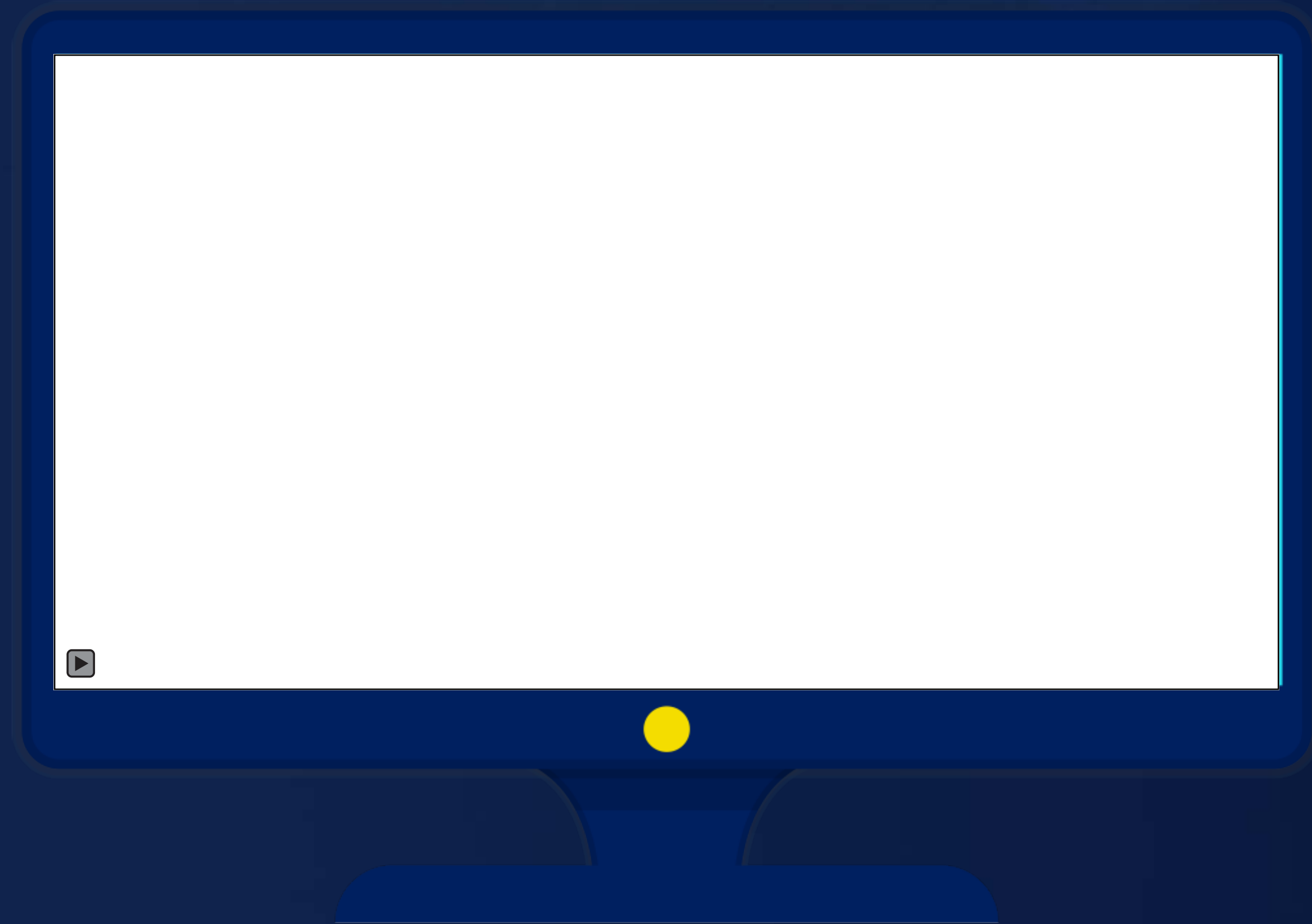
$M > T = 2^k$
表示范围小，精度好

— FPGA的网络选择

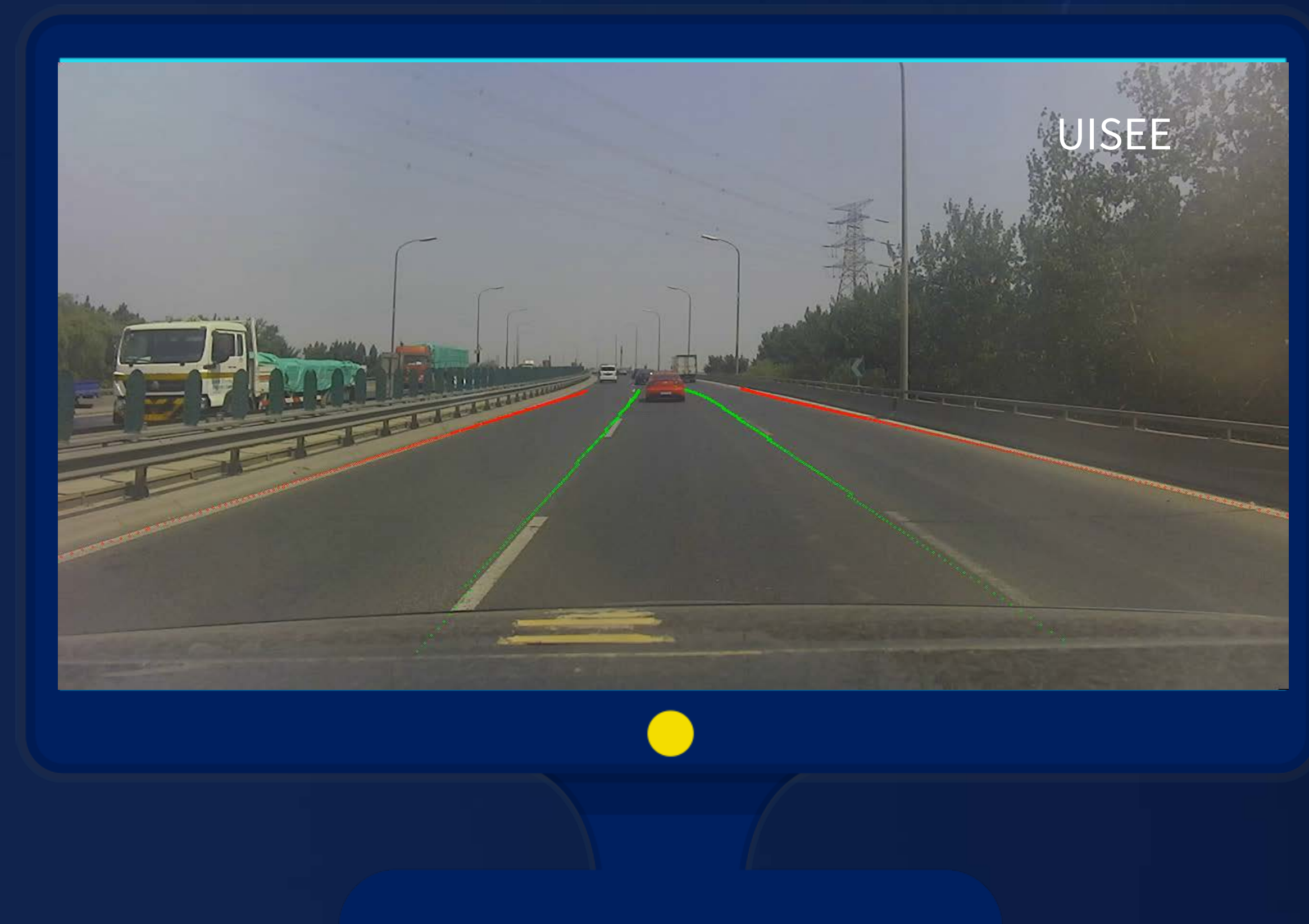
- MobileNet使用depth wise convolution+1x1 convolution
- 理论计算量低，同时精度很高
- GPU加速比较差，但适合CPU和定制计算设备



— MobileNet on FPGA



— UISEE Visual Perception demo video



— END 谢谢