

# 推理系统-模型小型化

# 模型小型化参数



# ZOMI



# Talk Overview

## 1. 推理系统介绍

- 推理系统与推理引擎
- 推理系统的工作流程
- 推理系统生命周期管理

## 2. 模型小型化

- NAS神经网络搜索
- CNN小型化结构
- Transform小型化结构

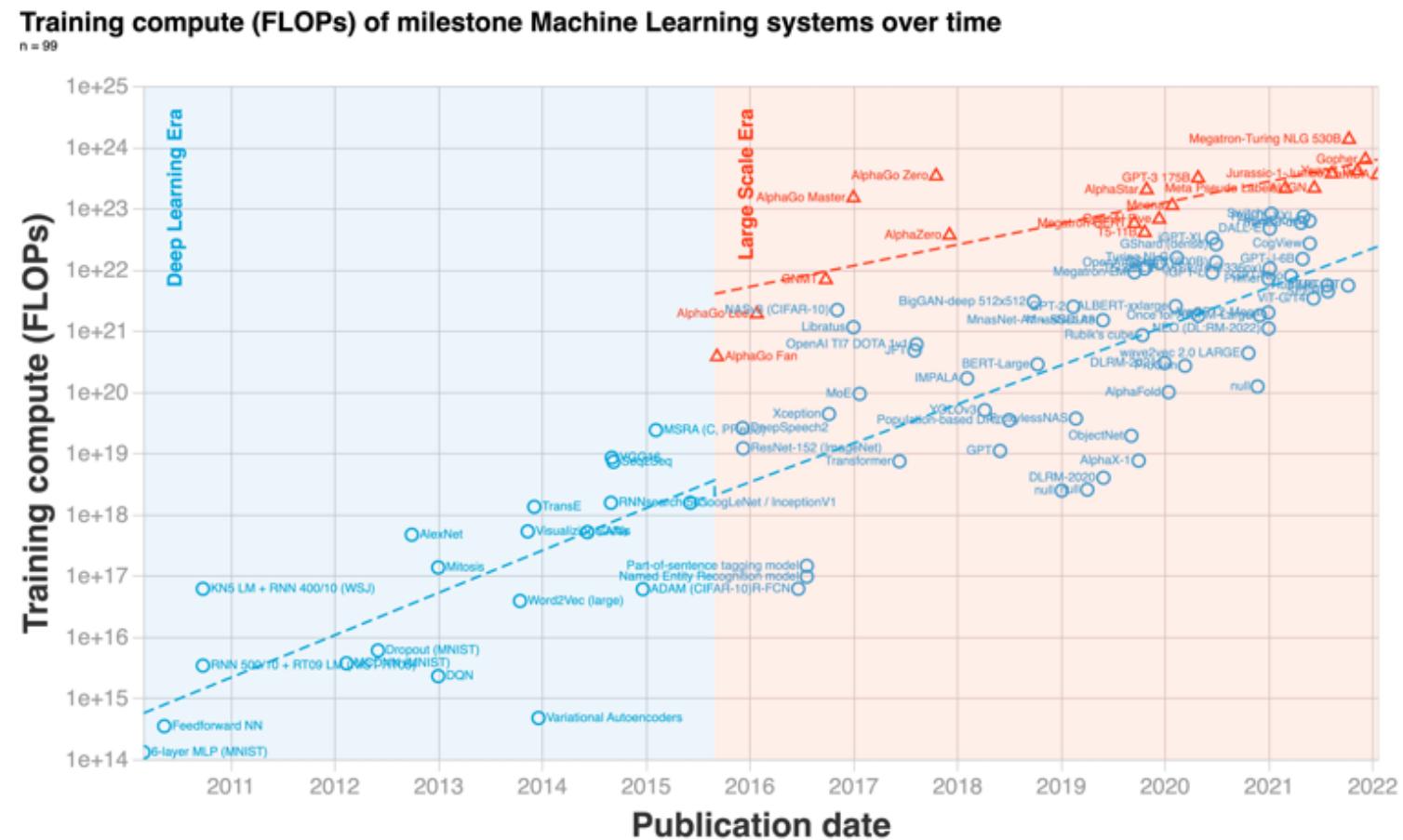
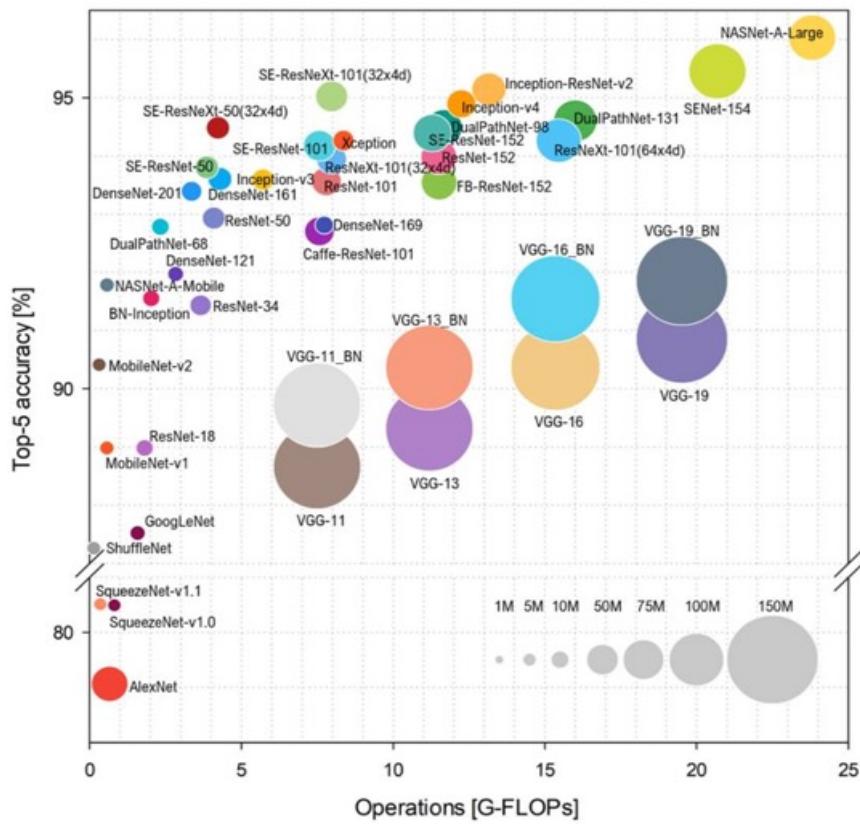
## 3. 离线优化压缩

- 低比特量化
- 二值化网络
- 模型剪枝
- 模型蒸馏

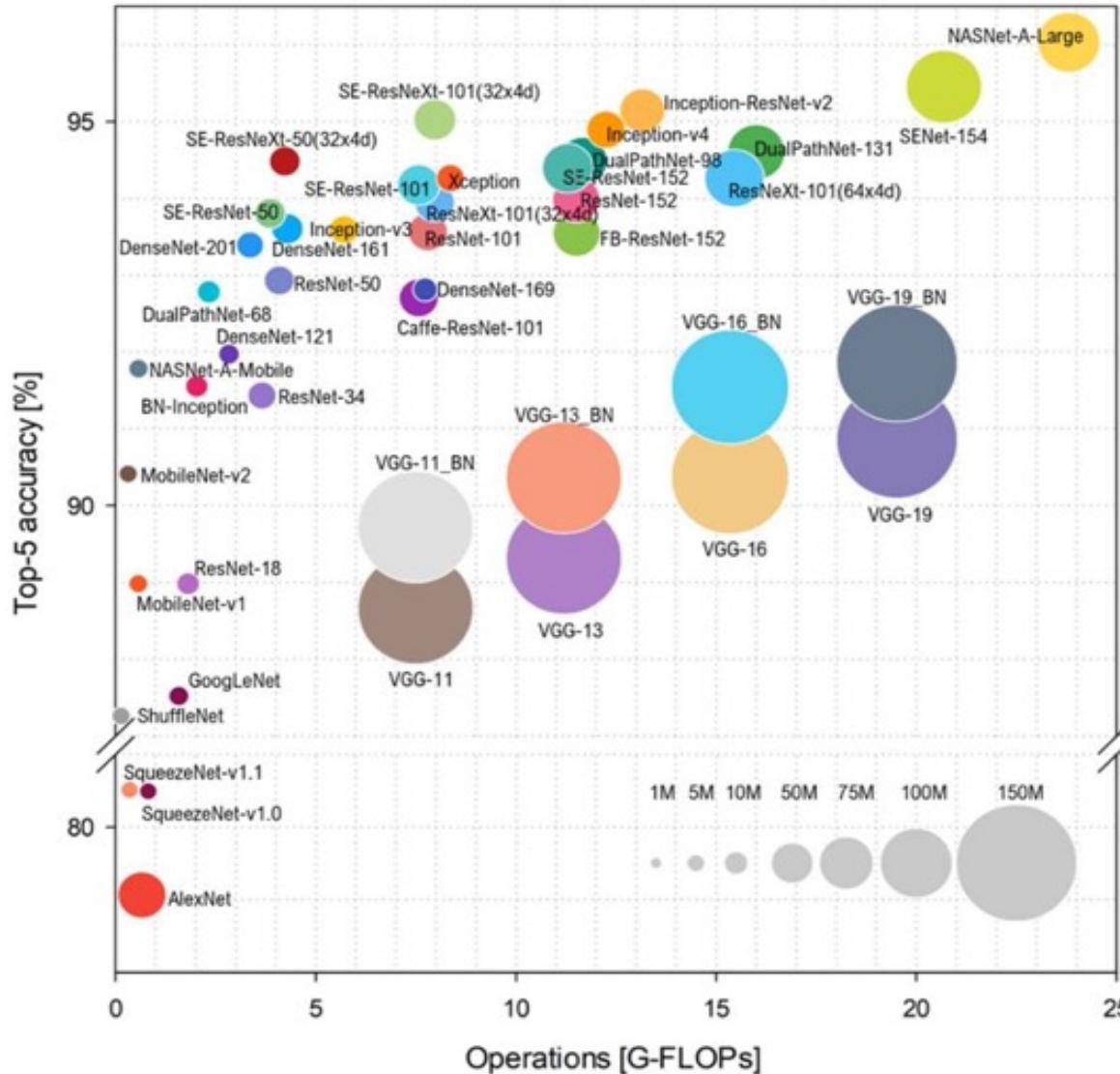
## 4. 部署和运行优化

- 图转换优化（算子融合/重排/替换）
- 并发执行与内存分配
- 动态batch与bin Packing

# 深度学习模型发展



# 深度学习模型发展



# 复杂度分析

## 1. FLOPs

浮点运算次数（ Floating-point Operation ），理解为计算量，可以用来衡量算法/模型时间的复杂度。

## 2. FLOPS

每秒所执行的浮点运算次数（ Floating-point Operations Per Second ），理解为计算速度，是一个衡量硬件性能/模型速度的指标，即一个芯片的算力。

## 3. MACCs

乘-加操作次数（ Multiply-accumulate Operations ）， MACCs 大约是 FLOPs 的一半，将  $w[0] \times x[0] \dots$  视为一个乘法累加或 1 个 MACC。

## 4. Params

模型含有多少参数，直接决定模型的大小，也影响推断时对内存的占用量，单位通常为 M，通常参数用 float32 表示，所以模型大小是参数数量的 4 倍。

# 复杂度分析

- **MAC**

内存访问代价（ Memory Access Cost ），指的是输入单个样本，模型/卷积层完成一次前向传播所发生的内存交换总量，即模型的空间复杂度，单位是 Byte。

- **内存带宽**

- 内存带宽决定了它将数据从内存（ vRAM ）移动到计算核心的速度，是比计算速度更具代表性的指标
- 内存带宽值取决于内存和计算核心之间数据传输速度，以及这两个部分之间总线中单独并行链路数量

# 典型结构对比

标准卷积层 Std Conv ( 主要贡献计算量 )

- Params :

$$k_h \times k_w \times c_{in} \times c_{out}$$

- FLOPs :

$$k_h \times k_w \times c_{in} \times c_{out} \times H \times W$$

# 典型结构对比

全连接层 FC ( 主要贡献参数量 )

- Params :

$$c_{in} \times c_{out}$$

- FLOPs :

$$c_{in} \times c_{out}$$

# 典型结构对比

Group Conv ( 主要贡献参数量 )

- Params :

$$(k_h \times k_w \times c_{in}/g \times c_{out}/g) \times g = k_h \times k_w \times c_{in} \times c_{out}/g$$

- FLOPs :

$$k_h \times k_w \times c_{in} \times c_{out} \times H \times W/g$$

# 典型结构对比

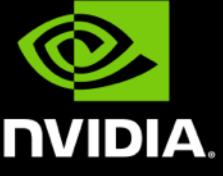
Depth-wise Conv ( 主要贡献参数量 )

- Params :

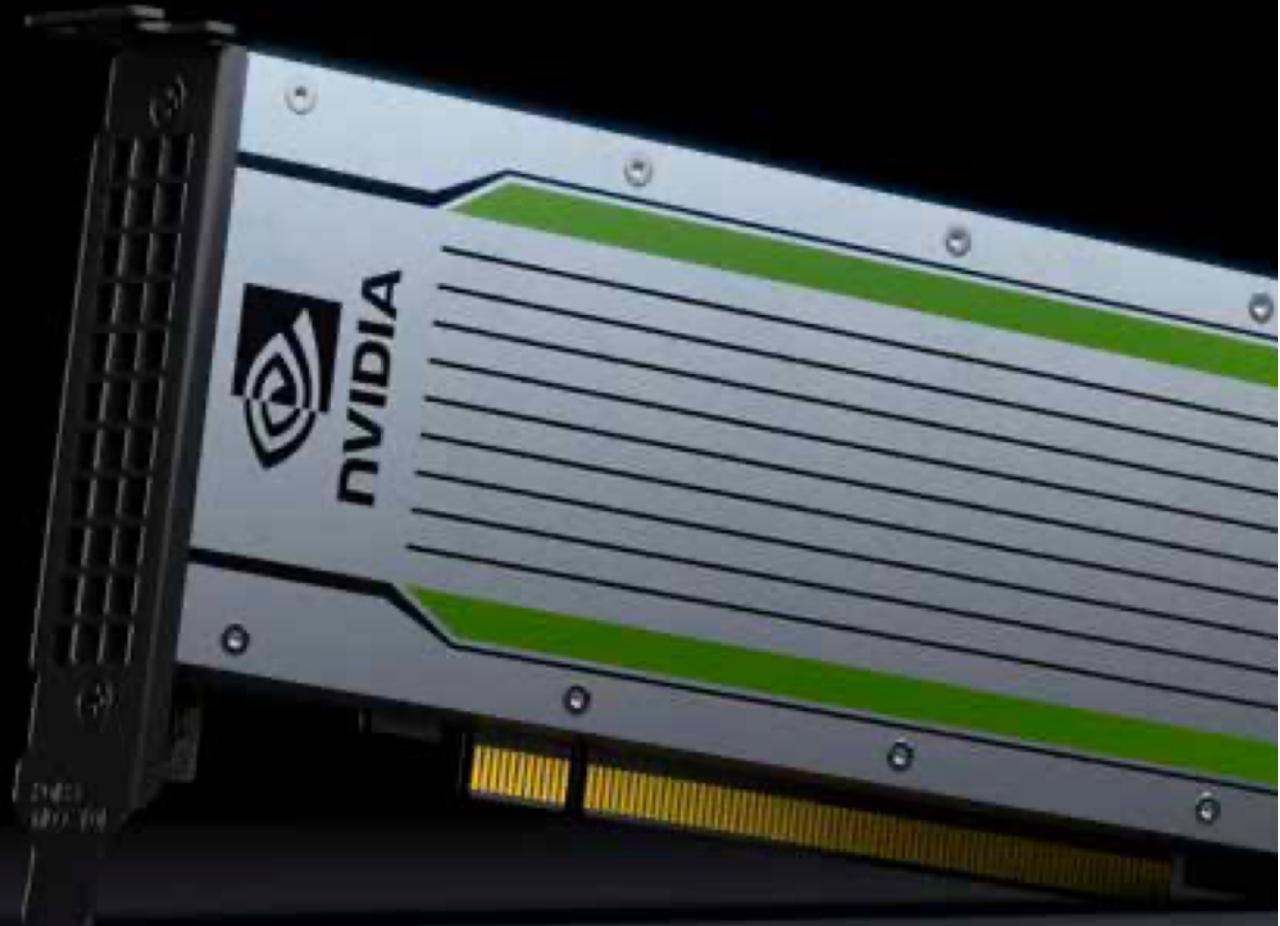
$$k_h \times k_w \times c_{in} \times c_{out} / c_{in} = k_h \times k_w \times c_{out}$$

- FLOPs :

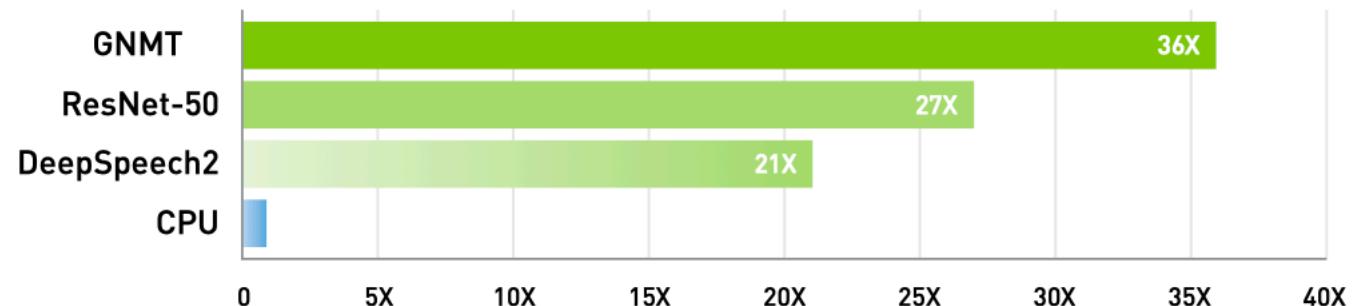
$$k_h \times k_w \times c_{out} \times H \times W$$



## NVIDIA T4 TENSOR 核心 GPU

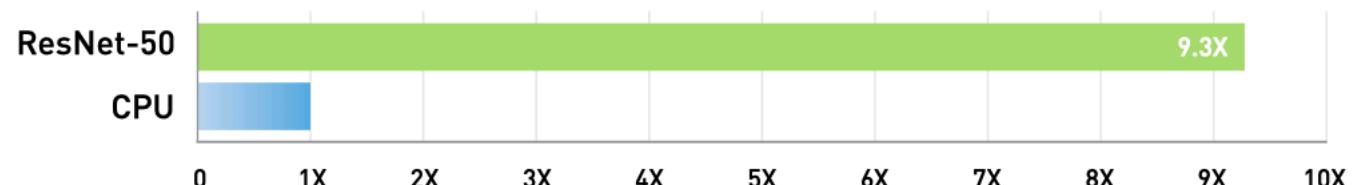


## 推理性能



一个 NVIDIA T4 GPU 与配双路至强 Gold 6140 CPU 的服务器进行对比

## 训练性能



两个 NVIDIA T4 GPU 与配双路至强 Gold 6140 CPU 的服务器进行对比

## 规格

GPU 架构	<b>NVIDIA Turing</b>
NVIDIA Turing Tensor Cores	<b>320</b>
核心数量	
NVIDIA CUDA® Cores	<b>2,560</b>
核心数量	
单精度	<b>8.1 TFLOPS</b>
混合精度 (FP16/FP32)	<b>65 TFLOPS</b>
INT8	<b>130 TOPS</b>
INT4	<b>260 TOPS</b>
GPU 显存	<b>16 GB GDDR6</b>
	<b>300 GB/s</b>
ECC	<b>支持</b>
互联带宽	<b>32 GB / 秒</b>
系统接口	<b>x16 PCIe Gen3</b>
外形尺寸	<b>PCIe 半高卡</b>
散热解决方案	<b>被动式</b>
计算 API	<b>CUDA</b>
	<b>NVIDIA TensorRT™</b>
	<b>ONNX</b>

## SPECIFICATIONS

GPU Architecture	<b>NVIDIA Turing</b>
NVIDIA Turing Tensor Cores	<b>320</b>
NVIDIA CUDA® Cores	<b>2,560</b>
Single-Precision	<b>8.1 TFLOPS</b>
Mixed-Precision (FP16/FP32)	<b>65 TFLOPS</b>
INT8	<b>130 TOPS</b>
INT4	<b>260 TOPS</b>
GPU Memory	<b>16 GB GDDR6</b>
	<b>300 GB/sec</b>
ECC	<b>Yes</b>
Interconnect Bandwidth	<b>32 GB/sec</b>
System Interface	<b>x16 PCIe Gen3</b>
Form Factor	<b>Low-Profile PCIe</b>
Thermal Solution	<b>Passive</b>
Compute APIs	<b>CUDA, NVIDIA TensorRT™, ONNX</b>



# BUILDING A BETTER CONNECTED WORLD

# THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.