NVIDIA 部分产品的详细分析

众所周知，GPU 的强大算力推动了 AI 的快速发展，变革了社会的各个行业。如何选择适合自己领域的 GPU 是公司采购时经常困惑的地方，本文的对比分析意在为此提供参考。

1. NVIDIA 部分产品目录

NVIDIA 的产品众多，涉及到游戏、专业视觉、AI 计算、自动驾驶等各个领域，不同领域对应多款产品，表 1-1 列出了部分产品对应的系列及架构。

表 1-1 NVIDIA 不同产品对应的系列及架构

架构 / 系列	Pascal	Volta	Turing
Tesla (高速计算)	P100 (SXM2,PCIE), P40, P4	V100(SXM2,PCIE)	T4
GeForce (游戏显卡)	GTX (1030,1050,1050Ti,1060, 1070,1070Ti,1080,1080Ti)		RTX (2070,2080, 2080Ti)
Quadro (专业绘图)	P400,P600,P620,P1000,P2000, P4000,P5000,P6000,GP100	GV100	RTX (5000,6000, 8000)
TITAN (通用)	X, Xp	V	RTX
Jetson (嵌入式系统)	TX2	AGX Xavier	

NVIDIA DGX 系列侧重于为 AI 行业提供端到端的解决方案。主要有：DGX station 采用 4 块 V100GPU，DGX-1 采用 8 块 V100GPU，DGX-2 采用 16 块 V100GPU。

NVIDIA HGX 系列侧重于为企业提供云平台服务，主要有: HGX-1 采用 8 块 V100GPU, HGX-2 采用 16 块 V100GPU。

思腾合力侧重于 AI 和高性能计算，因此本文重点介绍 Tesla 系列，GeForce 系列，TITAN 系列和 Jetson 系列产品。

2. Tesla 系列产品

利用 Tesla GPU 能够快速地处理要求最严格的高性能计算 (HPC) 和超大规模数据中心任务，数据科学家和研究人员可以在能源勘探、生物医疗、深度学习等应用场景解析 PB 级的数据，速度比传统 CPU 快几个数量级。此外，基于 Tesla GPU 组成的 HGX 系列和 DGX 系列产品分别为虚拟桌面和工作站提供超高性能和用户密度。

2.1 Tesla 系列产品的详细参数

表 2-1 NVIDIA Tesla GPU 参数

GPU / 指标	P100 (SXM2)	P100 (PCIE)	P40	P4	V100 (PCIE)	V100 (SXM2)	T4
CUDA 核	3584	3584	3840	2560	5120	5120	2560
Tensor core	NA	NA	NA	NA	640	640	320
FP64 峰值/ TFLOPS	5.3	4.7	NA	NA	7	7.8	NA
FP32 峰值/ TFLOPS	10.6	9.3	12	5.5	14	15.7	8.1
FP16 峰值/ TFLOPS	21.2	18.7	NA	NA	112	125	65
INT8 峰值/ TIOPS	NA	NA	47	22	NA	NA	NA
GPU 内存/ GB	16 HBM2	16/12 HBM2	24 GDDR5	8 GDDR5	32 HBM2	32 HBM2	8 GDDR6
内存带宽/ GB/s	732	732/549	346	192	900	900	320+
系统接口	NVLink + PCIe 3.0	PCIe 3.0	PCIe 3.0	PCIe 3.0	PCIe 3.0	NVLink + PCIe 3.0	PCIe 3.0 ×16
硬件加速视频引擎	---	---	1×解码引擎 2×编码引擎	1×解码引擎 2×编码引擎	---	---	1×解码引擎 2×编码引擎
功耗/W	300	250	250	50-75	250	300	70
发布时间	2016.4.5	2016.6.20	2016.9.13	2016.9.13	2017.6.21	2017.6.21	2018.9.12

*从 Volta 架构开始有 Tensor core，其存在极大的提升了半精度的计算峰值。

* FP64 表示双精度(double)，FP32 表示单精度(float)，分别在计算机存储中占 8，4 个字节。

* FP16 表示半精度，INT8 表示整型，分别在计算机存储中占 2，1 个字节。

* HBM2(High Bandwidth Memory)，基于 3D 堆栈工艺的高性能 DRAM，其存储器带宽较高。

* GDDR(Graphics Double Data Rate)，GDDR6 的带宽相较于 GDDR5 可提升一倍。

2.2 Tesla 系列产品的性能对比分析

由表 2-1 可知，P100,V100 的 CUDA 核心数较多，FP64,FP32,FP16 计算峰值较高，同时其内存采用 HBM2，内存带宽较大，但其功耗也大，因此适应于高性能计算和 AI 中的训练环节。P4,T4 的功耗较低，体积较小，P4 的 INT8 峰值较高，T4 的 FP16 峰值高，因此适合于超高效横向扩展服务器和 AI 中的推理部署环节。P40 的 CUDA 核数多，FP32，INT8 的峰值高，内存大，功耗大，适合于训练和高吞吐量的推理

3. GeForce 系列产品

GeForce GTX 10 系列和 RTX 20 系列产品能够提供强大的视觉特效和渲染技术，在高达 240 Hz 的刷新率及 HDR 等条件下，享受超级流畅，无画面撕裂的极致游戏体验。RTX 采用最新的 Turing 架构，同时为游戏引入了全新的实时光线追踪和 AI 技术。

3.1 GeForce 系列产品的详细参数

表 3-1 NVIDIA GeForce GPU 参数

GPU / 指标	1070	1080	1080Ti	2070	2080	2080Ti
CUDA 架构	Pascal	Pascal	Pascal	Turing	Turing	Turing
CUDA 核	1920	2560	3584	2304	2944	4352
FP32/TFLOPS	6.5	8.9	11.3	7.5	10.1	13.4
RTX-OPS	NA	NA	NA	42T	57T	76T
提升频率/MHz	1683	1733	1582	1620	1710	1545
显存速率/Gbps	8	10	11	14	14	14
GPU 显存/GB	8 GDDR5	8GDDR5X	11GDDR5X	8 GDDR6	8 GDDR6	11GDDR6
显存带宽/GB/s	256	320	484	448	448	616
功耗/W	150	180	250	175	215	250
发布时间	2016.6.10	2016.5.27	2017.3.5	2018.10	2018.9.20	2018.9.20

*RTX-OPS 指 GPU 在阴影、光线跟踪等操作中的平均性能，以及每秒千兆光线的测量结果。

3.2 GeForce 系列产品的性能对比分析

由表 3-1 可知，相较于 GTX 10 系列，RTX 20 系列产品 CUDA 核数增加，单精度的峰值更大，显存升级为 GDDR6，显存速率和带宽都相应提高，但其功耗增加不大，因此 RTX 20 的能耗比 GTX 10 的更高。

4. TITAN 系列产品

TITAN 系列产品意在打造运行速度更快的 PC 显卡，推动高性能计算和 AI 的外部极限，使研究人员快速运行其科学模型，在深度学习计算任务中，TITAN V 可达到 110TFLOPS 的浮点运算能力，TITAN RTX 更是具有 130TFLOPS 的性能。

4.1 TITAN 系列产品的详细参数

表 4-1 NVIDIA TITAN GPU 参数

GPU / 指标	TITAN X	TITAN Xp	TITAN V	TITAN RTX
CUDA 架构	Pascal	Pascal	Volta	Turing
CUDA 核	3072	3840	5120	4608
Tensor core	NA	NA	640	576
RT core	NA	NA	NA	72
FP32 峰值/TFLOPS	11	12	15	
提升频率/MHz	1075	1582	1455	1770
显存速率/Gbps	7	11.4	1.7	14
GPU 显存/GB	12 GDDR5	12 GDDR5X	12 HBM2	24 GDDR6
显存带宽/GB/s	336.5	547.7	652.8	672
功耗/W	250	250	250	280
发布时间	2016.8.2	2017.4.6	2017.12.7	2018.12.18

4.2 TITAN 系列产品的性能对比分析

由表 4-1 可知，TITAN X 和 TITAN Xp 的显存速率较高，显存容量较大，作为游戏显卡，能够提供强大的视觉特效和图片渲染效果；TITAN V 的 CUDA 核数量很大，同时拥有 640 个 Tensor Core，其计算能力强大，可作为 PC 级的 GPU 加速卡，提高深度学习任务的训练速度。TITAN RTX 不仅拥有非常高的显存速率，非常大的显存容量，添加了 72 RT core 用于增强光线追踪能力，而且其 CUDA 核数和 Tensor Core 数也很多，因此利用 TITAN RTX，可以任意挥洒创意。

5. Jetson 系列产品

NVIDIA Jetson 是业内领先的 AI 计算平台，面向移动嵌入式系统市场中的 GPU 加速并行处理。

Jetson 模块适用于计算密集型的嵌入式项目，非常适合低能耗和高计算性能的应用程序，使用者能够轻松上手并快速开发产品。例如，实时智能视频分析 (IVA) 系统助力创建更智能更安全的 AI 城市、无人机可协助检查手机信号塔、电线、风力涡轮机和其他基础设施、企业可以打造更高效且更具有可预见性的供应链和物流系统……

5.1 Jetson 模块的详细参数

表 5-1 NVIDIA Jetson 模块技术规格

模块 / 指标	**Jetson TX2**	**Jetson AGX Xavier**
GPU	256 Core Pascal @ 1.3GHz	512 Core Volta @1.37GHz 64 Tensor cores
深度学习加速器	---	(2×) NVDLA
视觉加速器	---	(2×)7-way VLIW Processor
CPU	6 core Denver and A57 @ 2GHz (2×) 2MB L2	8 core Carmel ARM CPU @ 2.26GHz (4×)2MB L2+4MB L3
内存	8GB 128bit LPDDR4 58.4GB/s	16GB 256-bit LPDDR4× @2133MHz 137GB/s
存储	32GB eMMC	32GB eMMC
视频编码	(2×) 4K @ 30 HEVC	(4×) 4Kp60/(8×) 4Kp30 HEVC
视频解码	(2×) 4k @ 30 12bit support	(2×) 8Kp30/(6×) 4Kp60 12 bit support
摄像头	12 lanes MIPI CSI-2 D-PHY 1.2 30Gbps	16 lanes MIPI CSI-2|8 lanes SLVS-EC D-PHY 40Gbps / C-PHY 109Gbps
PCIE	5 lanes PCIe Gen2 1×4 + 1×1 or 2×1 + 1×4	16 lanes PCIe Gen4 1×8 + 1×4 + 1×2 + 2×1
尺寸	50mm×87mm (400 pin connector)	100mm×87mm (699 pin connector)
功耗/W	7.5/15	10/12/30

5.2 Jetson 系列产品的性能对比分析

Jetson AGX Xavier 的 GPU 采用 Volta 架构，拥有 512CUDA 核和 64 个 Tensor core，深度学习加速器和视觉加速器，其并行计算能力大大提高，CPU 的核心数也增加至 8 个，主频增强至 2.26GHz，内存和存储也加大，视频编解码增强，摄像头增加至 16 路，因此 Xavier 产品的总体性能比 TX2 更强大，在深度学习、计算机视觉、工业机器人、车载设备等方面有很大的应用前景。

6. 深度学习 GPU 产品

深度学习中两个最重要的张量操作是矩阵乘法和卷积。矩阵乘法与显存大小和带宽密切相关，卷积受计算速度的约束，因此、显存大小，显存带宽，处理能力(FLOPS 和 Tensor Core 的组合)是深度学习任务中选择 GPU 的重要指标。其中 Tensor Core 是专用计算单元，可以加速计算，同时允许使用 16-bit 数字进行计算，在软件支持的情况下，采用 16-bit 输入进行乘法计算，相当于内存翻倍。

6.1 深度学习 GPU 的部分参数

表 6-1 NVIDIA 部分 GPU 产品的参数

指标 / GPU	显存容量/GB	显存带宽/Gbps	Tensor Core	FP32 峰值/TFLOPS
V100(SXM2)	32 HBM2	900	640	15.7
TITAN RTX	24 GDDR6	672	576	16.3
P100(SXM2)	16 HBM2	732	NA	10.6
TITAN V	12 HBM2	652.8	640	15
RTX 2080Ti	11 GDDR6	616	544	13.4
RTX 2080	8 GDDR6	448	368	10.1
RTX 2070	8 GDDR6	448	288	7.5
TITAN Xp	12 GDDR5X	547.7	NA	12
RTX 1080Ti	11 GDDR5X	484	NA	11.3
TITAN X	12 GDDR5	336.5	NA	11
GTX 1080	8 GDDR5X	484	NA	8.9
RTX 1070Ti	8 GDDR5	256	NA	8.1
RTX 1070	8 GDDR5	256	NA	6.5
RTX 1060	6 GDDR5	256	NA	4.4

6.2 深度学习 GPU 的性能对比分析

递归神经网络 RNN 使用大量的矩阵乘法，卷积神经网络 CNN 使用大量的卷积计算。图 6-1 显示了适合深度学习的大部分 GPU(没有 TITAN RTX 和 P100)的运算速度，纵坐标是 GPU 型号，从下至上 GPU 的 CNN 和平均运算速度增强；图 6-2 显示了不同 GPU 的性价比，即 GPU 的运算速度与价格之比。

对于如何选择深度学习 GPU，其建议如下：

对性能要求很严格：V100>TITAN RTX>TITAN V

对性能和价格要求都有要求（侧重于性能）：2070>2080>2080Ti>P100

对性能和价格要求都有要求（侧重于价格）： 1080Ti >1080>TITAN X>TITAN Xp

刚进入深度学习领域，对价格要求严格：1070Ti>1070>1060

购买时还要注意自己的数据集，如果一个 batch size 很大，一定要注意 GPU 的显存容量。

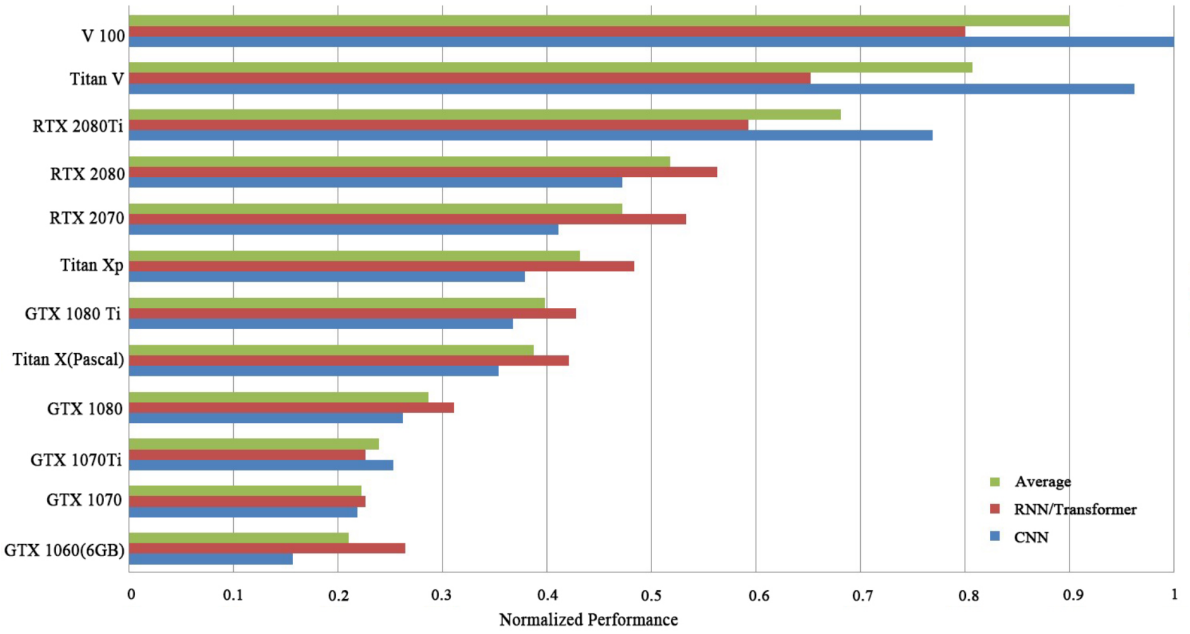

图 6-1 不同 GPU 的归一化运算速度

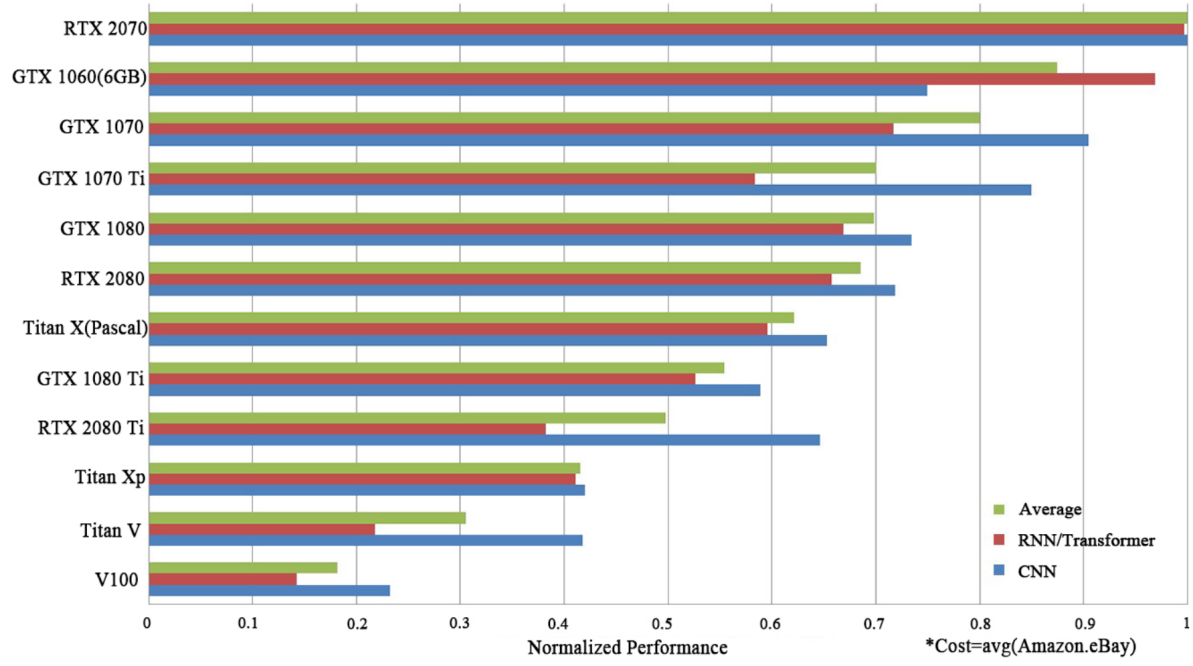

图 6-2 不同 GPU 的性价比

参考文献：

https://zhuanlan.zhihu.com/p/42809635
https://zhuanlan.zhihu.com/p/53667790
https://www.nvidia.com/zh-cn/
https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units#GeForce_700_Series