5.4. 谷歌TPU历史发展

5.4.1 为什么需要TPU

早在2006年，谷歌的内部就讨论过在自家的数据中心中部署图形处理器（GPU）、现场可编程门阵列（FPGA）或自研专用集成电路（ASIC）的可能性。但当时能够在特殊硬件上运行的少数应用程序可以几乎0代价的利用当前谷歌大型数据中心的过剩算力完成，那什么比免费的午餐更有吸引力呢？于是此项目并没有落地。但是在2013年，风向突变，当时谷歌的研究人员做出预测：如果人们每天使用语音搜索并通过深度神经网络（DNN）进行3分钟的语音识别，呢么当时谷歌的数据中心需要双倍的算力才能满足日益增长的计算需求，而仅仅依靠传统CPU来满足这种需求是非常昂贵的。[1]于是，在这个背景下，谷歌开始了TPU的设计。

谷歌不愧是谷歌，通常一个芯片的开发需要几年的时间，然而TPU从立项到大规模部署只用了15个月。TPU项目的领头人Norm Jouppi说到：“我们的芯片设计过程异常迅速，这本身就是一项非凡的成就。令人惊叹的是，我们首批交付的硅片无需进行任何错误修正或掩模的更改。考虑到在整个芯片构建过程中，我们还在同步进行团队的组建，紧接着迅速招募RTL（寄存器传输级）设计专家，并且急切地补充设计验证团队，整个工作节奏非常紧张。”[2]

5.4.2 历代TPU芯片与产品

5.4.2.1历代TPU芯片

以下表格是不同TPU芯片型号的具体参数和规格，我们的TPU系列会主要围绕v1, v2, v3, v4这一系统去展开。

TPU 比较	TPUv1	TPUv2	TPUv3	Edge TPU v1	Pixel Neural Core	TPUv4i	TPUv4	Google Tensor
推出日期	2016年	2017年	2018年	2018年	2019年	2020年	2021年	2021年
制程技术	28nm	16nm	16nm	-	-	7nm	7nm	-
芯片大小 (mm^2)	330	625	700	-	-	400	780	-
芯片内存 (MB)	28	32	32	-	-	144	288	-
时钟速度 (MHz)	700	700	940	-	-	1050	1050	-

TPU 比较	TPUv1	TPUv2	TPUv3	Edge TPU v1	Pixel Neural Core	TPUv4i	TPUv4	Google Tensor
内存	8 GiB DDR3	16 GiB HBM	32 GiB HBM	-	-	8GiB DDR	32 GiB HBM	-
内存带宽 (GB/s)	300	700	900	-	-	300	1200	-
热设计功耗 (W)	75	280	450	-	-	175	300	-
TOPS (Tera/Second)	92	45	123	4	-	-	275	-
TOPS/W	0.31	0.16	0.56	2	-	-	1.62	-

5.4.2.1历代TPU产品

在前文中，我们讨论了CPU的不同型号，现在让我们将注意力转向谷歌的TPU产品线。以下表格中除了芯片之外，随着芯片技术的不断进步，谷歌推出了TPU Pod，这是一种由众多TPU单元构成的超大规模计算系统，专为处理大量深度学习和AI领域的并行计算任务而设计。除了具有超强的算力之外，TPU Pod装备了高速的互联网络，保证了TPU设备之间无缝的数据传输以保证强大的数据、模型层的高效拓展性。

名称	时间	性能	应用
TPUv1	2016年	92Tops + 8GB DDR3	数据中心推理
TPUv2	2017年	180TFlops(浮点计算能力) + 64GB(HBM)	数据中心训练和推理
TPUv3	2018年	420TFlops + 128GB(HBM)	数据中心训练和推理
Edge TPU	2018年	可处理高吞吐量的稀疏数据	IoT 设备
TPUv2 Pod	2019年	11.5万亿次运算/秒，4TB (HBM)	数据中心训练和推理
TPUv3 Pod	2019年	>100万亿次运算/秒，32TB (HBM)	数据中心训练和推理
TPUv4	2021年	-	数据中心训练和推理
TPUv4 Pod	2022年	-	数据中心训练和推理

随着时间的推移，谷歌不仅在大型数据中心部署了先进技术，还洞察到将这些技术应用于消费电子产品，尤其是智能手机市场的巨大潜力。早在2017年，谷歌便在Pixel 2 和 Pixel 3上搭载了Pixel Visual Core，成为了谷歌针对消费类产品的首个定制图像芯片，而在2019年10月发布的Pixel 4上，谷歌基于Edge TPU的框架

研发了Pixel Visual Core的继任，Pixel Neural Core。而谷歌在 Pixel 产品线上对于TPU的依赖也一直延续到了今天。

在这个AI爆发的大时代，Google Silicon 高级总监 Monika Gupta 说到：“我们的合作与Tensor一直不仅仅局限于追求速度和性能这样的传统评价标准。我们的目标是推动移动计算体验的进步。在最新的Tensor G3芯片中，我们对每个关键的系统组件都进行了升级，以便更好地支持设备上的生成式人工智能技术。这包括最新型号的ARM中央处理器、性能更强的图形处理器、全新的图像信号处理器和图像数字信号处理器，以及我们最新研发的，专门为运行Google的人工智能模型而量身打造的TPU。”[3]

参考文献

1. In-Datacenter Performance Analysis of a Tensor Processing Unit
2. An in-depth look at Google’s first Tensor Processing Unit (TPU)
3. Google Tensor G3: The new chip that gives your Pixel an AI upgrade