QuantaFlow
量子流 AI 加速计算

QUANTAFLOW 量子流 AI 加速计算

PQ LABS

PQ LABS, INC. | 3754 Spinnaker Court, Fremont, California. 94538

QuantaFlow

量子流加速计算技术

QuantaFlow 技术是由品奇实验室（PQ Labs）研发的一项针对通用神经网络进行加速计算的技术方案。通过对类量子算法的模拟运行，实现了单卡相对于 NVIDIA Tesla V00 10 倍左右的性能提升。

量子计算和量子算法的背景介绍

量子力学的态叠加原理使得量子信息单元的状态可以处于多种可能性的叠加状态，从而导致量子计算天生适合处理大规模的复杂计算。

量子计算机的基本原理

就像经典计算架构是利用电子会从高电压流向低电压的基本物理特性一样，量子计算也是利用了处于量子态的微粒子的基本物理特性而设计出来的。

与经典计算系统一样，量子计算也是分为输入、计算、输出三个过程，但是与经典计算需要严格的设计计算过程不同，量子计算的计算过程是“自然演化”，即设计好输入信息，然后让量子比特在多个平行宇宙中并行演化，然后对演化结果进行观测，进而得到输出数据。

这里面和我们的主题相关的，主要有这么几个概念：

- **量子叠加态。**一个量子比特（qu-bit）可以同时处于多个可能结果的概率叠加状态，即一个量子比特可以同时代表多个数值，进而只需要很少的量子比特，即可处理非常多的概率叠加结果，而传统计算需要穷举运算每一个可能性；

- **并行演化。**正如前文提到，量子计算的“计算”过程是遵循物理规律自然演化的过程，而量子比特的每一个叠加状态都在不同的平行宇宙中并行的演化，互不干涉，同时计算。即使是对于经典计算架构非常复杂的运算，量子计算只需要非常少的计算即可实现。

这些特性使得量子计算在处理超大规模的复杂计算问题时，拥有传统计算架构无法比拟的性能优势。但是要真正实现有用的量子计算机，还存在巨大的技术鸿沟。

谷歌、IBM 的“量子门电路”方案，虽然能实现“量子霸权”，但只能跑特定的量子算法，QRAM(量子内存/量子随机存储器)也都没有，量子比特非常不稳定，容易坍塌。要将出错率降到可实用的程度，每个量子比特的背后需要消耗 1000 个量子比特来进行算法纠错。[1]

所以很遗憾，由于物理技术发展的限制，目前人类还没有制造出被广泛认可的通用量子计算机。但是基于量子图灵机的设计原理以及量子物理的数学特性而设计出的量子算法，已经可以在量子计算的模拟环境中运行起来。

量子计算机的争议

与谷歌、IBM 的“量子门电路”设计不同，加拿大 D-Wave 公司的方案看上去更像一个量子算法模拟器。2019 年 D-Wave 发布了“非主流”的 5000 个量子比特处理器，远远高于谷歌 53 个量子比特的量子霸权计算机。在一些专用领域 D-Wave 的量子算法比传统算法快了 1 亿倍（来自谷歌的一个研究小组展示的他们在 D-Wave 电脑上做的实验结果和论文）

[1] Indeed, with current experimental error rates in multi-qubit devices somewhat below 1%, large calculations like those needed for quantum chemistry might require 1,000 physical qubits per logical qubit, Preskill said.

https://cacm.acm.org/magazines/2019/10/239668-closing-in-on-quantum-error-correction/fulltext

但是，剩下一个问题就是这个 D-Wave 设备是否能评为“真正的量子计算机”？学术界对此存在争议，很多科学家 认为 D-Wave 不是“true quantum”，并没有利用到“纯物理意义”上的量子，而是用了仿真和量子算法。[2]

但这并不妨碍 D-Wave 成为第一个能用来解决实际问题的“量子计算机”。D-Wave 可以用来做药物分子的分类，数独游戏求解，优化酒店广告，图片着色，复杂网络问题，模拟天气模式和海啸，等等。

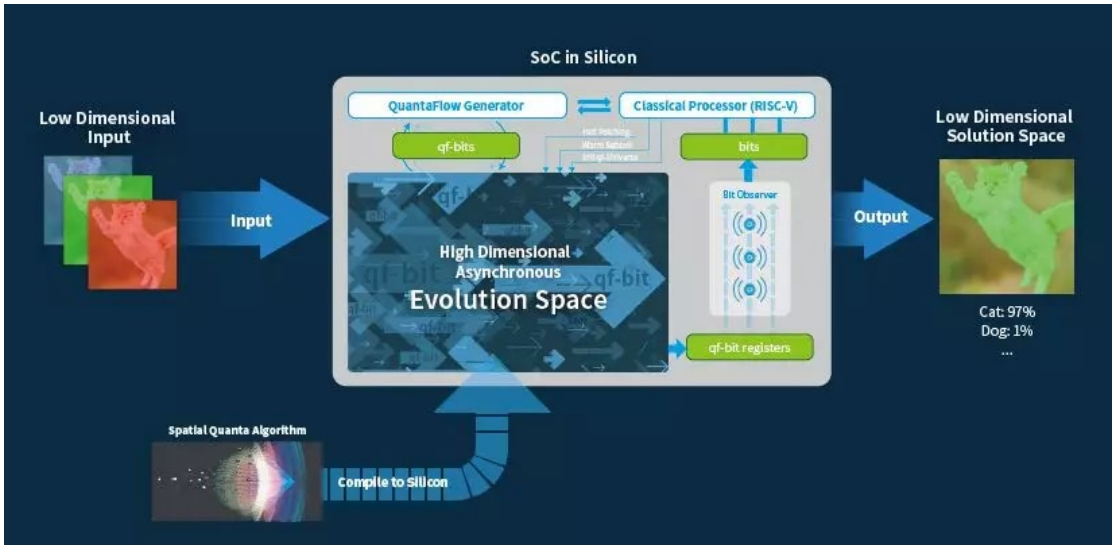
[2] https://www.theverge.com/2016/9/28/13057414/quantum-computer-d-wave-2000-qubit-chip

但是 D-Wave 的产品体积庞大，性价比不高，解决的问题类别依旧有限，在目前阶段妨碍了大规模应用。

QuantaFlow，量子流架构与人工智能的结合

品奇实验室尝试将量子算法的思想精髓引入到 AI 芯片架构中去，进而在现有的硅基计算平台上大幅提升并行计算算力。但是如何在现有的硅基计算平台上实现，则是个难题。

QuantaFlow，是来自品奇实验室的答案。QuantaFlow 技术在硅基芯片中模拟出一个量子比特流进行模拟演化的虚拟空间。在此基础上设计了基于经典架构的 RISC-V 处理器来实现对量子模拟空间的逻辑控制、结果观测等功能。

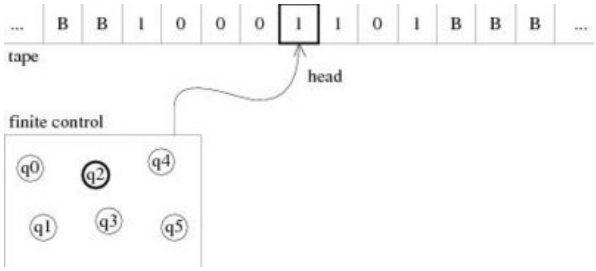


QuantaFlow 可以将低维空间中的输入信息映射到高维的量子演化算法空间中，进行大量异步、并行的演化计算，将结果保存在量子流比特寄存器（qf-bit registers）中，通过比特观察器（Bit Observer）对处于多个时间和空间中的量子流比特进行多重采样观测。通过 hot-patching 来动态改变演化空间内量子流的演化路径，当需要对演化空间做更大规模的 deformation 变动时，RISC-V 将控制演化空间进行热重启，这一切都在瞬间内完成。这一系列操作使得 QuantaFlow 架构可以以最快速度演算各种神经网

络模型，比如 ResNet-50 或谷歌最新的 EfficientNet 等。保持算法可扩展性，是通用人工智能算力的基石。目前大部分 AI ASIC 加速器只能优化某一类网络模型，换别的网络类别就力不从心，性能大幅下降。

QuantaFlow 将计算中间结果反馈至量子流生成器（QuantaFlow Generator），进行循环往复，定制不同的演进计算，直到网络结束，获得结果，输出到低维解空间，可用于分类，物体识别，语义分割等 AI 任务等。量子并行计算加速的一个主要特点就是计算/演化空间非常大，而“解”空间相对非常小，而人工智能领域的问题，一般都是这类问题。

QuantaFlow 相对于基于经典图灵机编写的算法，有很大的优势。



如上图所示，经典图灵机可做三种基本操作：

- Observe（读取观察头对应位置的内容）
- Modify (改写观察头对应位置的内容)
- Control（将纸带向左，或向右移动）

经典计算机和经典算法的实现离不开这个基本模型，比如在每个机器的时钟周期，寄存器都要能被 CPU Observe/Modify. 而电子在硅基芯片中的从 A 点到 B 点的移动速度是有限的，计算频率/次数也将受限。基于以上架构的硅基实现，会更早地被纯物理极限所制约。而靠堆叠 CPU 核心，并不能完全解决问题，因为各个核心之间需要通讯、共享资源、维持 cache coherence，本质上也受电子移动速度的物理极限影响。

在硅基计算平台上实现类量子算法，只是 QuantaFlow 技术对极致速度追求的第一步，在技术实现层面，品奇实验室有着更多的细节创新，比如对芯片的设计编写，设计空间搜索，仿真，综合和布局等，均用高级语言加算法来参与完成，使用传统的 verilog + EDA tool chain 流程，由于传统工具局限，并不能完成很多关键性指标。

QUANTAFLOW 是面对算力困境的解决方案

QunataFlow 是整个行业面对算力困境时，品奇实验室(PQ Labs)给出的一份答案。这一个勇敢的尝试，却也收获了令人赞叹的成绩。通过全新的类量子算法架构设计，大幅提升了 AI 推理的速度，完成了 10 倍级的飞跃。

		ResNet-50 Model, 93% Accuracy, Batch=1	
NVIDIA	Tesla V100	1X speed (reference)	955 images/sec
pqlabs.ai	QuantaFlow	10X speed	10338 images/sec

值得指出的是 ResNet (2015)虽然还在大规模使用，但是是比较早期的网络模型。QuantaFlow 架构对较新的网络模型,比如 MobileNet (2017), EfficientNet (2019) 会有更好的加速比，具体数据将会 2020 年上半年发布。而 GPU 架构对新型网络普遍存在“访存墙”瓶颈，性能反而下降，不能适应新型神经网络模型的演进。QuantaFlow 架构恰好是为新型网络应运而生。