研究现状

（加一段对冯诺依曼瓶颈的叙述？）

国内外对智能计算芯片与系统的研究可以分为三类。

第一类研究被神经科学所启发，主要目的是构建能够高精确度地模拟大规模生物神经系统结构的软硬件系统，从而更好地理解大脑的结构以及工作过程。尽管这类研究的一个重要目的是为神经科学的发展提供支持，但是其中的许多项目开拓出了与传统的超级计算机集群不同的体系结构。这类研究中比较突出的是英国曼彻斯特大学的SpiNNaker [Furber2013]。

SpiNNaker使用ARM968处理器为基本运算单元，整个系统中使用了1036800个处理器，峰值处理速度可以达到每秒228兆Dhrystone指令，而能量消耗可以控制在90kW内。SpiNNaker使用软件模拟神经元，能够灵活地支持不同的神经元模型。

第二类研究从现有的深度学习网络结构出发，针对特定类型的神经网络以及特定的训练算法开发专用的硬件来加速深度学习中的学习或训练过程。近些年深度学习方法在相当多的领域中取得了优异的成果，但是其训练过程在传统CPU结构上可能要花费数周甚至数月的时间，因此大多数高端的深度学习应用都会使用GPU来加速这个过程。

然而随着深度学习涉及的数据量以及应用使用量的增大，CPU与GPU的混合结构也越来越难以满足功耗等方面的要求。为了解决这类问题，一些公司和研究机构将目光转到了新的体系结构上。这类研究中，国外具有代表性的是Google公司的TPU [Jouppi2017]，在国内则有中科院计算所开发的“寒武纪”芯片[Chen2014]。

TPU是Google专门为TensorFlow深度学习框架设计的一款处理器，以加速深度学习中的矩阵运算部分为主要设计思路。与同时期的CPU和GPU相比，TPU可以将每瓦特的性能提高30-80倍。

中科院计算所从针对神经网络的指令系统出发，开发了“寒武纪”芯片。“寒武纪”芯片每秒能处理160亿个神经元和2.56万亿个突触运算，可达到每秒512G浮点运算速度，可广泛适应各种智能处理应用。

第三类研究受神经系统结构启发，参考神经系统的结构来开发新形态的硬件系统。这类系统参考生物的神经系统来组织自身结构，并利用数字电路或模拟电路来实现神经系统中的组成成分，如神经元、突触和灰质等。

这类研究与第一类研究不同，因为其目的并不是为了模拟神经系统，而是希望从神经科学中得到启发来构建低功耗高性能的新型体系结构；同时，这类研究也不同于第二类，因为这类研究并没有针对特定的一种或一类神经网络模型去进行设计。这类研究中具有代表性的工作有IBM开发的TrueNorth [Merolla2014]，以及浙江大学的“达尔文”[Shen2016]。

TrueNorth选用了注重速度的简化LIF神经元模型，最大支持100万脉冲神经元和2.5亿突触。“达尔文”同样采用了LIF神经元模型，本身作为协处理器辅助完成计算任务，最大支持2048个神经元和约420万个突触，以及15种突触延迟。“达尔文”的工作频率为70MHz，虽然频率很低，但处理延迟并未显著增加，并在多种分类算法上取得了高准确率的结果。

[Chen2014] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam, "DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning", in Proceedings of the 19th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'14), 2014.

[Furber2013] Furber, Steve B., David R. Lester, Luis A. Plana, Jim D. Garside, Eustace Painkras, Steve Temple, and Andrew D. Brown. "Overview of the SpiNNaker system architecture." IEEE Transactions on Computers 62.12 (2013): 2454¨C2467.

[Jouppi2017] Jouppi et al. "In-Datacenter Performance Analysis of a Tensor Processing Unit." In Proceedings of the 44th Annual International Symposium on Computer Architecture, 1-12. Toronto, ON, Canada: ACM. 2017.

[Merolla2014] Merolla, Paul A., et al. "A million spiking-neuron integrated circuit with a scalable communication network and interface." Science 345.6197 (2014): 668–673.

[Shen2016] Shen, Juncheng, et al. "Darwin: a neuromorphic hardware co-processor based on Spiking Neural Networks." Science China Information Sciences59.2 (2016): 1¨C5.