**研究背景**

人工智能的发展可以追溯至1956年夏在达特茅斯举行的“用机器模拟人类智能”研讨会。经典人工智能的一条路线主张用逻辑建立人工智能的统一理论体系，将智能视为符号处理过程，采用形式逻辑来实现智能，故称为“符号主义”或“逻辑主义”。经典人工智能的另一条路线是人工神经网络，这一路线认为智能活动是有大量简单的单元通过复杂连接后并行运行的结果，就如同产生人脑智能的神经网络一样，因而被称为“连接主义”。人工智能的第三个流派是“进化主义”，也称为行为主义，其思想来源于控制论，认为智能不仅来源于机器内，也来源于机器所处的环境以及机器与环境的相互作用。第四个流派将着眼点从“制造”智能转移到了“习得”智能上，通过模拟人类学习的行为来获得新的知识，这一流派从已知数据中构建模型，并运用模型进行预测与分析，因此又被称为“贝叶斯派”或“统计主义”。

近年来，这四种流派呈现出交叉融合的趋势，并诞生了AlphaGo等喜人的成果。AlphaGo的设计中集合了四种经典人工智能的技术：（1）采用符号主义的经典算法蒙特卡洛树缩小搜索空间；（2）利用卷积神经网络学习围棋盘面；（3）通过自我博弈棋局进行强化学习，训练得出估计盘面形式的“估值网络”；（4）利用统计机器学习中的局部特征匹配与线性回归方法加快搜索速度。AlphaGo将经典人工智能中的多种方法进行了有机的结合，从而在2016年3月战胜了九段围棋选手。

以AlphaGo为代表的人工智能技术在蓬勃发展的同时也暴露出了传统冯诺依曼结构的不足之处。传统CPU以通用为目的而设计，在处理深度学习等专门问题时性能与能效性等方面无法达到最优，因此主流的人工智能系统框架大多使用CPU+GPU的混合结构，但是这种混合结构也同样面临着能效性的问题。使用专门集成电路ASIC（Application Specific Integrated Circuits）可以取得较好的能效性，然而ASIC的设计周期可以长达数月至一年，远远长于算法更迭的周期，因而出现了基于FPGA或CGRA等可编程结构的设计方案。总体而言，以深度学习为主要框架的人工智能技术对结构具有足够灵活性的高能效体系结构提出了要求。

现如今相当多的人工智能系统在专项任务中取得了令人瞩目的成果，如超过人眼识别率的AlexNet（需再确认）及各种语音识别系统。但是我们仍需注意到一点：这些人工智能系统都是针对专门任务设计的，而不是针对更广泛的智能行为而设计，例如一个针对视觉识别而设计的人工智能系统未必能在语音识别上取得较好的成绩。然而早在1979年，Vernon Mountcastle就发现大脑用相同的原则来处理视觉、听觉、触觉等行为。他发现大脑新皮层的所有区域都执行相同的基础操作，会产生不同分工的皮层仅仅是由于它们从不同的来源接收到信号输入。这一事实说明，为了使人工智能有更进一步的发展我们仍有必要探索新的智能计算体系结构。

从计算机科学的角度来看，针对智能计算体系结构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。

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。

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。

**研究现状**

这一部分将根据三类不同的研究方向对国内外智能计算体系结构的发展成果进行一个简单的总结。

第一类研究中有代表性的包括欧盟2011年启动的BrainScaleS[Brüderle2011]，HP公司联合Boston大学开发的Cog Ex Machina[Snider2011]和曼彻斯特大学的SpiNNaker系统[Furber2013]。

2005年，欧盟启动了FACETS（Fast Analog Computing with Emergent Transient States）项目，由海德堡大学牵头研制基于模拟混合信号的神经形态芯片。BrainScaleS（Brain-inspired multiscale computation in neuromorphic hybrid systems）在2011年启动，作为FACETS的延续项目，目标是研发大规模并行类脑计算机。BrainScaleS中包括了4500万个可编程的动态突触，使用FACETS项目中开发的晶圆级别硬件系统作为组成单元。2016年3月，欧盟人类大脑计划宣布将BrainScaleS系统通过互联网对外开放使用，以以支持神经微回路模拟以及在机器学习和认知计算中应用类脑原理的相关研究.

Cog Ex Machina是一款低功耗、全数字平台的硬件系统，利用多线程和张量数据表示实现了层级结构、并行处理等生物大脑的特征，建立了可以与模拟真实环境实时交互的大脑模型。Cog Ex Machina系统尚未达到生物规模的模拟能力，但是仍为研究者提供了一条成本相对较低地快速建模验证新算法的途径。

曼彻斯特大学开发的SpiNNaker系统基于ARM 968处理器研发，整个系统中共使用了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，每个核可以模拟1000个神经元，突触则使用SDRAM进行模拟。2016年3月，SpiNNaker与BrainScaleS系统一同通过互联网对外开放使用。

第二类研究由企业主导的情况较多，其中具有代表性的成果有谷歌公司开发的TPU [Jouppi2017]、微软公司的Branwave[Burger2017]以及中科院计算所开发的“寒武纪”芯片[Chen2014]。

TPU是谷歌公司专门为TensorFlow深度学习框架所设计的一款处理器，主要加速深度学习中的矩阵运算部分。TPU使用脉动阵列结构对矩阵运算进行加速，核心思想是保持矩阵运算单元的忙碌状态。与同时期的CPU和GPU相比，TPU可以将每瓦特的计算性能提高30-80倍。

微软公司在2017年的Hot Chips会议上初次揭露了自己的深度学习加速平台Brainwave。Brainwave平台主要面向实时AI应用，未来将会在微软的云平台Azure和Bing搜索服务中投入使用。Brainwave以深度神经网络DNN为主要加速对象，可以将DNN网络映射到一个FPGA池中，在Intel的14nm Stratix 10 FPGA上可以达到29.5 Teraflops的运算速度。

中科院计算所采取了不同的研究思路，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。

第三类研究的成果包括IBM开发的TrueNorth神经网络芯片[Merolla2014]、斯坦福大学开发的Neurogrid[Benjamin2014]、浙江大学的“达尔文”芯片[Shen2016]以及清华大学的“天机”系列芯片[Shi2015]。

TrueNorth基本结构由硬件神经元和神经元之间的脉冲连接组成，神经元模型使用注重速度的LIF模型。TrueNorth具有4096个处理器核，每个核中包含256个硬件神经元，总计可以模拟100万个神经元与2.56亿个突触。

Neurogrid基于数模混合电路构建，利用晶体管的亚阈值工作区来实现突触和神经元的特殊功能，用以实时模拟大尺度神经网络。Neurogrid可以以极低的功耗实时模拟16x256x256个神经元与数十亿个突触构成的神经网络系统。

“达尔文”使用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，并在多种分类算法中取得了高准确率的结果。

清华大学的“天机”芯片可以支持多种神经网络模型，曾受邀在电子器件国际顶级会议IDEM2015上做了特邀报告。“天机”单个芯片中包含6个核，每个核中包含神经元、权值网络、路由和参数管理等模块，运行频率为100MHz，可达到每秒153.6G定点运算速度。

**研究任务**

人工智能技术发展到今天，智能计算系统已经可以在很多专门任务中胜过人类，但智能计算系统在这些任务中体现出的智能距离人类的智能仍有距离。从人工智能进一步发展角度来看，我们需要新的智能计算体系结构来让智能技术取得进一步的突破；从人工智能技术对计算平台提出的挑战来看，我们也需要新的智能计算体系结构来补充现有体系结构的不足之处，以更好地支持现有的多种智能应用。

为了在智能计算体系结构的关键技术领域中取得突破，计算机系统、电子器件与材料、神经科学等多个领域需要进行密切的相互协作。为了达到这一目的，需要进行以下几方面的研究：

1. **智能计算体系结构的基础理论研究。**为了准确地定义智能计算体系结构，我们需要解决若干问题，包括但不限于：新型体系结构的基本组成单元是什么，系统应该支持何种程度的可编程度，如何通过学习或训练来重新构建系统等等。
2. **面向智能计算的芯片及新型系统架构研究。**从计算单元设计、芯片研发、系统架构改进与创新等方面展开研究。对新型芯片和系统结构的研究一方面要注重适应现有的智能计算架构，另一方面也要结合神经科学、材料科学、器件工艺等多种领域的发展，利用其新的技术成果构建功能完整的新型系统
3. **配套基础软件的开发。**充分利用智能计算系统需要配套基础软件的支持。配套软件应该包括对非计算机从业人员友好的编程开发环境，和具备可视化功能、数据分析功能的调试软件等利于软硬件协同开发和系统交互的工具。
4. **智能计算应用形式的研究。**需要对运行在智能计算系统上的应用形式进行进一步的研究，与智能计算体系结构的基础理论研究相结合，界定出最为适合智能计算系统的应用范围，以充分利用计算资源。
5. **量化评价标准与评测技术。**发展智能计算系统需要对评测指标和基准进行研究。新型体系结构与系统的研发离不开对系统性能的准确评估，这需要能够准确反映出智能计算系统特性的基准程序。评测标准不仅应涵盖吞吐量、运算速度等计算机系统性能指标，还应该包括神经科学相关的基准程序。

[Benjamin2014] Benjamin, Ben Varkey, et al. "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations." Proceedings of the IEEE 102.5 (2014): 699–716.

[Brüderle2011] Brüderle, Daniel, et al. "A comprehensive workflow for general-purpose neural modeling with highly configurable neuromorphic hardware systems." Biological cybernetics 104.4-5 (2011): 263–296.

[Burger2017] Doug Burger, "Microsoft unveils Project Brainwave for real-time AI", August 22, 2017. Accessed at http://www.microsoft.com/en-us/research/blog/microsoft-unveils-project-brainwave/

[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.

[Shi2015] L. Shi et al., "Development of a neuromorphic computing system," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, 2015, pp. 4.3.1-4.3.4.

[Snider2011] G. Snider, R. Amerson, D. Carter, H. Abdalla, M. S. Qureshi, J. Léveillé, et al., "From synapses to circuitry: Using memristive memory to explore the electronic brain," Computer, vol. 44, p. 21, 2011.