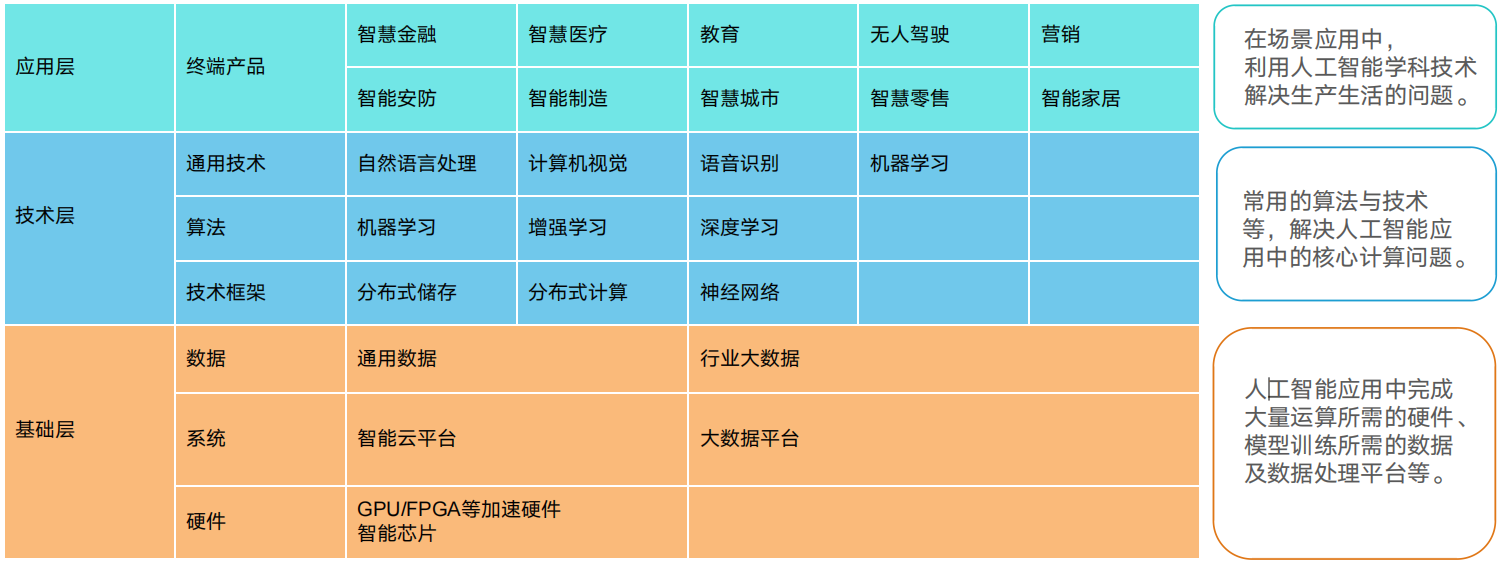
10.31阅读中国人工智能芯片行业研究报告，查找相关资料扩充。

一、中国人工智能芯片行业发展现状

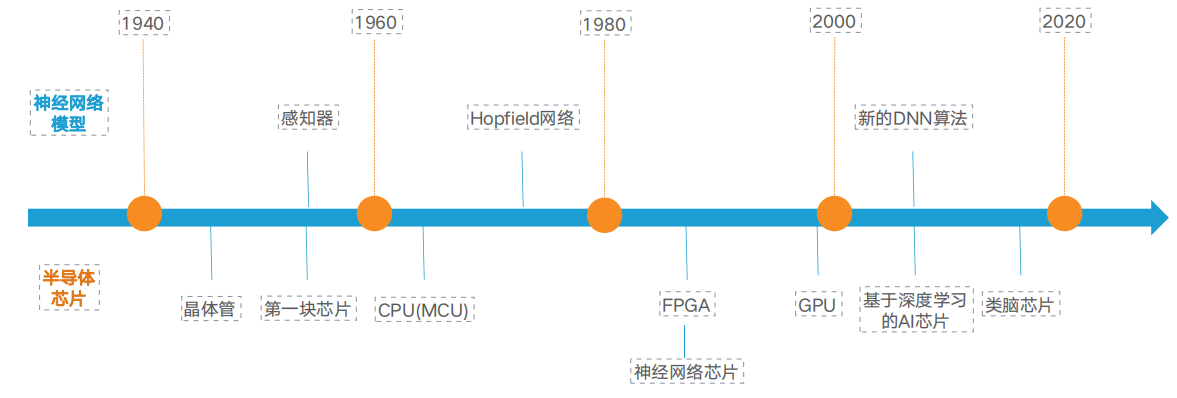
狭义的AI芯片即人工智能算法做特殊加速设计的四种主流芯片分别是GPU、ASIC、FPGA、类脑芯片以及系统级AI芯片技术。

首先介绍三个层次：基础层，技术层，应用层。



然后是AI芯片的发展历程：模仿人脑建立的模型和算法与半导体芯片发展交替进行。AI芯片的发展主要依赖两个领域：第一个是模仿人脑建立的数学模型与算法，第二个是半导体集成电路即芯片。人工智能于芯片的发展分为三个阶段：第一阶段由于芯片算力不足，神经网络算法未能落地；第二阶段芯片算力提升，但仍无法满足神经网络算法需求；第三阶段，GPU和新架构的AI芯片促进了人工智能的落地。目前，随着第三代神经网络的出现，弥合了神经科学与机器学习之间的壁垒，AI芯片正在向更接近人脑的方向发展。

下图是人工智能与半导体芯片的发展进程对照

 AI芯片的实现包含软件和硬件两个方面。既需研究高效率的智能算法，同时要研究如何将这些算法结合在半导体硅片上，形成最终的产品、

11.1 继续阅读研究报告

二、人工智能芯片解读

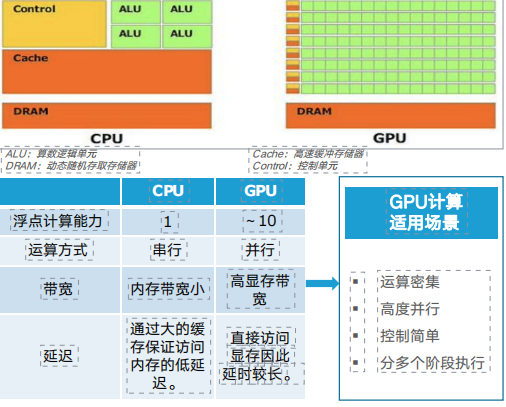
1技术层面

AI芯片根据其技术架构，可分为GPU、FPGA、ASIC及类脑芯片，同时CPU可执行通用AI计算。

以下是查找资料补充：

GPU是单指令、多数据处理，采用数量众多的计算单元和超长的流水线，主要处理图像领域的运算加速。GPU是不能单独使用的，它只是处理大数据计算时的能手，必须由CPU进行调用，下达指令才能工作。但CPU可单独作用，处理复杂的逻辑运算和不同的数据类型，但当需要处理大数据计算时，则可调用GPU进行并行计算。

Gpu图形处理器最初是一种专门用于图像处理的微处理器，因其计算能力 超过cpu的数十倍，成为通用计算机和超级计算机的主要处理器。下图是CPU与GPU的对比。



Gpu的开发环境常用的有四种：CG，CUDA，ATIStream，OpenCL。

* CG(CforGraphics)：为GPU编程设计的高级绘制语言，由NVIDIA和微软联合开发，微软版本叫HLSL，CG是NVIDIA版本。
* CUDA（ComputeUnifiedDeviceArchitecture，统一计算架构）：由NVIDIA所推出的一种集成技术，是首次可以利用GPU作为C-编译器的开发环境。可以兼容OpenCL或者自家的C-编译器。
* ATIStream：AMD针对旗下图形处理器（GPU）所推出的通用并行计算技术。利用这种技术可以充分发挥AMDGPU的并行运算能力，用于对软件进行加速或进行大型的科学运算。
* OpenCL(OpenComputingLanguage，开放计算语言)：为异构平台编写程序的框架，此异构平台可由CPU，GPU或其他类型的处理器组成。

对于未来GPU体系的发展趋势，有以下几种猜想：增加计算资源密度、提高存储体系性能和功能、增强通信能力和可靠性、降低功耗。

2半定制化芯片（FPGA）

FPGA:可编程逻辑门阵列，是一种“可重构”芯片，具有模块化和规则化的架构，主要包含可编程逻辑模块、片上储存器及用于连接逻辑模块的克重购互连层次结构。在较低的功耗下达到GFLOPS数量级的算力使之成为并行实现人工神经网络的替代方案。

FPGA适用于多指令，单数据流的分析，与GPU相反，因此常用于预测阶段，如云端。FPGA是用硬件实现软件算法,因此在实现复杂算法方面有一定的难度，缺点是价格比较高。与 GPU 不同，FPGA 同时拥有硬件流水线并行和数据并行处理能力,适用于以硬件流水线方式处理一条数据，且整数运算性能更高，因此常用于深度学习算法中的推断阶段。不过FPGA 通过硬件的配置实现软件算法，因此在实现复杂算法方面有一定的难度。将FPGA 和 CPU 对比可以发现两个特点,一是 FPGA 没有内存和控制所带来的存储和读取部 分速度更快,二是 FPGA 没有读取指令操作,所以功耗更低。劣势是价格比较高、编程复杂、整体运算能力不是很高。目前国内的AI 芯片公司如深鉴科技就提供基于 FPGA 的解决方案。

11.2阅读报告

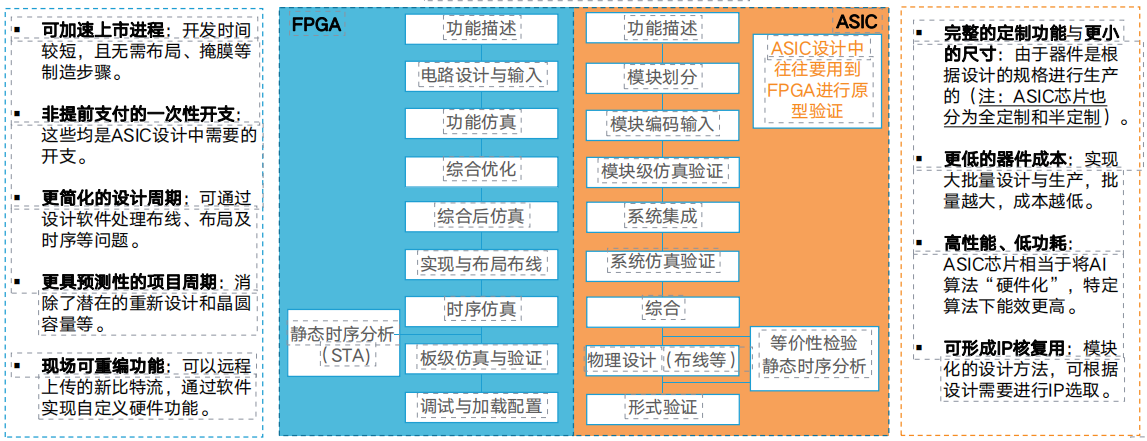
二、人工智能芯片解读

3全定制化芯片（ASIC）

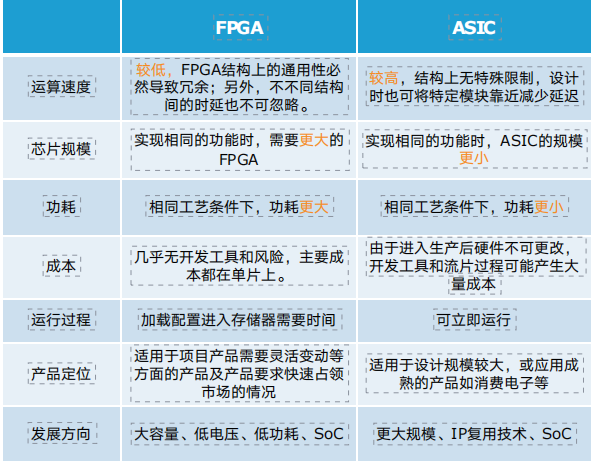
ASIC是指应特定用户要求和特定电子系统的需要而设计、制造的集成电路。ASIC从性能、能效、成本均极大的超越了标准芯片，非常适合AI计算场景，是当前大部分AI初创公司开发的目标产品。除了不能扩展以外,在功耗、可靠性、体积方面都有优势，尤其在高性能、低功耗的移动设备端。

定制的特性有助于提高 ASIC 的性能功耗比，缺点是电路设计需要定制，相对开发周期长，功能难以扩展。但在功耗、可靠性、集成度等方面都有优势，尤其在要求高性能、低功耗的移动应用端体现明显。谷歌的 TPU、 寒武纪的 GPU，地平线的 BPU都属于 ASIC芯片。谷歌的 TPU比 CPU和 GPU的方案快 30 至 80 倍，与 CPU和 GPU相比，TPU把控制电路进行了简化，因此减少了芯片的面积，降低了功耗。

下图是ASIC与FPGA芯片设计流程与设计优势对比



以及ASIC&FPGA总体对比



4类脑芯片

CPU/GPU/GFPGA/ASIC及SoC是目前用的较多的AI芯片，此类AI芯片大多是基于深度学习，也就是深度神经网络（DNN），以并行方式进行计算的芯片，此类AI芯片又被称为深度学习加速器。

如今，模仿大脑结构的芯片具有更高的效率和更低的功耗，这类基于神经形态计算，也就是脉冲神经网络（SNN)的芯片为类脑芯片。目前，部分企业产品已进入小批量试用阶段,类脑芯片最快将于2023年成熟，能效比有望较当前芯片提高2-3个数量级。

类脑芯片架构是一款模拟人脑的神经网络模型的新型芯片编程架构，这一系统可以模拟人脑功能进行感知方式、行为方式和思维方式。有人说，ASIC是人工智能芯片的一个主要发展方向，但真正的人工智能芯片未来发展的方向是类脑芯片。类脑芯片研究是非常艰难的，IBM、高通、英特尔等公司的芯片策略都是用硬件来模仿人脑的神经突触。

类脑芯片的硬件实现方式：忆阻器：在紧凑性、可靠性、耐用性、存储器保留期限、可编程状态和能效等方面的特性，有利用成为电子突触器件。自旋电子器件：其不易失性、可塑性及振荡和随机行为，允许创建模仿生物突触和神经元关键特征的组件。光子器件：利用半导体放大器中使用的非线性光学增益介质，实现神经形态计算所需的功能。电化学器件：具有高精度、现行和对称的电导效应，低开关能量和高扩展性，使用于SNN的内置定时机制。二维材料：包括过渡金属二硫族化合物、石墨烯等，可开发成为人工突触。

下图为DNN与SNN的对比



介绍完四类芯片后，总结四类芯片的优缺点以及应用场景。



若根据其在网络中的位置可以分为云端AI芯片、边缘及终端AI芯片；根据其在实践中的目标，可分为训练（training）芯片和推理（inference）芯片。云端主要部署训练芯片和推理芯片，承担训练和推理任务，具体指智能数据分析、模型训练任务和部分对传输带宽要求比高的推理任务；边缘和终端主要部署推理芯片，承担推理任务，需要独立完成数据收集、环境感知、人机交互及部分推理决策控制任务。

11.3阅读报告

5系统级芯片

在手机、可穿戴设备等端设备中，很少有独立的芯片，AI加速将由SoC上的一个IP实现。

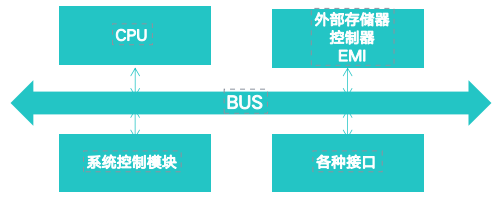
SoC(System-on-chip，片上系统）作为ASIC设计方法学中的新技术，始于20世纪90年代中期，是以嵌入式系统为核心，以IP复用技术为基础，集软、硬件于一体的集成芯片。在一个芯片上实现信号的传输、存储、处理和I/O等功能，包含嵌入软件及整个系统的全部内容。由于高集成效能，SoC已经成为微电子芯片发展的必然趋势。

SoC芯片制造流程



SoC具有的优势：降低耗电量：SoC多采用内部讯号的传输，大幅降低功耗。减少体积：数颗IC整合为SoC,有效缩小电路板上占用的面积。丰富系统功能：可整合更多的功能元件和组件，丰富系统功能。提高速度：内部信号传输距离缩短，信号传输效率提升。节省成本：IP复用，有效减少研发成本，降低研发时间。

SoC整体架构如图



现在用于深度学习的AI芯片（包括CPU、GPU、FPGA、ASIC）为了实现深度学习的庞大乘积累加运算和并行计算的高性能，芯片面积越做越大，带来了成本和散热等问题。AI芯片软件编程的成熟度、芯片的安全，神经网络的稳定性等问题也未能得到很好的解决，因此，在现有基础上进行改进和完善此类AI芯片仍是当前主要的研究方向。

最终，AI芯片将近一步提高智能，向着更接近人脑的高度智能方向不断发展，并向着边缘逐步移动以获得更低的能耗、更接近人脑、更靠近边缘的方向。

AI硬件加速技术已经逐渐走向成熟。未来可能更多的创新会来自电路和器件级技术的结合，比如存内计算，类脑计算；或者是针对特殊的计算模式或者新模型，比如稀疏化计算和近似计算，对图网络的加速；或者是针对数据而不是模型的特征来优化架构。

同时，如果算法不发生大的变化，按照现在AI加速的主要方法和半导体技术发展的趋势，或将在不远的将来达到数字电路的极限（约1到10TFlops/W），往后则要靠近似计算，模拟计算，甚至是材料或基础研究上的创新。

存内计算：核心问题：传统冯诺伊曼架构中，计算与内存是分离的单元，内存主要使用的DRAM方案性能提升速度远远慢于处理器速度，造成了阻碍性能提升的“内存墙”，直接在存储内做计算可有效解决。实现方法 1）改动存储模块电路：优势是容易和现有工艺进行集成，缺点是带来的性能提升有限；2）引入新的存储器件，实现在存储阵列内完成计算。

模拟计算：核心问题：传统模拟架构通过模数/数模转换器将模拟信号与数字表示形式进行相互转换，带来信号损耗、功率消耗和时延。实现方法：在AI芯片中使用模拟计算技术，将深度学习算法运算放在模拟域内完成，提高能效。

量子计算：核心问题：AI计算对大算力的需求。实现方法：完全新型的计算模式，理论模型为图灵机。从计算效率上，由于量子力学叠加性，配合量子力学演化的并行性，处理速度远超传统计算机，提供更强算力。

11.4阅读报告

1. 应用层面

1应用概况：算力向边缘侧移动，逐渐专注于特殊场景的优化

随着技术成熟化，AI芯片的应用场景除了在云端及大数据中心，也会随着算力逐渐向边缘端移动，部署于智能家居、智能制造、智慧金融等域；同时还将随着智能产品种类日渐丰富，部署于智能手机、安防摄像头、及自动驾驶汽车等智能终端，智能产品种类也日趋丰富。未来，AI计算将无处不在。

下图为AI芯片应用领域总结



当前云端仍是AI的中心，需更高性能计算芯片以满足市场需求。大多数AI训练和推理工作负载都发生在公共云和私有云中，云仍是AI的中心。在对隐私、网络安全和低延迟的需求推动下，云端出现了在网关、设备和传感器上执行AI训练和推理工作负载的现象，更高性能的计算芯片及新的AI学习架构将是解决这些问题的关键。

互联网是云端算力需求较旺盛产业，因此除传统芯片企业、芯片设计企业等参与者外，互联网公司纷纷入局AI芯片产业，投资或自研云端AI芯片。