创新技术与企业服务研究中心

寒武纪 (688256.SH) 买入（首次评级）

公司深度研究

市场价格（人民币）： 元

市场数据(人民币)

沪深 300 指数 4698

从端芯片到云系统的一站式 AI 龙头

公司基本情况(人民币)

项目	2018	2019	2020E	2021E	2022E	2023E
摊薄每股收益(元)	(0.11)	(3.27)	(1.28)	(0.95)	(0.66)	0.81
EPS 增长率 (%)	-89%	2772%	61%	25%	31%	223%
每股营收(元)	0.33	1.23	1.52	2.07	3.00	5.00
市盈率（倍）	NA	NA	NA	NA	NA	80
市销率（倍）	198	52	42	31	21	13
净资产收益率	-8%	-27%	-8%	-6%	-5%	5%

来源：公司年报、国金证券研究所

投资逻辑

- **一大潜力市场：**目前 GPU，FPGA+CPU 仍主导云端的深度学习训练和推理市场，但通用/专用 AI 芯片，因性能，成本及耗能优势将渗透部分传统型 AI 芯片在云，边，端的市场。我们预估全球 AI 云端半导体市场于 2019-2024 年复合成长率应有 36%，边缘运算及终端芯片市场于同期应有 55%增长，远超过全球市场的 7%，整体占全球份额从 2019 年的 3%到 2024 年的 11%。
- **云，边，端，软件一站式方案的核心竞争力：**寒武纪的核心技术是同时具备最底层的芯片设计，指令集及驱动器，基础系统软件，及加速卡，并使用相同的自研指令集，让开发者以各类算法完整云、边、端生态系平台的跨越。
- **寒武纪 AI 芯片对比 GPU 具有高效能，低成本，低耗能核心竞争力：**1. 低精度定点运算的优势；2. AI 推理的优势； 3. 算法演进可透过软硬件修改的优势； 4. 运算单元芯片面积较小，却有 2 倍以上高效能，50%低耗能优势。
- **弯道超车的机会：**美国商务部于 2020 年 5 月 15 日宣布限制海思在使用美国半导体 EDA 及设备技术来生产半导体，必须取得执照，但我们认为海思昇腾 AI 要申请得到执照有难度，这给了寒武纪一个绝佳机会来弯道超车的机会。

投资建议

- 在高毛利云及边缘运算端 AI 芯片新产品加持下，寒武纪未来五年营收将大幅增长 CAGR 超过 50%，2023 年扭亏转盈，给予“买入”评级。为了扩大事业版图，寒武纪在 IPO 发 4010 万新股，每股 64.39 元人民币，募集 25.8 亿的资金来补充现金流及用在云端训练及推理，边缘端 AI 芯片及系统软件的开发。

估值

- 首家有中科院支持的 AI 芯片公司，不但毛利将维持在 65%以上，又享受科创板溢价，比较可比新兴科技公司，平均 P/S 区间将达 40-60 倍，我们目前用 2022 年的 3.0 元每股营收给予 50 倍 P/S 给估值，一年的目标价为 150 元。

风险提示

- 终端 AI 处理器 IP 业务减少的风险，智能计算集群系统事业的风险，同业竞争的风险，现金流短缺的风险，进入实体清单的风险。

张纯 分析师 SAC 执业编号：S1130519100004
zhang_chun@gjzq.com.cn

郑弼禹 分析师 SAC 执业编号：S1130520010001
zhengbiyu@gjzq.com.cn

投资要件

■ **关键假设**

1. **云端及边缘运算 AI 芯片及加速卡是两大高毛利增长动能**：我们估计未来五年这两项业务将有 80-100%复合增长率的贡献，营收占比从 2019 年的 18%继续扩大。主要客户是关联方的中科曙光，并扩大到非关联方的江苏恒瑞通智能科技，浪潮，联想及北京金山云网络技术。
2. **智能计算集群系统事业不确定**：公司智能计算集群系统方面的在手订单包括横琴先进智能计算平台(二期) 的第二批供货硬件设备，授权软件，合同金额仅剩下 1.86 亿元，而上半年营收贡献连 20 万都没有，除非我们看到在下半年在手订单大幅回流，今年此业务营收贡献可能不到 60%，甚至连 50%占比都可能有问题。
3. **管理费用中的员工股权激励及研发费用的高低，决定亏损是否持续**：以 2019 年为例，员工股权激励（股份支付）费用高达 9.44 亿，而 5.43 亿的研发费用也是超过当年度营收，这些偏高的管理及研发费用，造成公司 2019 年营业亏损，也将决定亏损是否持续。

■ **我们区别于市场的观点**

市场的观点是寒武纪营收及获利会大幅增长；但我们的观点是寒武纪要投入大量研发，先建立庞大的产品，设计，制程工艺，客户壁垒，短中期获利不易。

■ **股价上涨的催化因素**

7nm 思元 290 云端训练芯片的推出（2021 年），5nm 新产品的研发进度，智能计算集群系统事业在手订单的回流，管理及研发费用的控制成效，海外重点客户的建立，及 AI 芯片定点运算的突破。

■ **估值和目标价格**

因为缺乏短，中期获利，用市销率 P/S 来评估寒武纪，较为适当，寒武纪是国内首家上市的 AI 芯片公司，又有中科院的支持及与政府补助，50%以上营收 CAGR 及 65%以上毛利率比大多数新兴科技公司高两倍以上，又享受科创板的溢价，所以其市销率应该比可比公司的 20 倍高出甚多，在闭锁期结束之前，我们认为公司平均 P/S 区间将达 40-60 倍，目前用 2022 年的 3.0 元每股营收给予 50 倍 P/S 的估值，并给予买入评级，一年的目标价为 150 元。

■ **投资风险**

终端 AI 处理器 IP 业务的减少会影响毛利率；智能计算集群系统事业要是在手订单没有回流，今年公司营收衰退可期待；很多同业有更多的研发资源投入，寒武纪将面临更大的价格竞争风险；在客户中科曙光，竞争者海思，依图，旷视，商汤，云从科技等 AI 算法公司，及科大讯飞，云天励飞等 AI 平台公司陆续进入美国商务部工业安全局的实体清单后，我们必须考量寒武纪被列入的风险。

内容目录

图表目录

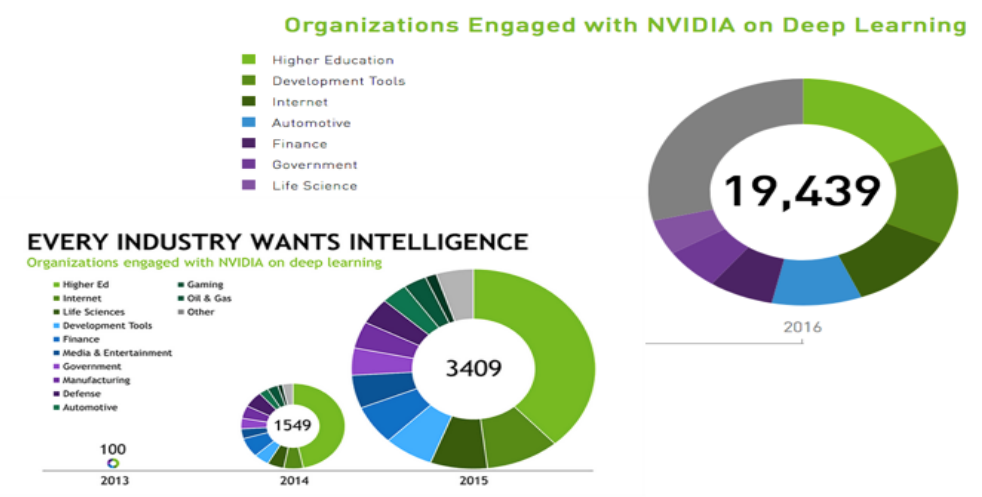
一、一大潜力市场

人工智能 AI 顾名思义就是想用高速计算机运算模式来模拟人脑的认知及推理，尤其是在收集大量原始数据后，再通过高速计算机利用特殊 AI 算法来训练 AI 的认知能力，其中包括视觉（图像，视频），听觉（语言，声音），嗅觉，当然还有味觉的酸甜苦辣，当 AI 高速计算机的认知能力训练完成后，推理算法才能帮助 AI 高速计算机进行，反应，推理，决定。

人工智能平台是工具,不是应用

人工智能平台（包括芯片，模组，软件）在一般人看起来像是一种新型应用，但在我们看来人工智能芯片在整合软硬件后将成为各种物联网应用的提升效能工具平台，这就像我们常用的微软 Office 软件，微软 Office 软件是我们在办公室应付各种应用的生财工具，因此人工智能平台除了被广泛利用在云端大数据的深度学习训练和推断外，我们认为人工智能平台也将出现在各式各样的应用端的边缘运算及终端，从去年英伟达公布的数字来看，早在 2016 年，公司就累计了 7 大领域（高等教育，发展工具，互联网，自驾车，金融，政府，生命科学）及 19,439 客户使用其深度学习的服务工具，配合软件和之前在云端大数据的深度学习训练和推断的数据成果库，来达到帮助使用者或取代使用者来执行更佳的智能判断推理。

图表 1：人工智能的多样性



来源：英伟达，国金证券研究所

虽然目前人工智能芯片仍多是传统型芯片，并以昂贵的图形处理器 (GPU)，或以现场可编程门阵列芯片配合中央处理器 (FPGA+CPU) 为主，来用在云端数据中心的深度学习训练和推理，但通用 / 专用型 AI 芯片，也就是张量处理器或特定用途集成电路 (ASIC)，主要是针对具体应用场景，固定算法及相同模型的 AI 将在样式类似，数量庞大的云，边缘运算及终端所需推理及训练设备遍地开花，及逐步渗透部分传统型 AI 芯片在云端，边缘运算，及终端的市场，成为人工智能芯片未来的成长动能，我们预估全球人工智能云端半导体市场于 2019-2024 年复合成长率应有 36%，边缘运算及设备端半导体市场于 2019-2024 年复合成长率应有 55% (请参考图表 2)，远超过全球半导体市场在同时间的复合成长率的 7%，整体约占全球半导体市场的份额从 2019 年的 3% 到 2024 年的 11%。

图表 2：人工智能云，边缘运算，终端半导体及行业市场营收预估

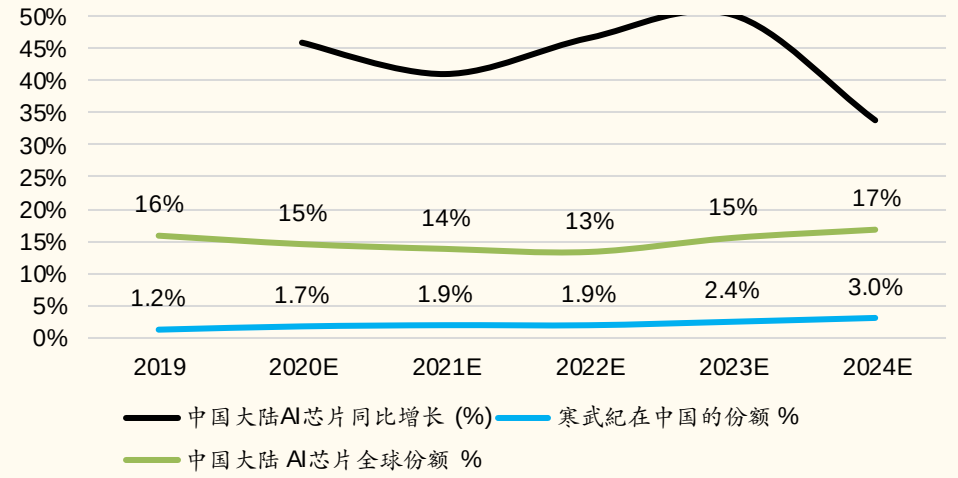
	2019E	2020E	2021E	2022E	2023E	2024E	CAGR
全球半导体市场 (US$bn)	410	410	472	495	535	588	7%
全球半导体市场 (同比)	-13%	0%	15%	5%	8%	11%	
AI 半导体 (US$bn)	11	18	26	40	51	63	42%

AI 半导体 (同比)	116%	59%	49%	52%	29%	24%	
AI IC 佔全球 IC 份额 (%)	3%	4%	6%	8%	10%	11%	
云端 AI 半导体 (US$bn)	8.1	12.3	17.6	23.5	30.0	37.0	36%
云端 AI 半导体 (同比)	69%	52%	43%	34%	28%	23%	
边缘及设备端 IC (US$bn)	2.9	5.2	8.4	16.0	21.0	26.0	55%
边缘及设备端 AI IC (同比)	190%	174%	62%	90%	31%	24%	
云端 AI 半导体占比 (%)	74%	70%	68%	59%	59%	59%	
边缘及设备端 IC 占比 (%)	26%	30%	32%	41%	41%	41%	
AI 行业销售额 (US$bn)	192	261	361	490	657	854	35%
AI 行业销售额 (同比)	34%	36%	38%	36%	34%	30%	

来源：Statista, Tractica，Frost & Sullivan, 国金证券研究所

如果把前瞻产业研究院对中国 AI 芯片市场五年 45%复合增长率及在 2024 年达到 785 亿人民币的规模预测拿来比较，中国大陆 AI 芯片市场的全球份额将从 2020 年的 15%，提升到 2024 年的 18%，而寒武纪在中国的份额，估计将从 2019 年的 1.2%，提升到 2024 年的 3.0%或更高。我们认为前瞻产业研究院预测的中国大陆 AI 芯片市场，应该有包括英伟达 Nvidia，赛灵思 Xilinx，英特尔 Intel AI 芯片在中国的销售金额。

图表 3：国内 AI 芯片同比增长率，全球份额，及寒武纪份额



来源：寒武纪招股说明书，Tractica, 前瞻产业研究院，国金证券研究所

二、两大核心竞争力

不同于华为海思虽然在人工智能云，边缘，终端芯片有着领先的地位，但除了靠华为内部集团的大量采购外，还有无冲突客户的采用，很多华为在 5G 通讯，服务器，手机及安防的直接竞争者，还是会以第三方公司像寒武纪设计的芯片为考量。比特大陆之前也投入大量资金发展云和端的人工智能芯片，但在吴忌寒与詹克团理念不合分家后，我们估计其人工智能芯片发展将因为资金短缺而从云转向人工智能在安防端及边缘运算的应用。但就寒武纪而言，公司始终在人工智能领域耕耘，亦步亦趋， 遥遥领先其他新进。我们以为寒武纪的两大核心竞争力为整体解决方案及定点算法的坚持。

1.从云，边缘运算，终端，及系统软件的一站式解决方案

寒武纪的主营业务是应用于各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的设计，为客户提供丰富的芯片产品与系统软件解决方案。公司的主要产品包括终端智能处理器 IP、云端训练及推理智能芯片及加速卡、边缘智能芯片及加速卡以及与上述产品配套的基础系统软件平台。自 2016 年 3 月成立以来，公司先后推出了用于终端场景的寒武纪 1A、1H、1M 系列芯片、还有基于台积电 16 纳米制程工艺的思元 100 云端推理和思元 270 云端推理训练芯片 及其 AI 加速卡系列产品，云端训练 AI 290 芯片及加速卡，以及基于思元 220 芯片的边缘智能加速卡。其中,寒武纪 1A、寒武纪 1H 应用于华为的麒麟手机芯片中，已集成于超过 1 亿台智能手机及其他智能终端设备中；思元系列产品也已应用于浪潮、联想，中科曙光，滴滴，及海康威视等多家服务器及其相关厂商的产品中，边缘智能芯片及加速卡的发布标志着公司已形成全面覆盖云端、边缘端和终端场景的系列化智能芯片产品布局，并广泛应用于手机，IOT、数据中心、云计算等诸多场景。对于已经建立庞大软（CUDA，应用软件 NGC）硬件(终端：Jatson Nano，TX2 Series；边缘运算端：Jet Xavier NX，Jet AGX Xavier Series，EGX；云端：Tesla，DGX A100，DGX-1/Station，HGX，NGC) 生态系的英伟达，及华为海思，在使用台积电的 7 纳米制程工艺后，在设计上靠着庞大资源也胜寒武纪一筹，**但美国商务部工业安全局 5/15/2020 宣布进一步限制华为海思在使用美国半导体设计软件 EDA 来设计半导体以及利用晶圆代工所使用的美国半导体设备来生产半导体，必须获得执照，但我们认为，手机及机顶盒芯片有机会获得执照，但海思的安防，昇腾 AI，鲲鹏伺服器 CPU，5G 基地站 ASIC 要申请得到执照可能有困难，这给了寒武纪一个绝佳机会来弯道超车的机会。**

我们认为有核心技术的人工智能通用芯片公司必须同时具备芯片（最底层的硬件物质载体包含高维张量 / 向量 / 传统算术逻辑计算部件），韧体的指令集及驱动器，基础系统软件（来管理，调用，控制智能芯片来运作），加速卡设计及测试的能力来完成完整的生态系，寒武纪使用相同的自研指令集与处理器架构，共用相同的基础系统软件平台，实现了云、边、端通用生态的跨越。而开发者可以研发各类人工智能算法、实现各类人工智能程序，最终实现机器视觉、 语音处理、自然语言处理以及推荐系统等多样化的人工智能功能。

图表 4：寒武纪的AI芯片技术

序号	技术大类名称	在主营业务及主要产品中的应用和贡献情况	专利或其他技术保护措施	成熟程度	技术来源
1	智能处理器微架构	公司迄今已自主研发了三代智能处理器微架构，是国内外在该技术方向积累最深厚的企业之一。公司在云端、边缘端、终端三条产品线的所有智能芯片和处理器核IP产品均基于自研处理器架构研制。	已取得专利 29 项（其中境外专利 8 项），PCT 专利申请 67 项	成熟稳定	自主研发
2	智能处理器指令集	指令集是处理器芯片生态的基石。公司是国际上最早开展智能处理器指令集研发的少数几家企业之一，迄今已自主研发了三代商用智能处理器指令集，形成了体系完整、功能完备、高度灵活的智能芯片指令集专利群。公司在云端、边缘端、终端三条产品线的所有智能芯片和处理器IP产品以及基础系统软件均构建于自研的 MLU 指令集基础之上。	已取得专利2项(其中境外专利1项)，PCT 专利申请 26 项	成熟稳定	自主研发
3	SoC 芯片设计	公司已掌握复杂 SoC 设计的一系列关键技术，有力支撑了云端大型 SoC 芯片（思元 100、思元 270 和思元 290）和边缘端中型 SoC 芯片（思元 220）的研发。	已取得专利 1 项，PCT 专利申请 8 项	成熟稳定	自主研发
4	处理器芯片功能验证	公司拥有成熟先进的处理器和 SoC 芯片功能验证平台，确保了智能处理器和 SoC 芯片逻辑设计按时高质量交付，有效保障了多款芯片产品的一次性流片成功。	已取得专利 2 项	成熟稳定	自主研发
5	先进工艺物理设计	公司已掌握 7nm 等先进工艺下开展复杂芯片物理设计的一系列关键技术，已将其成功应用于思元 100、思元 220、思元 270 及最新的思元 290 等多款芯片的物理设计中。	非专利技术	成熟稳定	自主研发
6	芯片封装设计与量产测试	应用于公司云端、边缘端和终端不同品类芯片产品的封装设计与量产测试过程，有效支撑了公司处理器芯片的研发。	相关专利正在申请中	成熟稳定	自主研发

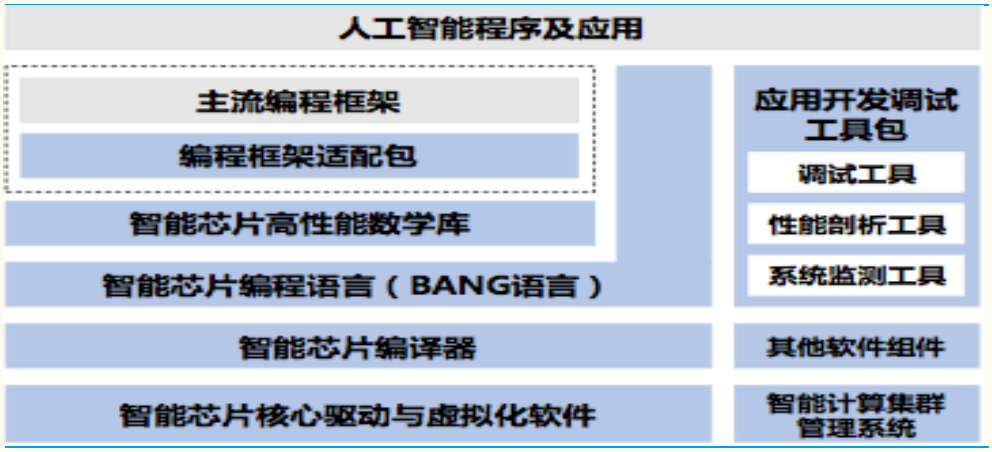
来源：寒武纪招股说明书，国金证券研究所

- **指令集的通用性：**针对特定场景乃至特定智能算法的加速芯片，这类芯片针对某个算法实施的硬件化开发，一般不具备指令集或指令集较简单。但寒武纪研发的通用型 AI 芯片，必须具备灵活的指令集，来覆盖人工智能领域多样化的应用场景（如视觉、语音、自然语言理解、传统机器学习等）。寒武纪智能芯片的设计思想是通过人工智能算法的计算特征和访存特征来降低数据搬运的延迟和功耗，支持多个处理器核之间高效并发协作，并针对性地设计更适用于智能算法的数百条处理器基本指令集，与处理器架构配合实现在人工智能领域内灵活通用的设计目标，不仅需要考虑当前各类智能算法的特点，也需要对智能算法未来发展的趋势进行预判，从而设计出完备高效的智能处理器指令集；通过高维张量、向量、逻辑指令等之间的灵活组合来覆盖对多样化的智能算法，实现人工智能领域内的通用性。举例来说，我们可以定义硬件的动作，00 是做加法(Add)，01 是做减法(Sub)，10 是做读取资料(Load)，11 是做存储资料（Store），而内部有两个指令暂存器 Instruction register a，register b，这样软件想做 C=A+B，则会变成读取 register a 的 A，读取 register b 的 B，加 register b 的 B 到 register a 中，再存储到 register a 的 C 中，而软件可以用高阶语言，让程式员只需要写 C=A+B，再透过编译器（Complier），转化成上面读取/加/存储的程式码，同理，AI 处理器的硬件也定义了一些指令集，其实就是如上面的一些简单基本动作，可以以软件的方式，组合出各式各样的功能，之后透过编译器，可以转换为更细碎的指令组合，而硬件就会依照指令的排列顺序，一个动作一个动作的完成。

- **处理器架构：**寒武纪智能处理器的主功能包含高维张量计算部件、向量计算部件、传统算术逻辑计算部件，分别用于处理各类智能算法的不同类型操作。其中高维张量计算部件可对智能算法中核心运算(如卷积运算)进行高效处理，提升整个处理器的能效。而向量运算部件与算术逻辑计算部件(尤其后者)则具有更强的灵活性，可对智能算法中频次不高且高维张量无法支持的运算(如分支跳转等)实现全面覆盖，有力保障了处理器架构的通用性。

- **基础系统软件 Cambricon Neuware (包含软件开发工具链等)：**无须繁琐的移植即可让同一人工智能应用程序便捷高效地运行在公司云，边，端系列化芯片与处理器产品之上。在 Cambricon Neuware 的支持下，程序员可实现跨云边端硬件平台的人工智能应用开发，以“一处开发、处处运行”的模式大幅提升人工智能应用在不同硬件平台的开发效率和部署速度，同时也使云，边，端异构硬件资源的统一管理、调度和协同计算成为可能。Cambricon Neuware 在开发应用时，用户既可以基于 TensorFlow，PyTorch，Caffe，MXNet 等主流编程框架接口编写应用代码，也可以使用公司预先优化的智能芯片高性能数学库对编程框架算子进行扩展或直接编写代码；用户同样可以通过智能芯片编程语言(BANG 语言)对算子进行扩展或直接编写代码；智能芯片编译器可以完成 BANG 语言到 MLU 指令的编译，并在智能芯片核心驱动的支持下使其高效地运行于公司各款芯片产品之上。在开发过程中，用户还可以通过应用开发调试工具包所提供的调试工具、性能剖析工具和系统监测 工具等高效地进行应用程序的功能调试和性能调优。此外，Cambricon Neuware 也可以通过智能芯片虚拟化软件为云计算与数据中心场景提供关键支撑。

图表 5：寒武纪 Neuware 软件架构



来源：寒武纪招股说明书，国金证券研究所

2.通用型智能芯片对比 GPU 有高效能，低耗电核心优势

寒武纪虽然在使用台积电的制程工艺上，明显落后于海思最高档 AI 昇腾 910 的 7nm＋EUV，AMD 超威 Radeon Instinct MI50 及 Nvidia 英伟达最新推出的 A100 的 7nm，但寒武纪 16nm 的思元 270 主要对标产品是英伟达价值 2,500-2,600 美元 12nm 的 Tesla T4 而不是上万美元的 7nm A100，思元 270 可支持 INT16/INT8/INT4 等多种定点精度计算，INT16 的峰值性能为 64TOPS[1]（64 万亿次运算），INT8 为 128TOPS，INT4 为 256TOPS。对比 Tesla T4，FP16 的理论峰值性能为 65 TFLOPS，INT8 为 130 TOPS，INT4 为 260 TOPS。思元 270 的功耗为 75w，与 Tesla T4 类似。但所谓的理论峰值在实测后通常有一定缩水。据阿里云早期核心技术研发人员曾经表示[1]，T4 在实测过程中，75w 功耗维持不了多久就降一半频率，而思元 270 就能维持相当的频率。() 我们估计在相同的效能下持续运作，T4 的耗能是思元 270 的 2 倍以上，在思元 270 的性能参数展示上，可以看到寒武纪有意强调其定点计算性能方面的优势，这应该是寒武纪在 AI 领域的低精度定点运算有突破，因为低精度计算的速度和能耗比优势一直受到业界密切关注。而寒武纪 7 纳米的思元 290，跟英伟达 V100 比较应该也具备 2 倍以上高效能，50%低耗能的优势，但

[1] 来源：寒武纪二代芯片发布在即，独家揭秘如何挑战英伟达. ChainNews

此低成本优势可能要等到寒武纪直接下单台积电，且大量出货达到经济规模才会展现出来（目前芯片出货量还是非常小，应该连 20k 都不到，是使用 Broadcom/Avago 博通的设计服务）。

除此之外，把 GPU 用在深度学习 AI 有几个缺点，第一个是深度学习包含训练和推理两个计算环节，GPU 在深度学习算法并行训练上非常高效，但只能对于一张输入图像进行推理，并行度的优势不能完全发挥；第二个是硬件结构固定不具备可编程性。深度学习算法还未完全稳定，若深度学习算法发生大的变化，GPU 无法像 FPGA 一样可以灵活的配置硬件结构，也无法像通用 AI 能够针对特殊应用来更改芯片设计；第三个是运算单元芯片面积过大，功耗及成本都较高。

图表 6：寒武纪 vs. 英伟达云端芯片加速卡价格差异比率

公司产品型号①	英伟达产品型号[1]②	价差③=(①-②)/②
思元100	Tesla P4加速卡	-17%~4%
思元270	Tesla T4加速卡	-10%~24%

来源：寒武纪问询函回复，国金证券研究所

定点运算与浮点运算是计算机计算中最为常用的两种运算表示法，其差异就体现在定点和浮点上，加减乘除运算都是一样的。定点表示法，即所有位都表示个位数字，小数点固定；而浮点表示法，则分成两部分，阶码和尾数，尾数就是数字部分，阶码表示乘幂的大小，也就是小数点位置。所以浮点数在做运算的时候，除了对尾数做加减乘除，还要处理小数点位置。基于两种不同的运算表示法规则，导致面对同样长度的定点和浮点运算，浮点计算模式更为复杂，需要消耗数倍多的功耗及更大的芯片去做运算。但浮点运算又有其不可取代性。首先，定点表示法运算虽然直观，但是固定的小数点位置决定了固定位数的整数部分和小数部分，不利于同时表达特别大的数或者特别小的数。而浮点运算的小数点位置可以移动，运算时不用考虑超出某种数据格式的范围，所以科学计算法一般都使用浮点。此外，具体到使用 GPU 做训练，业界通常更倾向于浮点运算单元，主要是因为只有浮点运算才能记录和捕捉到训练时很小的增量。由于训练的部分模块对精度要求比较高，所以通常必须是高精度的浮点运算，比如 FP32 / FP64 （32/64 位元的单精度浮点运算）才能搞定。虽然浮点运算相比定点运算在功耗、计算速度、性价比等方面都不占优势，但截止目前，浮点计算在云端的训练场景中仍占着主导地位，并且以高精度运算为主。

那么，如何在不增加芯片面积和功耗的前提下，如何大幅提升芯片做训练的运算能力就成为云端训练芯片的主要研究课题之一。参考计算过程相对简单的推理计算，目前该领域的 AI 芯片多采用通用 AI，ASIC 或低精度浮点运算 GPU，但面对计算过程更为复杂的训练计算，业界一直在尝试是否可能用性价比更高的定点运算器实现。如何以全部的定点单元（比如 INT8）代替浮点单元，或者以低精度定点单元配合少量的高精度浮点计算单元（比如 FP32）做更多的训练任务，目的是达到定点计算的快速度，同时实现接近高精度浮点计算的精度。目前看来低精度训练确实未必要是浮点数，只要能把数域表达好，0 附近的小量表达好，什么样的数据表示都可以。

总之，我们判断寒武纪之所以能够大幅度提升低精度训练阶段的计算功耗比，很有可能是大量采用以定点为主的低精度运算，但要能够成功的切入数据中心的 AI 高精度训练及推理市场，寒武纪除了要发展高精度的浮点运算外，一套完备成熟的软件生态也是其核心竞争力的重要体现，所以从 2016 年起，寒武纪逐步推出了 NeuWare 软件工具链，该平台终端和云端产品均支持，可以实现对 TensorFlow、Caffe 和 MXnet 的 API 兼容，同时提供寒武纪专门的高性能库。英伟达之所以能够在云端训练领域成为绝对主流，其 CUDA 软件生态的基础功不可没，所以目前 80% 以上的云端加速器是采用英伟达 GPU，而 AMD 的 GPU/CPU 及赛灵思的 FPGA 占据非常小的份额。

图表 7：人工智能云端推理及训练芯片在不同定点，浮点，精度，峰值比较

	思元 220	思元 100	思元 270-S4	思元 290	昇腾 910	Radeon Instinct MI50	Tesla T4	V100 GPU	A100 GPU
供应商	寒武纪	寒武纪	寒武纪	寒武纪	海思	超威	英伟达	英伟达	英伟达
应用	边缘运算	云端推理	云端推理训练	云端训练	云端训练推理	云端	云端推理训练	云端推理训练	云端推理训练
制程工艺	TSMC 16nm，94.8mm2	TSMC 16nm，326.5mm2	TSMC 16nm，369.6mm2	TSMC 7nm	TSMC 7nm+ EUV，456mm2	TSMC 7nm	TSMC 12nm	TSMC 210 亿晶体管 12nm CoWoS，815mm2	TSMC 540 亿晶体管 7nm CoWoS，826mm2
定点 INT8 1/4 精度理论峰值	8 TOPS	32 TOPS	128TOPS	512 TOPS	512TOPS	53 TOPS	130TOPS	60 TOPS	625 TOPS
定点 INT8 1/4 精度稀疏模式	32 TOPS	128TOPS	512TOPS						1248 TOPS
定点 INT16 半精度理论峰值	4 TOPS		64TOPS						
定点 INT16 半精度稀疏模式	16 TOPS		256TOPS						
浮点 FP16 半精度理论峰值		16 TFLOPS		256TFLOPS	256TFLOPS	26.5 TFLOPS	65TFLOPS	125 TFLOPS	310 TFLOPS
浮点 FP16 半精度稀疏模式		64 TFLOPS		1024TFLOPS	1024TFLOPS				620 TFLOPS
浮点 FP32 单精度理论峰值						13.3 TFLOPS	8.1TFLOPS	16 TFLOPS	160 TFLOPS
浮点 FP32 单精度稀疏模式									320 TFLOPS
浮点 FP64 双精度理论峰值		不支援	不支援	不支援		6.6 TFLOPS		8 TFLOPS	20 TFLOPS
最大耗电量	8.25W	75W/110W	70W	300W	310W	300W	70W	300W	400W

来源：寒武纪招股说明书及问询函回复，海思，超威，英伟达，国金证券研究所

三、三个挑战

作为一个国内最早发 AI 专用芯片公司，寒武纪不但要面对客户（华为海思）上下游整合及调用庞大的手机芯片设计及 IP 资源到人工智能芯片的挑战，以及英伟达在通用 AI GPU 加速芯片在制程工艺及 CUDA 软件的优势，寒武纪只能利用少数的资源及现金流，逐步从 IP 设计的点，到芯片设计，加速卡设计的线，一直跨足到系统整合的面，但未来 10 年，我们认为寒武纪将面对三个主要挑战，寒武纪如何克服这些挑战，将是我们建议客户持续关注的重点：1. 寒武纪如何靠着有限的现金流从市场较小的云端推理发展到云端训练？2. 寒武纪在一年内从 IP 及芯片设计公司，转变成智能计算集群系统公司，如何解决智能计算集群系统单一客户横琴新区管理委员会商务局占比过高及行业低毛利，高竞争，与客户如浪潮，联想，中科曙光争食市场的问题？3. 寒武纪如何缩短与海思升腾及英伟达 V100/A100 GPU 加速卡在 IP，设计，制程工艺，软件生态系，定点 INT8 1/4 精度理论峰值的技术差距？

1.如何从云端推理发展到云端训练

不同于华为海思及英伟达靠着过去强大本业的获利及现金流的积累，可以投入大量的研发资源，在人工智能领域持续的设计流片，及走到最后的系统整合。而寒武纪在持续亏损（2019 亏损高达 12 亿人民币，净亏损率达 266%）数年后，还要大力追赶，思元 270 虽然已经暂时掌握了云端推理用定点低精度的运算技术，但要顺利在明年推出云端训练用并具有浮点高精度运算的思元 290，寒武纪必须不断烧钱雇用人才，投入庞大研发及流片费用，这次 IPO 定增将筹措 28 亿人民币，当然对在手 5 亿人民币现金不到的寒武纪而言，具有稳定财务的作用，但要是年度亏损持续扩大而研发支出不放缓，两年后，我们不排除寒武纪将卷土重来市场融资，继续摊薄现有股权结构。

2. 扩大战场跟客户抢生意－一个横琴新区采购案占营收近 6 成

不同于海思跨入云端推理及训练人工智能加速卡，可以有华为这个超大客户在服务器行业的支持，但寒武纪的主营业务是应用于各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的设计，为客户提供丰富的芯片产品与系统软件解决方案，这其中包括终端智能处理器 IP、云端智能推理芯片及加速卡、边缘智能推理芯片及加速卡以及与上述产品配套的基础系统软件平台，并广泛应用于手机，IOT、数据中心、云计算等诸多场景；云端智能推理芯片及加速卡也已应用到国内主流服务器厂商（如浪潮）的产品中，并已实现量产出货；边缘智能推理芯片及加速卡的发布标志着公司已形成全面覆盖云端，边缘端，终端的系列化智能芯片产品布局。

但寒武纪为了进一步扩大其营收，于 2019 年首度跨入了智能计算集群系统，于一年内成为寒武纪营收占比 67%的主力事业群，就是间接跟直接的销售整机智能加速系统。从另一角度来看（参考图表），横琴新区管理委员会商务局于 2019 年就占了寒武纪近 61%的营收，这其中 14%营收是寒武纪销售 4000 块思元 100 加速卡给中科曙光，然后再出货给横琴新区管理委员会商务局，另外 47%营收是直接销售 1,300 台智能服务器及 48 台并行存储系统给横琴新区管理委员会商务局，这其中安装了 5,200 块思元 270 加速卡。

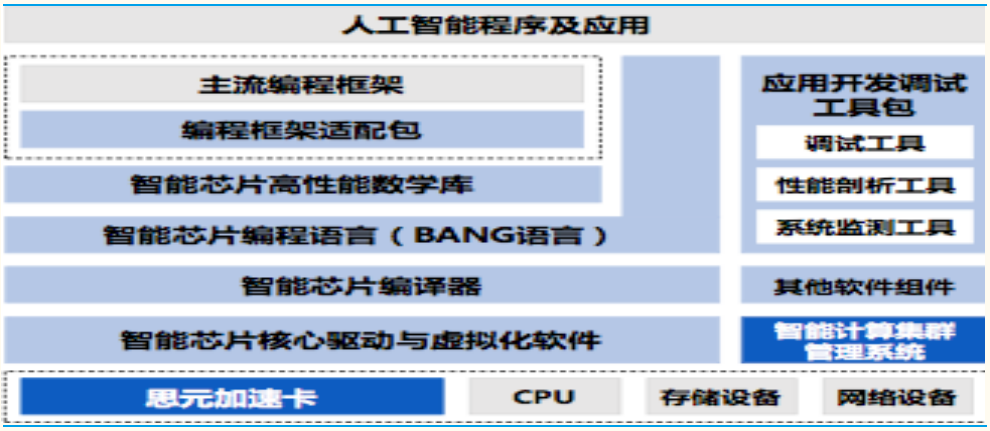
图表 8：横琴先进智能平台及其他 AI 集群系统采购细目整理

终端客户	西安沣东仪享	上海脑科学	横琴新区管理委员会商务局		
采购项目		类脑研究中心	横琴先进智能计算平台		
工期			一期	二期第一批	二期第二批
智能计算集群系统供应商	寒武纪	寒武纪	中科曙光	寒武纪	寒武纪
集群系统服务器供应商	浪潮信息	苏州超集信息	中科曙光	中科可控	中科可控
思元加速卡产品	思元 270	思元 100, 270	思元 100	思元 270	思元 270
思元加速卡数量	1,800.0	240.0	4,000	5,200	N/A
思元加速卡单价测算 US$	4,991	2,674	2,310	3,506	N/A
x86 服务器数量	225.0	30.0	1,000	1,300	N/A
每台服务器加速卡数	8	8	4	4	N/A
并行存储系统	N/A	N/A	N/A	48	N/A
寒武纪 Neuware 软件	包括	包括	包括	包括	包括
对寒武纪营收贡献 CNY$mn	81.1	3.6	63.8	207.1	185.7
对寒武纪营收占比 (%)	18.3%	0.8%	14.4%	46.6%	30.6%
时期	2019	2019	2019	2019	2020

来源：寒武纪招股说明书及回复函，国金证券研究所

虽然这事业群就跟使用英伟达 A100 芯片各种加速卡的服务器系统类似，但寒武纪总是冒着跟其系统客户抢生意的争议。这就像海思虽然在特定用途的人工智能云，边缘，终端芯片有着领先的地位，但除了靠华为内部集团的大量采购外（Atlas 智能计算集群系统），海思的产品要被其他智能计算集群系统客户的采用，就有相当的难度，很多华为在 5G 通讯，服务器，手机的直接竞争者，还是会以第三方公司像寒武纪设计的芯片及加速卡为考量，但当寒武纪也跨入其客户将积极发展的人工智能系统领域，跟客户如浪潮，联想，中科曙光抢生意的争议就逐步浮现。

图表 9：寒武纪智能计算集群系统的硬件，软件架构



来源：寒武纪，国金证券研究所

3.如何缩短设计，制程，软件生态系的差距

从寒武纪产品在图表 6 与海思及英伟达产品比较表中，我们很清楚的看到寒武纪在单及双精度浮点运算设计及制程工艺（16nm vs. 7nm）都有很大的改善空间，刚才提到寒武纪若要继续迈入云端训练用浮点高精度运算，就必须不断烧钱雇用人才，投入庞大研发及流片费用，进行 7nm 或更先进的设计。举例而言，寒武纪若要加快单及双精度浮点运算速度，决定从现在的 16nm 跳到 7nm 设计，但光是一个新产品设计流片光掩膜成本就会从 4-5 百万美元，增加到 1100 万美元， 这会让一个新产品流片成本暴增一倍以上，所以我们合理推断，寒武纪必须要等到其年度营业收入超过 10 亿人民币后或利用这次的上市公开发行新股融资来扩大其 5/7 纳米新产品研发动能。举例而言，公司 2019 年的测试化验加工费（主要系研发所用流片费）达 1.24 亿人民币，就占了当年营收的 28%。

而就制程工艺来看，台积电之前曾经公布过以 7nm 制程工艺设计的芯片比 16nm 设计的芯片，在耗能上减少达 60%，芯片速度增加 30%， 每固定单位的芯片面积可增加 233%的晶体管 (Transistors) ，简单来说，不管寒武纪在设计方面多有竞争力，如果长期要跟海思及英伟达在云端人工智能训练及推理芯片，加速卡产品的竞争，设计及制程工艺的演进到 7 纳米，甚至 5 纳米，就是不得不为的决定，所以寒武纪预期在 2021 年顺利推出与海思升腾 910 同等级的 7nm 云端 AI 训练芯片思元 290，要是能顺利量产，就显得非常重要。最后，在软件生态方面,英伟达凭借长久以来的经验积累以及产品推广已形成了较为完善的 CUDA 软件生态系，用户对其产品接受度较高，形成 了一定的用户习惯，而寒武纪的基础系统软件平台 Cambricon Neuware 的生态完善程度与英伟达相比仍有一定差距。

图表 10：光掩膜节点升级成本变化

	百万人民币	百万美金	% 节点升级成本
55 纳米	4.0	0.6	
40 纳米	7.0	1.0	75%
28 纳米	15.0	2.1	114%
16 纳米	32.0	4.6	113%
12 纳米	40.0	5.7	25%
10 纳米	59.5	8.5	49%
7/7+ 纳米	77.0	11.0	29%

来源：嘉楠耘智，国金证券研究所

图表 11：台积电制程工艺演进的效能变化比率

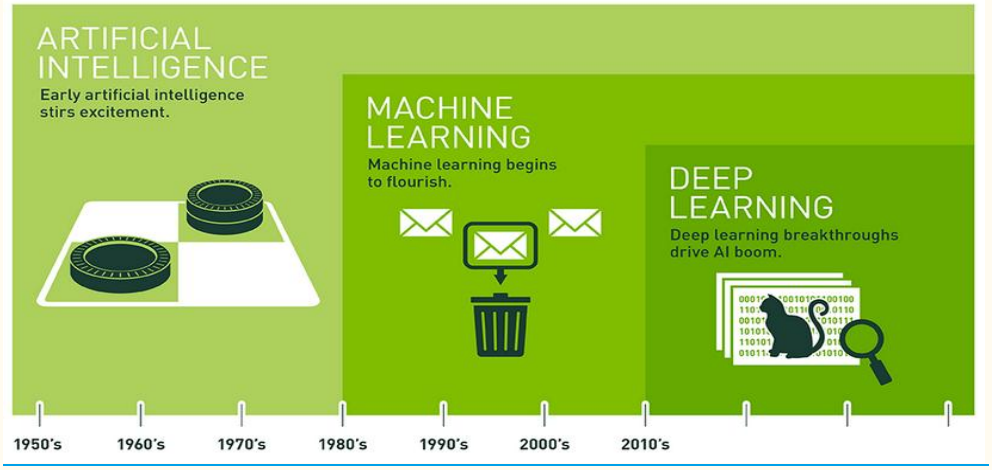
	16nmFF+ to 7nm	7nm to 5nm	5nm to 3nm
耗能减少比率	60%	20%	20-25%
速度增加比率	30%	15%	10-15%
晶体管密度增加比率	233%	45%	70%

来源：台积电，国金证券研究所

四、三种主流人工智能演算法

最早的人工智能出现及运用在 1950－1980 年代，接着转换到 1980-2010 年机器学习，从 2010 年以后，随着各种演算法 CNNs, RNNs, DNNs 等图影像视觉学习，辨识，推理的普及，让深入人工智能深入学习的突飞猛进。深度学习是人工智能和机器学习的一个子集，它使用多层人工神经网络在诸如对象检测，语音识别，语言翻译等任务中提供最先进的准确性。深度学习与传统的机器学习技术的不同之处在于，它们可以自动学习图像，视频或文本等数据的表示，而无需引入手工编码规则或人类领域知识。它们高度灵活的架构可以直接从原始数据中学习，并在提供更多数据时提高其预测准确性。人工智能的深度学习近来已经取得的许多突破，例如谷歌 DeepMind 的 AlphaGo 及更强大的 AlphaZero 陆续在围棋，西洋棋类比赛夺冠，谷歌 Waymo，英伟达的 Xavier，及 Intel/Mobileye 的 Eye 4/5 自动驾驶汽车解决方案，亚马逊的 Alexa，谷歌的 Google Assistant，苹果 Siri，微软的 Cortana，及三星的 Bixby 智能语音助手等等。借助加速的深度学习框架，研究人员和数据科学家可以显着加快深度学习培训，可以从数天或数周的学习缩短到数小时。当模型可以部署时，开发人员可以依靠人工智能芯片加速的推理平台来实现云，边缘运算设备或自动驾驶汽车，为大多数计算密集型深度神经网络提供高性能，低延迟的推理。

图表 12：人工智能技术工艺的演化

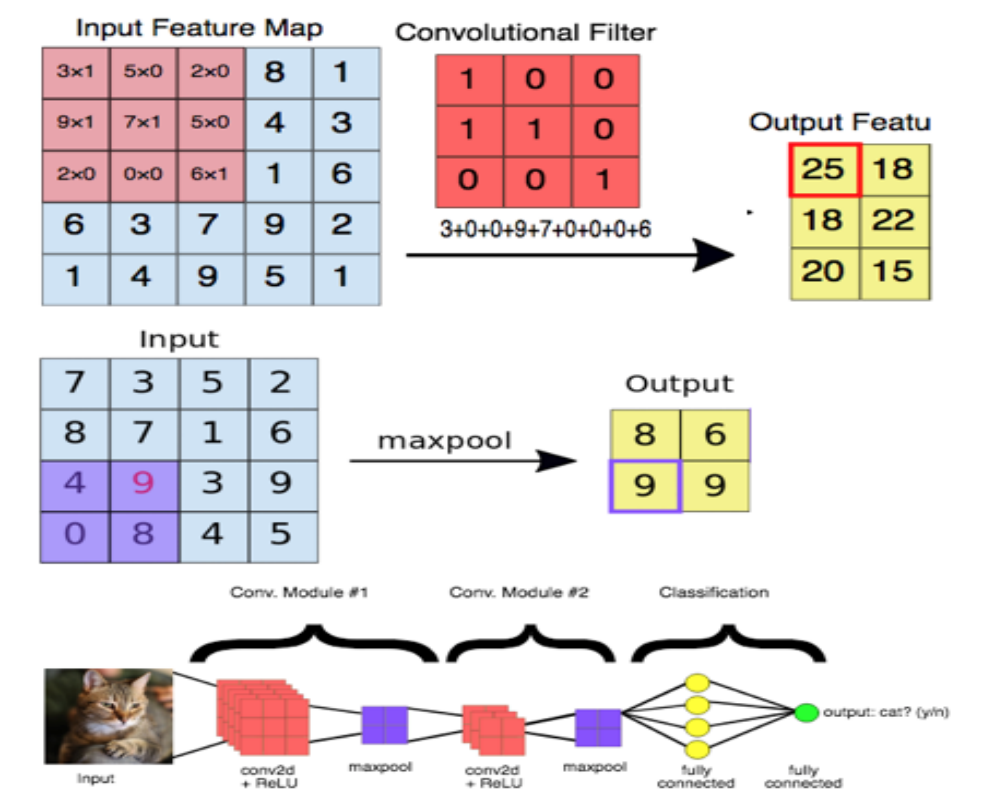


来源：英伟达，国金证券研究所

- **卷积神经网络 CNNs (Convolutional Neural Networks)**：卷积神经网络（CNN）是建立在模拟人类的视觉系统，并透过图影像分类模型的突破，也将是，主要来自于发现可以用于逐步提取图影像内容的更高和更高级别的表示。 CNN 是将图像的原始像素数据作为输入，并“学习”如何提取这些特征，并最终推断它们构成的对象。首先，CNN 接收输入特征图：三维矩阵，其中前两个维度的大小对应于图像的长度和宽度（以像素为单位），第三维的大小为 3（对应于彩色图像的 3 个通道：红色，绿色和蓝色）。CNN 包括一堆模块，每个模块执行三个操作。举例而言，卷积将 3x3 过滤贴图的 9 个条件（0，1）套用（先乘后求和以获得单个值）在 5x5 输入特征贴图的 9 个像素特征上，而得出 3x3 全新的卷积输出特征贴图。在每次卷积操作之后，会采用最大池演算法（Max pooling），CNN 对卷积特征贴图进行下采样（以节省处理时间），同时仍保留最关键的特征信息，最大池化是要从特征贴图上滑动并提取指定大小的图块 (2x2)，对于每个图块，最

大值将输出到新的特征贴图，并丢弃所有其他值。在卷积神经网络的末端是一个或多个完全连接的层，完全连接的层将一层中的每个神经元连接到另一层中的每个神经元。 它原则上与多层感知器神经网络（multi-layer perceptron neural network (MLP) 类似，他们的工作是根据卷积提取的特征进行分类，CNN 可以包含更多或更少数量的卷积模块，以及更多或更少的完全连接层，工程师经常试验要找出能够为他们的模型产生最佳结果的配置。总之，CNN 专门于图影像处理如自动驾驶汽车，安防，人脸辨识，及疾病图像辨识解决方案。

图表 13：卷积输入及输出特征贴图及最大池

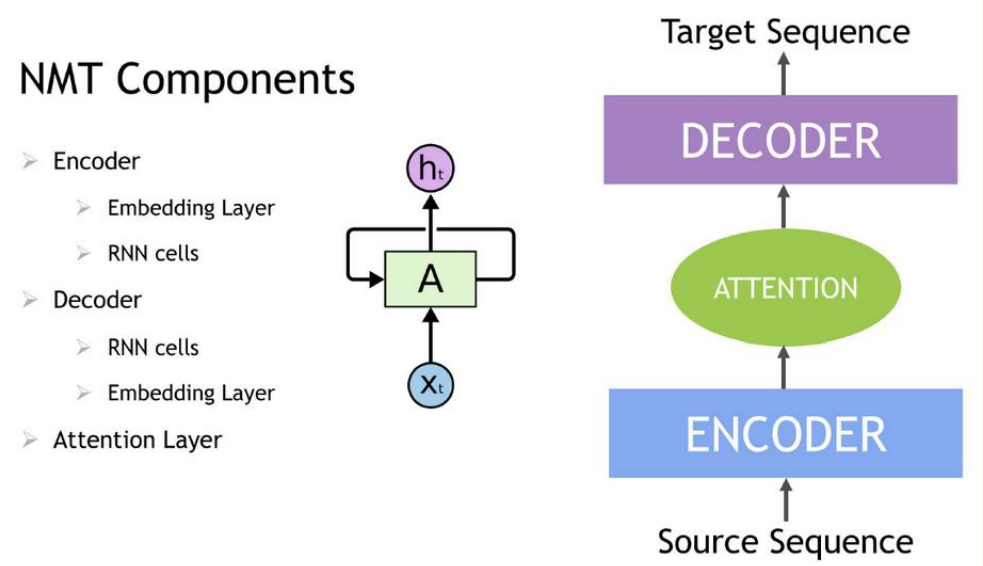


来源：谷歌，国金证券研究所

- **循环神经网络 RNNs (Recurrent Neural Network)**: RNN 是一类人工听觉及说话的神经网络，具有记忆或反馈回路，可以更好地识别数据中的模式。RNN 是常规人工神经网络的扩展，它增加了将神经网络的隐藏层送回自身的连接 - 这些被称为循环连接。循环连接提供了一个循环网络，不仅可以看到它提供的当前数据样本，还可以看到它以前的隐藏状态。具有反馈回路的循环网络可以被视为神经网络的多个副本，其中一个的输出用作下一个的输入。与传统的神经网络不同，循环网络使用他们对过去事件的理解来处理输入向量，而不是每次都从头开始。当正在处理数据序列以进行分类决策或回归估计时，RNN 特别有用，循环神经网络通常用于解决与时间序列数据相关的任务。不同于 CNN 专门于图影像处理，循环神经网络的应用包括自然语言处理，语音识别，机器翻译，字符级语言建模，图像分类，图像字幕，股票预测和金融工程。机器翻译是指使用机器将一种语言的源序列（句子，段落，文档）翻译成相应的目标序列或另一种语言的矢量。由于一个源句可以以许多不同的方式翻译，因此翻译基本上是一对多的，并且翻译功能被建模为有条件而非确定性。 在神经机器翻译（NMT）中，我们让神经网络学习如何从数据而不是从一组设计规则进行翻译。 由于我们处理时间序列数据，其中语境的上下文和顺序很重要，因此 NMT 的首选网络是递循环神经网络。 可以使用称为注意的技术来增强 NMT，这有助于模型将其焦点转移到输入的重要部分并改进预测过程。举两 RNN 的例子，为了跟踪你的自助餐厅主菜的哪一天，每周在同一天运行同一菜的严格时间表。如周一的汉堡包，周二的咖喱饭，周三的披萨，周四的生鱼片寿司

和周五的意大利面。使用 RNN，如果输出“生鱼片寿司”被反馈到网络中以确定星期五的菜肴，那么 RNN 将知道序列中的下一个主菜是意大利面（因为它已经知道有订单而周四的菜刚刚发生，所以星期五的菜是下一个）。另一个例子是如果我跑了 10 英里，需要喝一杯什么？人类可以根据过去的经验想出如何填补空白。由于 RNN 的记忆功能，可以预测接下来会发生什么，因为它可能有足够的训练记忆，类似这样的句子以“水”结束以完成答案。

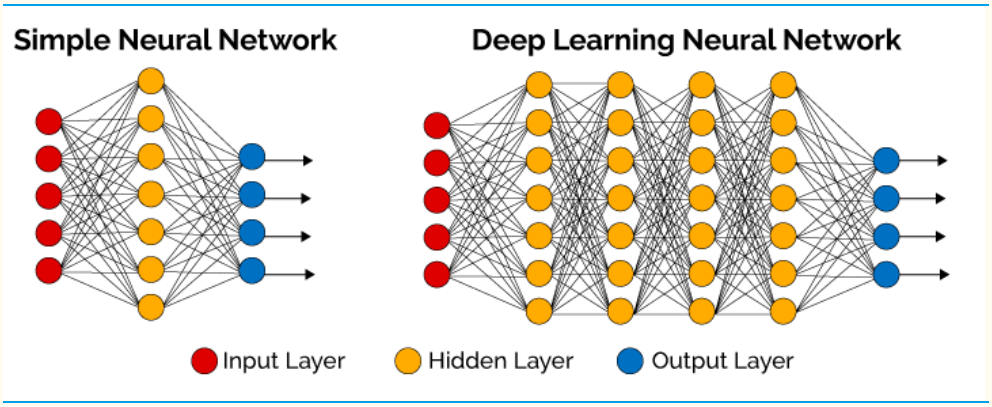
图表 14：循环神经机器翻译



来源：谷歌，国金证券研究所

- **深度神经网络 DNNs (Deep Neural Network)**：DNN 在视觉，语言理解和语音识别等领域取得了关键突破。为了实现高精度，需要大量数据和以后的计算能力来训练这些网络，但这些也带来了新的挑战。特别是 DNN 可能容易受到分类中的对抗性示例，强化学习中遗忘任务，生成建模中的模式崩溃的影响以及过长的运算时间。为了构建更好，更强大的基于 DNN 的系统，是能否有效地确定两个神经网络学习的表示何时相同？我们看到的两个具体应用是比较不同网络学习的表示，并解释 DNN 中隐藏层所学习的表示。设置的关键是将 DNN 中的每个神经元解释为激活向量，神经元的激活矢量是它在输入数据上产生的标量输出。例如，对于 50 个输入图像，DNN 中的神经元将输出 50 个标量值，编码它对每个输入的响应量。然后，这 50 个标量值构成神经元的激活矢量。因为深度神经网路的规模（即层数和每层的节点数），学习率，初始权重等众多参数都需要考虑。扫描所有参数由于时间代价的原因并不可行，因而小批次训练（微型配料），即将多个训练样本组合进行训练而不是每次只使用一个样本进行训练，被用于加速模型训练。而最显著的速度提升来自 GPU，因为矩阵和向量计算非常适合使用 GPU 实现。但使用大规模集群进行深度神经网路训练仍然存在困难，因而深度神经网路在训练并列化方面仍有提升的空间。

图表 15：深度神经网络

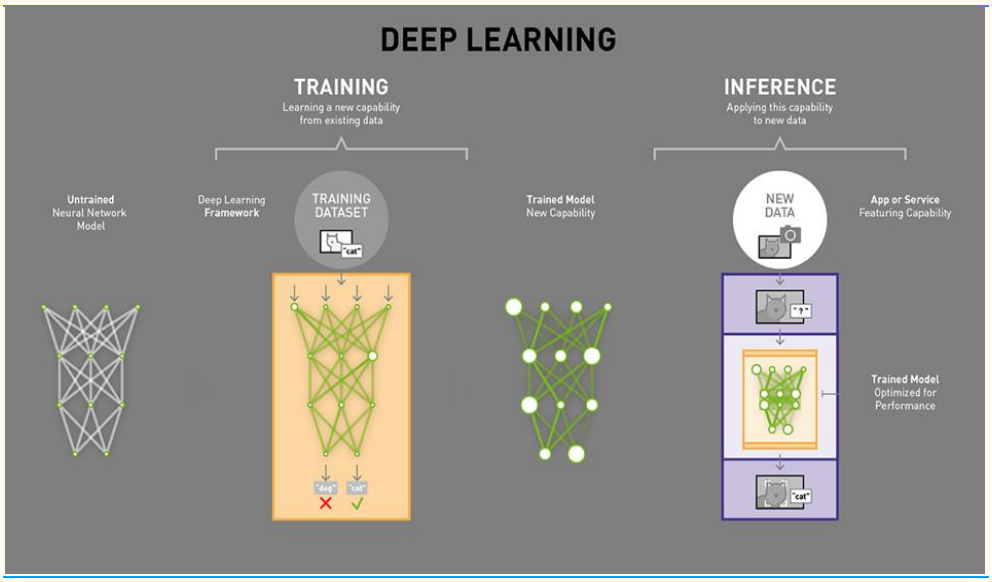


来源：Deep Cognition，国金证券研究所

五、寒武纪及谷歌的 AI 通用芯片将在边缘运算及终端渐成主流

深度学习是一种需要训练的多层次大型神经网络结构（请参考图表 16），其每层节点相当于一个可以解决不同问题的机器学习。利用这种深层非线性的网络结构，深度学习可以从少数样本展现强大的学习数据集本质特征的能力。简单来说，深度学习神经网络对数据的处理方式和学习方式与人类大脑的神经元更加相似和准确。谷歌的阿法狗也是先学会了如何下围棋，然后不断地与自己下棋，训练自己的深度学习神经网络，更厉害的阿法零 (AlphaZero) 透过更精准的节点参数，不用先进行预先学习就能自我演化训练学习。深度学习模型需要通过大量的数据训练才能获得理想的效果，训练数据的稀缺使得深度学习人工智能在过去没能成为人工智能应用领域的主流算法。但随着技术的成熟，加上各种行动、固定通讯设备、无人驾驶交通工具，可穿戴科技，各式行动、固定监控感测系统能互相连接与沟通的亿物联网，骤然爆发的大数据满足了深度学习算法对于训练数据量的要求。

图表 16：深度学习



来源：英伟达，国金证券研究所

训练和推理所需要的神经网络运算类型不同。神经网络分为前向传播（Forward algorithm）其中包括输入层，隐藏层，输出层和后向传播 (Backward algorithm)主要指的是梯度运算，两者都包含大量并行运算。训练同时需要前向和后向传播，推理则主要是前向传播。一般而言训练过程相比于推理过程计算量体更大。云端人工智能系统透过海量的数据集和调整参数优化来负责训练和推理，边缘运算终端人工智能设备负责推理。推理可在云端进行，也可以在边缘运算端或设备端进行。等待模型训练完成后，将训练完成的模型（主要是各种

通过训练得到的参数）用于各种应用。应用过程主要包含大量的乘累加矩阵运算，并行计算量很大，但和训练过程比参数相对固定，不需要大数据支撑，除在云端实现外，也可以在边缘运算端实现。推理所需参数可由云端训练完毕后，定期下载更新到应用终端。

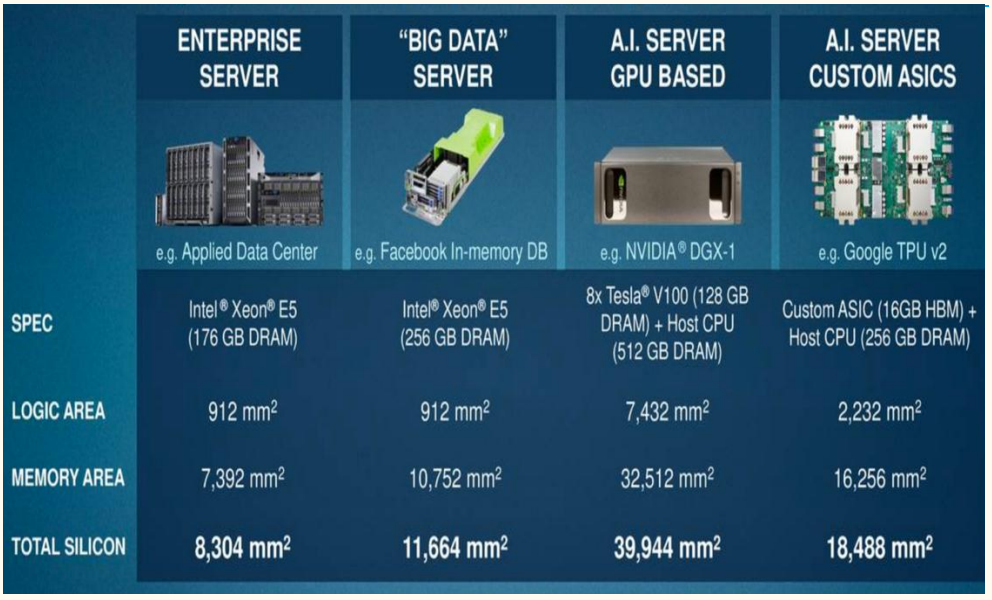
图表 17：各种人工智能半导体优缺点比较

	CPU	GPU	FPGA	TPU/ASIC
优点	通用性最高 一般用途 广泛应用于数据中心的云计算	适合深度学习训练 适用于云计算，图形渲染，科学计算 并行运算性能及通用性高	适合深度学习推理 较低功耗，低延迟，半定制 适合不断变化的工作量	可完全定制 耗电量低 性能高
缺点	不适合并行运算	耗电高，单位成本高	耗电高，浮点速度较 GPU 慢，布线不易	研发期长，初期开发成本高
代表企业	英特尔，超威	英伟达，超威	赛灵思，Lattice，英特尔/Altera	谷歌，亚马逊，华为海思，寒武纪，比特大陆，阿里巴巴，百度

来源：Frost & Sullivan, 国金证券研究所

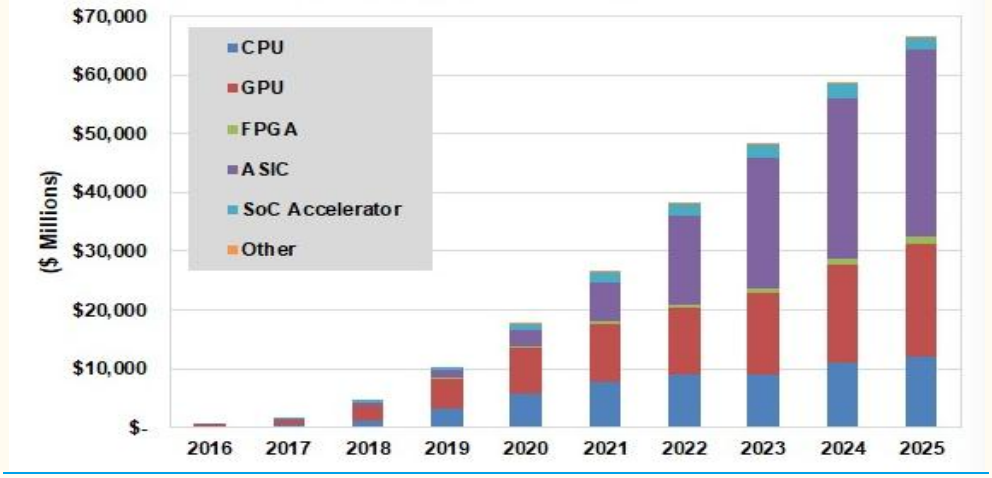
在深度学习半导体领域里，最重要的是数据和运算。谁的晶体管数量多，芯片面积大，谁就会运算快和占据优势。因此，在处理器的选择上，可以用于通用基础计算且运算速率更快的 GPU 迅速成为人工智能计算的主流芯片，根据美国应用材料的公开资料 (请参考图表 18)，英伟达的人工智能逻辑芯片配合英特尔的中央处理器服务器芯片面积达 7,432mm2，是不具人工智能的企业用和大数据服务器的八倍或谷歌专用张量处理器人工智能服务器的三倍多，存储器耗用面积 (32,512mm2) 是其他服务器的三倍以上。可以说，在过去的几年，尤其是 2015 年以来，人工智能大爆发就是由于英伟达公司的图形处理器，得到云端主流人工智能的应用。但未来因为各个处理器的特性不同，我们认为英伟达的图形处理器 GPU 和谷歌的张量处理器仍能主导通用性云端人工智能深度学习系统的训练，可编程芯片 FPGA 的低功耗及低延迟性应有利于主导云端人工智能深度学习系统的推理，而特殊用途集成电路 (ASIC) 未来将主导边缘运算及设备端的训练及推理，但因为成本，运算速度，及耗电优势，也会逐步侵入某些特殊应用人工智能云端服务器市场，抢下训练及推理运算的一席之地，以下就先列出各种处理器在云端人工智能系统的优缺点：

图表 18：人工智能云端系统图形处理芯片面积



来源：谷歌，微软，脸书，英伟达，应用材料，国金证券研究所

图表 19：人工智能半导体市场预测以不同芯片种类来分类



来源：Tractica，国金证券研究所

- **中央处理器 CPU**：X86 和 ARM 在内的传统 CPU 处理器架构往往需要数百甚至上千条指令才能完成一个神经元的处理，但对于并不需要太多的程序指令，却需要海量数据运算的深度学习的计算需求，这种结构就显得不佳。中央处理器 CPU 需要很强的处理不同类型数据的计算能力以及处理分支与跳转的逻辑判断能力，这些都使得 CPU 的内部结构异常复杂，现在 CPU 可以达到 64bit 双精度，执行双精度浮点源计算加法和乘法只需要 1～3 个时钟周期，时钟周期频率达到 1.532～3gigahertz。CPU 拥有专为顺序逻辑处理而优化的几个核心组成的串行架构，这决定了其更擅长逻辑控制、串行运算与通用类型数据运算，当前最顶级的 CPU 只有 6 核或者 8 核，但是普通级别的 GPU 就包含了成百上千个处理单元，因此 CPU 对于影像，视频计算中大量的重复处理过程有着天生的弱势。

图表 20：AI 芯片种类比较表

芯片类型		技术特点及对人工智能领域的适用性
传统芯片	CPU	➢ 通用性最强，可执行各种类型的计算机应用程序 ➢ 由控制单元、运算单元和片上存储等部件组成，运算单元占芯片面积比例较小，峰值运算性能有限 ➢ CPU 非常适合传统的控制密集型计算任务，但进行人工智能处理的性能和能效较低 ➢ 人工智能应用开发生态成熟，但性能已无法满足人工智能快速增长的计算能力需求 ➢ CPU 广泛应用于个人电脑、移动终端、传统服务器等领域 ➢ 代表厂商为 Intel、AMD 和 ARM
	GPU	➢ 最初为图形显示与渲染等任务专门设计，后逐步拓展至科学计算与人工智能领域，通用性较好 ➢ 为图形处理、科学计算等传统任务提供了良好的硬件支持，但也因此带来了显著的芯片面积开销 ➢ 运算单元占芯片面积比例很大，擅长数据级并行处理，其峰值运算性能高，但整体能耗较高 ➢ GPU 广泛应用于个人电脑、游戏机、工作站等领域；在人工智能领域，GPU 多用于服务器与数据中心，在终端应用较少 ➢ GPU 在云端具备成熟的应用开发生态，但在终端生态尚不成熟 ➢ 代表厂商为 Nvidia、AMD 和 ARM
	DSP	➢ 最初为数字信号处理任务设计，可用于传统的通信和音视频信号处理，常采用 VLIW 指令集 ➢ 编程开发的门槛较高，在云端应用较少，但在手机等终端设备中有一定生态基础 ➢ 代表厂商为 TI、CEVA 和 Cadence 等
	FPGA	➢ 在 IC 原型验证与仿真中有着广泛应用 ➢ FPGA 包含充裕的可重构逻辑单元阵列，可通过硬件重构方式灵活实现适合于人工智能应用的架构，但其成本和能效与实现相同架构的非 FPGA 芯片相比有很大差距 ➢ FPGA 开发和调试门槛较高 ➢ 代表厂商为 Xilinx
智能芯片	通用型智能芯片	➢ 针对人工智能领域内多样化的应用设计的处理器芯片，对视觉、语音、自然语言处理、传统机器学习技术等各类人工智能技术具备较好的普适性
		➢ 无需像 CPU 一样支持控制密集型计算任务，或者像 GPU 一样兼顾图形处理与科学计算任务，架构完全针对人工智能处理的实际需求所设计 ➢ 全新指令集完备高效，可覆盖各类智能算法所需的基本运算操作 ➢ 在指令集、处理器架构以及基础系统软件等方面具备较高的技术壁垒 ➢ 性能功耗比较传统芯片优势明显，可适应各种场景和规模的人工智能计算需求 ➢ 与传统芯片生态兼容，降低了程序员的开发难度 ➢ 架构灵活通用，可支撑其在云端、边缘端和消费类电子终端都获得广泛应用 ➢ 代表厂商为寒武纪和 Google（TPU）
	专用型智能芯片（ASIC）	➢ 面向特定的、具体的、相对单一的人工智能应用所设计的专用 IC ➢ 在架构层面对特定智能算法作硬化支持，指令集简单或指令完全固化 ➢ 常用于在低功耗、成本敏感的终端上支撑特定的智能应用，芯片架构相对简单，技术门槛相对较低

来源：寒武纪招股说明书，国金证券研究所

- **图形处理器 GPU 仍主导云端人工智能深入学习及训练**：最初是用在计算机、工作站、游戏机和一些移动设备上运行绘图运算工作的微处理器，但其海量数据并行运算的能力与深度学习需求不谋而合，因此，被最先引入深度学习。GPU 只需要进行高速运算而不需要逻辑判断。GPU 具备高效的浮点算数运算单元和简化的逻辑控制单元，把串行访问拆分成多个简单的并行访问，并同时运算。例如，在 CPU 上只有 20-30%的晶体管(内存存储器 DRAM dynamic random access memory，缓存静态随机存储器 Cache SRAM，控制器 controller 占了其余的 70-80% 晶体管) 是用作计算的，但反过来说，GPU 上有 70-80%的晶体管是由上千个高效小核心组成的大规模并行计算架构 (DRAM 和微小的 Cache SRAM，controller 占了剩下的 20-

30% 晶体管)。大部分控制电路相对简单，且对 Cache 的需求小，只有小部分晶体管来完成实际的运算工作，至于其他的晶体管可以组成各类专用电路、多条流水线，使得 GPU 拥有了更强大的处理浮点运算的能力。这决定了其更擅长处理多重任务，尤其是没有技术含量的重复性工作。不同于超威及英特尔的 GPU 芯片，英伟达的人工智能芯片具有 CUDA 的配合软件是其领先人工智能市场的主要因素。CUDA 编程工具包让开发者可以轻松编程屏幕上的每一个像素。在 CUDA 发布之前， GPU 编程对程序员来说是一件苦差事，因为这涉及到编写大量低层面的机器码。CUDA 在经过了英伟达的多年开发和改善之后，成功将 Java 或 C++这样的高级语言开放给了 GPU 编程，从而让 GPU 编程变得更加轻松简单，研究者也可以更快更便宜地开发他们的深度学习模型。因此我们认为目前英伟达配备 8 颗 A100 的 DGXA100 系统（199,000 美元），配合其 CUDA 软件及 NVLink 快速通道，以 16bit 的半精度浮点性能来看，能达到近 2,480 兆次 (2.4 petaFLOPS) 深入学习的浮点运算训练速度，若以 32bit 的单精度浮点性能来看，可达到 1,280 兆次浮点运算（1.28 petaFLOPS），目前仍然是云端人工智能深入学习及训练的最佳通用型解决方案。

图表 21：英伟达云端人工智能芯片 A100 及系统 DGX A100 规格比较表

	適用 NVLink 的NVIDIA A100
FP64 最佳效能	9.7 TF
FP64 Tensor 核心最佳效能	19.5 TF
FP32 最佳效能	19.5 TF
TF32 Tensor 核心最佳效能	156 TF | 312 TF*
BFLOAT16 Tensor 核心最佳效能	312 TF | 624 TF*
FP16 Tensor 核心最佳效能	312 TF | 624 TF*
INT8 Tensor 核心最佳效能	624 TOPS | 1,248 TOPS*
INT4 Tensor 核心最佳效能	1,248 TOPS | 2,496 TOPS*
GPU 記憶體	40 GB
GPU 記憶體頻寬	1,555 GB/s
互連	NVIDIA NVLink 600 GB/s PCIe Gen4 64 GB/s
多執行個體 GPU	最高到七個 5GB 不同容量的執行個體
尺寸規格	NVIDIA HGX™ A100 使用 4/8 SXM
最大 TDP 功耗	400W

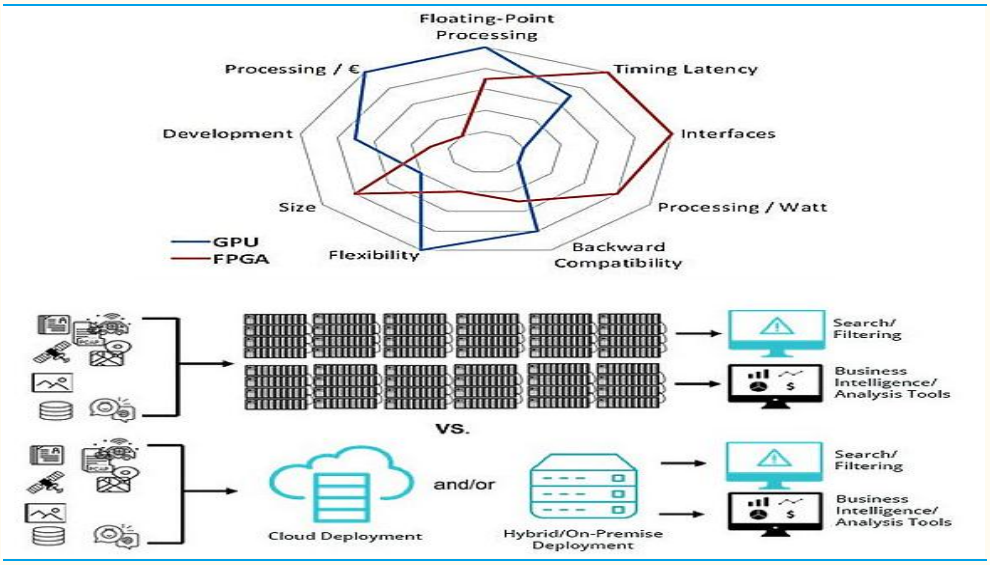
SYSTEM SPECIFICATIONS	
GPUs	8x NVIDIA A100 Tensor Core GPUs
GPU Memory	320 GB total
Performance	5 petaFLOPS AI 10 petaOPS INT8
NVIDIA NVSwitches	6
System Power Usage	6.5kW max
CPU	Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)
System Memory	1TB
Networking	8x Single-Port Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 1x Dual-Port Mellanox ConnectX-6 VPI 10/25/50/100/200Gb/s Ethernet
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives
Software	Ubuntu Linux OS
System Weight	271 lbs (123 kgs)
Packaged System Weight	315 lbs (143kgs)
System Dimensions	Height: 10.4 in (264.0 mm) Width: 19.0 in (482.3 mm) MAX Length: 35.3 in (897.1 mm) MAX
Operating Temperature Range	5°C to 30°C (41°F to 86°F)

来源：英伟达，国金证券研究所

- **现场可编程门阵列芯片 FPGA 的优势在低功耗，低延迟性**：CPU 内核并不擅长浮点运算以及信号处理等工作，将由集成在同一块芯片上的其它可编程内核执行，而 GPU 与 FPGA 都以擅长浮点运算著称。FPGA 和 GPU 内都有大量的计算单元，它们的计算能力都很强。在进行人工智能神经网络 (CNN, RNN, DNN) 运算的时候，两者的速度会比 CPU 快上数十倍以上。但是 GPU 由于架构固定，硬件原来支持的指令也就固定了，而 FPGA 则是可编程的，因为它让软件与应用公司能够提供与其竞争对手不同的解决方案，并且能够灵活地针对自己所用的算法修改电路。虽然 FPGA 比较灵活，但其设计资源比 GPU 受到较大的限制，例如 GPU 如果想多加几个核心只要增加芯片面积就行，但 FPGA 一旦型号选定了逻辑资源上限就确定了。而且，FPGA 的布线资源也受限制，因为有些线必须要绕很远，不像 GPU 这样走 ASIC flow 可以随意布线，这也会限制性能。FPGA 虽然在浮点运算速度，增加芯片面积，及布线的通用性比 GPU 来得差，却在延迟性及功耗上对 GPU 有着显着优势。英特尔斥巨资收购 Altera 是要让 FPGA 技术为英特尔的发展做贡献。表现在技术路线图上，那就是从现在分立的 CPU 芯片+分立的 FPGA 加速芯片，过渡到同一封装内的 CPU 晶片

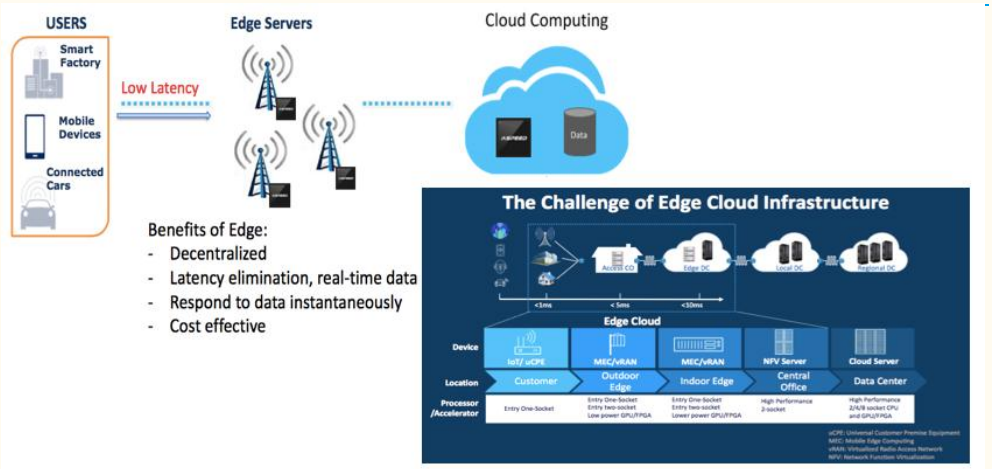
+FPGA 晶片，到最终的集成 CPU+FPGA 系统芯片。预计这几种产品形式将会长期共存，因为 CPU 和 FPGA 的分立虽然性能稍差，但灵活性更高。目前来看，用于云端的人工智能解决方案是用 Xeon CPU 来配合 Nervana，用于云端中间层和边缘运算端设备的低功耗推断解决方案是用 Xeon CPU 来配合 FPGA 可编程加速卡。赛灵思（Xilinx）于 2018 年底推出以低成本，低延迟，高耗能效率的深度神经网络（DNN）演算法为基础的 Alveo 加速卡 (Alveo U25，U50，U200，U250，U280 ,采用台积电 16nm 制程工艺的 UltraScale FPGA。

图表 22：赛灵思／BlackLynx 与 GPU 在机器学习推理解决方案的比较



来源：赛灵思，国金证券研究所

图表 23：5G 带动不同延迟的人工智能边缘运算的需求

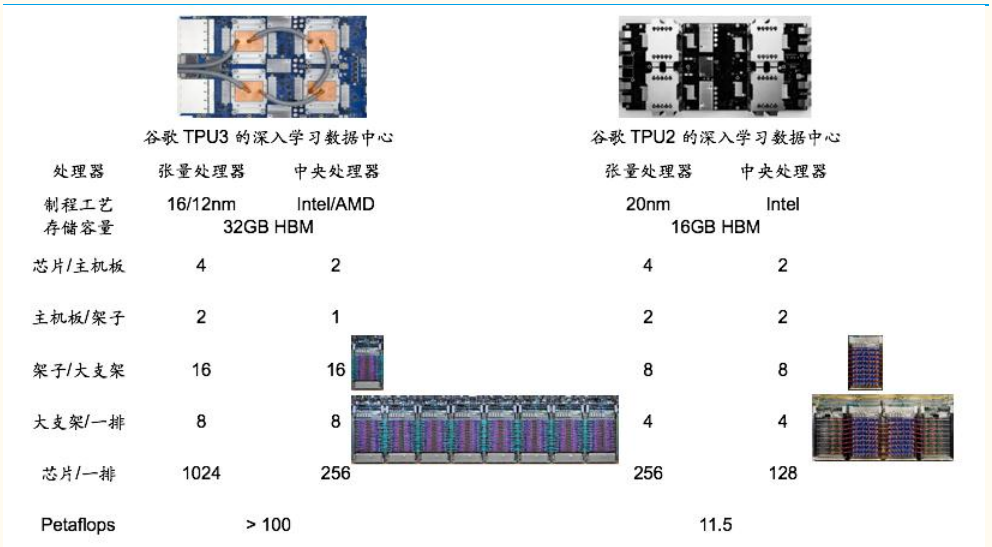


来源：信骅，国金证券研究所

- **谷歌张量处理器 TPU 3，寒武纪 AI 芯片及 ASIC**：因为它能加速其人工智能系统 TensorFlow 的运行，而且效率也大大超过 GPU，谷歌的深层神经网络就是由 TensorFlow 引擎驱动的，其第三代张量处理器 (TPU v3, Tensor Processing Unit, 大约 420 Tera FLOPS/hp-16bit) 是专为机器学习由谷歌提供系统设计，博通提供芯片设计服务及智财权专利区块，台积电提供 16/12 纳米制程工艺量身定做的，执行每个操作所需的晶体管数量更少，自然效率更高。TPU 每瓦能为机器学习提供比所有商用 GPU 和 FPGA 更高的量级指令。TPU 是为机器学习应用特别开发，以使芯片在计算精度降低的情况下更耐用，这意味每一个操作只需要更少的晶体管，用更多精密且大功率的机器学习模型，并快速应用这些模型，因此用户便能得到更正确的结果。以谷歌子公司深度思考的阿尔法狗及零 (AlphaGo,

AlphaZero/DeepMind) 利用人工智能深度学习训练和推理来打败世界各国排名第一的围棋高手，世界排名第一的西洋棋 AI 程式 Stockfish 8，世界排名第一的日本棋 Shogi AI 专家，但我们估计 AlphaZero 系统使用至少近 5 大排人工智能主机，5,000 个张量处理器，1,280 个中央处理单元而让云端的设备异常昂贵。Google TPU 和寒武纪智能芯片的相同点都是通过对人工智能领域的计算特征和访存特征进行分析和抽象,设计出的通用型智能芯片，其指令集、运算器架构和存储层次都非常适合智能算法，从而让智能应用上的能效超过了传统 CPU 及 GPU。Google TPU 和寒武纪智能芯片的不同点是在处理器架构上采用了不同的路线。Google TPU 的核心是经典的脉动阵列机技术，脉动阵列本身对于卷积类运算的效率较高，但是对于相对低频的部分运算操作(如全连接运算、激活运算)的效率不高。对于后者，Google TPU 引入了额外的硬件单元作为补充。而寒武纪的芯片架构，则直接将算法的基本操作区分为高维张量运算、向量运算、算数逻辑运算，并在处理器中分别通过高维张量、向量、传统算术逻辑等计算部件予以处理。高维张量计算部件可高效支持卷积运算、全连接运算，而向量计算部件则可以支持激活等运算，传统算术逻辑计算部件则可以支持分支跳转等。即使研发期长，初期开发成本高，但国内芯片业者因缺乏先进 x86 CPU, GPU, 及 FPGA 的基础设计智慧财产权（IPs），因此可完全客制化，耗电量低，性能强的 AI 通用 IC 设计就立刻成为国内进入人工智能云端及边缘运算及设备端芯片半导体市场的唯一途径。目前除了比特大陆的算丰(SOPHON) BM1682/1684 安防及大数据边缘运算人工智能推理系列产品已经上市之外，华为海思使用台积电 7 纳米制程工艺设计的升腾 Ascend 910 系列，号称在 16bit 半精度下能达到 256 兆次的浮点运算，只有小幅低于英伟达目前最先进的 A100 GPU 的 310 兆次的浮点运算解决方案(台积电 7 纳米) 及谷歌之前推出的张量处理器 3 的 420 兆次的浮点运算解决方案（台积电 16/12 纳米）。而国内搜寻引擎龙头百度的昆仑芯片(818-300 采用 Samsung 14 纳米，在 Xilinx FPGA 的架构下及 150 瓦的功耗下提供 260TOPS 性能)，除了以上这些公司产品外，阿里巴巴的 Ali-NPU, 及亚马逊的 Inferentia (128TOPS) 目前都还没有提供更确实的芯片速度，耗电量，应用，价格，量产时点，及软件框架规格让我们做出更好的比较图表

图表 24：谷歌张量处理器 TPU3 vs. TPU2



来源：谷歌, TIRIAS, 国金证券研究所

六、公司介绍

1.基本资料

中科寒武纪科技股份有限公司为中科计算所的博士生导师陈天石于北京成立于 2016 年 3 月 15 日，陈天石目前担任董事长暨总经理，负责把控公司整体的技术方向、业务进程以及战略发展方向，并牵头开展学术研究和产业化工作。截至 2019 年 12 月 31 日，公司总员工数 858 人，拥有研发人员 680 人，占员工总人数的 79.25%；拥有硕士及以上学历人员 546 人，占员工总人数的 63.64%。截至 2020 年 2 月 29 日，公司已获得授权的专利共计 65 项，其中境内专利共计 50 项，境外专利共计 15 项。

公司自成立以来，在短短时间内推出了手机 AI 用系列 IP， 寒武纪 1A，1H，1M， 并被海思于 2018 年大量采用在其麒麟芯片系列，但在海思于 2019 年自主研发其手机用 AI IP 后，寒武纪又能成功的推出云端 AI 推理芯片思元 100 及云端 AI 推理训练芯片 270，加速卡，及设计并外包生产销售智能计算集群系统，让去年营收同比增长达 279%。而今年再叠加思元 220 边缘运算芯片及加速卡的贡献，而明年有思元 290 云端训练芯片及加速卡的营收贡献。公司是目前国内少数几家能够掌握智能芯片及其基础系统软件研发和产品化核心技术的企业之一，能提供云端，边缘端，终端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件来处理图像，视频，语音识别与合成和自然语言。

图表 25：寒武纪主要产品介绍

产品类型	寒武纪主要产品	推出时间
终端智能处理器 IP	寒武纪 1A 处理器	2016 年
	寒武纪 1H 处理器	2017 年
	寒武纪 1M 处理器	2018 年
云端智能芯片及加速卡	思元 100（MLU100）芯片及云端智能加速卡	2018 年
	思元 270（MLU270）芯片及云端智能加速卡	2019 年
	思元 290（MLU290）芯片及云端智能加速卡	芯片样品测试中
边缘智能芯片及加速卡	思元 220（MLU220）芯片及边缘智能加速卡	2019 年
基础系统软件平台	Cambricon Neuware 软件开发平台（适用于公司所有芯片与处理器产品）	持续研发和升级，以适配新的芯片

来源：寒武纪招股说明书，国金证券研究所

除陈天石外，首席运营官王在是中国科学技术大学计算机应用技术博士学历，2016 年至 2018 年就职于中科院计算所从事科研工作。2016 年作为公司创始团队成员加入公司，现任公司董事、副总经理兼首席运营官。而梁军是中国科学技术大学通信与信息系统硕士学历，于 2003 年至 2017 年，就职于华为技术有限公司基础业务部、并在海思半导体历任工程师、高级工程师、主任工程师、技术专家、高级技术专家。2017 年起为公司服务，现任公司副总经理兼首席技术官，总体负责公司的研发工作，总体领导研发团队完成芯片设计与实现、芯片与板卡产品量产、基础系统软件研发等。

图表 26：主要产品核心研发领导

主要产品或业务	涉及的主要核心技术	主要研发人员
终端智能处理器 IP	智能处理器微架构、智能处理器指令集、编程框架适配与优化、智能芯片编程语言、智能芯片编译器、智能芯片高性能数学库	由陈天石、梁军总体指导，刘少礼带领架构、验证及软件相关工程师团队开展研发和产品化
云端智能芯片及加速卡	智能处理器微架构、智能处理器指令集、SoC 芯片设计、处理器芯片功能验证、先进工艺物理设计、芯片封装设计与量产测试	由陈天石、梁军总体指导，刘少礼带领芯片、架构、验证、软件及产品相关工程师团队开展研发和产品化
边缘智能芯片及加速卡	智能处理器微架构、智能处理器指令集、SoC 芯片设计、处理器芯片功能验证、先进工艺物理设计、芯片封装设计与量产测试	由陈天石、梁军、刘道福总体指导，陈煜带领芯片、架构、验证、软件、产品相关工程师团队开展研发和产品化

来源：寒武纪招股说明书，国金证券研究所

2.股权结构

在本次 IPO 增资 11%股数发行前（从 3.6 亿股增加到 4.0 亿股），公司控股股东陈天石直接持有公司 33.19%的股份，并作为艾溪合伙的执行事务合伙人控制艾溪合伙持有公司 8.51%的股份,陈天石合计控制公司 41.71% 的股份，为公司的实际控制人，第二大股东中科算源（为中科院计算所 100%持有的子公司）在 IPO 增资后持有 16.4%的寒武纪，而第四大股东苏州工业园区古生代创业投资主要是南京智子集成电路产业投资企业（65.63%）及江苏金财投资有限公司（32.77%）持有的，而第五大股东国投科技成果转化创投基金企业主要系国家开发投资集团（21%），国家科技风险开发事业中心（20%），宁波梅山保税港区干平涌顺 / 珞佳熙明投资管理合伙企业（19.25% / 19.25%），上海科技创业投资（10%）所持有。

寒武纪控股股东，实际控制人陈天石，艾溪合伙，艾加溪合伙都承诺在寒武纪股票上市后，三年内不能转让持股，但其他主要股东如中科算源，古生代创投，国投基金，南京招商，宁波瀚高等，其闭锁期都是一年，而公司董事，监事，高级管理人员及核心技术人员都承诺，原则上公司获利后闭锁期为一年，假如持续亏损，闭锁期最长为三年，因为我们预期公司在未来 2-3 年内要扭亏转盈相对困难，所以整体经营团队应该都是在三年闭锁期的规定之下，但其他策略投资股东的卖压应该在一年闭锁期之后逐步涌现。

图表 27：寒武纪前 10 大股东 IPO 前后持股变化

序号	股东名称/姓名	本次发行前		本次发行后	
		持股数（股）	占比（%）	持股数（股）	占比（%）
1	陈天石	119,497,756	33.19	119,497,756	29.87
2	中科算源（SS）	65,669,721	18.24	65,669,721	16.41
3	艾溪合伙	30,645,870	8.51	30,645,870	7.66
4	古生代创投	14,151,905	3.93	14,151,905	3.54
5	国投基金	14,124,730	3.92	14,124,730	3.53
6	南京招银	13,002,264	3.61	13,002,264	3.25
7	宁波瀚高	12,339,146	3.43	12,339,146	3.08
8	深圳新芯	8,571,090	2.38	8,571,090	2.14
9	艾加溪合伙	8,485,379	2.36	8,485,379	2.12
10	阿里创投	6,975,170	1.94	6,975,170	1.74

来源：寒武纪招股说明书，国金证券研究所

3.募资投入研发

寒武纪预期在 IPO 用 4010 万股，募集将近 25.8 亿人民币的资金，每股定价在 64.39 元人民币，略低于招股说明书预期的 28 亿资金募集，这其中近 7 亿或 27.1%主要用在云端训练芯片及系统软件三年新产品的开发，另外 6 亿或 23.2%用在云端推理芯片及系统软件的开发，还有 6 亿或 23.3%用在边缘运算端 AI 芯片及系统软件的开发，当然剩下的 6.8 亿现金是用来补充现金流，这对去年底手上只剩 3.8 亿现金的寒武纪而言，确实需要。如果看公司目前还在开发阶段的 8 个研发项目，总金额达 7.55 亿，都还是在募资计划的可控范围。当然除了思元 290 云端 AI 训练芯片在明年推出外，寒武纪现阶段也已经投入 5nm 制程工艺产品的研发设计，提供支持布局布线、物理验证、静态时序分析等全流程设计支持，覆盖所有环节设计需求。保障最新工艺芯片流片一次成功与量产良率。

图表 28：寒武纪原始募资使用计划

序号	项目名称	总投资额（万元）	使用募集资金投入金额（万元）	建设期
1	新一代云端训练芯片及系统项目	69,973.07	69,973.07	3 年
2	新一代云端推理芯片及系统项目	60,016.97	60,016.97	3 年
3	新一代边缘端人工智能芯片及系统项目	60,072.47	60,072.47	3 年
4	补充流动资金	90,000.00	90,000.00	
合计		**280,062.51**	**280,062.51**	

来源：寒武纪招股说明书，国金证券研究所

图表 29：寒武纪研发项目及进展

序号	项目名称	项目介绍	研发目标	所处阶段	经费投入（万元）
1	智能处理器架构	本项目旨在持续研发一系列先进的智能处理器架构和 IP，支撑内部智能芯片研发和外部终端智能处理器 IP 销售	持续提高智能处理器架构的先进性，提高智能处理器 IP 的性能和能效。给公司所有产品线的提供核心竞争力支撑	开发阶段	14,484.81
2	边缘智能芯片	本项目旨在研发面向边缘推理的智能加速芯片，用于各种边缘场景的小尺寸边缘智能加速卡	面向边缘智能处理低延时、低功耗以及部署环境的小尺寸要求，研发一款高性能、低功耗、小尺寸的边缘智能芯片；同时要求支持主流的边缘场景应用接口，比如 EMMC、GMAC，以支撑各种应用场景部署	开发阶段	6,914.61
3	基础系统软件（推理）	本项面向人工智能推理任务，研发（并持续升级迭代）适用于公司各类芯片/处理器产品的基础系统软件，支撑开发者基于该软件平台开发推理应用	提供云边端一体化的应用开发环境，支持跨云边端硬件平台的应用开发；支持业界主流人工智能编程框架，提供完备的开发、调试、性能调优工具链	开发阶段	4,830.41
4	PCIe 加速卡硬件产品	本项目旨在基于云端推理芯片，研发适用于数据中心服务器的，易于部署的 PCIe 加速卡硬件产品	符合标准 PCIe 加速卡规范，兼容主流服务器；研发不同功耗规格的，面向不同场景的硬件加速卡	开发阶段	4,431.64
5	基础系统软件（训练）	本项面向人工智能训练任务，研发适用于公司各云端芯片的基础系统软件，支撑开发者基于该软件平台开发训练应用	为云端的人工智能训练任务提供高效、灵活的应用开发平台，在单机单卡、单机多卡和多机多卡等不同场景下达到优异的性能；支持业界主流人工智能编程框架，提供完备的开发、调试、性能调优工具链	开发阶段	1,240.91

6	硬件平台（训练）	本项目基于公司云端芯片产品，研发适用于各类训练服务器、易于部署的硬件加速卡产品与硬件底板	用于人工智能训练的加速卡兼容业界主流训练服务器板卡接口，硬件底板支持多卡间互联	开发阶段	1,572.50
7	高档云端智能芯片	本项目面向云端的人工智能训练场景，研发性能和能效出色的云端芯片	单芯片具备充裕的峰值运算能力，支持多芯片间互联，以支持分布式训练；芯片适用于多样化的人工智能训练任务	开发阶段	17,738.21
8	中档云端智能芯片	本项目面向云端推理任务以及云端相对简单的训练任务，研发性价比与能效出色的云端芯片	芯片的能效与计算能力密度（单位面积提供的计算能力）具有竞争力；芯片适用于多样化的人工智能推理应用	开发阶段	24,279.17

来源：寒武纪招股说明书，国金证券研究所

4.核心客户及供应商的变化

2017 年、2018 年和 2019 年,公司前五大客户的销售金额合计占营业收入比例分别为 100.00%、99.95%和 95.44%, 表示客户集中度的风险还是相当高，而且在 2017 及 2018 年海思的占比竟高达 98.34%及 97.63%， 自从海思自主研发人工智能 IP 及芯片后，2019 年的占比大幅下降到只有 14.34%的营收。而智能计算集群系统营收从 2018 年的零贡献，一下增加到 2019 年的 67%，这主要是从珠海市横琴新区管理委员会商务局（46.65%）及西安沣东仪享科技服务有限公司（18.26%）开展的智能计算集群系统项目而来。而第三大客户关联方中科曙光（14.38%）主要系寒武纪云端 AI 推理加速卡思元 100（采购 4,000 块思元 100 加速卡，安装在 1,000 台中科曙光的服务器中，然后供给横琴先进智能计算平台一期项目）的主要客户。中科曙光（603019 CH）是寒武纪第二大股东中科算源（IPO 后持有寒武纪股份 16.41%）控制达 20.63%的企业。而我们预期 2020 年边缘运算 AI 芯片及加速卡客户及基础系统软件客户将占有一席之地， 但智能计算集群系统客户的比重会降低。

图表 30：寒武纪 2017-2019 年前五大客户销售金额及比重变化（万元）

年份	序号	客户名称	销售金额	占营业收入比例
2019 年	1	珠海市横琴新区管理委员会商务局	20,708.35	46.65%
	2	西安沣东仪享科技服务有限公司	8,108.46	18.26%
	3	中科曙光	6,384.43	14.38%
	4	华为海思	6,365.80	14.34%
	5	上海脑科学与类脑研究中心	801.34	1.81%
	合计		**42,368.37**	**95.44%**
2018 年	1	华为海思	11,425.64	97.63%
	2	杭州博雅鸿图视频技术有限公司	141.51	1.21%
	3	厦门星宸科技有限公司	99.06	0.85%
	4	江苏恒瑞通智能科技有限公司	20.04	0.17%
	5	北京的卢深视科技有限公司	10.67	0.09%
	合计		**11,696.92**	**99.95%**
2017 年	1	华为海思	771.27	98.34%
	2	中科院院士上海浦东活动中心	4.85	0.62%
	3	南京航空航天大学	4.80	0.61%
	4	南开大学	3.40	0.43%
	合计		**784.33**	**100.00%**

来源：寒武纪招股说明书，国金证券研究所

2017 年-2019 年，公司向前五名直接供应商合计采购的金额占比分别为 92.64%、82.53%和 66.49%，占比虽然高，EDA 工具及 IP（主要使用新思科技，上海国际科学技术有限公司代理 Cadence，世芯 Alchip 的 EDA 设计工具及 IP，及 ARM 的 IP）以及晶圆代工（主要应该是透过深圳市朗华供应链服务有限公司来采购台积电的晶圆代工共乘服务（CyberShuttle prototyping service）及采购博通/Avago 及泰科源所代理的各种电子元器件及半导体），根据产业链了解，寒武纪是透过博通来进行芯片的设计服务及提供部分高速通讯 IP，而晶圆代工主要就是以台积电为 16nm 及 7nm 为其主力，芯片 IP 及 EDA 工具主要向 Cadence、Synopsys 和 ARM 等采购，日月光、Amkor 和长电科技是寒武纪的主要封装测试商。

图表 31：寒武纪 2017-2019 年前五大供应商采购金额及比重变化（万元）

年份	序号	供应商名称	采购金额	主要采购内容	占采购总额比例
2019 年	1	深圳市朗华供应链服务有限公司	15,502.74	晶圆、电子元器件	28.42%
	2	中科可控信息产业有限公司	8,110.00	服务器	14.87%
	3	新思科技有限公司	4,797.79	EDA 工具、IP	8.80%
	4	上海国际科学技术有限公司	4,371.35	EDA 工具、IP	8.01%
	5	安谋科技（中国）有限公司	3,489.29	IP	6.40%
	合计		36,271.17		66.49%
2018 年	1	上海国际科学技术有限公司	14,310.96	EDA 工具、IP	58.14%
	2	深圳市朗华供应链服务有限公司	3,035.41	晶圆、电子元器件	12.33%
	3	上海协进电脑科技有限公司	1,483.26	服务器、电脑	6.03%
	4	北京晟图瑞德科技有限公司	887.02	服务器、电脑	3.60%
	5	北京联创芯源科技有限公司	598.84	芯片、电子元器件	2.43%
	合计		20,315.49		82.53%
2017 年	1	上海国际科学技术有限公司	876.41	EDA 工具、IP	57.08%
	2	东方科仪控股集团有限公司	246.20	电子设备	16.04%
	3	上海协进电脑科技有限公司	193.43	服务器、电脑	12.60%
	4	北京志翔科技股份有限公司	67.50	软硬件设备	4.40%
	5	深圳芯力电子技术有限公司	38.73	芯片	2.52%
	合计		1,422.28		92.64%

来源：寒武纪招股说明书，国金证券研究所

七、盈利预测及假设

1.寒武纪营收/获利的历史数据及预测的假设基础

- 终端 AI 处理器 IP 业务逐步淡出：在 2017－2019 年，寒武纪的终端 AI 处理器 IP 业务主要是让华为海思内建其 IP 到其手机麒麟芯片，并占到公司终端 AI 处理器 IP 授权业务销售收入比例的 100.00%、97.94%和 92.56%，而 2019 年终端 AI 处理器 IP 授权业务收入相较于 2018 年下滑 41.23%，主要系华为海思自研 AI 芯片，未与寒武纪继续合作，加上公司短期内难以开发同等规模的大客户，因此 2020 年公司终端智能处理器 IP 授权业务收入将继续下滑，公司预期全年 1A 加 1H IP 授权收入为 600-800 万元，第三代 IP 1M 授权收入为 1,000 万元，除了海思外，目前有杭州博雅鸿图视频技术，厦门星宸科技，展讯通信，北京智芯微电子科技等潜在客户，但营收占比可能连 5 个点占比都没有，我们预期未来这些业务占比将从 2017-2018 的 98-100%，降低到去年的 15%，到未来几年低于 4%的营收。

■ **云端及边缘运算 AI 芯片及加速卡是两大高毛利增长动能：**从 2019 年开始，寒武纪的云端 AI 推理思元 100 芯片及 100 / 270 加速卡，就贡献了 7888 万营收（主要是思元 100 加速卡占比 98%，思元 100 芯片及思元 270 加速卡各占 1%），占整体营收比 18%，公司预期今年上半年将贡献了近 6300-6500 万营收，虽然同比衰退仍近 1.62-4.65%，但营收占比大幅提升到 76%，而公司去年底才推出的云端 AI 边缘运算芯片思元 220 及其加速卡，公司就预期今年上半年将贡献近 440-530 万营收，占比将达到 6%。再加上 2021 年将量产的云端 AI 训练芯片思元 290 及加速卡，我们估计未来五年这两项业务将可能有 80-100%复合增长率的贡献。寒武纪去年云端 AI 推理芯片及加速卡的主要客户是关联方的中科曙光（占比 80.9%）及非关联方的江苏恒瑞通智能科技，浪潮，联想，新华三，宝德，技嘉，长城飞腾及北京金山云网络技术。

图表 32：寒武纪云端智能芯片及加速卡的适配及认证

厂商名称	通过认证产品	该厂商认证的标准	该厂商认证需履行的程序	认证有效期限
联想	思元 100/270	厂商内部标准	板卡第三方认证审核 通过板卡引入测试：结构/PI/SI/OS 兼容/可靠/性能 加入服务器部件可选 BOM	无期限限制
浪潮	思元 100/270	厂商内部标准	板卡第三方认证审核 通过板卡引入测试：结构/PI/SI/OS 兼容/可靠/性能 加入服务器部件可选 BOM	无期限限制
新华三	思元 100 认证完成，思元 270 正在认证中	厂商内部标准	板卡第三方认证审核 通过板卡引入测试：结构/PI/SI/OS 兼容/可靠/性能 加入服务器部件可选 BOM	无期限限制
曙光	思元 100/270	厂商内部标准	板卡第三方认证审核 通过板卡引入测试：结构/PI/SI/OS 兼容/可靠/性能 加入服务器部件可选 BOM	无期限限制
宝德	思元 100 认证完成，思元 270 正在认证中	厂商内部标准	通过板卡引入测试：结构/PI/SI/OS 兼容/可靠/性能 加入服务器部件可选 BOM	无期限限制
技嘉	思元 100	厂商内部标准	通过板卡引入测试：结构/PI/SI/OS 兼容/可靠/性能 加入服务器部件可选 BOM	无期限限制
长城飞腾	思元 100	厂商内部标准	通过板卡引入测试：结构/PI/SI/兼容/可靠/性能 加入服务器部件可选 BOM	无期限限制

来源：寒武纪招股说明书，国金证券研究所

■ **智能计算集群系统事业不稳定，密切关注其风险：** 2019 年，寒武纪智能计算集群系统业务收入主要来源于与珠海市横琴新区管理委员会商务局的采购案、西安沣东仪享科技服务有限公司开展的智能计算集群系统招标项目，该等项目占公司智能计算集群系统业务收入比例为 97.3%，而且在一年之内从 0%占比到占总收入的大头达到 67%（主要贡献 2019 年下半年）。以横琴先进智能计算平台 (二期)采购项目来看，其实就是寒武纪销售整机智能加速系统，而这些检测好的系统是委托给中科可控代工将 5,200 块思元 270 云端推理及训练加速卡及软件整合到透过中科可控代工的 1,300 台智能服务器及 48 台并行存储系统。但以招股书的最新资料显示，公司智能计算集群系统方面的在手订单包括横琴先进智能计算平台 (二期) 的第二批供货硬件设备，授权软件，合同金额 仅剩下 1.86 亿元，而上半年营收贡献连 20 万都没有(上海脑科学与类脑研究中心项目优化服务收入在一季度贡献 6.4 万)，除非我们看到在下半年寒武纪拿到新的地方政府新基建案及其他在手订单大幅回流，今年此业务营收贡献可能不到 60%，甚至连 50%占比都可能有问题。而就长期而言，寒武纪这事业群直接面对终端使用者客户，如地方数据中心，行业企业和科研机构等，就需要庞大分散的客户群来稳定在手订单及营收，否则将带给投资人相当大的营收及获利上下大幅波动的风险。之前也有提到，寒武纪于 2019 年跨

入了智能计算集群系统（简单来说，就是自己采购并整合从中科可控来的服务器设备），但寒武纪总是冒着跟其系统客户抢生意的争议。这就像海思虽然在特定用途的人工智能云，边缘，终端芯片有着领先的地位，但除了靠华为内部集团的大量采购外，海思的产品要被其他智能计算集群系统客户的采用，就有相当的难度，很多华为在 5G 通讯，服务器，手机的直接竞争者，还是会以第三方公司像寒武纪设计的芯片及加速卡为考量，但当寒武纪也跨入其客户擅长的人工智能系统设计及外包代工领域，跟客户如浪潮，联想，中科曙光抢生意的疑虑就逐步浮现。

图表 33：横琴先进智能平台及其他 AI 集群系统采购细目整理

终端客户	西安沣东仪享	上海脑科学	横琴新区管理委员会商务局		
采购项目		类脑研究中心	横琴先进智能计算平台		
工期			一期	二期第一批	二期第二批
智能计算集群系统供应商	寒武纪	寒武纪	中科曙光	寒武纪	寒武纪
集群系统服务器供应商	浪潮信息	苏州超集信息	中科曙光	中科可控	中科可控
思元加速卡产品	思元 270	思元 100, 270	思元 100	思元 270	思元 270
思元加速卡数量	1,800.0	240.0	4,000	5,200	N/A
思元加速卡单价测算 US$	4,991	2,674	2,310	3,506	N/A
x86 服务器数量	225.0	30.0	1,000	1,300	N/A
每台服务器加速卡数	8	8	4	4	N/A
并行存储系统	N/A	N/A	N/A	48	N/A
寒武纪 Neuware 软件	包括	包括	包括	包括	包括
对寒武纪营收贡献 CNY$mn	81.1	3.6	63.8	207.1	185.7
对寒武纪营收占比 (%)	18.3%	0.8%	14.4%	46.6%	30.6%
时期	2019	2019	2019	2019	2020

来源：寒武纪招股说明书及问询函回复，国金证券研究所

- **>50%营收复合增长率：**受到终端 AI 处理器 IP 业务大幅衰退（从 2019 年 15%的营收贡献到 2020 年小于 4%的营收贡献），加上智能计算集群系统在手订单不足的影响（从 2019 年 67%的营收贡献到 2020 年小于 60%的营收贡献），我们预估寒武纪 2020 年营收同比增长 37%到 6 亿（公司在第二轮审核问询函之回复报告中预期 6—9 亿营收），但未来 5 年将逐步缩短与英伟达在 AI 芯片技术上的差距下，及其主要竞争者海思因为受到海思条款的管制（美国商务部工业安全局于 5/15/2020 宣布进一步限制华为海思在使用美国半导体设计软件 EDA 来设计半导体以及使用晶圆代工所使用的美国半导体设备来生产半导体，必须获取执照），我们预期寒武纪的云端及边缘运算端推理及训练芯片及加速卡反而有机会打入替代海思的华为供应链而大幅成长，从而带动寒武纪 2020-2024 年的营收复合增长率超过 50%。

图表 34：寒武纪产品营收，同比增长，占比变化图表的历史数据及预测

百万人民币	2018	2019	2020E	2021E	2022E	2023E	2024E	20-24 CAGR (%)
终端 AI 处理器 IP	116.66	68.77	17.0	20.0	30.0	40.0	50.0	31%
云端 AI 芯片及加速卡		78.88	90.0	180.0	290.0	600.0	1,000.0	83%
边缘 AI 芯片及加速卡			50.0	100.0	200.0	450.0	800.0	100%
智能计算集群系统		296.18	300.0	350.0	500.0	650.0	1000.0	35%
基础系统软件			50.00	60.00	75.00	110.00	150.00	32%
其他	0.36	0.10	100	120	105	150	250	26%
营业总收入（百万人民币）	**117.0**	**443.9**	**607.0**	**830.0**	**1,200.0**	**2,000.0**	**3,250.0**	52%
y/y 同比增长(%)	1391%	279%	37%	37%	45%	67%	63%	
各产品同比增长 (%)								

终端 AI 处理器 IP	1412%	-41%	-75%	18%	50%	33%	25%
云端 AI 芯片及加速卡			14%	100%	61%	107%	67%
边缘 AI 芯片及加速卡				100%	100%	125%	78%
智能计算集群系统			1%	17%	43%	30%	54%
基础系统软件				20%	25%	47%	36%
营收占比(%)							
终端 AI 处理器 IP	100%	15%	3%	2%	3%	2%	2%
云端 AI 芯片及加速卡	0%	18%	15%	22%	24%	30%	31%
边缘 AI 芯片及加速卡			8%	12%	17%	23%	25%
智能计算集群系统	0%	67%	49%	42%	42%	33%	31%
基础系统软件			8%	7%	6%	6%	5%

来源：国金证券研究所

- **毛利率的变化：**2017 到 2019 年度,公司综合毛利率分别为 99.96%、99.90%及 68.19%, 毛利率下降原因系 2019 年公司拓展了较低毛利的云端智能芯片及加速卡、智能计算集群系统业务。但就长期而言，我们估计 99.8% 毛利率的终端 AI 处理器 IP 及 50-60%毛利率的智能计算集群系统营收比重会持续下降，而 70-80%毛利率的云端 / 边缘 AI 芯片及加速卡，及基础系统软件营收比重会持续提升，这些因素应该会让寒武纪整体毛利率维持在 65-70%。而因为 AI 芯片可比标的不多，还有寒武纪仍处于扩大研发支出增加 IP 组合及扩大员工股权激励时期，所以我们拿英伟达及 CEVA 的毛利率做同业比较，而在智能计算集群系统业务方面，我们拿浪潮及中科曙光做比较，寒武纪高达 58%的系统毛利率应该无法长期维持。

图表 35：寒武纪各产品线毛利率比较

各产品毛利率 (%)	2018	2019	2020E	2021E	2022E	2023E	2024E
终端 AI 处理器 IP	100%	100%	99%	99%	99%	99%	99%
云端 AI 芯片及加速卡		78%	80%	79%	75%	75%	75%
边缘 AI 芯片及加速卡			75%	73%	70%	70%	70%
智能计算集群系统		58%	57%	55%	53%	51%	50%
基础系统软件			80%	80%	80%	80%	80%
综合毛利率 (%)	99.9%	68.2%	68.8%	68.0%	65.5%	66.5%	66.3%

来源：寒武纪，国金证券研究所

图表 36：寒武纪与相关同业毛利率比较

毛利率(%)	**2017**	**2018**	**2019**	**2020E**	**2021E**
寒武纪	100%	100%	68%	69%	68%
英伟达	60%	62%	62%	66%	67%
CEVA	81%	70%	77%	81%	90%
浪潮	11%	11%	12%	12%	12%
中科曙光	17%	18%	22%	22%	22%

来源：各公司财报，国金证券研究所

- **管理费用中的员工股权激励及研发费用的高低，决定亏损是否持续：**我们从销售（1,044.77 万），管理（6,326.88 万），研发（2.83 亿）费用中的职工薪酬中，可以初步计算出公司 2019 年每人及每位研发员工的平均薪资为 41-42 万元，跟国内半导体设计行业水平相当，但我们认为管理费用中的员工股权激励（股份支付）费用及研发费用的高低，决定公司营业亏损是否持续数年，我们以 2019 年为例，员工股权激励（股份支付）费用高达 9.44 亿，甚至超过当年度营收的 4.44 亿，还好的是员工股权激励（股份支付）费用不会是每年都发的经常性费用。但 2019 年研发费用就

是经常性费用，横跨各种新技术研发如边缘智能芯片，基础系统软件（推理及训练），云端硬件训练平台，云端 AI 训练及推理芯片，而 5.43 亿的研发费用也是超过当年度营收，这些偏高的管理及研发费用，造成公司 2019 年营业亏损达 11.8 亿，营业亏损占营收比率达 265%，我们估计未来两年持续亏损，到 2023 年才有机会扭亏转盈。

图表 37：寒武纪各营业费用比率及营业利润率预测

	2018	2019	2020E	2021E	2022E	2023E	2024E
税金及附加/营收	1%	1%	0.8%	0.7%	0.7%	0.7%	0.7%
销售费用/销售	5.3%	4.3%	4.8%	4.5%	4.5%	4.5%	4.5%
管理费用/销售	38%	238%	96%	66%	50%	25%	14%
研发费用/销售	205%	122%	82%	66%	50%	30%	22%
财务费用- 净额/销售	-2%	-1%	-1.6%	-1.3%	-1.5%	-1.4%	-1.4%
资产减值损失	0%	1%	0.3%	0.4%	0.3%	0.4%	0.3%
其他收益 (+)	59%	8%	8%	8%	8%	5%	3%
投资收益 (+)	53%	23%	16%	12%	8%	5%	3%
公允价值变动(+)	-1%	0%	-0.4%	-0.2%	-0.3%	-0.2%	-0.3%
营业费用率	247%	365%	182%	137%	104%	59%	40%
其他营业收益率	112%	31%	25%	21%	16%	10%	6%
营业利润率 (%)	**-35%**	**-265%**	**-89%**	**-48%**	**-23%**	**17%**	**33%**

来源：寒武纪招股说明书，国金证券研究所

- **政府补助扮演重要角色：**寒武纪拿到的政府补助主要有二种，一是与资产相关（政府文件规定用于购建或以其他方式形成长期资产的政府补助划分为与资产相关的政府补助）的是逐年上升，从 2017 年的 124 万，暴增到 2018 及 2019 年的 1,228 万及 1,701 万；二是与收益相关（与收益相关的政府补助，用于补偿以后期间的相关成本费用或损失的，确认为递延收益，在确认相关成本费用或损失的期间，计入当期损益或冲减相关成本）的，2018 年最高有 5,686 万，但降到 2019 年的 1,685 万。整体补助占营收比从 2017 年的 105%，2018 年的 59%，降到 2019 年的 7.6%。

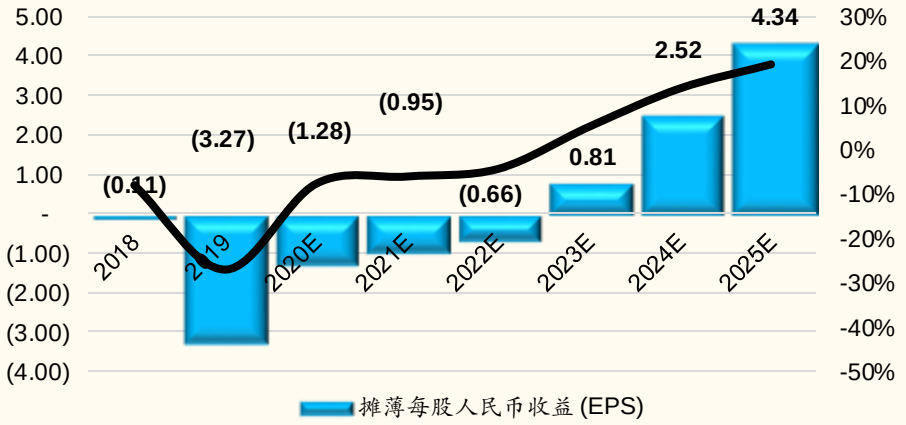
图表 38：政府补助（万元）

项目	2019 年度		2018 年度		2017 年度	
	金额	比例	金额	比例	金额	比例
与资产相关的政府补助	1,701.41	50.23%	1,228.06	17.76%	124.22	15.07%
与收益相关的政府补助	1,684.99	49.74%	5,685.95	82.23%	699.47	84.89%
代扣个人所得税手续费返还	0.90	0.03%	0.46	0.01%	0.32	0.04%
合计	**3,387.31**	**100.00%**	**6,914.47**	**100.00%**	**824.01**	**100.00%**

来源：寒武纪招股说明书，国金证券研究所

- **亏损应还会持续数年：**虽然寒武纪于 2020 年会因为科创板 IPO 增加近 11%的增资新股上市，但因为持续亏损，所以反而减少 2020 年摊薄每股损失，但估计未来 2023 年 EPS 将扭亏转盈达 0.81，2024 年达 2.52，ROE 也将从 2019 年的－27%，回升到 2024 年的 14.2%。

图表 39：寒武纪 EPS 与 ROE 比较表



来源：寒武纪招股说明书，国金证券研究所

2. 给予买入评级及 150 元目标价

不像其他的半导体设计行业在 IP，布线，制程工艺，市场，韧体指令集，软件，应用都相对成熟，人工智能芯片设计公司宛如新兴科技行业，未来 5 年全球 AI 芯片市场复合增长率高达 33%，国内 AI 芯片市场复合增长率高达 41%，而寒武纪在国内市场份额目前连 3%都不到（目前主要系英伟达，英特尔，赛灵思等传统芯片商在把持），全球份额连 1%都没有，而且必须不断投入大量研发费用在建立各种新智能算法 IP，召募大量研发软件及硬件的设计人才，5/7 纳米制程工艺产品研发，及多样光掩膜产品流片，大量使用 EDA 设计工具并采买各种现有 IP 组合，发展自己的韧体指令集及从云，边缘，到端的生态系软件，这些都让寒武纪必须不断的扩大研发经费，及利用各种员工股权激励计划来吸引业界人才，当然数年之后的技术壁垒叠加将让新进者困难重重。

所以短，中期用摊薄每股人民币收益 (EPS)，每股净资产 (Book value per share)及 ROE 来评估寒武纪都很困难。我们认为用价格对每股营收 （Sales per share）比或市销率来评估寒武纪，较为适当， 而寒武纪因为毛利率又比大多数市场上新兴科技公司高数倍，所以其 Price to sales 应该比市场高出甚多。我们以现在股价超过 1000 美元的电动车龙头 Tesla 特斯拉为例， 当时在 2010 年上市时，五年平均营业亏损率达 71%，毛利率只有 20-25%， P/S 平均达 20-30 倍，最高达 50 倍。谷歌在初上市的前五年，其平均 P/S 达 14 倍，五年最高平均达 19 倍，平均毛利率在 60%。英伟达在进入 AI 芯片后的 2017-2019 年，其平均 P/S 达 17 倍，最高平均达 24 倍，而平均毛利率为 61%。亚马逊在初上市的前三年，其平均 P/S 达 45 倍，五年最高平均达 80 倍，平均毛利率在 20%，营业亏损率达 23%。

图表 40：寒武纪与新兴科技公司利润率及市销率比较

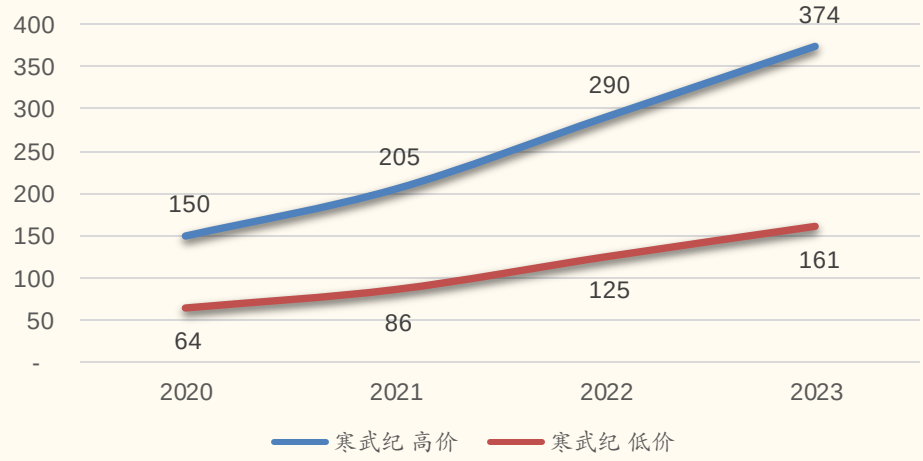
	特斯拉	谷歌	英伟达	亚马逊	寒武纪
年份	2010-2014	2005-2009	2017-2019	1997-1999	2017-2020E
平均毛利率 %	23%	60%	61%	20%	84%
营业利润率 %	-71%	33%	31%	-23%	-182%
P/S 高点	50.5	18.7	23.7	79.7	N/A
P/S 平均	28.8	13.8	17.1	45.0	N/A
P/S 低点	7.0	8.9	10.5	10.3	N/A

来源：寒武纪招股说明书，彭博社，国金证券研究所

除了可预期的未来营收大幅增长，寒武纪也是国内首家上市的人工智能芯片公司，又有中科院的支持及与资产及收益相关的政府补助，不但毛利率将维持在 65%以上，又享受科创板公司股数流通少及闭锁期的溢价，所以我们**给予买入评级**，我们认为在未来 12 个月的闭锁期结束之前，公司平均 P/S 区间将达 40-60 倍，我们目前用 2022 年的 3 元每股营收给予 50 倍 P/S 给估值，我们

给寒武纪**一年的目标价为 150 元**，但在少数股东一年及经营团队持股三年闭锁期结束后，P/S 区间将逐年向下调整。但对于成长性较高的半导体设计行业而言，投资人给估值多看长期营收获利增长趋势，至少看未来 1-2 年的营收及获利，而至于成长性更高的人工智能芯片设计行业而言，投资人给估值至少看未来 2-3 年的营收及获利，这是为什么我们用 2022 年的每股营收预期来定目标价。

图表 41：寒武纪股价高低区间预测



来源：中微半导体，国金证券研究所

八、主要行业及公司面对的风险

1.终端 AI 处理器 IP 业务减少的风险：刚才提到过从 2019 年开始华为海思自研 AI 芯片，未与寒武纪继续合作，而公司短期内难以开发同等规模的大客户，加上国际巨头英伟达，英特尔，高通，联发科，ARM 的竞争，因此我们预期其 2020 年公司终端智能处理器 IP 授权业务收入将继续下滑，预期未来这些业务占比将从 2017-2018 的 98-100%，降低到去年的 15%，到未来几年低于 4%的营收。

2.密切关注智能计算集群系统事业风险：在寒武纪招股书的最新资料显示，公司智能计算集群系统方面的在手订单包括横琴先进智能计算平台(二期) 的第二批供货硬件设备，授权软件，合同金额 仅剩下 1.86 亿元，而上半年营收贡献连 20 万都没有，除非我们看到在下半年在手订单大幅回流，今年此业务营收贡献可能不到 60%，甚至连 50%占比都可能有问题。且就长期而言，寒武纪这事业群直接面对终端使用者客户，如地方数据中心，行业企业和科研机构等，就需要庞大分散的客户群来稳定在手订单及营收，否则将带给投资人相当大的营收及获利上下大幅波动的风险。

3.竞争风险：寒武纪在云端 AI 芯片除了目前英伟达的 V100, A100，海思的昇腾 910 系列的竞争之外，英特尔，超微，赛灵思，比特大陆也陆续推出相关产品，而提供 AI 处理器 IP 的厂商也从 ARM， 扩大到 CEVA, Cadence 等等。而这些竞争都会带来价格及毛利率的压力。

4.现金流短缺风险：寒武纪在持续亏损数年后，还要在设计及制程工艺大力追赶海思及英伟达，寒武纪必须不断烧钱雇用人才，投入庞大研发及流片费用，举例而言，寒武纪若要加快单及双精度浮点运算速度，决定从现在的 16nm 跳到 7nm 设计，但光是一个新产品设计流片光掩膜成本就会从 4-5 百万美元，增加超过一倍到 1100 万美元。这次 IPO 定增将筹措 25.8 亿人民币，当然对在手 5 亿人民币现金不到的寒武纪而言，具有稳定财务的作用，但要是年度亏损持续，两年后，我们不排除寒武纪将卷土重来市场融资，继续摊薄现有股权结构。而且公司在审核问询函之回复报告书中也确认，未来三年内除募集资金外，仍需 30-36 亿投入研发项目。

5.进入实体清单的风险：自从美国商务部将人工智能相关公司科大讯飞，依图，旷视，商汤，云天励飞，云从等放入采购美国技术需要美国商务部许可的实体

清单后，我们担心上市后的寒武纪（持续采用美国 Synopsys，Cadence 的 EDA 工具）将成为下一波的目标之一。

附录：三张报表预测摘要

损益表（单位：百万元）	2018	2019	2020	2021	2022	2023
营业总收入	**117.03**	**443.94**	**607.00**	**830.00**	**1,200.00**	**2,000.00**
营业成本	0.12	141.23	189.67	265.90	414.30	670.90
营业毛利	116.91	302.71	417.33	564.10	785.70	1,329.10
营业费用	158.24	1,481.03	956.46	965.60	1,061.81	987.80
营业利润	**-41.33**	**-1,178.32**	**-539.13**	**-401.50**	**-276.11**	**341.30**
折旧	6.65	22.61	38.72	96.90	142.20	173.55
摊销	10.35	30.85	40.00	50.00	58.00	68.00
折旧前净利	-24.33	-1,124.86	-460.41	-254.60	-75.91	582.85
利息	0.00	0.00	0.00	0.00	0.00	0.00
其他	0.28	-0.81	-0.26	-0.53	-0.40	-0.47
投资收入	0.00	0.00	0.00	0.00	0.00	0.00
特例	0.00	0.00	0.00	0.00	0.00	0.00
税前利润总额	-41.05	-1,179.13	-539.39	-402.04	-276.50	340.83
所得税/少数股东损益	0.00	-0.14	-26.97	-20.10	-13.83	17.04
净利润	-41.05	-1,178.99	-512.42	-381.94	-262.68	323.79
优先股	0.00	0.00	0.00	0.00	0.00	0.00
净利润	**(41.0)**	**(1,179.0)**	**(512.4)**	**(381.9)**	**(262.7)**	**323.8**
获利能力比率(%)						
毛利率	100%	68%	69%	68%	65%	66%
营业利润率	-35%	-265%	-89%	-48%	-23%	17%
折旧前净利率	-21%	-253%	-76%	-31%	-6%	29%
税前利润率	-35%	-266%	-89%	-48%	-23%	17%
净利率	-35%	-266%	-84%	-46%	-22%	16%
净资产收益率	-8%	-27%	-8%	-6%	-5%	5%
投入资本回报率	-7%	-27%	-8%	-6%	-5%	5%
同比成长率(%)						
营业总收入	1392%	279%	37%	37%	45%	67%
营业毛利	1391%	159%	38%	35%	39%	69%
营业利润	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
折旧前净利	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
税前利润总额	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
净利润	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
收益评估(CNY$)						
每股营业收入	0.33	1.23	1.52	2.07	3.00	5.00
每股折旧前净利	(0.11)	(3.27)	(1.35)	(1.00)	(0.69)	0.85
每股利润	(0.11)	(3.27)	(1.28)	(0.95)	(0.66)	0.81
每股账面净值	1.42	12.10	16.06	15.11	14.45	15.26
每股经营现金流量	1.68	(16.12)	(0.66)	(0.08)	0.50	0.78
每股自由现金流量	1.44	(16.63)	(1.19)	(0.68)	(0.16)	0.06
资产负债表（单位：百万元）	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
货币资金	1,354.37	383.31	2,413.47	2,140.40	2,076.27	2,100.92
应收票据及应收账款	32.64	64.61	88.34	113.70	147.95	246.58
存货	5.15	51.07	63.22	88.63	118.37	167.73
其他流动资产	1,561.78	3,952.26	3,793.75	3,608.70	3,428.57	3,809.52

流动资产合计	**2,953.94**	**4,451.24**	**6,358.78**	**5,951.43**	**5,771.16**	**6,324.74**
固定资产	44.90	92.96	210.45	277.57	307.60	314.88
其他非流动资产	42.61	122.91	140.00	167.00	200.00	240.00
长期股权投资	-	1.36	1.36	1.36	1.36	1.36
非流动资产合计	**87.51**	**217.23**	**351.81**	**445.93**	**508.96**	**556.24**
资产总计	**3,041.45**	**4,668.47**	**6,710.59**	**6,397.36**	**6,280.12**	**6,880.98**
短期借款	-	-	-	-	-	-
应付票据及应付账款	22.40	124.91	129.91	145.70	198.64	275.71
其他流动负债	2,425.37	113.06	154.59	207.50	300.00	500.00
流动负债合计	**2,447.77**	**237.98**	**284.50**	**353.20**	**498.64**	**775.71**
长期借款	-	-	-	-	-	-
其他长期借款	83.03	74.02	-	-	-	-
非流动负债合计	**83.03**	**74.02**	**-**	**-**	**-**	**-**
负债合计	**2,530.81**	**311.99**	**284.50**	**353.20**	**498.64**	**775.71**
实收资本(或股本)	1.13	360.00	400.10	400.10	400.10	400.10
资本公积金	932.37	4,851.12	7,393.06	7,393.06	7,393.06	7,393.06
未分配利润	(422.86)	(854.64)	(1,367.06)	(1,749.00)	(2,011.68)	(1,687.88)
少数股东权益	-	-	-	-	-	-
所有者权益合计	**510.64**	**4,356.48**	**6,426.09**	**6,044.16**	**5,781.48**	**6,105.27**
负债和所有者权益总计	**3,041.45**	**4,668.47**	**6,710.59**	**6,397.36**	**6,280.12**	**6,880.98**
资本总额比率(%)						
负债比率	0%	0%	0%	0%	0%	0%
所有者权益比率	100%	100%	100%	100%	100%	100%
债务股本比	0%	0%	0%	0%	0%	0%
净债务股本比	Net Cash	Net Cash	Net Cash	Net Cash	Net Cash	Net Cash
经营比率(x)						
应收账款月数	3.3	1.7	1.7	1.6	1.5	1.5
存货月数	527.4	4.3	4.0	4.0	3.4	3.0
存货周转率	0.0	2.8	3.0	3.0	3.5	4.0
固定资产周转率(x)	2.6	4.8	2.9	3.0	3.9	6.4
总资产周转率(x)	0.0	0.1	0.1	0.1	0.2	0.3
现金流量表（单位：百万元）	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
净利润	-41.05	-1,179.13	-512.42	-381.94	-262.68	323.79
折旧费用	6.65	22.61	38.72	96.90	142.20	173.55
摊销费用	10.35	30.85	40.00	50.00	58.00	68.00
其它	-65.37	794.82	0.00	0.00	0.00	0.00
现金流量净额	**-89.42**	**-330.85**	**-433.70**	**-235.03**	**-62.48**	**565.35**
营运资金变动	33.93	129.05	169.14	202.98	261.58	-251.86
经营活动现金净流	**-55.49**	**-201.80**	**-264.56**	**-32.05**	**199.10**	**313.49**
资本开支	-74.85	-156.21	-156.21	-164.02	-172.23	-180.84
其他	1.29	-28.67	-57.09	-77.00	-91.00	-108.00
投资	-1,147.85	-2,283.60	0.00	0.00	0.00	0.00
投资活动现金净流	**-1,221.41**	**-2,468.48**	**-213.30**	**-241.02**	**-263.23**	**-288.84**
债权募资	0.00	0.00	0.00	0.00	0.00	0.00
其他长期债权募资	0.00	0.00	-74.02	0.00	0.00	0.00
股权募资	2,405.07	1,699.95	2,582.04	0.00	-0.00	-0.00
筹资活动现金净流	**2,405.07**	**1,699.95**	**2,508.02**	**0.00**	**-0.00**	**-0.00**

现金净流量	**1,127.18**	**-971.07**	**2,030.16**	**-273.07**	**-64.13**	**24.65**
自由现金净流量	-1,276.90	-2,670.28	-477.86	-273.07	-64.13	24.65

来源：公司年报、国金证券研究所

市场中相关报告评级比率分析

日期	一周内	一月内	二月内	三月内	六月内
买入	0	0	0	0	0
增持	0	0	0	0	0
中性	0	0	0	0	0
减持	0	0	0	0	0
评分	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

来源：朝阳永续

市场中相关报告评级比率分析说明：

市场中相关报告投资建议为“买入”得 1 分，为“增持”得 2 分，为“中性”得 3 分，为“减持”得 4 分，之后平均计算得出最终评分，作为市场平均投资建议的参考。

最终评分与平均投资建议对照：

1.00 =买入； 1.01~2.0=增持 ； 2.01~3.0=中性
3.01~4.0=减持

投资评级的说明：

买入：预期未来 6—12 个月内上涨幅度在 15%以上；
增持：预期未来 6—12 个月内上涨幅度在 5%—15%；
中性：预期未来 6—12 个月内变动幅度在 -5%—5%；
减持：预期未来 6—12 个月内下跌幅度在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证，对由于该等问题产生的一切责任，国金证券不作出任何担保。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整。

本报告中的信息、意见等均仅供参考，不作为或被视为出售及购买证券或其他投资标的邀请或要约。客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告反映编写分析员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，且收件人亦不会因为收到本报告而成为国金证券的客户。

根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；非国金证券 C3 级以上（含 C3 级）的投资者擅自使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

此报告仅限于中国大陆使用。

上海

电话：021-60753903

传真：021-61038200

邮箱：researchsh@gjzq.com.cn

邮编：201204

地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 7 楼

北京

电话：010-66216979

传真：010-66216793

邮箱：researchbj@gjzq.com.cn

邮编：100053

地址：中国北京西城区长椿街 3 号 4 层

深圳

电话：0755-83831378

传真：0755-83830558

邮箱：researchsz@gjzq.com.cn

邮编：518000

地址：中国深圳福田区深南大道 4001 号时代金融中心 7GH