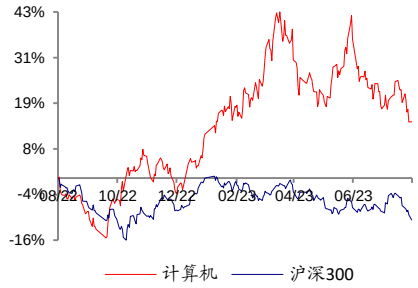
计算机行业

独立自主的 AI 系统级计算平台是国产 AI 芯片构建生态壁垒的关键

行业评级	买入
前次评级	买入
报告日期	2023-08-22

相对市场表现



核心观点：

- **系统级 AI 计算平台是提升 AI 芯片算力利用率，培养用户生态的关键。**影响 AI 芯片计算能力的因素除了硬件层面的芯片制程、内存、带宽等，还包括调用各硬件资源的系统级软件计算平台。AI 芯片厂商开发的系统计算平台不仅仅有效提升各家 AI 芯片产品的算力利用率，还为各类 AI 应用开发提供了丰富的函数库，提供开发者简便易用的开发环境。
- **英伟达的 CUDA 计算平台是主流 AI 应用开发平台。**通过对比各公司开发的 AI 计算平台，我们发现英伟达的 CUDA 开发时间最早，积累的开发者数量最多。英伟达一方面依靠 AI 芯片优异的硬件性能快速获客，另一方面持续拓展 CUDA 的算法覆盖面，不断巩固客户群体。英伟达通过“滚雪球”式的软硬件协同创新，将其 AI 芯片的市场份额不断扩大，并构建起了深厚的生态壁垒。
- **寒武纪的 Neuware 对 AI 算法覆盖面全面，兼容性强。**Neuware 不仅支持 CV、NLP、智能推荐等主流商用 AI 算法，还兼容第三方的训练框架（例如百度飞桨 Paddle Lite），在向互联网公司等商业客户拓展用户生态时更具优势。依托于 Neuware，开发者可完成从云端到边缘端、从模型训练到推理部署的全部流程，提升 AI 芯片的算力利用率。
- **在中美科技领域竞争日益激烈的背景下，自主研发架构更具优势。**在英伟达 AI 芯片对中国出口前景具有不确定性的背景下，以互联网为代表的 AI 计算下游厂商适配和采用国产化的系统软件（寒武纪的 Neuware，华为的 CANN）的动力将大大增加。我们认为，独立于 CUDA 的自研计算平台对于我国芯片产业的长远发展至关重要。国产 AI 芯片厂商独立研发、自主迭代的 AI 计算平台更加具备长久的持续发展能力。
- **寒武纪的中立属性在行业格局中具有独特价值，更有利于其用户生态的拓展。**寒武纪提供的产品以 AI 算力基础设施为主，不涉足 AI 应用领域，与行业参与者更多构成的是互补关系而不是竞争关系。寒武纪的中立属性使其保持智能化升级中赋能者的定位，与产业链上下游形成合作共赢的关系，这更加有利于其 Neuware 计算平台用户生态的拓展。
- **投资建议：**推荐拥有自研 AI 计算平台的寒武纪，长期生态壁垒较高。
- **风险提示：**科技巨头在 AI 计算平台领域长期积累，生态壁垒较高，国产 AI 芯片公司中短期突破难度较大；AI 计算平台对于 AI 算法覆盖面要求较高，前期投入较大与生态培育不及预期的风险；AI 芯片存在供应链不稳定的风险。

分析师： 刘雪峰
SAC 执证号：S0260514030002
SFC CE No. BNX004
021-38003675
gfliuxuefeng@gf.com.cn

分析师： 周源
SAC 执证号：S0260523040001
0755-23948351
shzhouyuan@gf.com.cn

相关研究：

计算机行业:积极配置正当时	2023-08-20
计算机行业:积极配置穿越下游景气度周期的优质板块个股	2023-08-13
计算机行业:金融 IT 只是开始，基本面托底迭加外部环境因素催化将持续推动行业表现	2023-08-06

联系人：许晟榕 021-38003800
xushengrong@gf.com.cn

重点公司估值和财务分析表

股票简称	股票代码	货币	最新收盘价	最近报告日期	评级	合理价值（元/股）	EPS(元)		PE(x)		EV/EBITDA(x)		ROE(%)	
							2023E	2024E	2023E	2024E	2023E	2024E	2023E	2024E
寒武纪-U	688256.SH	CNY	151.40	2023/05/02	增持	251.75	-1.70	-1.00	-	-	-	-	-15.90	-10.40

数据来源：Wind、广发证券发展研究中心

备注：表中估值指标按照最新收盘价计算

目录索引

图表索引

投资要点：

系统级AI计算平台是提升AI芯片算力利用率，培养用户生态的关键。影响AI芯片计算能力的因素除了硬件层面的芯片制程、内存、带宽等，还包括调用各硬件资源的系统级软件计算平台。AI芯片厂商开发的系统计算平台不仅仅有效提升各家AI芯片产品的算力利用率，还为各类AI应用开发提供了丰富的函数库，提供开发者简便易用的开发环境。

英伟达的CUDA计算平台是主流AI应用开发平台。通过对比各公司开发的AI计算平台，我们发现英伟达的CUDA开发时间最早，积累的开发者数量最多。英伟达一方面依靠AI芯片优异的硬件性能快速获客，另一方面持续拓展CUDA的算法覆盖面，不断巩固客户群体。英伟达通过“滚雪球”式的软硬件协同创新，将其AI芯片的市场份额不断扩大，并构建起了深厚的生态壁垒。

CANN（Compute Architecture for Neural Networks）是华为针对AI场景推出的异构计算架构。CANN构建了从上层深度学习框架到底层AI芯片的桥梁，提供多层次的编程接口，全面支持昇思MindSpore、飞桨PaddlePaddle、PyTorch、TensorFlow、Caffe等主流AI框架，提供900多种优选模型覆盖众多典型场景应用，兼容多种底层硬件设备，提供异构计算能力，支持用户快速构建基于昇腾平台的AI应用。

寒武纪Cambricon Neuware是针对其云、边、端的AI芯片打造的软件开发平台。为了加快用户端到端业务落地的速度，减少模型训练研发到模型部署之间的繁琐流程，Neuware整合了训练和推理的全部底层软件栈，包括底层驱动、运行时库（CNRT）、算子库（CNNL）以及工具链等，将Neuware和深度学习框架Tensorflow、Pytorch深度融合，实现训推一体。依托于Cambricon Neuware，开发者可完成从云端到边缘端、从模型训练到推理部署的全部流程，提升AI芯片的算力利用率。

海光DCU全面兼容ROCm GPU计算生态。ROCm（Radeon Open Compute Platform）是AMD基于开源项目的GPU计算生态系统，支持多种编程语言、编译器、库和工具，以加速科学计算、人工智能和机器学习等领域的应用。ROCm还支持多种加速器厂商和架构，提供了开放的可移植性和互操作性。

选择兼容CUDA平台的AI芯片虽然短期在产品推广方面具有一定便利性，但是难以形成长期的生态壁垒。CUDA作为英伟达AI计算平台已经得到广泛应用，兼容CUDA平台的AI芯片具有强大的生态系统，并且可获得社区支持，利用现有的AI平台和共享计算资源，有利于其AI芯片产品的推广。但是长期来看，仍然存在以下问题：

1. 软硬件适配度不够：兼容CUDA的AI计算平台存在与其自研的AI芯片适配度不够，导致AI芯片算力利用率不足甚至性能衰减的问题。

2. 迭代速度较慢：CUDA更新节奏较快，选择兼容CUDA的AI计算平台存在由于投入研发资源不充足导致其迭代跟不上CUDA的更新节奏从而影响芯片性能的问题。

3. 客户粘性不足等问题：由于其开发环境与CUDA的相似度较高，开发者难以通过长期使用形成对其的粘性。

在中美科技领域竞争日益激烈的背景下，自主研发架构更具优势。在英伟达AI芯片对中国出口前景具有不确定性的背景下，以互联网为代表的AI计算下游厂商适配和采用国产化的系统软件（寒武纪的Neuware，华为的CANN）的动力将大大增加。我们认为，独立于CUDA的自研计算平台对于我国芯片产业的长远发展至关重要。国产AI芯片厂商独立研发、自主迭代的AI计算平台更加具备长久的持续发展能力。

自主研发的计算平台具有选择AI芯片技术路线的灵活性，长期发展空间更宽广。选择兼容CUDA的AI芯片虽然在短期内可以获得产品快速推广的机会，但在长期却失去了可以自由选择其他技术路线实现性能突破的机会。我们认为，在以英伟达为代表的GPU在国内市场受到限制的情况下，各厂商或加快探索除了GPU以外的其他技术路线。在此背景下，不依赖于GPU技术路线的自主研发的AI计算平台具有自由探索和选择其他技术路线的灵活性，长期发展空间更宽广。

国产AI计算平台持续迭代和推广，用户生态建设已具有一定基础。华为的昇腾芯片和寒武纪的思元系列芯片是国内较早开始推广，并在商业落地上取得一定领先优势的产品。国产AI芯片的计算平台正在依托各地AIDC的算力在各行业的应用而实现用户生态的快速拓展。各地建成的人工智能计算中心的AI算力租用给当地企业使用，在这一过程中，华为的CANN和寒武纪的Neuware的用户生态得到快速拓展。我们认为，最先成功商业化的公司将会扩大对追赶者的优势，因为最终用户不大会接受同时采用诸如四五种以上不同的芯片计算平台体系。我们看好此前有相关经验的华为昇腾CANN和寒武纪Neuware生态的发展前景。

华为提供全栈AI解决方案，在部分场景与AI应用公司构成竞争关系。华为在AI产业链中扮演的角色不仅仅作为底层软硬件基础设施提供商，还针对部分场景开发了具体的AI应用。以AI大模型为例，华为不仅提供底层算力（昇腾AI芯片）、训练框架（Mindspore昇思）和基础大模型（盘古大模型），还开发了行业级大模型（盘古金融大模型、盘古制造大模型等）以及针对场景的AI应用（先导药物筛选、传送带异物检测等）。这与部分AI应用提供商构成同业竞争的关系，其发展会受到一定限制。我们认为，华为在各场景中提供全栈AI解决方案的战略会影响其基础AI算力产品以及计算平台CANN的商业拓展。

寒武纪的中立属性在行业格局中具有独特价值，更有利于其用户生态的拓展。寒武纪提供的产品以AI算力基础设施为主，不涉足AI应用领域，与行业参与者更多构成的是互补关系而不是竞争关系。Neuware不仅支持CV、NLP、智能推荐等主流商用AI算法，还兼容第三方的训练框架（例如百度飞桨Paddle Lite），在向互联网公司等商业客户拓展用户生态时更具优势。寒武纪的中立属性使其保持智能化升级中赋能者的定位，与产业链上下游形成合作共赢的关系，这更加有利于其Neuware计算平台用户生态的拓展。

投资建议：推荐拥有自研AI计算平台的寒武纪，长期生态壁垒较高。

风险提示：科技巨头在AI计算平台领域长期积累，生态壁垒较高，国产AI芯片公司中短期突破难度较大；AI计算平台对于AI算法覆盖面要求较高，前期投入较大与生态培育不及预期的风险；AI芯片存在供应链不稳定的风险。

一、AI 芯片的系统计算平台是用户生态培育的关键

系统级AI计算平台是提升AI芯片算力利用率，培养用户生态的关键。影响AI芯片计算能力的因素除了硬件层面的芯片制程、内存、带宽等，还包括调用各硬件资源的系统级软件计算平台。AI芯片厂商开发的系统计算平台不仅仅有效提升各家AI芯片产品的算力利用率，还为各类AI应用开发提供了丰富的函数库，提供开发者简便易用的开发环境。以英伟达为例，其开发的CUDA平台，自2007年推出后持续更新，已吸引了大量AI应用开发者使用，形成了庞大的用户生态。此篇报告将深入分析各厂商开发的AI计算平台的功能、效果并前瞻分析未来的发展趋势。

表 1：各公司系统级AI计算平台发展历史对比

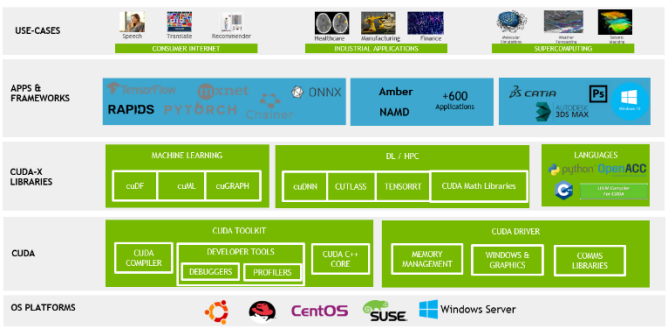
名称	开发厂商	推出时间	截止2023年8月的版本号	主要函数库
CUDA	英伟达	2007.02	V12.2	并行计算库、深度学习库、图像视频库
CANN	华为	2018.09	7.0.RC1.alpha002（社区版）CANN 6.3.RC2(商用版)	运行时库、视觉预处理和人工智能预处理库、华为集合通信库等
Neuware	寒武纪	2017.11	-	AI加速库、通信库、视觉库
ROCm	AMD	2016.11	AMD ROCm Platform 5.6.0	数学库、并行计算库、图像视频库、深度学习库、通信库

数据来源：各公司官网，广发证券发展研究中心

（一）CUDA：释放英伟达 GPU 算力的系统级 AI 计算平台

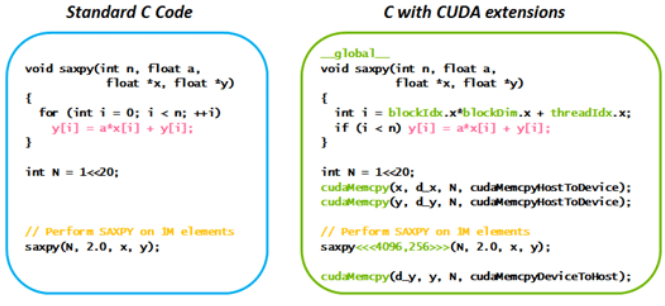
英伟达开发的CUDA系统计算框架构建了GPU和开发者之间的桥梁。CUDA（Compute Unified Device Architecture）是英伟达公司于2007年推出用于释放GPU并行计算能力和增强通用性的系统级计算平台。CUDA直接对接GPU的物理层，将海量数据分配给多个线程上分别处理，再调用GPU的多核心（计算单元）进行并行计算。为方便开发者更好的调用GPU的计算能力，CUDA也提供了一系列封装好的函数库和API，可在芯片物理层上实现指令级和算子的直接调用。总体而言，CUDA一方面可高效利用底层AI芯片的算力，另一方面给予开发者便捷的开发环境，满足了开发者高效利用AI底层算力的需求。

图 1：英伟达CUDA在IT系统中的功能定位



数据来源：英伟达官网，广发证券发展研究中心

图 2：基于 CUDA 函数库的 C 语言界面对比



数据来源：英伟达官网，广发证券发展研究中心

CUDA不仅仅是AI算法开发的工具链，还是调用底层计算资源的系统平台。与一般的软件工具不同，CUDA是更加贴近芯片物理层的系统平台，其提供的封装函数可以实现对于内存、计算单元（算术逻辑单元）、数据传输速率（带宽）等底层算力资源的调用。因此，CUDA在设计之初的产品定位是给程序员提供对于硬件性能优化和调试的功能。后续，随着CUDA版本的升级迭代，其对于底层硬件资源调用能力持续增强。例如，CUDA 5.0版本中新增的动态并行技术，可以根据数据处理量在内核中动态调用多条线程，减少单一线程上的工作负载，从而保证了不同线程上的负载均衡。

表 2：英伟达CUDA对于硬件资源加速的扩展功能

功能支持	计算能力（版本）								
	1.0, 1.1	1.2, 1.3	2.x	3.0	3.2	3.5, 3.7, 5.x, 6.x, 7.0, 7.2	7.5	8.x	9.0
线程束表决函数	否	是							
线程束表决函数	否		是						
内存栅栏函数									
同步函数									
曲面函数									
线程块的三维网格									
线程束统筹函数	否			是					
统一内存编程									
线程访问通道转换	否				是				
动态并行性	否					是			
统一数据路径	否						是		
硬件加速异步复制	否							是	
硬件加速异步到达/同步等待栅栏									
对缩减操作的线程束级支持									
二级高速缓冲存储器驻留管理									
用于加速动态编程的 DPX 指令	否								是
分布式共享内存									
线程块群集									
张量储存加速器（TMA）单元									

数据来源：英伟达官网，广发证券发展研究中心

CUDA提供了易用友好的开发环境。CUDA提供了丰富的库函数和工具，方便程序员对于各类AI算法进行开发。经过多年的拓展，CUDA不仅兼容主流的AI训练框架（Tensorflow、Pytorch等），对各类AI算法（DLRM、Resnet-50、BERT等）的覆盖面也更加广阔。通过CUDA，程序员可以高效利用GPU的大规模并行计算能力来加速各种计算密集型任务，包括图像和视频处理、物理模拟、金融分析、生命科学等领域。我们认为，CUDA经过长期积累可提供对于各类AI算法开发的函数库和工具链更加丰富，对各类算法覆盖面更加广泛，在易用性方面具有一定优势。

表 3：英伟达CUDA-X的函数库

函数库类型	函数库	功能介绍
数学库	cuBLAS	基本线性代数库
	cuFFT	用于快速傅里叶变换的GPU加速库
	CUDA Math Library	标准数学函数库
	cuRAND	随机数生成函数库
	cuSOLVER	密集和稀疏求解器
	cuSPARSE	稀疏矩阵函数库
	cuTENSOR	张量线性代数库
	AmgX	用于模拟和隐式非结构化方法的线性求解器
并行算法库	Thrust	该库包含常用的C++并行算法和数据结构
图像和视频库	nvJPEG	用于JPEG格式数据解码的函数库
	NVIDIA Performance Primitives	提供GPU加速的图像、视频和信号处理功能
	NVIDIA Video Codec SDK	一套完整的API、示例和文档，用于执行硬件加速的视频编码和解码
	NVIDIA Optical Flow SDK	用于计算图像之间像素的相对运动
通信库	NVSHMEM	根据OpenSHMEM标准的并行计算扩展包，可有效扩展多个GPU互联时的内存
	NCCL	用于多GPU、多节点通信的开源代码库，可在保持低延迟的同时更大限度地增加带宽
深度学习库	NVIDIA cuDNN	针对常用深度神经网络算法的标准函数库和工具
	NVIDIA TensorRT	用于生产部署的高性能深度学习推理优化器和运行时
	NVIDIA Riva	用于各种情景下互动性强的AI对话应用的函数库
	NVIDIA DeepStream SDK	实时流分析工具包，用于基于AI的视频数据和多传感器数据的处理
	NVIDIA DALI	开源代码库，用于解码和增强图像和视频数据，并加速相关深度学习应用
合作开发的库	OpenCV	开源代码库，用于计算机视觉、图像处理和机器学习
	FFmpeg	开源代码多媒体框架，包含用于音频和视频处理的插件库
	ArrayFire	开源代码库，用于矩阵、信号和图像处理
	MAGMA	在异构的环境中提供线性代数相关算法处理工具
	IMSL Fortran Numerical Library	开源代码库，包含数学、信号和图像处理以及统计专用的函数
	Gunrock	图形处理函数库
	CHOLMOD	稀疏求解器函数库
	Triton Ocean SDK	在游戏、仿真等应用中应用于海洋与水体仿真的函数库
	CUVIlib	用于加速医学、工业等领域成像应用的函数库

数据来源：英伟达官网，广发证券发展研究中心

CUDA与英伟达AI芯片强绑定，随着AI芯片迭代而持续升级。英伟达在每一代芯片架构升级的过程中，添加了一些新的特性来提升对于AI算法的计算效率。针对这些新的特性，CUDA也不断丰富SDK中的函数库从软件层面进一步对AI算法进行加速。例如，英伟达在2017年推出Volta架构AI芯片产品的时候首次引入了Tensor Core，其将单一维度的数字运算扩展到二维度的矩阵运算，从而提升单次运算能力。在软件层面，CUDA 9.0版本则新增了各类矩阵运算操作符，对于矩阵的加载、相乘、累加都有很好的处理效果。因此，用户可以通过CUDA更好的发挥硬件层面的新特性，从而扩展产品的应用场景。

表 4：英伟达不同版本的CUDA支持历代架构AI芯片的对应关系

CUDA SDK 版本	Tesla	Fermi	Kepler (early)	Kepler (late)	Maxwell	Pascal	Volta	Turing	Ampere	Ada Lovelace	Hopper
1.0	1.0 – 1.1										
1.1	1.0 – 1.1+x										
2.0	1.0 – 1.1+x										
2.1 - 2.3.1	1.0 – 1.3										
3.0 - 3.1	1.0 –	2.0									
3.2	1.0 –	2.1									
4.0 - 4.2	1.0 –	2.1+x									
5.0 - 5.5	1.0 –			3.5							
6.0	1.0 –			3.5							
6.5	1.1 –				5.x						
7.0 - 7.5		2.0 –			5.x						
8.0		2.0 –				6.x					
9.0 - 9.2			3.0 –				7.0				
10.0 - 10.2			3.0 –					7.5			
11.0				3.5 –					8.0		
11.1 - 11.4				3.5 –					8.6		
11.5 - 11.7.1				3.5 –					8.7		
11.8				3.5 –							9.0
12.0 - 12.2					5.0 –						9.0

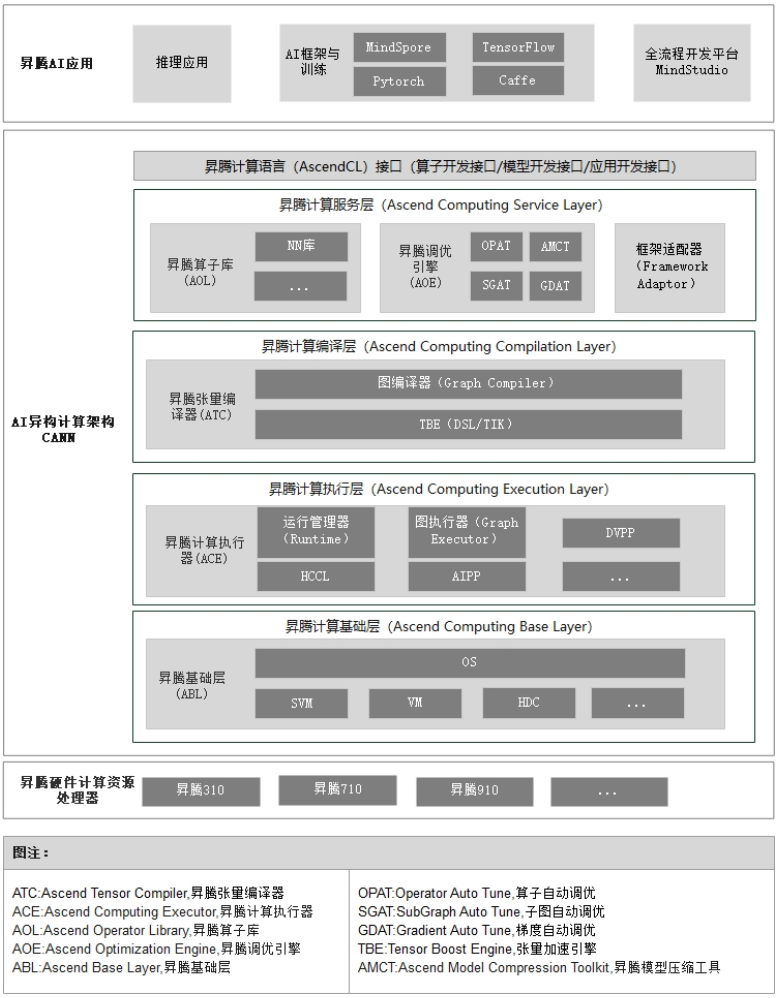
数据来源：英伟达官网，广发证券发展研究中心

CUDA构建了英伟达长而深的生态护城河。英伟达针对数据中心场景的大数据和AI功能的开发起步早，积累深厚。自2007年，英伟达推出CUDA以来，至今已迭代了12个版本。在多年市场推广下，CUDA已成为AI算法开发主流的系统平台，具有较高的生态壁垒。截止2023年4月，海内外主要科技公司超过百万的开发人员都是基于CUDA开发AI算法。硬件层面的架构升级吸引用户采购新一代AI芯片更新换代，软件层面丰富的工具和易用的开发环境则培养了用户粘性。在长期的积累下，CUDA形成的生态壁垒较好的巩固英伟达的市场份额和龙头地位。

（二）CANN：华为拓展昇腾 AI 芯片生态的关键

CANN（Compute Architecture for Neural Networks）是华为针对AI场景推出的异构计算架构。CANN构建了从上层深度学习框架到底层AI芯片的桥梁，提供多层次的编程接口，全面支持昇思MindSpore、飞桨PaddlePaddle、PyTorch、TensorFlow、Caffe等主流AI框架，提供900多种优选模型覆盖众多典型场景应用，兼容多种底层硬件设备，提供异构计算能力，支持用户快速构建基于昇腾平台的AI应用。

图 3：CANN架构图



数据来源：昇腾社区，智能计算芯世界，广发证券发展研究中心

计算架构方面，CANN被抽象为五大层级，分别为计算语言接口、计算服务层、计算编译引擎、计算执行引擎和计算基础层，共同构建高效而简捷的计算平台。CANN的优势是兼容性较强，可在不同的硬件、OS和AI开发框架的异构环境中发挥较好的计算性能，实现端边云多设备的协同，赋能各场景的AI开发。

表 5：CANN各层级功能介绍

架构层	功能
昇腾计算语言接口	昇腾计算语言（Ascend Computing Language，AscendCL）接口是昇腾计算开放编程框架，是对低层昇腾计算服务接口的封装。它提供 Device（设备）管理、Context（上下文）管理、Stream（流）管理、内存管理、模型加载与执行、算子加载与执行、媒体数据处理、Graph（图）管理等 API 库，供用户开发人工智能应用调用。
昇腾计算服务层	本层主要提供昇腾计算库，例如神经网络（Neural Network，NN）库、线性代数计算库（Basic Linear Algebra Subprograms，BLAS）等；昇腾计算调优引擎库，例如算子调优、子图调优、梯度调优、模型压缩以及 AI 框架适配器。
昇腾计算编译引擎	要提供图编译器（Graph Compiler）和 TBE（Tensor Boost Engine）算子开发支持。前者将用户输入中间表达（Intermediate Representation，IR）的计算图编译成 NPU 运行的模型。后者提供用户开发自定义算子所需的工具。
昇腾计算执行引擎	本层负责模型和算子的执行，提供如运行时（Runtime）库（执行内存分配、模型管理、数据收发等）、图执行器（Graph Executor）、数字视觉预处理（Digital Vision Pre-Processing，DVPP）、人工智能预处理（Artificial Intelligence Pre-Processing，AIPP）、华为集合通信库（Huawei Collective Communication Library，HCCL）等功能单元。
昇腾计算基础层	本层主要为其上各层提供基础服务，如共享虚拟内存（Shared Virtual Memory，SVM）、设备虚拟化（Virtual Machine，VM）、主机-设备通信（Host Device Communication，HDC）等

数据来源：华为昇腾官网，广发证券发展研究中心

CANN是系统级计算平台，位于物理层和基础软件层之间。CANN根据应用于不同场景中具体的算法需求，为开发者提供了可调用的计算资源以及可操作的功能模块，具体包括超过1200个算子、统一编程接口AscendCL、ModelZoo模型库以及图编译器等。CANN提供了从底层算子、模型开发再到上层应用全流程的开发工具，可覆盖全场景应用，方便开发者快速开发各类算法。作为华为昇腾AI基础软硬件平台的核心，CANN在面向底层硬件资源的调用、面向开发者的工具模块以及面向生态伙伴的接口等方面都有较好设计和提升，其具体特点包括:

1. 简便开发：针对多样化应用场景，统一编程接口AscendCL适配全系列硬件，助力开发者快速构建基于昇腾平台的AI应用和业务。

2. 性能优化：通过自动流水、算子深度融合、智能计算调优、自适应梯度切分等核心技术，软硬件协同优化，提升AI芯片的算力利用率。

3. 开放生态：丰富的高性能算子库和优选ModelZoo模型库，吸引各领域的开发者共建生态。

表 6：CANN逻辑层以及各层级组件功能介绍

抽象层	作用	组件	组件功能
应用层	包括基于 Ascend 平台开发的各种应用，以及 Ascend 提供给用户进行算法开发、调优的应用类工具	推理应用	基于 AscendCL 提供的 API 构建推理应用
		AI 框架	包括 Tensorflow、Caffe、 Mindspore 以及第三方框架
		模型小型化工具	实现对模型进行量化，加速模型
		AutoML 工具	基于 MindSpore 自动学习工具，根据昇腾芯片特点进行搜索生成亲和性网络，充分发挥昇腾性能
		加速库	基于 AscendCL 构建的加速库 （当前支持 Blas 加速库）
		MindStudio	提供给开发者的集成开发环境和调试工具，可以通过 MindStudio 进行离线模型转换、离线推理算法应用开发调试、算法调试、自定义算子开发和调试、日志查看、性能调优、系统故障查看等
芯片使能层	实现解决方案对外能力开放	AscendCL 昇腾计算语言库	开放编程框架，提供 Device/Context/Stream/内存等的管理、**模型及算子的加载与执行、媒体数据处理、Graph 管理**等 API 库，供用户开发深度神经网络应用
	图优化和编译（统一的 IR 接口对接不同前端，支持 TensorFlow/Caffe/MindSpore 表达的计算图的解析/优化/编译，提供对后端计算引最优化部署能力）	Graph Engine	图编译和运行的控制中心
		Fusion Engine	管理算子融合规则
		AICPU Engine	AICPU 算子信息管理
		HCCL	HCCL 算子信息管理
	算子编译和算子库	TBE	编译生成算子及算子开发工具
		算子库	神经网络加速库
	基于计算图的业务流的控制和运行	数字视觉预处理	实现视频编解码(VENC/DEC)、JPEG 编解(JPEGD/E)、 PNG 解码(PNGD)、VPC(预处理）
		Runtime	为神经网络的任务分配提供资源管理通道
		Task Scheduler	计算图 Task 序列的管理和调度、执行
计算资源层	计算设备	AI Core	执行 NN 类算子
		AI CPU	执行 CPU 算子
		DVPP	视频/图像编解码、预处理
	通信链路	PCIe	芯片间或芯片与 CPU 间高速互联
		HCCS	实现芯片间缓存一致性功能
		RoCE	实现芯片内存 RDMA 功能

数据来源：华为昇腾官网，广发证券发展研究中心

CANN提供算子层面多种开发方式，开发者对AI芯片功能拓展更具灵活性。算子通常是AI芯片的核心部件，其包含各种不同类型的运算操作符，如矩阵乘法、卷积、池化、非线性激活等。CANN提供开发者在算子层面可编程的能力。针对不同算法特点，开发者可以从更加底层修改资源调度方式，从而降低神经网络的计算复杂度和时间开销，提高模型的训练速度和精度。

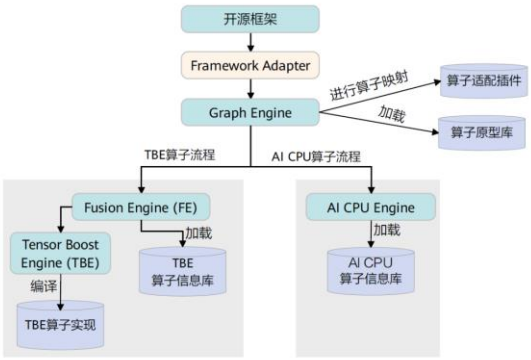
表 7：CANN自定义算子开发对比分析

算子开发方式	语言	计算单元	优点	适用场景
DSL	Python	AI Core	接口高度封装，开发者无需感知硬件内部逻辑，入门较快	适用于通用场景下的逻辑运算，开发效率较高:对于特殊场景支持度不足。
TIK	Python	AI Core	开发者可自行控制数据搬运和调度过呈，熟悉 AI 处理器架构的开发者，可以快速开发出高效的算子。	适用各类算子的开发，对复杂计算场景支持度好
AI CPU	C++	AI CPU	提供原生 C++接口，具备 C++程序开发能力的开发者入门较快。无需感知硬件内部复杂逻辑。	AI CPU 算子性能较低，算子无法通过 AI Core 方式实现或者需要临时快速打通网络的场景下使用。

数据来源：华为昇腾官网，广发证券发展研究中心

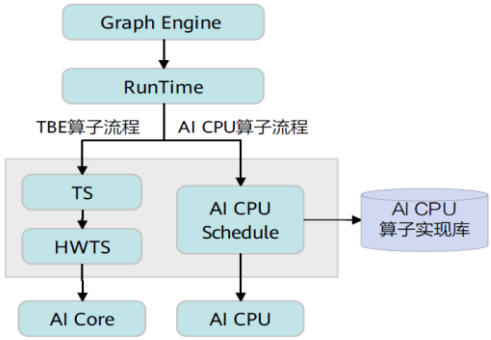
CANN提供的高性能算子库有效提高训练和推理阶段的计算效率。算子库在AI模型的训练和推理阶段都有重要功能和作用。在AI训练过程中，卷积算子、全连接算子、批量归一化算子等对于神经网络的训练过程需要大量的矩阵乘法和复杂的数学运算有很好性能满足，可以显著提高训练速度和效率。在AI推理过程中，卷积算子、池化算子、激活算子等则可用于加速神经网络的推断，减少响应时间。开发者基于CANN提供的支持包括TensorFlow、Pytorch、Mindspore、Onnx框架在内超过1200个高性能算子，帮助开发者有效提升训练和推理的计算效率。

图 4：CANN算子的编译逻辑架构



数据来源：华为昇腾官网，广发证券发展研究中心

图 5：CANN算子的运行逻辑架构



数据来源：华为昇腾官网，广发证券发展研究中心

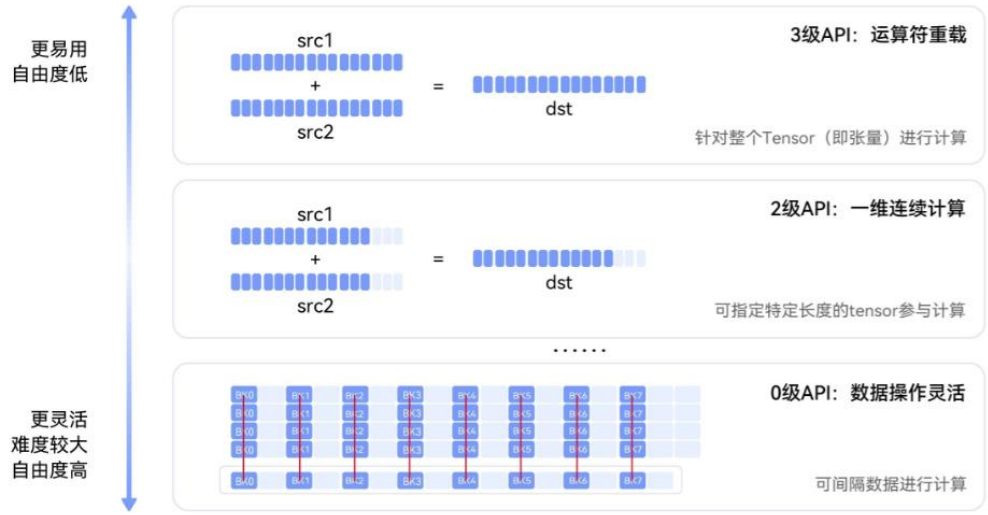
表 8：CANN算子编译运行中的推理和训练场景

	推理场景	训练场景
算子编译	使用 ATC 模型转换工具将原始网络模型转换为适配昇腾 AI 处理器的离线模型	CANN 内部实现逻辑会先将开源框架网络模型下发给 Graph Engine 进行图编译
算子运行	使用 ATC 模型转换工具将原始网络模型转换为适配昇腾 AI 处理器的离线模型后，开发 AscendCL 应用程序，加载转换好的离线模型文件进行模型推理	内部实现逻辑将开源框架网络模型下发给 Graph Engine 进行图编译后，后续的训练流程会进行算子的调用执行

数据来源：华为昇腾官网，广发证券发展研究中心

CANN为开发者提供的可调用的API更加灵活。针对场景复杂度不同，CANN提供的API接口分为多个层级。多层级的API的设计使得开发者在高效和易用之间有可选择的灵活度。API级数越低，自由度越高，更易于表达复杂场景所需功能；级数越高，接口的封装度越高，更易于表达常用语义，使用更简单。此外，华为针对算子开发场景自研了Ascend C编程语言，通过多层接口抽象使得算子的开发过程更加简洁和高效，自动并行计算则充分利用了硬件的并行计算能力，提升了算子的计算性能，而孪生调试技术为开发者提供了方便的调试环境，帮助用户更快地发现和解决问题。通过这些关键技术，Ascend C助力AI开发者在低成本下完成算子开发和模型调优部署。它使得算子开发过程更加高效和便捷，为AI开发者提供了强有力的工具和支持，让他们能够更专注于算法和模型的优化，从而取得更好的成果。

图 6：CANN的多层级API



数据来源：昇腾官网，广发证券发展研究中心

（三）Neuware：寒武纪实现训练推理一体化的 AI 计算平台

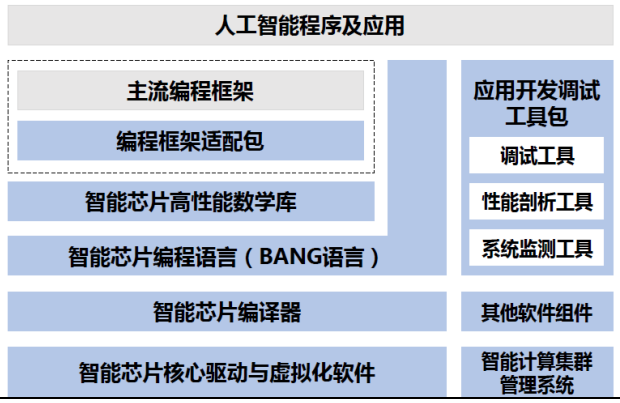
寒武纪Cambricon Neuware是针对其云、边、端的AI芯片打造的软件开发平台。为了加快用户端到端业务落地的速度，减少模型训练研发到模型部署之间的繁琐流程，Neuware整合了训练和推理的全部底层软件栈，包括底层驱动、运行时库（CNRT）、算子库（CNNL）以及工具链等，将Neuware和深度学习框架Tensorflow、Pytorch深度融合，实现训推一体。依托于Cambricon Neuware，开发者可完成从云端到边缘端、从模型训练到推理部署的全部流程，提升AI芯片的算力利用率。

图 7：寒武纪基础软件平台Cambricon Neuware在IT系统中的定位

数据来源：寒武纪官网，广发证券发展研究中心

Neuware提供了全面的AI算法开发工具。Neuware包括编程框架适配包、智能芯片高性能数学库、智能芯片编程语言、智能芯片编译器、智能芯片核心驱动、应用开发调试工具包和智能芯片虚拟化软件等关键组件。在开发应用时，用户既可以基于TensorFlow和PyTorch等主流编程框架接口编写代码，也可以通过公司自研的BANG编程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，使得AI算法各项指令高效地运行于思元系列AI芯片上。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，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 8：寒武纪基础软件平台Cambricon Neuware架构图



数据来源：寒武纪招股说明书，广发证券发展研究中心

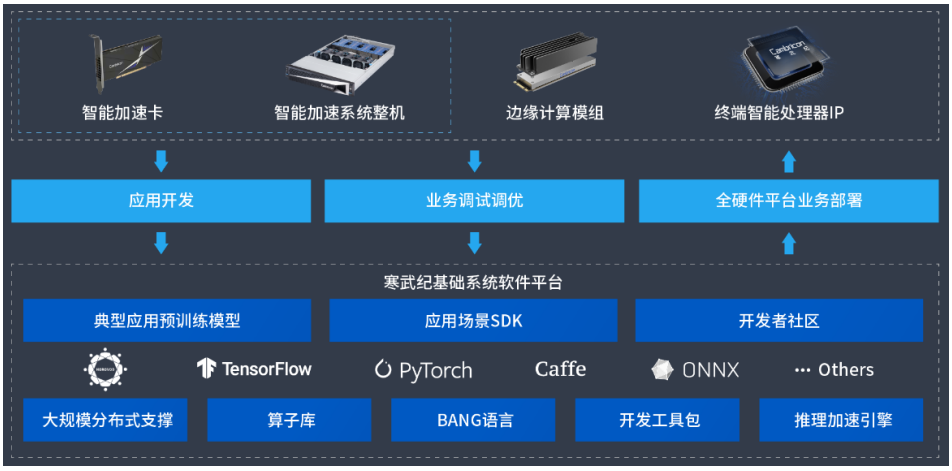
训练任务方面，Neware的训练软件平台拥有多项强大特性，为用户提供高效且灵活的训练环境。

（1）首先，平台支持主流开源框架原生分布式通信方式以及Horovod开源分布式通信框架，使用户能够轻松实现从单卡到集群的分布式训练任务。多种网络拓扑组织方式的支持，使得用户可以根据需求灵活地选择适合的分布式训练方式，包括数据并行、模型并行和混合并行的训练方法。

（2）其次，训练软件平台提供丰富的训练任务支持，涵盖图形图像、语音、推荐以及NLP等多个领域。用户可以在一个统一的平台上完成各类训练任务，极大地简化了训练流程，提高了开发效率。另外，通过底层算子库CNNL和通信库CNCL，训练软件平台在实际训练业务中达到了业界领先的硬件计算效率和通信效率。这意味着用户可以获得更快的训练速度和更高的计算性能，从而加速模型的训练过程。

（3）最后，训练软件平台提供了模型快速迁移方法，帮助用户快速完成现有业务模型的迁移。这为用户节省了大量的时间和工作，让他们能够更快地将已有模型应用到新的平台上，提高了平台的易用性和适配性。

图 9：寒武纪训推一体开发和部署流程



数据来源：寒武纪官网，广发证券发展研究中心

推理任务方面，寒武纪自研的MagicMind推理引擎对主流推理场景应用加速效果较好。2021年底，公司将Neuware架构升级了一个新的模块，MagicMind推理引擎。MagicMind推理引擎支持跨框架的模型解析、自动后端代码生成及优化，可帮助用户在MLU、GPU、CPU训练好的算法模型上，降低用户的研发成本，减少将推理业务部署到寒武纪AI加速卡产品上。此外，MagicMind和深度学习框架Tensorflow、Pytorch深度融合，使得用户可以无缝地完成从模型训练到推理部署的全部流程，进行灵活的训练推理业务混布和潮汐式的业务切换，可快速响应业务变化，降低运营成本。MagicMind的特点包括：

1. 训练到推理的无缝衔接：MagicMind和人工智能框架TensorFlow，PyTorch深度融合，模型训练到推理一键部署。

2. 多种计算精度支持：支持FP32、FP16、INT16、INT8等多种计算精度，支持用户指定不同层级计算精度以及定义量化方法细节。

3. 原生支持动态张量输入：具有完备动态张量表达能力，原生支持任意数据规模的动态张量输入。

4. 丰富的调试调优工具：丰富的调试调优工具以及相应的文档和指引，便利的调试调优体验。

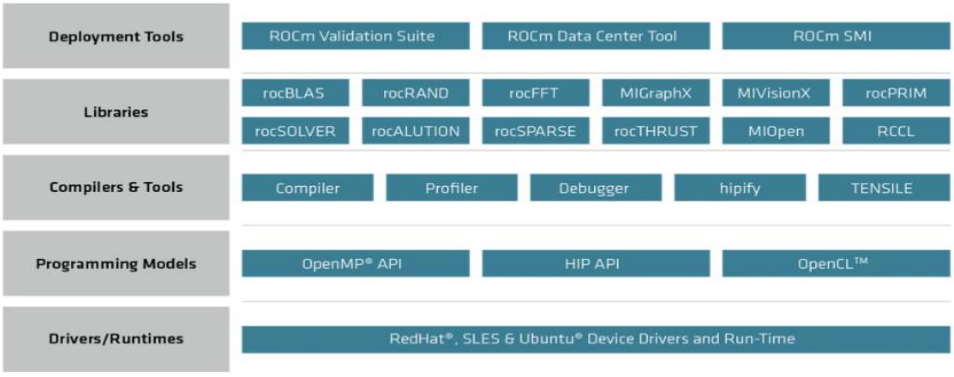
图 10：寒武纪MagicMind推理引擎的架构

数据来源：寒武纪官网，广发证券发展研究中心

（四）ROCm：为海光 DCU 提供高兼容性的 AI 计算平台

海光DCU全面兼容ROCm GPU计算生态。ROCm（Radeon Open Compute Platform）是AMD基于开源项目的GPU计算生态系统，支持多种编程语言、编译器、库和工具，以加速科学计算、人工智能和机器学习等领域的应用。ROCm还支持多种加速器厂商和架构，提供了开放的可移植性和互操作性。海光的DCU兼容ROCm生态的特性使得其得到国际主流商业计算平台生态系统和社区的支持，可以利用现有的AI平台和共享计算资源，快速实现模型训练和推理的性能提升，短期内有利于其DCU产品的推广。

图 11：AMD ROCm 5平台架构

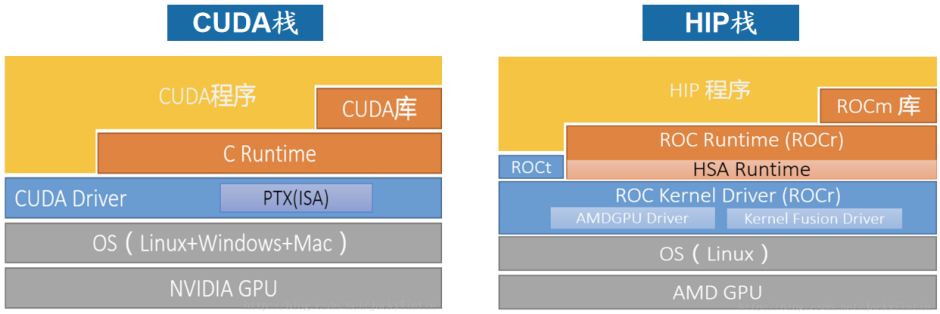


数据来源：AMD 官网，广发证券发展研究中心

在架构层面，ROCm与CUDA相似度较高。ROCm和CUDA在生态、编程环境等方面具有高度的相似性，两者能很好地兼容兼容，因此**ROCm也被称为“类CUDA”**。

ROCm为了更好的兼容CUDA，其实现了源码级的对CUDA程序的支持。AMD团队不仅推出了与CUDA API高度类似的“HIP”工具集（Heterogeneous-compute Interface for Portability），使得AI算法工程师在编写ROCm的代码风格上与CUDA尽量贴近，还提供了Rocblas（类似于Cublas）、Hcsparse（类似于Cusparse）等一系列CUDA生态函数库的替代版本。CUDA用户可以以较低代价快速迁移至ROCm平台。

图 12：CUDA和ROCm技术栈对比



数据来源：CSDN，广发证券发展研究中心

ROCm已实现包括函数接口、编译器和函数库等各方面对CUDA的兼容。API函数接口方面，开发者可以在HIP里得到与CUDA类似的编程语法和大量API指令集，以类似CUDA的风格为AMD GPU编程。函数库方面，ROC库提供了实现常用AI算法的功能，允许开发人员使用类似于CUDA的函数，便捷开发支持ROCm的AI应用。最后在编译环节，HCC（Heterogeneous Compute Compiler）也是对应CUDA的NVCC的编译器。ROCm实现了对CUDA的全面兼容，使得原本为CUDA编写的代码可以在ROCm平台上重新编译和运行，从而在AMD GPU上实现GPU加速计算。

表 9：CUDA和ROCm各项对比

结构	CUDA	ROCm	备注
API函数接口	CUDA API	HIP	C++ 扩展语法
编译器	NVCC	HCC	编译器
函数库	CUDA函数库	ROC库、HC库	
并行算法执行	Thrust	Parallel STL	HCC 原生支持
性能测试工具	Profiler	ROCm Profiler	
debug工具	CUDA-GDB	ROCm-GDB	
总线接口	nvidia-smi	rocm-smi	
远程内存访问	DirectGPU RDMA	ROCn RDMA	peer2peer
SDK（软件开发工具包）	TensorRT	Tensile	张量计算库
软件集成	CUDA-Docker	ROCm-Docker	

数据来源：CSDN，广发证券发展研究中心

英伟达的CUDA计算平台是主流AI应用开发平台。通过对比各公司开发的AI计算平

台，我们发现英伟达的CUDA开发时间最早，积累的开发者数量最多。CUDA推出的的时间是2007年，相较于其他厂商早了十年左右。2012年，以Alexnet为代表的识别类AI技术取得突破后带来的AI算法开发的初期阶段，CUDA即取得了先发优势，在AI算法开发群体中快速推广使用。之后，英伟达一方面依靠AI芯片优异的硬件性能快速获客，另一方面持续拓展CUDA的算法覆盖面，不断巩固客户群体。英伟达通过“滚雪球”式的软硬件协同创新，将其AI芯片的市场份额不断扩大，并构建起了深厚的生态壁垒。

表 10：各公司系统级AI计算平台兼容性和算法覆盖面对比

名称	开发厂商	硬件兼容性	软件兼容性	支持的编程语言	算法覆盖面
CUDA	英伟达	仅兼容英伟达 AI 芯片，随着英伟达 AI 芯片的迭代而持续升级	兼容包括 Caffe2, Chainer, Keras, MATLAB, MxNet, PaddlePaddle, PyTorch, and TensorFlow 等主流 AI 开发框架	支持 C, C++, C#, Fortran, Java, Python 等主要编程语言	支持 C, C++, C#, Fortran, Java, Python 等主要编程语言
CANN	华为	主要兼容昇腾 310P、910、910P AI 处理器	兼容 Pytorch 和 TensorFlow，ONNX，飞桨，Jittor，OpenLab，OpenCV 等主流 AI 开发框架	自研 Ascend C 语言，主要支持 C、C++原生编程	具有图编译技术，以及超过 1200 个高性能算子，其中 Ascend 神经网络加速库内置丰富算子，用于支撑神经网络训练和推理加速
Neuware	寒武纪	主要兼容 MLU 云端一体芯片	兼容 TensorFlow、Caffe、Caffe2、MXNet、ONNX 等编程接口，间接通过 CNML 调用 CNRT 进行软件编程，也可以直接调用 CNRT，运行上述过程所生成的离线模型	自研 Bang 语言，提供 C、C++语言编程接口	构建常见的前向和反向算子，包括神经网络算子、数值运算算子、图像增强融合算子、NLP 算子和自定义算子接口，满足调用和自定义算子开发需求
ROCm	AMD	主要兼容 AMD Radeon GPU	兼容包括 Caffe2,CNTK，Torch, Caffe，MxNet, PaddlePaddle, PyTorch, and TensorFlow 等主流 AI 开发框架	利用可移植异构计算接口（HIP）全面兼容 CUDA 编程语言	ROCm 优化的库目前包括 BLAS、FFT、RNG、Sparse、NCCL (RCCL) 和 Eigen，并不断改进和扩展这些库，以帮助在 AMD 加速器上运行功能更出色的高性能计算程序代码，从而不断提升性能

数据来源：各公司官网，广发证券发展研究中心

二、自主研发的 AI 计算平台有利于长期生态的构建

（一）短期来看，兼容 CUDA 的 AI 计算平台在产品推广上具有便利性

市面上主要有两种计算平台：一种是类似于海光DCU的兼容CUDA框架的计算平台，另一种是华为、寒武纪等自主研发的计算平台。兼容CUDA框架的计算平台提供的函数库和软件工具与英伟达的CUDA平台相似度较高，从而降低了开发者改变开发习惯的难度和迁移的成本，有利于产品中短期的推广。此外，CUDA作为英伟达AI计算平台已经得到广泛应用，兼容CUDA平台可获得较好的生态系统和社区支持。

表 11：兼容CUDA架构的计算平台的优势

优势	描述
GPU 加速优势	充分利用 NVIDIA GPU 的并行计算能力，实现 GPU 加速，提高深度学习和高性能计算任务的计算速度。
成熟生态系统	CUDA 拥有庞大的生态系统和社区支持，主流深度学习框架如 TensorFlow 和 PyTorch 都已兼容 CUDA，有丰富的预训练模型和优化算法，加快开发效率。
跨平台支持	虽然主要支持 NVIDIA GPU，但 NVIDIA GPU 广泛应用于各种服务器、工作站和云平台，使团队能够在多种硬件平台上部署和运行模型，增加代码的可移植性和灵活性。
开发成本和时间	减少自研 GPU 加速技术的开发成本和时间，直接利用 CUDA 技术快速获得 GPU 加速优势。
稳定性和可靠性	CUDA 作为 NVIDIA 开发的官方 GPU 加速计算平台，具有较高的稳定性和可靠性，降低系统崩溃和错误的风险。
生态厂商的支持	CUDA 作为英伟达 AI 计算平台已经得到广泛应用，兼容 CUDA 平台可获得较好的生态系统和社区支持

数据来源：广发证券发展研究中心

兼容CUDA的AI计算平台长期面临挑战。兼容CUDA虽然能够给AI芯片短期内带来推广上的便利，但是长期来看仍然存在软硬件适配度不够、迭代速度较慢以及客户粘性不足等问题。具体来看：

1. 软硬件适配度不够：CUDA是由英伟达开发的，其版本迭代都是根据其自身AI芯片性能特点而进行优化。选择兼容CUDA的AI芯片厂商虽然在软件层面可以实现提供类似CUDA的函数库、API接口以及编译环境给AI算法开发人员带来便利，但是硬件层面其自研的AI芯片实现与英伟达AI芯片的内核、架构、制造工艺等方面相似的难度较大。因此，兼容CUDA的AI计算平台存在与其自研的AI芯片适配度不够，导致AI芯片算力利用率不足甚至性能衰减的问题。

2. 迭代速度较慢：相较于一般的AI芯片公司，英伟达可投入CUDA更新迭代的研发资源较多，使得CUDA的更新速度很快。此外，由于底层架构的差异，部分复杂的CUDA代码仍需要进行适当的修改和优化才可在“类CUDA”平台上运行。这对兼容CUDA的架构造成了较大的技术研发压力，其存在由于投入研发资源不充足导致选择兼容CUDA的AI计算平台迭代跟不上CUDA的更新节奏从而影响芯片性能的问题。

3. 客户粘性不足：兼容CUDA的AI计算平台难以通过提升开发环境的易用性和便捷性，培养开发者的使用习惯，让开发者对平台产生粘性。长期来看，相应的AI芯片在缺少生态壁垒的情况下，则需较大的研发投入，实现历代产品在硬件性能上的突破才可与英伟达AI芯片进行竞争。

表 12：兼容CUDA架构的计算平台的潜在问题

缺点	说明
硬件限制	类似 CUDA 的结构通常依赖于特定的 GPU 硬件平台，因此在其他厂商的 GPU 上可能无法兼容或无法获得最优性能。这限制了代码在不同硬件平台上的可移植性
依赖第三方技术	类似 CUDA 的结构通常需要依赖特定厂商提供的技术和工具，比如英伟达提供的 CUDA 工具。这使得团队在使用这些结构时对特定厂商的技术有较强的依赖性，可能受限于特定厂商的技术发展和支持
兼容性问题	并非所有 CUDA 代码都能在“类 CUDA”平台上无缝运行。复杂的 CUDA 代码可能需要进行调整和优化，以适配不同的硬件平台，增加了开发和维护的复杂性
技术风险	类似 CUDA 的结构可能在未来面临技术风险，特别是如果硬件平台或厂商策略发生变化。如果团队已经大量投入了类似 CUDA 的技术，可能需要花费额外的时间和资源来适应可能出现的变化。
移植性差	类似 CUDA 的结构可能降低代码的可移植性，使得代码在不同平台上的移植变得更加复杂。这可能会限制团队在未来切换硬件平台或扩展到其他领域的能力

数据来源：广发证券发展研究中心

独立开发的AI计算平台与AI芯片软硬件协同力更强，且减少了对于外部技术的依赖。这种框架通常由独立的技术团队或组织开发，旨在为开发者提供灵活、定制化的解决方案，使其能够根据特定的业务需求和数据特点开发定制化的算法和模型。独立开发的AI计算平台提高模型在不同硬件环境上的适配能力，在软硬件协同方面的效果较好，对于部分AI算法的加速效果可以体现出优异的性能和效果。此外，AI芯片公司独立开发的AI计算平台的指令集和函数库都掌握在自家手中，拥有更多的技术自主权，在针对特定场景开发AI算法时可提供更加灵活和个性化的解决方案。

表 13：独立开发框架的优势

优势	描述
定制性和灵活性	独立开发框架可以根据团队的具体需求进行自主定制和开发。团队可以根据特定的业务需求和数据特点设计和优化算法，从而实现更精确和高效的模型。
独特竞争优势	自主研发过程使团队能够创造独特的算法和模型，这有可能带来独特的竞争优势。与其他团队使用通用框架相比，独立开发的框架可以为团队提供更多的创新潜力，从而在市场上脱颖而出。
长期发展性能和效果	独立开发框架允许团队在技术上不断创新和改进。随着时间的推移，团队可以持续优化和更新自己的框架，实现更好的性能和效果，适应不断变化的市场需求。
减少硬件依赖	独立开发框架相对于兼容特定硬件的框架，具有更强的硬件平台适配能力。团队可以在不同硬件环境上部署和运行模型，减少对特定硬件平台的依赖。
知识产权和保密性	自主研发的框架使团队拥有对自身技术的知识产权，无需依赖第三方技术或开源代码。这有助于保护公司的技术资产和知识产权，同时保持竞争优势。
解决特定问题	独立开发框架可以专注于解决特定的问题和场景，针对行业中独特的挑战和需求提供更加专业和个性化的解决方案。

数据来源：广发证券发展研究中心

OpenCL是独立开发框架的代表。OpenCL（Open Computing Language）是一种开放的并行计算编程框架，由Khronos Group开发和维护。它旨在提供一个统一的编程接口，使开发人员能够在各种不同类型的硬件上进行并行计算，包括CPU、GPU、FPGA等。OpenCL可以在不同平台和设备上运行，并利用这些设备的计算能力来加速各种计算密集型任务，使得OpenCL成为一个灵活且可移植的解决方案，特别适用

于需要在不同硬件平台上运行的应用。

表 14：CUDA与OpenCL对比

	CUDA	OpenCL
硬件	主要用于 NVIDIA GPUS	是 AMD GPUS 的主要图运算架构，在 Intel 和 NVIDIA 的 GPU 上都能用但表现一般；还可以用于 CPU、FPGA 和 ASIC（未来主要的发展方向）
操作系统	Windows, Linux, and MacOS，但只能用 NVIDIA 的硬件	OpenCL 的应用能在几乎所有操作系统和硬件上运行，包括 FPGAs 和 ASICs
软件和社区平台	NVIDIA 致力于 CUDA 平台的商业化和发展。NVIDIA 开发的工具包括 CUDA 工具包、NVIDIA 性能原语(NPP)、视频 SDK 和 Visual Profiler，并与 Microsoft Visual Studio 和其他流行平台集成。CUDA 拥有广泛的第三方工具和库生态系统。最新的 NVIDIA 硬件功能在 CUDA 工具包中得到快速支持	AMD 的社区活动更为有限。AMD 构建了 CodeXL 工具包，它提供了全面的 OpenCL 编程工具
编程模型	CUDA 不是一种语言或 API。它是一个并行计算的平台和编程模型，利用 gpu 加速通用计算。开发人员仍然可以用 C 或 C++编写软件，并通过使用 CUDA 关键字来实现并行化	OpenCL 不支持用 c++编写代码，但可以在类似于 C 编程语言的环境中工作，并直接使用 GPU 资源

数据来源：run.ai，广发证券发展研究中心

类似OpenCL的独立开发框架具有强大的泛用性和跨平台能力，但在推行过程中会遇到较多阻碍。一方面，由于其较低级别的编程模型和相对复杂的调试过程，部署成本可能相对较高；另一方面，由于OpenCL并非主流框架，在主流市场中适配度较低，缺乏广泛的应用和支持。此外，开发和维护成本、缺乏成熟生态系统、技术和算法限制、风险和不稳定性和缺乏社区支持等因素都成为了OpenCL进一步发展的主要障碍。

表 15：独立开发框架的劣势

劣势	描述
开发和维护成本	独立开发框架通常需要投入大量的时间、资源和人力进行设计、开发和维护。相比于使用成熟的开源或商业框架，自研框架的开发成本更高，可能需要更多的技术专家和工程师参与。
缺乏成熟生态系统	自研框架通常无法与那些拥有庞大成熟生态系统的开源框架相媲美。例如，像 TensorFlow 和 PyTorch 这样的框架已经拥有丰富的文档、社区支持、预训练模型和优化算法，而自研框架可能无法轻易与之相提并论。
技术和算法限制	自研框架可能无法追赶那些经过多年优化和演进的开源框架。这些成熟框架在技术上和算法上已经经历了许多实践和改进，而自研框架可能受制于限定的资源，难以达到相同的性能和效果。
缺乏跨平台支持	自研框架通常会优先支持特定的硬件平台，导致在其他平台上的适配能力相对较差。这可能限制了应用程序的移植性和灵活性。
风险和不稳定性	自研框架在初期可能面临不稳定性和漏洞的风险。相比于成熟框架，可能需要更多的测试和优化来确保自研框架的稳定性和可靠性。
缺乏社区支持	成熟的开源框架通常有庞大的社区支持，开发者可以从社区中获取帮助、分享经验和学习最佳实践。而自研框架可能缺乏这样的社区支持，开发者可能会面临更多的困惑和障碍

数据来源：广发证券发展研究中心

短期来看，选择兼容CUDA平台的AI芯片在产品推广方面具有一定便利性。CUDA作为英伟达AI计算平台已经得到广泛应用，兼容CUDA平台的AI芯片具有强大的生态系统，并且可获得社区支持，利用现有的AI平台和共享计算资源，有利于其AI芯片产品的推广。独立开发框架由于缺乏成熟生态系统和社区支持，开发和部署可能需要更多的时间和资源，在短期内可能面临一些挑战。

长期来看，兼容CUDA架构的AI芯片难以形成生态壁垒。对于兼容CUDA的AI芯片，由于其开发环境与CUDA的相似度较高，开发者难以通过长期使用形成对其的粘性。长期来看，AI芯片难以通过软件层面的提升形成开发者的生态壁垒，则在硬件层面需要更多的研发投入来提升芯片本身的性能和功能从而吸引用户使用。而在目前英伟达作为AI行业龙头，地位稳固的情况下，单纯通过硬件性能的提升，实现对其每一代AI芯片产品的持续超越的难度较大。

（二）长期来看，自主研发的 AI 计算平台有利于长期生态的构建

美国针对高端芯片及其产业链上下游对中国实施出口限制措施。2022年8月26日，美国政府通知英伟达公司，美国政府对其未来出口到中国和俄罗斯的A100和H100等高端AI芯片实施了许可证要求。AMD证实其也收到类似通知，其用于AI计算的GPU等产品线的出口也受到了类似限制。2022年10月7日，美国商务部工业与安全局宣布修订《出口管理条例》，加强限制中国获得先进计算芯片、开发超级计算机以及制造先进半导体的能力，针对高端芯片及相关终端产品、制造设备等产业链上下游升级对华出口管制措施。

美国高端AI芯片出口管制政策变化存在不确定性，与英伟达AI芯片强绑定的CUDA生态或出现变化。在美国对华高端芯片出口管制政策影响下，以英伟达为代表的部分芯片厂商或选择调整产品配置，以兼顾客户对产品性能要求和出口管制标准。未来，在高端芯片供给不确定性增加的情况下，在部分场景中，AI应用的开发或受到一定影响。**我们认为，在英伟达AI芯片对中国出口前景具有不确定性的背景下，以互联网为代表的AI计算下游厂商适配和采用国产化的系统软件（寒武纪的Neuware，华为的CANN）的动力将大大增加。**

在中美科技领域竞争日益激烈的背景下，AI产业链自主可控建设节奏有望加快。在芯片设计环节，国产高端AI芯片在过去几年性能有了较大提升，以华为昇腾910和寒武纪思元370为代表的国产AI芯片已经具备和英伟达高端AI芯片直接竞争的技术基础。长期来看，美国对华实施的高端芯片出口管制措施预计将催化国内高端AI芯片产业链的国产化进程，加快自主可控的建设节奏。在这一背景下，芯片制造和计算平台领域的国产化建设预计将迎来快速发展的机遇。

1. 在芯片制造环节，先进制程芯片是实现AI芯片算力提升的关键。目前，中国大陆

晶圆厂的先进制程芯片规模化量产能力与国际一流厂商仍有一定差距。在AI芯片供应链整体呈现不稳定的背景下，国产芯片制造企业对于先进制程芯片制造工艺的研发动力大大增强。

2. AI计算平台有效提升AI芯片算力利用率并提供开发环境，是AI芯片整体产业链不可或缺的一环。相较于芯片制造环节而言，AI计算平台的自主可控建设主要涉及开发环境的迁移和用户开发习惯的改变，难度相对较低。我们认为，独立于CUDA的自研计算平台对于我国芯片产业的长远发展至关重要。国产AI芯片厂商独立研发、自主迭代的AI计算平台更加具备长久的持续发展能力。

自主研发的计算平台具有选择AI芯片技术路线的灵活性，长期发展空间更宽广。在AI芯片领域，虽然英伟达的GPU不断升级架构并持续推出新品，是主流的技术路线，但是也存在潜在竞争产品。近年来，针对于特定深度学习算法设计的专用芯片ASIC已成为包括谷歌、英特尔、华为、寒武纪等科技公司采用的技术路线。在数据中心，针对神经网络训练特定设计的ASIC类芯片专用性更强，对于部分算法的加速效果有望超过GPU。

选择兼容CUDA的AI芯片虽然在短期内可以获得产品快速推广的机会，但在长期却失去了可以自由选择其他技术路线实现性能突破的机会。我们认为，在以英伟达为代表的GPU在国内市场受到限制的情况下，各厂商或加快探索除了GPU以外的其他技术路线。**在此背景下，不依赖于GPU技术路线的自主研发的AI计算平台具有自由探索和选择其他技术路线的灵活性，长期发展空间更宽广。**

表 16：各类型AI芯片优劣势对比

AI芯片类型	代表厂商	定制化程度	优势	劣势
GPU	英伟达、AMD	通用型	由于其多线程结构，GPU拥有较强的并行计算能力；相较于FPGA和ASIC，其通用性更强	价格和功耗比FPGA和ASIC要高，并行运算能力在推理任务上无法完全发挥
FPGA	AMD、英特尔、百度昆仑芯	半定制化	可对芯片硬件层进行灵活编译，具有低延迟高吞吐的优势，且功耗小于CPU和GPU	硬件编程语言难以掌握，单个单元计算能力较弱，电子管冗余，功耗可进一步压缩
ASIC	谷歌、华为、寒武纪	全定制化	针对专门的任务进行架构层的优化设计，可实现PPA最优化设计、量产后成本最低	初始设计投入大，可编程架构设计难度较大，针对性设计会限制芯片通用性

数据来源：广发证券发展研究中心

各地建成的人工智能计算中心的AI算力租用给当地企业使用的同时也拓展了国产AI芯片的用户生态。以武汉人工智能计算中心（AIDC）为例，自2021年5月投运以来，为300多家科研机构和企业、高校提供算力服务，日均算力使用超过90%。（数据来源：武汉晚报）武汉AIDC的算力在广泛应用于制造业、交通管理、生物工程等领域的同时也间接推广了华为的CANN计算平台。例如，武汉纳思系统公司的业务是AI赋能电力巡检，通过武汉AIDC的算力支持和华为CANN提供的软件工具开发出的AI

算法让镜头在极微弱的光线环境下也能辨认隐患，实现快速监测。我们认为，国产AI芯片的计算平台正在依托各地AIDC的算力在各行业的应用而实现用户生态的快速拓展。

表 17：2021-2023年华为中标的人工智能计算中心项目

项目投运时间	项目名称	算力规模
2021年5月	武汉人工智能计算中心	200 PFLOPS
2021年9月	未来人工智能计算中心	300P FLOPS
2022年4月	南京鲲鹏-昇腾人工智能计算中心	40 PFLOPS
2022年5月	成都智能计算中心	300 PFLOPS
2022年8月	沈阳人工智能计算中心	一期100 PFLOPS，远期300 PFLOPS
2022年12月	济南人工智能计算中心	首期100P，远期400P
2022年11月	长沙人工智能创新中心	首期200 PFLOPS，2025年1000 PFLOPS
2023年2月	青岛人工智能计算中心	100 PFLOPS
2023年1月	宁波人工智能超算中心	一期100 PFLOPS（FP16），5 PFLOPS（FP64）; 二期300 PFLOPS（FP16），15 PFLOPS（FP64）
2023年2月	北京昇腾人工智能计算中心	一期100 PFLOPS；远期1000 PFLOPS
2023年3月	天津市人工智能计算中心	300 PFLOPS
2023年5月	重庆人工智能创新中心	一期400 PFLOPS

数据来源：华为官网，各地级市政府网站，各地级市招商局官网，各地方公共资源交易平台，各地方新闻网，新华网，人民网，中国日报网，广发证券发展研究中心

国产AI计算平台持续迭代和推广，用户生态建设已具有一定基础。华为的昇腾芯片和寒武纪的思元系列芯片是国内较早开始推广，并在商业落地上取得一定领先优势的产品。在具有一定客户基础后，其持续迭代和推广自研的计算平台，积极拓展用户生态。以寒武纪为例，其于2018年推出用于支撑AI应用跨平台开发、便捷高效运行基础系统软件Neuware。在过去几年，Neuware持续迭代升级，不仅通过对于底层硬件资源的优化提升算力利用率，降低用户的研发成本，还提升了用户粘性，打造用户生态圈。我们认为，最先成功商业化的公司将会扩大对追赶者的优势，因为最终用户不大会接受同时采用诸如四五种以上不同的芯片计算平台体系。我们看好此前有相关经验的华为昇腾CANN和寒武纪Neuware生态的发展前景。

表 18：2019-2023年寒武纪中标的人工智能计算中心项目

中标时间	采购单位	中标项目名称	中标金额（亿元）
2019年4月	珠海市横琴新区管理委员会商务局	横琴先进智能计算平台(一期)采购项目	0.6
2019年6月	西安沣东仪享科技服务有限公司	西咸新区沣东人工智能计算创新中心项目	0.9
2019年11月	珠海市横琴新区管理委员会商务局	横琴先进智能计算平台(二期)采购项目	4.4
2020年12月	南京市科技创新投资有限责任公司	南京智能计算中心项目(一期)智能计算设备采购	3.0
2021年11月	江苏昆山高新技术产业投资发展有限公司	昆山市智能计算中心基础设施建设项目	5.1
2022年1月	南京市科技创新投资有限责任公司	南京智能计算中心项目（二、三期）智能计算设备（二期）采购	5.0
2023年6月	沈阳市大东区城市建设局	沈阳市汽车城新型基础设施建设项目-智能计算中心	1.6
2023年8月	台州市黄岩置成物产管理有限公司	浙东南数字经济产业园数字基础设施提升工程(一期)	5.3

数据来源：寒武纪招股说明书，寒武纪中标公告，招标网，广发证券发展研究中心

寒武纪Neuware计算平台具有较好的兼容性。与寒武纪的Neuware相比，华为CANN的算子库数量较多，算法覆盖面更广泛。CANN不仅覆盖计算机视觉、自然语言处理、智能推荐等商业环境常用的AI算法，还为科学计算和大模型相关算法的开发提供支持。但是现阶段科学计算在大部分AI商业化落地的场景中应用空间有限，因此，我们判断，在商业环境的应用中，CANN的算法丰富度并不会使得华为拉开和其竞品明显差距。另一方面，在对AI训练框架的兼容性上，CANN以兼容华为自研的昇思Mindspore为主，而寒武纪Neuware选择兼容第三方的训练框架（例如百度飞桨Paddle Lite）。具有更为广阔的兼容性使得寒武纪的AI芯片向以互联网公司为代表的商业客户拓展用户生态时更具优势。

表 19：华为CANN和寒武纪Neuware计算平台对比

	华为 CANN	寒武纪 Neuware
算子	具有自定义算子开发体系，多种开发框架，以及与昇腾芯片硬件高兼容的图编译技术。高性能算子超过 1200 个，其中 Ascend 神经网络加速库内置丰富算子，用于支撑神经网络训练和推理加速。	构建常见的前向和反向算子，包括神经网络算子、数值运算算子、图像增强融合算子、图像加速算子、NLP 算子和自定义算子接口，满足调用和自定义算子开发需求。
算法覆盖面	具有丰富的开源仓储备，在大模型（Transformer、推荐领域、生成任务扩散模型、低参微调算法）、科学计算（AI 电磁仿真、计算生物、流体仿真、量子计算）、领域套件与拓展包（计算机视觉（CV）、自然语言处理（NLP）、人脸识别、通用搜索、贝叶斯学习和深度学习融合、强化学习。	具有 AI 加速库（MagicMind、CNNL）、通信库（CNCL）、视觉库（CNCV）和自研 BANG 语言（BANGC、BANGPy），满足计算机视觉（CV）、自然语言处理（NLP）、搜索、推荐等深度学习算法需求。
AI 框架兼容性	主要兼容自研的昇思 MindSpore 框架，包括分布式训练原生、AI4S 融合计算框架和全场景快速部署。	以兼容第三方 AI 框架为主。2020 年 5 月，与百度飞桨 PaddleLite 完成适配，可以基于百度飞桨框架高效稳定运行。
开发组件	ABL 基础层、ACE 计算执行器、AOL 算子库、AOE 调优引擎、AscendCL 计算语言。	CNRT 异构计算、CNNL 算子库、Neuware 运算框架、CNStream 视频结构化框架、Bang 编程语言。
迭代版本数	从 2018 年发布开始，已迭代到 CANN 6.0，最新商用版本为 6.3.RC2。	自 2017 年推出，寒武纪的 Neuware 已经迭代到 03 系列，当前 SDK 最新版本为 1.13.0。
应用场景和领域	应用在视频超分、通用目标识别、分子动力学、蛋白质结构预测等领域，覆盖医疗、运营商、能源、交通、人工智能、互联网、金融、机器人、制造和平安城市等行业。	应用在视觉、语音、NLP、搜索、推荐等领域，覆盖互联网、金融、交通、能源、运营商、制造、教育和自动驾驶等行业。

数据来源：华为官网，寒武纪官网，广发证券发展研究中心

华为提供全栈AI解决方案，在部分场景与AI应用公司构成竞争关系。华为在AI产业链中扮演的角色不仅仅作为底层软硬件基础设施提供商，还针对部分场景开发了具体的AI应用。以AI大模型为例，华为不仅提供底层算力（昇腾AI芯片）、训练框架（Mindspore昇思）和基础大模型（盘古大模型），还开发了行业级大模型（盘古金融大模型、盘古制造大模型等）以及针对场景的AI应用（先导药物筛选、传送带异物检测等）。这与部分AI应用提供商构成同业竞争的关系，其发展会受到一定限制。我们认为，华为在各场景中提供全栈AI解决方案的战略会影响其基础AI算力产品以及计算平台CANN的商业拓展。

图 13：华为盘古大模型产品矩阵

数据来源：华为云官网，广发证券发展研究中心

寒武纪的中立属性在行业格局中具有独特价值，更有利于其用户生态的拓展。寒武纪提供的产品以AI算力基础设施为主，不涉足AI应用领域，与行业参与者更多构成的是互补关系而不是竞争关系。寒武纪作为AI产业链的上游，其下游客户可涵盖云计算公司、智能化升级的科技公司以及AI初创公司等各种类型的AI应用的开发者和提供商。寒武纪的中立属性使其保持智能化升级中赋能者的定位，与产业链上下游形成合作共赢的关系，这更加有利于其Neuware计算平台用户生态的拓展。

三、风险提示

（一）科技巨头在AI计算平台领域长期积累，生态壁垒较高，国产AI芯片公司中短期突破难度较大

（二）AI计算平台对于AI算法覆盖面要求较高，前期投入较大与生态培育不及预期的风险

（三）AI芯片存在供应链不稳定的风险

广发计算机行业研究小组

刘 雪 峰：首席分析师，东南大学工学士，中国人民大学经济学硕士，1997 年起先后在数家 IT 行业跨国公司从事技术、运营与全球项目管理工作。2010 年就职于招商证券研究发展中心负责计算机组行业研究工作，2014 年加入广发证券发展研究中心。

李 傲 远：资深分析师，重庆大学金融学硕士，曾任职于国泰君安、安信基金，2020 年加入广发证券发展研究中心。

吴 祖 鹏：资深分析师，中南大学材料工程学士，复旦大学经济学硕士，曾先后任职于华泰证券、华西证券，2021 年加入广发证券发展研究中心。

李 婉 云：资深分析师，西南财经大学金融学硕士，2022 年加入广发证券发展研究中心。

雷 棠 棣：资深分析师，哈尔滨工业大学软件工程硕士，悉尼大学商科硕士（金融学与商业分析方向），注册会计师非执业会员。2020 年加入广发证券发展研究中心。

周 源：高级分析师，慕尼黑工业大学硕士，2021 年加入广发证券，曾任职于 TUMCREATE 自动驾驶科技公司，负责大数据相关工作。

许 晟 榕：研究员，香港大学金融科技硕士，2023 年加入广发证券发展研究中心。

广发证券—行业投资评级说明

买入：预期未来 12 个月内，股价表现强于大盘 10%以上。

持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-10%～+10%。

卖出：预期未来 12 个月内，股价表现弱于大盘 10%以上。

广发证券—公司投资评级说明

买入：预期未来 12 个月内，股价表现强于大盘 15%以上。

增持：预期未来 12 个月内，股价表现强于大盘 5%-15%。

持有：预期未来 12 个月内，股价相对大盘的变动幅度介于-5%～+5%。

卖出：预期未来 12 个月内，股价表现弱于大盘 5%以上。

联系我们

	广州市	深圳市	北京市	上海市	香港
地址	广州市天河区马场路 26 号广发证券大厦 47 楼	深圳市福田区益田路 6001 号太平金融大厦 31 层	北京市西城区月坛北街 2 号月坛大厦 18 层	上海市浦东新区南泉北路 429 号泰康保险大厦 37 楼	香港德辅道中 189 号李宝椿大厦 29 及 30 楼
邮政编码	510627	518026	100045	200120	-
客服邮箱	gfzqyf@gf.com.cn				

法律主体声明

重要声明

广发证券股份有限公司及其关联机构可能与本报告中提及的公司寻求或正在建立业务关系，因此，投资者应当考虑广发证券股份有限公司及其关联机构因可能存在的潜在利益冲突而对本报告的独立性产生影响。投资者不应仅依据本报告内容作出任何投资决策。投资者应自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或者口头承诺均为无效。

本报告署名研究人员、联系人（以下均简称“研究人员”）针对本报告中相关公司或证券的研究分析内容，在此声明：（1）本报告的全部分析结论、研究观点均精确反映研究人员于本报告发出当日的关于相关公司或证券的所有个人观点，并不代表广发证券的立场；（2）研究人员的部分或全部的报酬无论在过去、现在还是将来均不会与本报告所述特定分析结论、研究观点具有直接或间接的联系。

研究人员制作本报告的报酬标准依据研究质量、客户评价、工作量等多种因素确定，其影响因素亦包括广发证券的整体经营收入，该等经营收入部分来源于广发证券的投资银行类业务。

本报告仅面向经广发证券授权使用的客户/特定合作机构发送，不对外公开发布，只有接收人才可以使用，且对于接收人而言具有保密义务。广发证券并不因相关人员通过其他途径收到或阅读本报告而视其为广发证券的客户。在特定国家或地区传播或者发布本报告可能违反当地法律，广发证券并未采取任何行动以允许于该等国家或地区传播或者分销本报告。

本报告所提及证券可能不被允许在某些国家或地区内出售。请注意，投资涉及风险，证券价格可能会波动，因此投资回报可能会有所变化，过去的业绩并不保证未来的表现。本报告的内容、观点或建议并未考虑任何个别客户的具体投资目标、财务状况和特殊需求，不应被视为对特定客户关于特定证券或金融工具的投资建议。本报告发送给某客户是基于该客户被认为有能力独立评估投资风险、独立行使投资决策并独立承担相应风险。

本报告所载资料的来源及观点的出处皆被广发证券认为可靠，但广发证券不对其准确性、完整性做出任何保证。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价。广发证券不对因使用本报告的内容而引致的损失承担任何责任，除非法律法规有明确规定。客户不应以本报告取代其独立判断或仅根据本报告做出决策，如有需要，应先咨询专业意见。

广发证券可发出其它与本报告所载信息不一致及有不同结论的报告。本报告反映研究人员的不同观点、见解及分析方法，并不代表广发证券的立场。广发证券的销售人员、交易员或其他专业人士可能以书面或口头形式，向其客户或自营交易部门提供与本报告观点相反的市场评论或交易策略，广发证券的自营交易部门亦可能会有与本报告观点不一致，甚至相反的投资策略。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且无需另行通告。广发证券或其证券研究报告业务的相关董事、高级职员、分析师和员工可能拥有本报告所提及证券的权益。在阅读本报告时，收件人应了解相关的权益披露（若有）。

本研究报告可能包括和/或描述/呈列期货合约价格的事实历史信息（“信息”）。请注意此信息仅供用作组成我们的研究方法/分析中的部分论点/依据/证据，以支持我们对所述相关行业/公司的观点的结论。在任何情况下，它并不（明示或暗示）与香港证监会第 5 类受规管活动（就期货合约提供意见）有关联或构成此活动。

权益披露

(1)广发证券（香港）跟本研究报告所述公司在过去 12 个月内并没有任何投资银行业务的关系。

版权声明