**随着GPU生产商NVIDIA的利润大增，股价大涨，资本市场中的机构投资者也开始密切关注GPU，希望能够从中预测行业的变化趋势，发现潜在的应用领域以及新的投资机会。**

**GPU在以后的硬件计算领域会有多大的市场空间呢？ GPU是一种图形处理器，它的主要任务是完成CPU与显示器的衔接。GPU在过去近二十年的飞速发展，主要得益于游戏产业发展对于高性能计算的需求；设计栩栩如生的游戏画面需要计算机图形学的支持，计算机图形学的运算非常适合单指令多数据/线程(SIMD/SIMT)的并行化，这种并行化正是GPU最为擅长的事情，换一点直白的语言，SIMD/SIMT就是一个领导指挥一群人做同样的事情。GPU是为了并行运算设计的，最开始只能进行浮点数数计算；后来开始支持整数运算，这使得GPU开始具备通用计算能力。对于可并行，无数据依赖的运算，GPU会加速其运算速度。为什么乍看起来，历经几十年发展众多人智慧的CPU运算速度会比不了GPU？其实就单个核心来说，CPU的运算速度远远超过GPU，CPU的主频有3~4G，而GPU只有1G左右，仅此项就决定CPU的运算速度高于GPU，所以在单个核心上，CPU的运算能力更强。但是乱拳打死老师傅，拥有数千个核心的GPU，在某些特殊的计算场景中会拥有更强的整体性能。那么CPU为什么不采用GPU的架构设计计算模块呢,也采用数千个核心？这是由CPU负担的工作决定的。CPU的指令集更加全面，可以支持所有的计算场景，可以处理非对齐以及需要大量回写的数据，这些数据场景决定了CPU无法采取众多核心的GPU设计架构。所以说GPU目前主要的业务场景是某些可以单个并行处理的数据，比如图形渲染，基于CNN的深度学习等。对于CPU负担的大多数的通用计算领域，GPU目前不会，将来也不会取代CPU。因为目前的CPU设计已经是经过众多科学家、工程师优化的智慧结晶。但是随着人工智能等领域的蓬勃发展，专用计算的应用不断膨胀，GPU的需求肯定会越来越大。GPU在整个计算领域占得市场份额，会逐渐加大，这个增大不是来源于占领CPU的市场份额，而是来源于人工智能计算的市场不断扩大。比如一台用于深度学习任务的顶配服务器，可以安装五十块GPU，但是这台服务器的CPU数量不会减少。GPU与CPU的关系，如同跑车与SUV的关系，SUV是全地形车辆，可以胜任任何路况，跑车在特定路面下具有极高的速度，但是跑车或者SUV哪个卖得好，还是由市场需求决定的。**

**随着人工智能的蓬勃发展，与此紧密相关的行业都蕴藏着无限的机会。比如智能手机的发展就带动了移动互联网的爆发，与智能手机行业紧密相关的公司都获得了非常巨大的成长。那么人工智能的发展，能给已经进入或者打算进入人工智能计算领域的公司带来怎样的影响呢？**

**为了更好地了解目前做人工智能计算的公司，我们首先科普一下集成电路：光刻机在硅基上刻蚀出复杂的晶体管组合，利用这些设计精妙的晶体管组合完成控制、计算的电路单元，就是集成电路。集成电路也叫芯片，集成电路中晶体管的组织方式不同，电路的功能也不同，GPU与CPU就是设计不同的集成电路。现在人工智能计算领域的产品有NVIDIA的K80，Google 的TPU，IBM的truenorth等。国内，中科曙光与寒武纪也在着手开发人工智能芯片。TPU的诞生与GPU的诞生有着几乎相同的原因。图形处理是一项重复以及繁重的工作，为了减轻CPU的负担，同时提升图形处理的性能，工程师针对图形处理的数据特点设计了专用的集成电路GPU。人工智能的研究者发现深度学习的数据也非常适合用GPU处理，所以GPU应用于人工智能计算领域，并承担了绝大部分的计算量，CPU再一次退居二线主要负责调度的职责。但是GPU最初不是为了深度学习设计的，在架构上还有提升的空间，所以Google为了提升他们人工智能算法的效果以及速度，就开发出了专门用于深度学习计算的集成电路TPU，TPU的性能相对于GPU在人工智能计算领域更强，当然可以应用的场景更少。TPU与GPU的关系类似于F1赛车与跑车的关系，F1主要跑在赛车场内，TPU同样主要用于深度学习，在GPU擅长的图形渲染方面，TPU估计无能为力了。TrueNorth是IBM重新设计架构的一种集成电路，由DARPA资助，它是突破冯.诺依曼计算架构的一次全新的实验，主要目的是降低功耗，提升速度。TrueNorth是针对深度学习的数据特点以及计算方式重新设计的集成电路，数据不再需要存储在冯.诺依曼体系计算机的内存中，而是存在集成电路的内部。TrueNorth更像是一部绑着航空发动机的实验车辆Bloodhound SSC，目的是探索人工智能计算领域计算架构的边界。国内寒武纪的处理器类似于Google的TPU，是一种针对深度学习重新设计的集成电路。**

**人工智能算法、应用的发展与人工智能的硬件发展实际上是相辅相成的。硬件的发展，可以提高算法的计算速度，加快算法的研究进程；可以提升应用的体验效果，扩大人工智能的市场份额。反过来，不断扩大的市场，激励着硬件开发者设计出更快更好地芯片。目前GPU是深度学习领域的通用计算硬件，可以适用于各种各样的深度模型，是一款成熟的，占据主导地位的人工智能硬件产品，稳定的开发生态决定其在一两年之内不会受到任何有效的挑战。TPU是以及寒武纪是针对深度学习设计的专用集成电路（ASIC），不具备通用性，它不会像GPU一样在其它的特定并行计算领域占有一些市场份额，更不会占据CPU的市场份额，它们主要是瞄准的是深度学习计算市场。TPU已经在Google应用了一段时间了，应该是随时可以推向市场的产品。寒武纪的芯片好像只是停留在模拟试验阶段，离市场化的产品还有一段距离，真实的产品性能还没有得到实际产品的检验。TrueNorth 是一款实验性的集成电路，目前还不具备大规模应用的基础。**

**我们现在讲的人工智能计算硬件，实际上是专用集成电路的一个特例。专用集成电路是指针对某种用途单纯设计的集成电路，比如针对信号处理设计的DSP。专用集成电路已经有很长时间的历史了，它们主要负责承担一部分专门的计算任务，但是之前没有表现出取代CPU的趋势。目前GPU的特性使它在多个领域大显身手，可能给许多人一种计算会从CPU分离的错觉。这种特定场景下的计算分离，实际上是为了解决特定需要的过渡技术，不会成为CPU设计的趋势。如果CPU需要经常性的负责某一项计算，或者CPU在某项计算的性能不能满足我们的要求，我们就可以考虑设计专用集成电路，满足这一部分计算的需求。将部分计算从CPU中分离出来又会提高整个系统的设计复杂度，所以是否设计专用集成电路，将某部分计算分离出来，还需要折衷考虑。在人工智能场景中，把计算从CPU中分离出来，是出于提升算法计算速度的考虑。这是一个目前硬件条件下提高算法计算速度的最好解决方案。**