



宋继强, 英特尔中国研究院院长, 主要研究方向为先进移动计算等

智能时代的芯片技术演进

宋继强

英特尔中国研究院, 北京 100190

摘要 集成电路发明至今60年来, 半导体产业依循摩尔定律的发展, 带动整个信息产业达到了今天的规模。分析了芯片技术的演进, 得出在互补金属氧化物半导体(CMOS)微缩仍然可以继续推进的趋势下, 通过有效整合新材料和新封装技术、异构设计和数据处理新功能, 摩尔定律的经济效益将继续存在。

关键词 智能时代; 芯片技术; 集成电路

集成电路发明以来, 芯片对推动人类社会发展到起到了举足轻重的作用。最初被用于帮助提高生产率, 到使人在移动中处理各种事情。现今, 人类社会已经进入了智能计算的时代, 通过智能设备可以更多地了解人类。

除了戈登·摩尔提出的摩尔定律之外, 行业内另外一个非常有名的定律是由戈登·贝尔提出的“贝尔定律”。贝尔定律是指计算设备的主流类别大约每隔10年有一次大的变迁, 尺寸会有一个数量级的缩小, 用户量至少会有一个数量级的增长。

20世纪70—90年代, 贝尔定律一直准确地与实际状况相吻合; 而2007年iPhone以智能手机的形态和全新的交互方式, 通过智能互联设备这一使用模式, 使智

能手机成为了主流的计算设备。

然而随着计算设备发展到现在, 智能计算的时代已经开始, 芯片向着小体积、低功耗、无线通信、无线充电的方向发展, 并逐渐进入无人驾驶车、智能家居等众多产品, 使人们整个生活充满了智能互联设备。在万物智能互联的时代, 海量数据的流动和分析则需要端到端的芯片支持。

Intel公司的目标是通过芯片提供端到端的数据流动支持, 而并非只关注PC芯片领域。

1 新材料和新封装技术

10年之前, Intel就新材料、新制程做了一次路线图

预测,目前看来,预测中大部分的方向是准确的。在推进摩尔定律的过程中,80%的工作都是基于材料的改革。不光要研究怎样将芯片做小,还要研究如何用不同的方式来做这些器件。另外20%的工作基本上都是在寻求化学工艺方面的进步,如原子层沉积、原子层蚀刻等技术。

经过多次尝试之后,收获的同时也发现了很多错误:石墨烯晶体管没有如期出现;许多带有具体日期的预测也都普遍失准;随着工艺越来越小,越来越难以控制和生产半导体芯片。

从摩尔定律来讲,互补金属氧化物半导体(CMOS)的微缩还将继续,但会通过材料、化学工艺等不同方法来实现,更重要的是,还可以通过三维设计将晶体管堆叠起来。除此之外,一些新的功能和电路控制方式,可以让摩尔定律继续推进下去。

不过未来的应用种类非常多,这些是远远不够的,现今已进入了7 nm时代,制程更迭的速度变慢,要想快速应对很多种不同的应用,需要采用异构集成方式及人工智能(AI)等新的数据处理算法。

总结近些年半导体器件发展,Intel认为现在并没有达到器件物理上的极限。Intel已经做出了2 nm宽度的晶体管,也做出了5 nm宽度的连线,但是只把元器件做小还不够,怎样才能在精准的工艺下同时生产数以千亿的晶体管才是最重要的问题;还需要应对多样的需求,了解如何通过各种集成方式以及各种新架构、新数据处理方式来快速解决这些问题,而不是完全依靠CMOS的微缩技术。摩尔定律还将继续演进,只不过是将以不同的形态、不同的方式往前推进。

如果将Intel过去的研究映射到产品线上,可以看到很多不同节点上的产品,比较直接的例子是45 nm节点时的高K金属栅极,就是同时利用全新的化学工艺和全新的材料制造全新结构的元器件。22 nm节点时的3D FinFET晶体管也是同样的例子,目前已经对5 nm节点时的晶体管制造有了清楚的认知。

通过Intel联合产业界及学术界共同追踪半导体工艺的演进,进一步发现如何更好地在电路中使用晶体管,如何将新式晶体管与新架构、新功能相结合。Intel得出的结论是,CMOS功耗和性能表现要优于大部分半导体元器件。至少在最近的10年里,还是要以CMOS为主来制造芯片,其他的新技术可以与CMOS混合使用以提高性能、降低功耗或降低价格。

令人疑惑的是,有些看起来很好的技术却没能产业领域获得成功。从Intel对摩尔定律的观点来看,可以用“用户价值三角”来解释这一情况。用户价值三角指出,经济上的效益要强过技术上,也就是说经济的驱动力会更强,Intel在推动摩尔定律的过程中,也通过许多不同的方式来解决经济学效益的问题。

例如,在整个计算机系统的发展过程中,随着中央处理器(CPU)速度的快速提升,内存遇到了容量不够大、带宽不够高、延迟不够低等问题,虽然业界早已弄清降低内存延迟的方法,但是所需的成本非常高,远不如直接提高内存的容量密度,因此从经济学效益来讲,业界最终选择了重点解决容量问题。

另外一方面,3D内存是一项难度很高的技术,在经济学的驱动下,人们最终成功研发出了3D NAND技术和3D XPoint技术。所以与产业界及学术界伙伴合作,可以解决遇到的许多技术与经济效益之间的矛盾。

2 异构设计和数据处理新功能

在异构集成方面,Intel也有许多技术成果,例如嵌入式多芯片互连桥接技术(embedded multi-die interconnect bridge, EMIB),又称为“混搭”异构设计。可以将不同节点上做出的芯片以2D/2.5D/3D的方式封装起来,保证其互联带宽并降低功耗,这也是Intel持续推动摩尔定律的一项关键技术。

终端设备的多样性已经近在眼前,但终端设备受限于尺寸、功耗、价格,可以发挥的空间不是很大,而环境计算或边缘计算则可以更好地部署提高计算能力。同时,云计算规模逐渐增大,即便其中的一小块的算力都非常强大,哪怕是0.1%的需求也足以支撑一颗定制芯片的研发,因此阿里巴巴和谷歌都在自主研发云计算处理器。

Intel目前正在同时推动异构设计和数据处理新功能的开发,这么多技术同时向前推进可以带来非常多样性的未来。

对于一个智能系统来讲,同时具备强的感知能力、学习能力、抽象能力和推理能力才是一套真正智能的系统。目前通过深度学习只是把系统的感知能力和学习能力提升到比较高的水平,而其抽象能力和推理能力远远不够。最高阶段是整合4方面的能力,通过硬件支撑起这样一套智能系统。

以集感知、学习、抽象和推理能力于一体的“智能自主系统”——无人驾驶为例,该系统要能看到环境中的路标,接受各种信号指示;还要能够根据观察到的东西和自身知识来做出行动指导,实时且准确地选择出口、匝道、交叉路口等。这些能力可以抽象为感知、决策、行动三大类。感知层以多路视觉为主,可能还会有多线3D雷达,因此要求强大的并行计算能力;决策层以感知层处理过的抽象信息为主,需要结合知识和规则做推理,因此需要强大的串行计算能力;而行动层必须在确定的时间点完成执行过程,因此需要强大的实时处理能力。这三部分需要不分先后的同时工作,如果希望这一系统具有学习能力和适应能力,还要能够将更多未知的世界状态变为已知。因此想要实现智能计算这一远大目标,芯片行业还有许多事情要做。

从已有的思路进行鉴定式的研究,认为可重构计算是一项必须要做的事。通用计算和串行计算都可以通过CPU来处理,但并行计算需要有专门的器件来做,以现场可编程门阵列(FPGA)为例,可以通过FPGA实现高度并行运算的硬件加速。因而可以通过结合通用和定制硬件来为未来的多样性应用提供加速。

具备这样的硬件基础后,还要考虑如何让用户使用。很多新芯片的失败并非因为性能和技术,而是因为用户用不好甚至不会用;还要和系统软件对接,让系统可以无缝切换到加速处理器上;最重要的是要有更强的安全性。

而如果跳出之前的渐进式思路,换用变革性的思想来解决智能计算又该怎样?首先要转变计算模式。

传统的计算模式是根据事先画好的流程图编程,靠人的思维来解决问题并编写程序,程序员是这一阶段中最大的价值所在;而现在在做感知任务时,程序员已经不知道如何去描述感知计算的过程,也无法画出流程,但如果大量经过标注的清晰数据,便可以通过深度学习模型训练出一套计算过程,在这一阶段,数据

科学家和算法工程师发挥价值最大。

再往前一步,如果想真正像人脑一样搞定感知、学习、抽象和推理等多种计算,需要研究神经拟态计算。神经拟态计算可以模仿人的大脑结构,使多种计算过程同时进行,并且可以与外界交互,通过观察和反馈来继续学习。

更厉害的是量子计算。量子计算可通过纠缠在一起的量子位来进行高度并发的大量计算,但目前的问题在于量子纠缠的状态极不稳定,计算过程容易出错,未来要解决的是量子计算错误率的问题。

在这几个方面,Intel都有相应的工作在进行。例如神经拟态芯片,这是一个非冯·诺依曼架构的芯片,完全把存储和计算单元融合在了一起,模拟了神经元和神经元之间的连接,是一种异步控制的芯片。神经拟态芯片可以在片上进行自学习,支持无监督学习、监督学习、自监督学习和强化学习模式,目前Intel已经有14和10 nm的神经拟态芯片样片,并在国内大学和企业间开展合作,推动神经拟态芯片的发展。

而在量子计算方面,Intel正在进行两个方向的研究,其一是目前学术界应用比较多的基于超导量子位的传统计算方式,Intel已经在7、17、49三种量子比特节点上进行了大量实验。同时Intel还在观察,当新技术出现后,怎样才能让用户产生非用不可的意愿,是什么阻挡了用户从早期接纳者转变为大众使用者。

3 结论

集成电路发明至今60年来,半导体产业依循摩尔定律的发展,带动整个信息产业达到了今天的规模。有理由相信,未来在CMOS微缩仍然可以继续推进的趋势下,通过有效整合新材料和新封装技术、异构设计和数据处理新功能,摩尔定律的经济效益将继续存在。

(责任编辑 刘志远)