从芯片行业发展历史的角度

展望芯片行业未来十年

【主要观点摘要】

1. 可穿戴设备将会在下个十年全面取代智能手机，可穿戴智能眼镜将会对芯片性能功耗比提出比智能手机芯片高两个数量级的需求。
2. 持久性内存的推出具有划时代的意义，芯片级的存算一体将会实现。
3. 云厂商会通过自己设计硬件架构和芯片来取得竞争优势，类似于上个十年的智能手机芯片。
4. 端云融合将会是未来十年的一个趋势，伴随着人工智能兴起带来的端对数据处理需求的增长以及端有限的存储空间和电量的不足，以及 5G 通讯带来的通畅带宽和云可伸缩的海量计算容量。
5. 随着摩尔定律走到尽头，通用的空分架构将突破冯诺伊曼架构瓶颈实现通用加速。

【全文约 6900 字，阅读约需 14 分钟】

虽然没有可以预知未来科技发展的水晶球，但科技历史和其它科学史一样都有其内在发展的规律可循，计算机发展的历史相对比较短，最早的计算机大概可以追溯到 1935 年左右，在这辞旧迎新之际静下心来回首计算机行业这 85 年来的演变，期待知古可以鉴今，或许可以从历史视角展望下个十年的行业走势。

芯片行业从美中贸易战到全球芯片荒

2019 科技行业的大背景是中美贸易战，5 月份美国全面封锁供应商对华为的供给，从安卓软件到芯片硬件。华为加大自研力度，接连发布鸿蒙操作系统和加大芯片研发的力度。尽管从操作系统到编译器等系统软件都已有准备可以自力更生，底层芯片架构也解禁了对华为的支持，但是中间的芯片设计自动化工具链一时还难以国产替代，萌芽中的开源 OpenROAD 也是美国国防部 ERI 支持的开源计划。华为美研所和经验丰富的美籍雇员也被迫重新调整，美籍的海思主要架构师移师阿里。阿里达摩院从中受益，颇有点类似于当年的 Xerox PARC，受益于 70 年代美国的反越战情绪，很多在军方项目支持下工作的大牛转投了 Xerox PARC。

美国方面，高通收购 NXP 受阻导致高通赔钱又裁员，美国的 5G 发展现已滞后。从 1878 年发明电话的贝尔实验室，到 1983 年推出大哥大的摩托罗拉，美国领导有线和无线通信的时代已然成为历史。2019 年，苹果 5G 手机芯片 Intel 无货可供，后来整个业务部门被苹果收购。高通 5G 芯片也没有占到先机，但基本上却是苹果眼下的唯一选择，四月份苹果无奈之下也选择赔钱和高通专利纠纷和解。

反观 60 年代的美国，当时的两大霸权美苏冷战正酣，美国太空技术被苏联赶超，美苏核武战争一触即发，美国通过立法成立了 NASA 和 DARPA，使命就是保持美国太空和军事科技较其他的潜在敌人更为尖端。在成立后的 10 多年，NASA 和 DARPA 都取得

了突出的成绩，1969 年 NASA 成功送美国宇航员阿姆斯特朗登上月球，DARPA 支持的研究在 60 年代就奠定了现代互联网的基石。最近几年 DARPA 认识到美国在高科技芯片行业的差距和中国正在缩小，特别是高性能计算，5G 和人工智能芯片，部署了电子复兴计划（ERI）以保持美国的领先优势。

在美国的国家意志下，可以预见中美在高科技芯片行业的较量将会长时间存在，美国在高科技芯片行业的较量甚至连欧洲也不例外，欧洲提出了欧洲芯片计划（European Processor Initiative）防患于未然。中国在芯片制造领域已上升到了国家的高度，在芯片设计领域还需要更多国家层面的大战略规划和统筹协调，不仅仅是大基金来扶持投资现有芯片设计企业，大基金的目标是弥补和替代而非赶超，还需要类似于 DARPA 和 SIA/SRC 等的战略机构来统筹支持全国芯片类的研究才有希望 2030 赶超美国。

2020 中美贸易的战火在芯片行业进一步扩散，从芯片设计蔓延到了芯片制造业，以中芯国际为代表的多家本土科技公司受到美国方面的出口限制。中美贸易的战火也让欧洲进一步受到了威胁，欧洲进一步推进欧洲芯片计划并签署了欧洲电子芯片和半导体产业联盟计划，计划未来三年扩大投资近 1450 亿欧元于欧洲芯片行业。随着英国脱离欧盟，大国之间的科技较量将会在可以预见的将来持续加深。

2021 芯片行业受全球新冠疫情的影响，芯片的供应链交付周期持续变长，引发了今年的全球“芯片荒”。受芯片供给端的影响，汽车行业新车的交付周期持续变长至一年多，导致二手车价格顺势走高，为了应对芯片供给对整车交付的影响，大众，奔驰等一线车厂已投入芯片研发。汽车行业芯片的先行者特斯拉，也继首款 FSD (Full Self-Driving)自动驾驶（含人工智能推理）芯片于 2019 年问世之后，今年八月强势发布了更强大的自动驾驶人工智能训练芯片 Dojo。受“芯片荒”和“英国汽油荒”影响，电动汽车销售增长强劲，必将持续推高一线车厂芯片研发的力度和汽车芯片的垂直整合。

端业务从智能手机滞涨到元宇宙开启

2007 年苹果推出 iPhone 标志着智能手机和移动计算时代的到来，智能手机业务也几乎疯涨了 10 多年，苹果的市值也超过万亿美元成为美国市值最高的科技公司。智能手机业务的滞涨这几年已趋于明显，三星在 2019 年 11 月份停止了其手机 CPU 设计业务，300 多人的美国 Austin 设计团队被解散；前年高通也停止了其手机 CPU 芯片设计业务，同样也是近 400 人的团队被遣散，一部分转做 Arm 服务器芯片后来也被遣散。

此消彼长，智能手机可能正在被进化中的可穿戴设备取代，可穿戴手表已成为现实，可穿戴眼镜还在紧锣密鼓的研发中。不过历史总是惊人的相似，2000 年代前 20 年智能手机逐渐取代个人电脑，1970 年代个人电脑逐渐取代大型机和小型机。不变的是人机介面的更替，电脑体积变得更小便于携带，功能更加集成一机多用，人机交互更加方便，从个人电脑的键盘鼠标，到智能手机的触摸屏，再到现在基于人工智能的语音交互以及人脸和姿势识别；变的是计算机作为工具进化的速度加快了，从 30 年缩减到 20 年甚至 10 年。2019 年，科技行业的领导者几乎都在研发迭代增强现实眼镜，比如 Google Glass, Facebook AR Glass, Microsoft HoloLens, Apple AR Glass, Amazon Smart Glass 等, 可以期待下个 10 年可穿戴设备将会全面取代智能手机。可穿戴眼镜将会对芯片性能功耗比提出更高的需求，毫不夸张的说，在性能功耗比上必须比现有手机芯片

提升两个数量级，这就要求系统的集成度更高，比如计算和存储集成在同一芯片上减少数据移动的功耗以及提升性能。

随着智能手机业务的滞涨和摩尔定律的放缓，通过手机软件高利润补贴手机硬件低利润的模式正悄然失去平衡，引爆了 2020 年腾讯和华为关于手机游戏利润分成的矛盾。一方面是硬件平台低利润需要通过软件来补贴，摩尔定律放缓使得趋势对硬件厂商更加不利，因为成本并未按摩尔定律的规律随着时间下降。游戏行业利用摩尔定律通过游戏软件补贴硬件由来已久，这就是为什么首发的游戏硬件需要预订而且很快就卖断货，因为首发的硬件基本不赚钱，所以数量肯定是供不应求。2020 年 11 月新发布的 Xbox X 和 PS5 游戏机硬件即是如此，通过压低硬件利润和售价吸引来更多用户，然后通过游戏软件的高利润补贴硬件的低利润，小米手机正是沿着这个模式率先在国内推出千元智能手机。可以预见原有的内容与平台平衡将被打破，新的平衡将会建立。

2020 年 11 月值得一提的是苹果发布 M1 系列处理器将处理器垂直集成从手机端稳步带入手提电脑端，受此消息鼓舞，2021 年 1 月高通通过收购初创公司 Nuvia 回到个人电脑芯片市场。

2021 年 10 月 28 日，Facebook 公司宣布更名为 Meta，标志着 Metaverse 也即元宇宙的愿景逐渐变得清晰，元宇宙融合游戏和虚拟/增强现实，将持续推进对芯片性能功耗比提升两个数量级的需求和进展。比肩一线大厂，Oppo 也于两周前顺势推出智能眼镜 Air Glass。

数据中心业务存传管算百花齐放

越来越多的行业公司在 2019 年认识到云端数据处理能力的重要性。三家 SmartNIC 公司在 2019 年被收购整合，Mellanox 三月被英伟达以 69 亿美元收购，SolarFlare 四月被 Xilinx 收入旗下，Barefoot 六月则被 Intel 纳入麾下，还有一家 Netronome 差点也被 Facebook 收购。多家人工智能初创公司和大厂推出了自己的云端人工智能训练芯片，11 月份的 SC’19 会议上，Groq, SambaNova, Cerebras 和 Graphcore 落地了人工智能训练的芯片。国内华为，阿里，百度等大厂，以及腾讯支持的燧原，在 2019 年也纷纷推出了人工智能训练芯片和系统。2019 年底，又一家人工智能芯片公司被收购，Intel 以 20 亿美元收购了以色列的 Habana 公司，Habana 在 2019 年推出了云端人工智能训练芯片，2018 年推出了云端人工智能推理芯片。可以预计未来几年将会带来更多的人工智能芯片行业整合。

数据中心的核心在数据的存传管算，除了在传和算上做文章，集成减轻服务器负担、提高服务器利用率、用网卡做计算的 SmartNIC；以及在算上做文章，针对人工智能推理和训练专业应用的加速器芯片，存储器件和服务器芯片在过去一年也有划时代的发展。

2019 年 4 月 2 日，Intel 推出了基于 3DXP 的持久性内存，单个内存 DIMM 容量达 0.5TB，内存进入了 TB 时代。从历史的角度看，持久性内存的推出是具有划时代意义的，1949 年王安发明磁心内存，很快取代了当时流行的威廉姆斯电子管，然而好景也就 20 年；1968 年 IBM 发明 DRAM，很快在 70 年代就取代了磁心内存，一直使用至

今。也就是说，内存其实在过去的 50 年没有革命性的变化，有的只是从 1996 年开始由 JEDEC 主导的标准化和性价比的不断优化，直到 2019 年推出持久性内存。持久性内存和 DRAM 相比在 PAPER 参数上(Performance 性能，Area 面积，Power 功耗，Endurance 可擦写次数，Retention 数据持久性)没有绝对的优势，但是当年磁心内存也比威廉姆斯电子管性能差，由于在稳定性上秒杀威廉姆斯电子管而取代之。今天缓慢增长的内存性能相比快速增长的处理器性能已让内存处于限制系统性能增长的尴尬地位（“Memory Wall”），这个情况和 1950 年是惊人的相似，当时基于真空电子管的处理器很快但基于磁鼓/磁带甚至打孔卡的存储速度却很慢，后来在处理器和存储中间加入了内存来缓解数据访问瓶颈，比如，威廉姆斯电子管或者水银延迟线存储器。随着时间的推移，1965 年 SRAM 推出，1968 年 IBM 即在 System 360/85 计算机系统中率先引入基于 SRAM 的单级缓存，介于处理器和磁心内存之间。今天的持久性内存至少将会在 DRAM 和 SSD 之间取得一席之地，并且随着系统存算一体整合（Monolithic 3D）在不久的将来有可能在一些系统里取代 eNOR Flash 甚至 DRAM，其它更快的持久性内存 ReRAM 或 MRAM 甚至有可能取代片上 SRAM，比如可穿戴和物联网设备将会带来更近一步的系统整合需求和机会。

2019 年 12 月 3 日，AWS 在赌城拉斯维加斯推出了基于 Arm 服务器的 Graviton2, 标志着云端服务器不再是 Intel 一家独大。上一个十年智能手机时代，苹果，三星，华为，高通，德州仪器，飞思卡尔纷纷设计智能手机芯片，历史总会惊人的相似。下一个十年，计算机架构方面也会带来深刻的变革，从基于 CPU 的通用计算到集成加速器的异构计算，各大云厂商都会有自己设计的硬件架构以取得竞争优势，亚马逊，微软，谷歌，脸书，百度，阿里，腾讯，华为等将无一例外，今天所有这些互联网公司都从前端的应用扩展到了后端的计算平台，即使还没有公开加入定制芯片业务的腾讯和微软也都有自己的 FPGA 业务，并且分别押注了人工智能芯片初创公司燧原和 Graphcore。

正是看好数据中心业务的增长，2020 芯片行业进一步整合，Nvidia 收购 Arm 整合 Arm CPU 和 Nvidia GPU，AMD 收购 Xilinx 整合 AMD CPU, GPU 和 Xilinx FPGA，可以预见的未来数据中心芯片业务将呈现 Nvidia, AMD 和 Intel 三足鼎立的局面。更多的互联网云计算公司在 2020 也纷纷加入 Google TPU 和 Amazon Graviton 自研芯片的潮流，微软，百度，阿里和腾讯据报道都在研发人工智能芯片或通用服务器芯片。除了算力，数据中心成本很大一块在内存，Intel 继续推动 3DXP 持久性内存在数据中心的应用，整合持久性内存和数据中心芯片的协同，同时剥离非核心 NAND 业务（作价 90 亿美元出售予 SK Hynix）以专注 3DXP 和数据中心协同业务。2020 年 10 月，Arm 成立 Cerfe Labs 公司专注推进持久性内存 CeRAM 和 FeRAM 的产业化。可以预见存储将是继算力和网络芯片之后芯片行业新一轮并购的热点，如果持久性内存可以取代 DRAM，对计算机架构和软件的影响是深远的，将会加深存储和算力的整合。

不管是自研加速器还是服务器芯片，数据中心为了提高性价比的另一个趋势是资源分离和共享来提高资源的利用率，比如计算、存储和加速器分离与共享，通过提高资源利用率来降低云计算成本。随着存算分离，数据移动对互连带宽和功耗需求必将增加，存内计算可以缓解存算分离对互连的影响。2021 年三星发力存内计算隆重推出 Aquabolt-XL 系列产品，将 PIM 融合到不同接口的内存产品中。同时，以亚马逊 Aqua 为代表的计算存储（Computational Storage）逐渐崭露头角。

从端云之争的起源到端云融合

1962 年 11 月，两大霸权美苏冷战正酣，五年前苏联成功发射了人类第一颗人造卫星"伴侣号" (Sputnik)，标志着苏联在太空领域打败了美国。美国总统艾生豪威尔将军责成国会成立了 NASA 和 ARPA (现 DARPA)，国会要求高级研究计划局（ARPA）承担保持美国军事科技较其他的潜在敌人更为尖端的使命。仅一个月之前的 1962 年 10 月，古巴导弹危机差点引起了一次毁灭地球的核战争，美军认识到这次差点误打起来的核战争主要是由通讯不畅引起，不仅是美苏两大霸权之间的通讯不畅，也是白宫和五角大楼与美军在海上执勤的军舰之间的通讯不畅。时任高级研究计划局(ARPR)信息处理技术办公室(IPTO)主任的J.C.R. Licklider主管着大笔科研经费将要投入通讯技术领域，在回华盛顿的火车上说服了同行的 MIT 计算机系教授 Robert Fano 进行大型机时分多用（time-sharing）研究（Project MAC），用今天的通俗说法就是云计算。正是沿着这个时分多用思想，在接下来十年支持的研究项目里，诞生了包含 Packet Switching, TCP/IP, ARPANET 等的通讯技术，以及 MULTICS 操作系统并深深影响 UNIX（segfault, dynamic linking 即来源于此），成为了今天互联网和云计算的基石。1962 年接下来的十年，时分多用用户感受到的性能一直不佳，直到 1973 年施乐公司(Xerox PARC)的 Robert Metcalfe 和 David Boggs 提出以太网来组建局域网实现后才逐渐成为现实。过去的半个世纪通讯技术持续突飞猛进，从有线局域网到无线互联网，伴随着无线数据网络性能的蒸蒸日上和云端设备的集成化小型化互联化，云计算已经卷土重来。

在 1962 年这个岔道口，1962 年之前的计算机都是大型机，但并不是所有人都认同时分多用思想来让大型计算机为更多人所共享使用，MIT 林肯实验室教授 Wesley Clark 就是其中的代表，他认为小型机或个人计算机将是未来发展的方向，同年受 NIH 资助研制出 LINC，后被认为是历史上第一台个人计算机。值得一提的是，十年后的 1972 年克拉克（Wesley Clark）成为尼克松总统访华后首批访问中国的五位美国计算机科学家之一。过去的 60 年，个人计算机越来越集成化小型化，人机交互越来越方便，计算机与人的关系从工具的功能转换到服务的功能，从人使用工具到计算机主动为人服务，比如自动驾驶，机器人等。人类的需求在过去的 60 年也几乎没有什么本质的变化，70 年代个人计算机兴起的电脑游戏/娱乐依旧，只是现在的 VR 游戏更加身临其境；社交网络，信息/新闻的获取/共享，出行导航这些需求依旧，只是人机介面比原来更友好，比如现在可以通过自然语言人机对话，甚至计算机自动驾驶完全把人从操作工具的繁琐中解放出来。

伴随着人工智能兴起带来的端对数据处理需求的增长以及端有限的存储空间和电量的不足，以及 5G 通讯带来的通畅带宽和云可伸缩的海量计算容量，端云融合互通将会是未来十年的一个趋势。

展望未来十年

十多年前的 2007 年 Dennard Scaling 失效，计算机芯片设计走向多核设计，而不能通过提高 CPU 主频率让应用软件的性能通过硬件升级来跟着提升，过去的 10 多年软件多线程化的速度依旧非常缓慢，归咎于多核编程的复杂性，很多软件现今还是单线程，

比如著名的商业软件 Redis 内存数据库，虽然多核 CPU 可以运行同一应用的多个实例，但是单一应用性能的提升十分有限。2015 后计算又迎来了摩尔定律放缓带来的挑战，晶体管的数量停滞不前，需要更有效的硬件架构设计来提升系统性能，时分多用的 CPU 架构虽然节省空间，但是却花费很多时间和能耗来读取指令，受单位面积能耗的限制，空分架构(Spatial Architecture) 可以很好的避免这个浪费，提升单位面积单位能耗的性能。这就类似于从车上卸载一车西瓜到仓库，如果有十个人同时搬西瓜，每个人都来回奔波于车和仓库，每次每个人来回都得转几个弯，每个弯都得左右看看是不是可以安全通过，这个额外的来回奔波劳碌和转弯判断消耗会很大，就好比并行多线程运行在多核 CPU 上；另外一个方案是，如果十个人都可以一字在空间排开，每个人无需跑动，只做一个动作从上一个人那里接住西瓜传给下一个人就好了，没有来回奔波于路径和路径上重复判断，这样流水线操作效率一下就会提高很多，甚至达到一到两个数量级的性能提升。

计算机的冯诺伊曼架构从 1945 年被提出后的 75 年不曾大变，冯诺伊曼架构的瓶颈在冯诺伊曼架构被提出 32 年后的 1977 年图灵奖获奖致辞中就已明确被当年的图灵奖得主 John Backus 指出，这之后的 40 年 CPU 缓存和分支预测在很大程度上缓解了冯诺伊曼架构瓶颈痛点的出现。伴随着云计算的崛起和云计算应用的多元化，日益增长的计算需求和进步放缓的应用性能提升将会使得冯诺伊曼架构瓶颈痛点越来越明显，即使上文提到的异构计算也还是大体沿着冯诺伊曼架构的老路走，而且专有硬件加速器也只能加速特定应用，放在计算机历史的长河里注定只能是一个当下权宜之计的解决方案。2020 年离计算机诞生的百年已不远，是时候重新审视冯诺伊曼架构瓶颈了，也许下一个十年空分架构将突破冯诺伊曼架构瓶颈，并逐渐在冯诺伊曼架构诞生后的百年成为主流。

2021 年 11 月 15 日标志着以英特尔 4004 为代表的通用微处理器诞生 50 周年，半个世纪微处理器架构经历了复杂指令集微码（Microcode），流水线(Pipelining)，精简指令集（RISC），超标量（Superscalar），向量计算（Vector Processing），乱序执行（Out-of-Order Execution），多核/多线程（Multicore/Multithreading）等架构更新并持续提高终端应用性能。过去的 50 年中也不乏几轮专有架构和通用架构之争（Makimoto’s Wave），历史证明随着时间的推移大多专有处理器的功能往往逐渐被吸纳到通用处理器中（CPU Centre of Gravity），伴随着应用软件渐进移植，逐步提高终端应用运行的性能并取代专有加速器。尽管过去 5 年人工智能专有架构兴起，但是通用架构由于其软件移植友好性，且伴随着通用处理器对人工智能加速的支持，终端应用感受到的性能逐步提升，最终会取代专有架构。空分架构可以有效平衡加速和通用，已崭露头角成为了新通用加速架构。

王伟草于 2020.1.1，更新于 2021.1.1，更新于 2022.1.1