黄仁勋提AI“iPhone时刻” 欲推计算光刻革命？

本报记者 李玉洋 上海报道

3月21日晚，英伟达（NASDAQ:NVDA）召开的GTC开发者大会犹如“深水炸弹”，在AI领域掀起巨浪。而让人印象最深刻的，是英伟达创始人兼CEO黄仁勋提出的全新概念：“我们正处于AI的‘iPhone时刻’。”

黄仁勋所谓AI的“iPhone时刻”，即AI技术正在迎来爆发式增长，将成为数十年来最有前途的技术领域之一。在此次GTC 2023大会上，英伟达发布了专为ChatGPT这样的大语言模型设计的GPU H100 NVL、AI超级计算服务DXG Cloud等多款针对AI的最新技术。

值得注意的是，英伟达还发布了让计算光刻变得更加“聪明”的软件库cuLitho。尽管英伟达此前针对生物制药、化学、气候预测、量子计算等领域也发布过一些中间件和软件库，但都属于常规操作，这次针对芯片制造工艺——计算光刻的举动显得有点特别。

有受访者告诉《中国经营报》记者，cuLitho是一个用于运算式微影函式库，可以缩短先进制程芯片的光罩时程、拉升良率并大幅减低晶圆制作的能耗，英伟达此举意义重大。“英伟达的计算光学加速，确实对先进节点的光刻有所帮助。”一家国内排名靠前的IC设计公司研发人员表示，常规的光学近场修复耗时耗力，尤其是5nm节点以下工艺挑战很大，用AI加速能够分担很大一部分工作量。

AI的“iPhone时刻”

黄仁勋指出，人工智能发展至今，对社会的影响可能像苹果iPhone打开智能手机市场那样。

在本次大会上，黄仁勋多次提及生成式AI，将ChatGPT称为AI的“iPhone时刻”。2022年11月底，OpenAI发布聊天机器人ChatGPT，迅速引发全球旋风，仅用两个月全球独立访问用户便过亿。

黄仁勋认为，生成式AI是一种新型计算机，一种可以用人类语言进行编程的计算机，每个人都可以命令计算机来解决问题，这之前是只有程序员才能接触的领域，而现在每个人都能是程序员。正如此前的互联网一样，生成式AI也将重塑每个行业。

针对部署像ChatGPT这样的大语言模型（LLM），英伟达发布了AI重磅产品H100，它将英伟达的两个H100 GPU拼接在一起。“当前唯一可以实际处理ChatGPT的GPU是英伟达HGX A100。与前者相比，现在一台搭载四对H100和双NVLINK的标准服务器速度能快10倍，可以将大语言模型的处理成本降低一个数量级。”黄仁勋说。

英伟达还介绍，会把由8块旗舰版A100或H100芯片集成的DGX超级AI计算系统通过租赁的方式开放给企业，每月租金为37000美元，以推动加速这轮大语言模型引领的AI繁荣。

“我们在欧美与云服务提供商合作，提供英伟达的DGX系统AI超级计算机的能力。在中国，我们有特别定制的Ampere和Hopper芯片。这些会通过中国云提供商，比如阿里巴巴、腾讯、百度这些企业提供落地的能力，我完全相信他们有能力去提供顶级的系统服务，对于中国初创公司一定会有机会来开发自己的大语言模型。”黄仁勋在接受媒体采访时表示。

此外，英伟达还推出云服务NVIDIA AI Foundations，提供语言、数据和生物学模型的定制服务，与Adobe、GettyImages、Shutterstock等进行合作。

“人工智能的iPhone时刻已经开始。”黄仁勋指出，人工智能发展至今，对社会的影响可能像苹果iPhone打开智能手机市场那样。

用软件做建模的计算光刻

光刻图案未来将一步步走向模糊，或者说没有很高的保真度。

据黄仁勋介绍，所谓计算光刻就是为芯片生产制作光掩模（photomask）的技术，掩膜是一种平面透明或半透明的光学元件，上面有芯片加工所需的图案，并通过曝光将图案转移到光刻胶层上。

光刻加工过程开始后，通过控制光刻机的曝光和开关操作，可以将光束根据掩膜上的图案进行分割和定位，使得光束只照射到需要曝光的区域，从而将芯片上的图案转移到光刻胶层上，实施芯片光刻。

“其实，光刻就像是用‘光刀’在晶圆上‘雕刻’一样。而雕刻需要刻出特定图案。这个图案首先要呈现在光掩膜上。掩膜板就像是漏字板，激光一照，通过镜头，漏字板上的图案也就落到了硅片上。”长期关注半导体行业发展的资深观察人士黄烨锋生动地解释光刻原理。

事实上，晶体管、器件、互联线路都需要经过这样的光刻步骤。因为每种芯片都要经历多次曝光，所以光刻中使用的掩膜数量不尽相同。“实际生产要复杂得多，比如现在的芯片上下很多层，不同的层就需要不同的光刻和掩膜板，且某些层如果器件间距很小，就可能需要多次光刻。”黄烨锋说。比如，NVIDIA H100（台积电4N工艺，800亿晶体管）需要89张掩膜，英特尔（Intel）的14nm CPU需要50多张掩膜。

有业内人士表示，英伟达针对计算光刻发布的cuLitho，此举意义重大。 视觉中国/图

黄烨锋指出，光刻过程其实很反常识，比如要在晶圆上光刻一个类似“+”的图案，那么掩膜板要做成形似二维码样的图案。对此，英伟达先进技术副总裁Vivek Singh解释说，半导体经过几十年的发展，晶体管互联间距变得越来越小，但“大概30年前，晶体管的尺寸变得比（光刻机所用的）激光波长还要小，于是衍射效应就产生了，晶体管成像就会变得模糊。”

“对于相机而言，当光圈小到某种程度以后，照片受到衍射效应的影响就会显著增大，导致画面解析力的大幅下降。实际上，超高像素（或小像素）也受制于衍射效应。”黄烨锋表示，尽管光刻机所用光源也发生过几次大的迭代，比如目前讨论最多的DUV（深紫外线）和EUV（极深紫外线），但哪怕是波长显著变小的EUV极紫外光刻，其波长与器件间距之间的差异，也变得比过去更小，“换句话说，光刻图案未来将一步步走向模糊，或者说没有很高的保真度。”

因此，计算光刻得以切入，借助计算光刻缓解衍射效应所带来的像差对芯片制造的不良影响。据黄烨锋介绍，此前ASML中国就曾提起过计算光刻，计算光刻已是ASML（阿斯麦）的“铁三角”业务之一。“ASML说计算光刻是通过软件对整个光刻过程来做建模和仿真，对工艺流程做优化，比如说形貌优化、掩膜板修正等。”他说。

GPU通用计算加速的又一方向

GPU加速后，生产光掩模的计算光刻工作用时可以从两周减少到8小时。

随着晶体管和互联线宽的持续微缩，掩膜板的复杂度越来越高，相应的对计算光刻的算力要求也变高。

“按照过去15年的趋势，如果某个foundry（集成电路代工厂）现有3座数据中心，那么未来10年内就要100座这样的数据中心。”Vivek Singh说，“功耗方面，45兆瓦可能还能接受，但如果是45千兆瓦，问题就比较大了。对此，英伟达给出的回答是全新的AI加速技术cuLitho。”

Vivek Singh还提到，包含于计算光刻中的OPC（光学临近效应修正）含有大量矩阵乘法运算，这种运算很适用于GPU加速。说到底，计算光刻也是GPU通用计算加速的某个应用方向，站在这个角度看，就不难理解英伟达发布cuLith加速库了。

此前，这种“精雕细琢”的计算光刻依赖CPU（图形处理器）服务器集群。据介绍，英伟达cuLitho可以实现在500套DGX H100（包含4000颗Hopper GPU）完成与4万颗CPU运算服务器相同的工作量，但速度快40倍，功耗低9倍。

可见，GPU加速后，生产光掩模的计算光刻工作用时可以从两周减少到8小时。台积电可以通过在500个DGX H100系统上使用cuLitho加速，将功率从35MW降至5MW，从而替代用于计算光刻的4万台CPU服务器。也就是说，传统的计算模式两周才能完成的任务，同时使用英伟达的芯片和软件就能在一夜之间搞定。

值得注意的是，黄仁勋还提到了cuLitho在台积电2nm工艺中的使用。借助cuLitho，台积电可以缩短原型周期时间，提高晶圆产量，减少芯片制造过程中的能耗，并为2nm及以上的生产做好准备。据悉，台积电将于6月开始对cuLitho进行生产资格认证，并会在2024年对2纳米制程开始风险性试产，2025年开始量产。

“可以大胆预测，对于IC设计中的版图设计，这个AI工具（指cuLitho）应该也会有很大帮助。”前述国内IC设计公司研发人员表示。不过，对于cuLitho是否运用AI，Vivek Singh并没有明确回答。

对于是否支持旧工艺的问题，Vivek Singh表示，cuLitho是用作提升解析力、消除衍射的计算光刻，那么对于存在这种需求的工艺来说都适用。“不过，我认为这还涉及到成本和foundry的选择问题，包括cuLitho究竟能多大程度提升良率。”黄烨锋说。

英伟达方面表示，除了台积电之外，还正在与ASML、Synopsys合作，将这种技术推向市场。Synopsys表示，正将cuLitho软件库整合到EDA软件中，而ASML则是“与英伟达在GPU和cuLitho方面紧密合作，计划将GPU支持整合到所有的计算光刻软件产品中”。

比亚迪入股昆仑芯引关注

本报记者 谭伦 北京报道

《中国经营报》记者日前从天眼查上注意到，由百度作为最大控股股东的昆仑芯（北京）科技有限公司（以下简称“昆仑芯”）发生了工商变更，新增投资方中出现了比亚迪的名字。

变更信息显示，昆仑芯注册资本由1767.77万元增至1785.25万元，比亚迪股份有限公司、深圳市创启开盈商务咨询合伙企业（有限合伙）等新进为其股东。其中比亚迪出资额为5.8万元，完成此次工商变更后，持有昆仑芯的股权比例增至0.33%，而昆仑芯最大股东仍为百度，后者持有70.87%的股权。

据公开信息，昆仑芯成立于2011年6月，前身为百度智能芯片及架构部。2021年4月，昆仑芯以130亿元的估值完成首轮独立融资。在国内，昆仑芯是最早布局AI加速领域，并在体系架构、芯片实现、软件系统和场景应用均有积累的AI芯片企业。

对于此次入股细节，《中国经营报》记者联系了昆仑芯方面相关人士，截至发稿暂未获得回复。但有业内人士告诉记者，虽然此次入资规模很小，但释放了比亚迪进军车芯领域的信号，双方后续或许会有更深入的合作。

百度之“芯”

昆仑芯缘何被比亚迪相中，原因或许在于这家号称“百度之芯”的子公司所致力的目标。

“昆仑芯正研发的第三代AI芯片，是针对高阶自动驾驶系统，未来会考虑推出定制的车规高性能的SOC系统级芯片。”在2022年举行的百度Apollo Day技术开放日上，昆仑芯科技CEO欧阳剑如此介绍。

此前，昆仑芯AI芯片已经迭代两代，其第一代产品诞生于2018年，这一采用100%全自研XPU架构、三星14纳米工艺、算力达到260Tops的AI芯片最初定位是云端全场景。发布两年后量产实现2万片。

“昆仑芯首代产品的汽车色彩还没有那么浓。”CHIP全球测试中心中国实验室主任罗国昭向记者表示，彼时包括BAT在内互联网巨头都在布局并推出AI芯片，但由于AI的商用场景非常多，百度对于昆仑芯初代的产品定位还未明晰。

自2019年始，国内智能与电动汽车产业开始爆发式增长，而百度也开始挺进造车领域。2021年1月，百度宣布正式组建智能汽车公司，以整车制造商的身份进军汽车行业。同年，百度智能芯片及架构部从集团独立出来，成立了如今的昆仑芯。

同年8月，昆仑芯二代问世，这一支持NLP、视觉、语音等各种类型算法的迭代版本被视为百度入局车芯的标志。百度方面也于随后宣布昆仑芯二代已完成无人驾驶场景端到端性能适配，第三代将于2024年初量产。而据欧阳剑在2022年底披露，昆仑芯二代已经出货数万片，并实现亿元级商业化收入。

半导体产业分析师季维认为，相比于一代产品，昆仑芯二代在支持自然语言处理、计算机视觉、语音以及传统机器学习等车规芯片的功能性上已经越来越明显，加上百度纯自研的架构，因此，将是百度未来主打的AI芯片产品。

值得注意的是，在日前举行的备受关注的文心一言发布会上，李彦宏介绍，在文心一言的四层AI架构的布局中，昆仑芯正是位于芯片层。“这也表明百度未来会在昆仑芯上投入重金。”罗国昭表示。

比亚迪的半导体版图

比亚迪入股昆仑芯的意图，在其董事长兼总裁王传福抛出“新能源汽车的上半场是电动化，下半场是智能化”的观点后，便已进一步明晰。

“汽车智能化，本质还是复杂场景下由车机系统处理各类问题的计算能力。”季维表示，因此汽车智能化水平的比拼，比拼的其实还是算力，这也是业内认为半导体将主导未来智能汽车时代的主要原因。

比亚迪的半导体版图也由此开始构建。2020年4月，在对比亚迪微电子等子公司进行重组后，比亚迪成立了比亚迪半导体有限公司，业务范围覆盖功率半导体、智能控制IC、智能传感器和光电半导体的研发、生产和销售。

“2020年开始的缺芯潮或许加重了比亚迪的忧虑。”季维表示，很多国内造车新势力在那一轮缺芯潮中被迫减产，尤其是高端芯片，因此给很多车企造成了很大困难，比亚迪也许也是其中之一。

记者注意到，进入2021年后，比亚迪明显加快了在芯片领域的布局。在投资宁波、广东、长沙、西安、济南等5家半导体子公司后，比亚迪又投资了进芯电子和中科昊芯两家DSP（数字信号处理器芯片）厂商，受到业界关注。

同年，比亚迪重磅宣布了与国内车规芯片独角兽地平线以及GPU巨头英伟达达成商业合作。按照计划，2023年上半年起，比亚迪将在部分新车上搭载英伟达DRIVE Hyperion平台，实现车辆智能驾驶和智能泊车，而搭载地平线面向L4高等级自动驾驶的第三代车规级产品征程5的比亚迪车型最早将于2023年中上市。

窥探比亚迪半导体版图的投资策略，罗国昭分析称，从智能座舱、操作系统、视觉算法到各类传感器，比亚迪几乎是全领域涉足，而且基本是与智能汽车的产业核心需求同步，因此，此次投资昆仑芯，应该也是看中了前者未来在高端智能驾驶系统上的投入。

但鉴于当前比亚迪在昆仑芯的持股比例，罗国昭认为，比亚迪不太可能直接参与昆仑芯的研发环节，但或许会作为合作伙伴，提出自身的场景需求。

据不完全统计，截至目前，比亚迪及旗下公司投注的半导体企业接近20家，涵盖设计、材料、设备等产业链各个环节。

车企巨头抢滩“造芯”

比亚迪的入局，也让目前已成红海的汽车芯片市场，竞争更加白热化。

记者注意到，除比亚迪外，目前包括小米、蔚来、小鹏、理想等造车新势力，以及大众、上汽、东风、长城等传统车企，都已经开始自研或者合作参与到这场“造芯”运动之中。

“和比亚迪一样，芯片短缺带来的焦虑是推动车企下场造芯的原因之一。”季维援引市研机构Gartner在去年发布的一份报告指出，由于芯片短缺趋势，全球前10大车商中的半数将自行设计芯片，以期主导车芯的供应链。

同时，需求则是另一大拉动力。据中国电动汽车百人会副理事长兼秘书长张永伟在2022年底举行的行业会议上介绍，2022年中国汽车智能化渗透率超过30%，到2030年，这一比例将达70%，届时中国汽车芯片的规模约300亿美元，芯片需求量可达1000亿~1200亿颗/年。

而在巨大的需求之中，系统级SoC芯片也在占据高端汽车芯片的需求榜首。财通证券分析师张益敏指出，智能汽车时代，将CPU与GPU、FPGA、ASIC等通用/专用芯片异构融合、集合AI加速器的系统级芯片应运而生。这也正是比亚迪与昆仑芯都看重的领域。

但是，车企下场能否成功，业内却依然持谨慎态度。罗国昭认为，芯片是长周期的资金与技术密集型行业。“长期投入与耐心等待产出非常重要，很多车企也许坚持不到回报产生，就被市场淘汰了。”他指出。

与此同时，季维指出，汽车芯片是专业性很强的功能性芯片，这意味着使用场景较为单一。“如果只是企业自给自足，而没有足够的商业化需求，那是非常不划算的事情。”他表示，巨大成本与商业模式的不清晰，会是未来国内车厂造芯的主要挑战。