

英伟达用软件“捕获”元宇宙



本报记者 李佳师

如果说图形芯片、AI芯片、数据中心芯片供应商英伟达是一家软件公司,可能会被很多人笑话“不专业”。但事实上,英伟达正在成为一家靠软件平台驱动来赚钱的芯片公司。内部人士透露,英伟达75%的员工为研发人员,而研发人员中一半为软件开发人员,这样的人员比例足以看出软件平台战略在公司的地位。英伟达也一直声称自己是“计算平台”公司而非芯片企业。

英伟达创立时市场上有20多家图形芯片企业,三年后市场上的图形芯片玩家是70多家。英伟达之所以能够一骑绝尘,离不开15年前开发的CUDA编程模型和数据中心GPU平台,正是这个平台把英伟达带到通用计算芯片市场。

靠着定制的软件平台、开发工具、应用工具,英伟达构建了“由软件、大学、初创公司和合作伙伴组成的强大而稳健且自给自足的生态系统”(福布斯杂志评价)。

通过研发特定软件平台,英伟达不断拓展疆土,超越英特尔、AMD、三星、台积电,成为全球市值最高的半导体公司。

在前几天举行的英伟达GTC大会上,身着皮衣、又一次在自家厨房“谈话”的英伟达创始人兼CEO黄仁勋,同样没有大谈坚硬的芯片,而是细数了几大新软件平台,其中的每一个平台都有可能给英伟达创造数以百亿美元的新增市场机会。

Omniverse 捕获元宇宙百亿美元市场?

黄仁勋总能撩动产业与资本的神经,因为他的芯片策略永远是梦想与平台开路,当开发者如期来到他的特定平台,自然带来芯片的大卖。其梦想与平台先行战略屡试不败,现在正呈现“雪球效应”。在这次GTC大会上,黄仁勋透露,目前英伟达的开发者近300万,在过去15年里,CUDA的下载量是3000万次,仅去年一年,下载量达到700万次。

现在,英伟达将平台战略复制到元宇宙市场。今年元宇宙炙手可热,而英伟达早在两年前就开始了它的元宇宙平台研发。去年12月,英伟达元宇宙平台Omniverse正式上线公测,黄仁勋没有像Meta(脸书)那样将这个元宇宙平台吹得神乎其神,而称其为全球4000多万工程师、设计师的实时仿真与协作平台。但这个协作平台被各界看好。

前几天,英伟达的市值超过伯克希尔哈撒韦成为全美排名第7的公司,富国银行将英伟达的股票价格预期从245美元调高到320美元,给出的理由为“英伟达将是元宇宙的硬件入口”。未来5年,元宇宙将为英伟达提供100亿美元的市场增值机会。而这个能够带来100亿美元增长的引擎,正是Omniverse。

Roundhill的研究副总裁Mario Stefanidis认为,英伟达的Omniverse平台是虚拟现实且兼容的,允许用户实时合作,用真实的光线创建精确的物理模拟。“这种类型的基础设施将是虚拟世界的关键,因为用户会寻求身临其境的体验,将现实世界和虚拟世界连接起来。”

关于英伟达的元宇宙策略,HTC中国区总裁汪丛青认为,英伟达是提供PC/服务器/云图形解决方案和内容创建工具解决方案的基础设施技术提供商,拥有许多基于云的XR解决方案/工具,用于将工作处理转移到网络;英伟达与业内许多参与者合作,试图实现他们的计划,英伟达的元宇宙战略会为行业带来很高的价值。“当一家公司明确其在整个元宇宙故

事中扮演一个角色,而不是成为整个元宇宙的所有者时,他们将有更大的成功机会。”汪丛青说。

就在几天前,《时代》杂志将2021年的最佳发明奖颁给了Omniverse,评价是“虚拟世界不仅仅是为游戏而打造——它对于道路、建筑等基础设施的规划来说也有帮助,并且可以用于测试自动驾驶汽车。极度逼真的虚拟空间有助于现实世界目标的实现,英伟达的Omniverse项目正在让创造过程变得更简单。该平台结合了最新的NVIDIA GPU的光线追踪技术与一系列开源工具,允许在真实感十足的3D世界中进行实时协作。Foster建筑事务所的建筑专家正在使用该技术,在更早的设计阶段进行建筑细节的可视化;宝马公司则使用它建造了一个工厂的数字孪生,以便测试更高效的流水线的可能性”。

Meta元宇宙的目标客户是茶余饭后进行娱乐游戏的人群,英伟达的客户则是企业工程师与设计人员。元宇宙对于大众消费者或许是“锦上添花”,但对于企业用户和创业人员则是创造生产力的“必需品”。

事实上,英伟达的竞争对手并不是Meta,而是AMD与英特尔。就在英伟达举行GTC的11月9日,AMD与Meta宣布达成合作,Meta将采用AMD的数据中心芯片,这条消息让AMD的股价创下历史新高。同为芯片厂商,AMD触达上层的元宇宙应用生态需要借助Meta的平台,而英伟达自己就有平台,这才是英伟达在资本市场扶摇直上的根本原因。

Omniverse 能力不断丰满

Omniverse去年年底开始上线测试,现在拥有700多家企业级用户,包括宝马、爱立信、航空航天制造商洛克希德·马丁、索尼影视动画等。

此前,根据英伟达介绍,Omniverse分为三部分。第一部分是数据库引擎Omniverse Nucleus,用户可在此连接并进行3D资产和场景描述的交换,需要建模、布局、阴影、动画、照明、特效或渲染设计师可以一起协作创建场景。第二部分是合成、渲染和动画引擎——虚拟世界的模拟,比如英伟达的图形技术可以实时模拟每条光线如何在虚拟世界中反射。第三部分是NVIDIA CloudXR,它包括客户端和服务端软件,可用于将扩展现实内容从OpenVR应用程序传输到Android和Windows设备,允许用户进出Omniverse。

今年的GTC上,英伟达对Omniverse进行了一系列性能更新,并发布了最新工具,还举办了Omniverse技术专场。黄仁勋表示:“有了Omniverse,我们就有了创建全新3D世界或对物理世界进行建模的技术。而如何使用Omniverse模拟仓库、工厂、物理和生物系统、5G边缘、机器人、自动驾驶汽车,甚至是虚拟形象的数字孪生,是一个永恒的主题。”

英伟达的软件平台与开发工具、应用开发工具之所以赢得大量“开发粉”,当然不是因为黄仁勋能忽悠。从PC时代开始,英伟达对工具重视,对开发者体贴,现在这一传统在元宇宙发扬光大。

在最新发布的Omniverse工具中,有几个工具备受关注。

其一是Omniverse Avatar(阿凡达),这是一个用于创建虚拟人的开发工具,基于Omniverse平台以及世界上最大的定制语言模型Megatron530B。该工具可以让开发者便捷地开发出AI驱动交互式虚拟形象,可使此角色具备视觉和语言能力,能够就各类主题进行交谈,并理解自然语言中的深层含义。

为了让开发者见识英伟达阿凡达的能力,黄仁勋在GTC现场演示了用阿凡达创造的自己的虚拟形象“Toy-Me”,还带来了帮人点餐的虚拟人“蛋壳人”,以及可以嵌入会议软件的虚拟人。

其二是Omniverse Replicator,这是一种合成数据生成引擎,目的是帮助构建更好的数字孪生体。Replicator应用使开发者能够以人类无法做到的方式引导AI模型,填补现实世界的数据空白,并标记真实的数据。

在GTC的主题演讲中,英伟达还展示了Omniverse的一系列新功能,包括Showroom、Farm以及Omniverse AR等。Omniverse VR将率先推出完全图像,实时光线追踪VR环境。

在各个风口 布设关键平台

作为市值节节攀升的平台型半导体公司,英伟达当然不会只守着元宇宙这一亩三分地。在包括自动驾驶、机器人、地球气候建模、大模型、数据中心等领域,英伟达没有错漏一个,统统进行了布局,推出相关的平台、框架、工具以及算法模型等。

自动驾驶开启的是万亿美元的市场,黄仁勋认为:“到2024年,绝大多数新电动车都将具备真正的自动驾驶功能。”针对自动驾驶,英伟达不可能缺席,推出了NVIDIA DRIVE,这是英伟达的自动驾驶汽车全栈开放式平台,而Hyperion8是英伟达最新的完整硬件和软件架构。

在这次大会上,黄仁勋详细介绍了Hyperion内置的几项新技术,包括用于DRIVE Sim的Omniverse Replicator。自动驾驶领域的诸多领先企业、创业团队都采用了英伟达的端到端自动驾驶解决方案,包括Lotus、百度、威马汽车、高合汽车、蔚来、轻重智行等,甚至特斯拉早期同样采用了英伟达的GPU以及系统平台。

机器人是又一个兵家必争之地,英伟达同样在此重兵布局。英伟达的机器人生态系统VIDIA Isaac已有700多家公司和合作伙伴,这个数字在过去4年里增长了5倍。黄仁勋在GTC大会上宣布,NVIDIA Isaac机器人平台可轻松集成入机器人操作系统(ROS)。这是一套应用广泛的软件库和工具,适用于机器人应用。黄仁勋解释道,建立在Omniverse之上的Isaac Sim是有史以来最逼真的机器人模拟器。

大模型被视为通往通用AI的一条可能路线,也是AI的战略要地,目前IT头部企业均在此布局。

针对大模型,黄仁勋发布了用于训练大型语言模型的Nemo Megatron。他表示,此模型“将成为有史以来最大的主流HPC应用”。同时,英伟达还发布了NVIDIA Modulus,用于构建和训练物理学机器学习模型,这些模型可以学习并充分遵守物理学法则。

此外,黄仁勋还宣布推出全新的Clare Holoscan平台,以创建更好的医疗健康服务;推出量子计算模拟平台,显著提升量子模型性能,将于12月上线,构建Earth Two(E-2)模拟和预测气候变化。“为了创建Earth Two,我们需要使用迄今为止发明的所有技术。”黄仁勋说。

不过,对于英伟达是否是一家软件公司,汪丛青认为:“Not really,英伟达是工具和基础设施提供商。他们不制作内容,而是提供一些底层技术/工具,让其他人能够制作。”

在数字经济时代,在软件正在定义一切的时代,我们用软件构建一切,而这一切又怎么能离开芯片,这是英伟达打造计算平台、开发工具、应用工具的真正逻辑。

“十四五”规划提出构建愿景 6G准备工作箭在弦上

本报记者 姬晓婷

11月16日,工信部发布《“十四五”信息通信行业发展规划》(以下简称《规划》),将开展6G基础理论及关键技术研发列为移动通信核心技术演进和产业推进工程,提出构建6G愿景、典型应用场景和关键能力指标体系,鼓励企业深入开展6G潜在技术研究,形成一批6G核心研究成果。

根据“每十年更新一代”移动通信技术的产业规律,6G技术的准备工作已经箭在弦上。在6GANA(6G Alliance of Network AI)第二次会议上,中国工程院院士邬贺铨指出,每当新一代移动通信技术开始商用,更新一代的移动通信研究就要开始启动,这个过程一般需要十年,经过需求提出、标准提出、技术准备、试验等阶段才能走到商用,因此现在启动6G研究正当时。

6G需要什么样的 传输性能指标?

0.1~1Gbps、毫秒级端到端时延、500+km/h移动性等是对5G提出的性能指标需求。对此,有专家称,上述特征依然是6G需要关注的指标,但6G的需求绝不限于此。

高通技术市场总监Danny Tseng在接受《中国电子报》记者采访时表示:“无线通信技术的基础特性将持续提升,6G将带来全新水平的速率、时延、覆盖范围和移动性,6G也将智能化、用户体验、可扩展性和易于部署等维度上引入全新特性。”

通信数据是通信系统最重要的表征指标之一。紫光展锐首席技术专家(标准研究)潘振岗在接受《中国电子报》记者采访时表示:“在移动通信从语音为主转为数据为主之后,3G/4G/5G每代系统的数据速率指标都是以1个数量级进行提升,6G也是这样。”一般情况下,这些指标特指通信系统的空口速率。而实际上,用户端到端的速率体验会在这一水平基础上大打折扣。普通用户在日常状态下享受流畅的全高清视频通话的机会并不多,空口的极限速率需要特定条件的支撑。

提升数据传输速率需要频谱资源的支持,需要更高的频段、更宽的频谱。“当前讨论的太赫兹传输,基本上是0.1THz以上的频谱,是当前使用的毫米波频段(20GHz~80GHz)的节奏合理的技术延伸。”潘振岗在接受《中国电子报》记者采访时提出了这样的观点。他同时也提到,高频段的传输距离的覆盖极其受限,这就要求明确把握太赫兹在未来通信系统中的定位。

6G需要什么样的 通信芯片?

6G需要性能更优的芯片吗?从当前市面上5G手机普遍搭载的5G通信芯片来看,7nm已成为芯片厂商首选。高通骁龙865Plus、华为海思麒麟990 5G、联发科天玑1000+均采用7nm工艺制程。根据中国移动发布的2020年《5G芯片评测》,华为麒麟990平均下载速率已经达到800Mbps,高通骁龙865、联发科天玑1000+平均下载速率也达到700Mbps。

潘振岗认为,以目前半导体领域对未来10年或15年的技术发展展望而言,当前研发的芯片的能力能够继续支持移动通信系统的发展需求。潘振岗表示,极高数据速率考验的是数字基带芯片的计算能力和密度,5G时代使用的是12nmFinFET,预计7nm、5nm将延续较长的时间,之后将进入3nmGAA、2nmGAA时代。6G时代将始于1.5nmGAA,在与1nm和0.7nmGAA等效后实现大发展。

Gartner研究副总裁盛陵海在接受《中国电子报》记者采访时表示,要进一步提升传输速率,高耗能是需要解决的关键问题:“这不是通信技术的问题,而是物理问题。在一定的工艺下,传输的数据量更大,这势必带来功耗的提升。所以需要两条腿走路,一边提升速度,一边降低功耗。但现在半导体芯片已经接近物理极限,英特尔规划的20A已经是原子的物理尺寸级别,在此基础上再如何实现提升,便成了问题。”

6G有哪些典型 应用场景?

赛迪智库无线电管理研究所曾编撰《6G概念及愿景白皮书》,针对6G的未来应用,提出了人体数字孪生、空中高速上网、基于全息通信的XR、新型智慧城市群、全城应急通信抢险、智慧工厂PLUS、网联机器人和自制系统七大应用场景。

关于6G的未来应用场景,潘振岗在接

受《中国电子报》记者采访时指出,6G应当满足“无界”,即打破各行业、各领域的专业壁垒,实现深度渗透融合。根据用户的不同需求,6G要实现对频谱、计算、存储资源的灵活调度。在6G网络中,频谱接入的趋势是以低频段为基础,高频段按需开启,实现低频段、毫米波、太赫兹和可见光多频段共存与融合组网,在覆盖、速率、安全等方面满足不同的用户需求。计算任务应该在终端、边缘节点、云端等实现灵活、实时的分割、下放和卸载,以匹配不同节点的计算能力和能效需求。

此外,潘振岗将“AI”视为6G技术服务的核心场景。

例如在医疗行业,AI辅助诊疗/癌症诊断、医疗影像智能识别等需要高质量的影像数据传输;医疗机器人需要低时延高可靠性的传输数据。在交通行业,AI技术应用在智慧桥梁健康监测与运维管理、智能航道技术、智慧码头、无人驾驶以及无人机配送等方面;基于大量传感器和机器人的检测数据收集、事故问题定位以及移动硅基预测是这些领域的突出需求。对于仿真设计,基于VR、AR技术进行数据传输时,需要高质量的数据传输通道。对于个性化生产和质量监控,工业互联网积累的大量有价值的生产数据,需要低时延高可靠性的传输通道。

而要满足上述场景需求,潘振岗认为,几个领域的新指标尚需重新定义,包括数字孪生等级、空天覆盖能力、计算能力、智能级别、感官指标、存储能力级别等。

此外,也有专家认为,引领6G发展的核心场景不仅限于支持AI应用。Danny Tseng在接受《中国电子报》记者采访时表示:“AI和机器学习赋能的端到端通信、频谱扩展和共享、新的无线空口设计可扩展网络架构,通信系统的弹性以及物理世界、数字世界与虚拟世界的融合将成为引领6G之路的重点研究领域。”

6G要实现 空天地海一体化?

更高频率、更高传输速率带来的直接结果便是覆盖面积受限。而物联网、AI等新技术的应用,必然对通信的全域覆盖效果提出更高的要求。

在潘振岗看来,“全域覆盖”实际上包括了两个不同方向的通信覆盖能力的延伸。一是向外的延伸,即空天地海一体化;二是向内的延伸,即地域网、新的人机交互等方面。

“要实现空天地海一体化,卫星、高空平台等的引入是必要条件。”潘振岗在接受《中国电子报》记者采访时表示,“近20年,地面移动通信系统走过了3G、4G、5G,这深刻影响且改变人们的生活。相比之下,卫星通信系统的历史更为悠久。最近几年,随着oneweb、starlink的兴起,卫星通信又一次进入大发展前夜。”

盛陵海在接受《中国电子报》记者采访时表示,卫星通信将可能成为6G发展的重要方向。现阶段,很多偏远地区都没法布局5G、4G基站,要实现“全域覆盖”,卫星通信可能是解决之道。但若布局卫星通信,又会面临组网和维护成本、维修难度、轨道密集度、国际合作等在内的新问题。

“整合卫星通信接收地面站、布局检查系统都会比4G、5G复杂得多。再者,组网问题如何解决,全球这么多运营商是否需要各自发送卫星,都是将要面临的问题。”盛陵海在接受《中国电子报》记者采访时,提出了对卫星布网的疑问。

潘振岗表示,6G需要重点考虑的是,如何将两条不同发展轨道的技术融为一体。最彻底的融合模式是全面融合,即从组网到空口,完全实现无缝对接。简单的形式是网络各自独立发展,通过多模终端完成多系统支持。全域覆盖对终端侧工程实现能力的挑战相对较小。对于基带处理能力而言,由于卫星通信的速率相对地面系统较低,因此不存在额外挑战。射频方面,需要考虑毫米波及更高频段、大带宽的支持能力,另外还需要考虑大规模相控阵天线涉及的TR组件。

“全域覆盖对系统侧的影响非常大。需要考虑星上计算、存储、星间通信链路/组网,还需要考虑宇航级的高性能处理器、大规模存储器、星间激光/THz通信等。”潘振岗这样介绍道。

上述领域目前都存在技术瓶颈,可以从两个方面来解决:一是增强芯片自身的抗辐射、抗高低温能力,需要从材料、工艺、封装等方向解决。二是在设计上考虑极大的冗余、纠错方案,任何一种方案都决定了星上系统的处理能力无法跟地面系统相比拟,进而也会约束卫星通信系统的部分系统性能指标定义。

此外,为了利用好频谱,提升多频段利用效率,也要重新规划6G系统的架构,促进各频谱资源各司其职、合理使用。例如提升射频器件的多频段支撑、灵活可配能力,设立更加智能的频谱感知、动态共享的规则、机制等。