人工智能企业研究报告

let's go !!!

HUAWEI

amazon

Alibaba Group
阿里巴巴集团

WOW

Microsoft

Baidu 百度

Google

NVIDIA

ByteDance
字节跳动

为什么是英伟达?

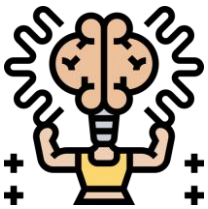
从20世纪信息技术革命以来，人类进入了网络时代，我们经历了PC互联网、移动互联网，未来还将进入物联网、车联网、以及人工智能时代。你会发现网络时代有个非常明显的特征，比如，计算机、手机是由硬件层（如CPU）、软件层（操作系统）以及上层的应用层组成（社交、电商、出行等），所以物联网、车联网以及人工智能同样具备类似的逻辑，而且我们也会发现不同层次孕育出了不同的公司，如PC互联网：英特尔、英伟达、苹果、微软、腾讯、Facebook、阿里巴巴、谷歌等；移动互联网：ARM、高通、谷歌（安卓操作系统、IOS操作系统）、苹果、微信、推特、YouTube、哔哩哔哩、字节跳动等等。所以未来的投资逻辑就是照着当前的发展情况寻找不同层次未来有可能成为行业头部的公司，此外，还应该加上这些公司的周边公司。

而英伟达则是人工智能时代的领头羊，随着数据越来越庞大，非结构化的数据越来越多，需要的算力即计算能力越来越大，例如2023年大火的ChatGPT，它是一个大型的语言模型，需要大量的数据进行训练，而训练所需的算力是惊人的，而GPU相比CPU更加适合大量的并行计算，他的计算时间会缩短很多，用一个熟知的例子解释就是：2012年吴恩达领衔谷歌大脑从1000万张图中识别一只猫，整个过程动用了1000台电脑和16000个CPU。而之后用英伟达的GPU代替了英特尔的CPU，仅用了16台电脑和64个GPU就完成了同样的识别工作。所以未来GPU可能会取代CPU成为AI算力的核心。

说到这里，按照上面的逻辑，那么英伟达不就成为了AI时代的硬件层吗？对，正是这样，但是还不止这些，如果看看PC互联网时代各个层次公司的股价就会发现，英特尔尽管雄霸全球计算机的CPU市场，但是它的市值远远低于微软的市值，而微软正是因为它控制了PC的操作系统。在信息技术行业有个规律叫做安迪比尔定律，即what Andy gives，bill takes away。比尔的操作系统吃掉了英特尔硬件的利润。所以做硬件还不够，如果能做到一统AI时代的操作系统，也许就会有统治整个AI时代的机会。

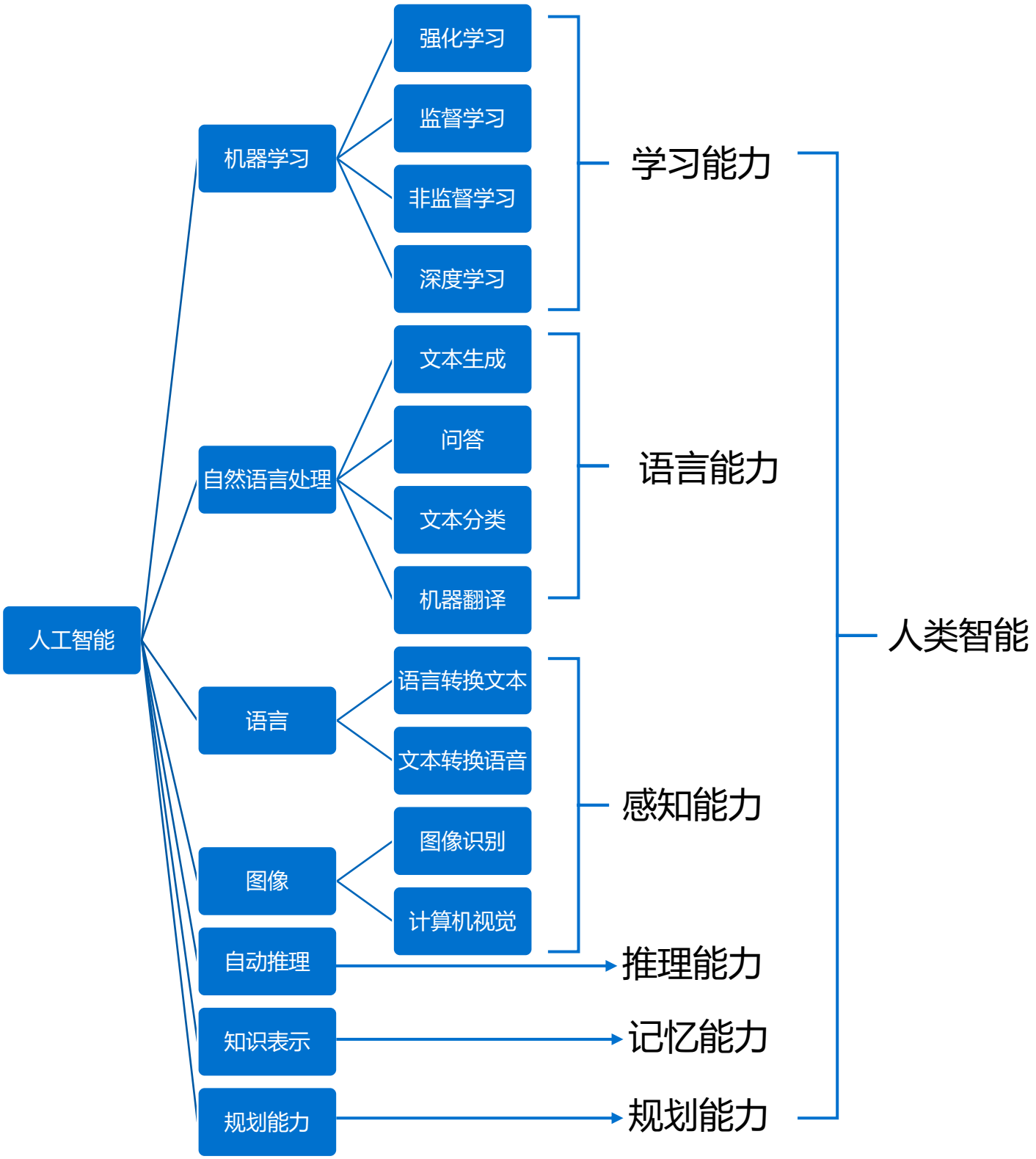
什么是人工智能?

首先我们需要介绍一些简单的人工智能的知识：人工智能的意思是让机器模拟人类的决策方式，具备人的一些能力。如图所示：



什么是人工智能?

那么为了模拟人的决策方式，还需要具备一些能力模拟人的决策，例如，人是通过感官接收外部的信息（视觉，味觉，嗅觉，触觉等等），并将信息传递给大脑，由大脑分析后做出决策，因为，机器也需要具备此类能力，这也就变成了人工智能的研究方向：



人工智能研究框架

正如前面所指出的，与PC互联网和移动互联网一样，人工智能同样分为三个层次，即**基础层、技术层和应用层。**

基础层包括芯片、传感器、算法、云计算、大数据等等；技术层包括上页所说的语音识别、计算机视觉等等；应用层就是人工智能技术在各行业的应用，比如最近大火的chatGPT。中国最强的就是应用层，在基础层的差距较大。

清楚了三个框架，那么做投资的时候就非常清晰了，就对照三个层次寻找不错的公司即可。实际上，chatGPT就是人工智能行业发展的一个转折点，从今年2月以来英伟达和微软的股价一直在涨，英伟达因为提供算力芯片，所以尽管很多行业外的人不懂，但是股价已经反映了一切，而且在2023年3月23日超过伯克希尔哈撒韦的市值，成为美国第五大上市公司。国内的人工智能公司也受此影响，股价纷纷大涨，例如科大讯飞等等。

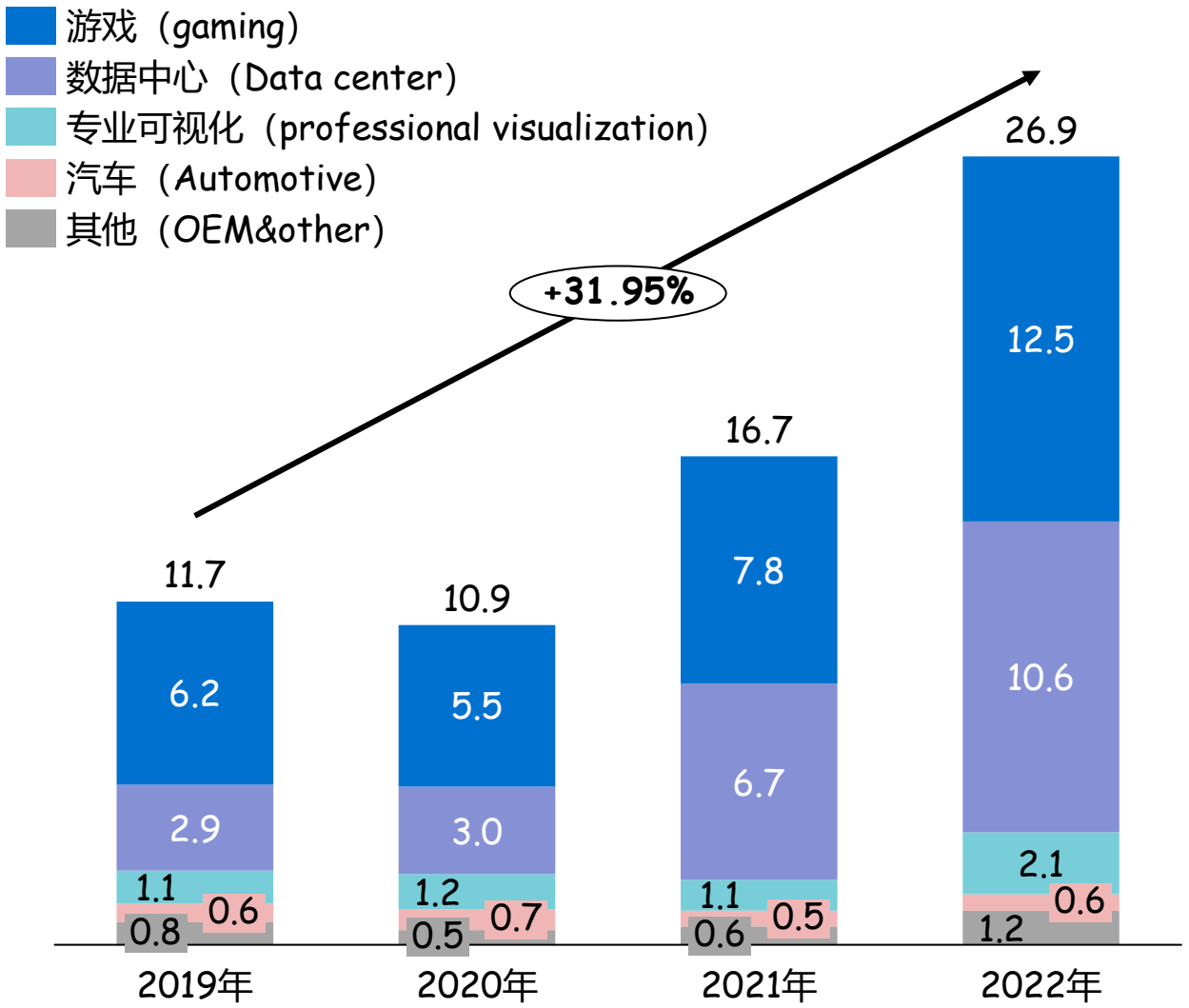
当然，英伟达不仅仅是做芯片而已，聊天机器人、推荐系统（这个时代，推荐引擎会替代搜索引擎）、自动驾驶、元宇宙等等一切跟人工智能相关的内容它都在做，这也是我从去年8月开始研究英伟达的原因，但是我想目前的了解还是不够的，还需要更深入的内容。我会继续，后续的报告会按照三个层次来研究不同的人工智能公司，为大家提供一些不同的视角，但是此份报告就先到这里，因为拖得时间太长了，对于投资来讲，时间就是一切。同时，也可以对照美股的公司研究国内的公司，例如GPU方面，国内一些公司与英伟达的差距是一到两代的差距，不过我们的举国体制下，也许有超越的机会，不是吗？

本篇报告的意义在于给大家展示一个更全面的英伟达，但是对于其商业上的分析不足，更多是技术上。

注：报告中的数据来自年报、wind、公开信息，图片来自官网，icon来自 https://www.flaticon.com/authors/darius-dan，如需复制文字等，请转成PPT使用

英伟达—人类未来科技的心脏

首先需要明确的是，英伟达是做什么的？确定一家公司做什么业务以及以什么为商业模式，最好的办法就是看其年报的收入结构，因此，我们从其近4年的年报中获取其收入情况。



数据来源：NVIDIA Quarterly Revenue Trend，单位：billion

以上的数据来自于英伟达官网的NVIDIA Quarterly Revenue Trend，与其年报收入有较小的差距，但是也基本反映了其收入结构。从收入结构图中可以看到，其游戏收入占比最高，其次是数据中心收入，再次是专业可视化部分收入，最后是其他收入部分。以2022年为例，游戏收入部分主要还是其GeForce GPU的销售；数据中心部分收入主要来自于其Ampere架构的GPU，用于云计算和AI；专业可视化的收入来自于Ampere架构产品和3D设计、AI等的需求增加。

英伟达发展历史梳理

接下来的篇幅，将主要来分析英伟达目前在做什么，以及做到了什么样的程度，这对于我们的启示在于，由于英伟达是以GPU为核心进而拓展出了人工智能、元宇宙等更新的业务，我们可以通过了解英伟达的发展现状，从而理解目前这些领域最前沿的技术是什么？这些技术可以做什么？并且这一发展过程绝不仅仅是英伟达单打独斗，他是一家生态型企业，各部分业务都涉及诸多群体，我们也可以挖掘更多的公司，他们将是人工智能时代的明星。例如，英伟达的自动驾驶业务所需要的HD mapping companies(即高精度地图行业)，这真是个有趣的发现，因为要发展自动驾驶业务的话，对于高精度地图的需求必不可少，例如高德地图等公司，所以当你研究到这步的时候，不得不佩服阿里、腾讯这些公司，他们早已布局，而且你会发现高德地图衍生了非常多的商业模式，it's amazing。带着这样的好奇心，那我们就先看看英伟达究竟是一家什么样的公司，以及他最终构建的究竟是什么样的商业版图。

首先，沿着时间的脉络，来了解这家公司的发展历程：

1993 黄仁勋和其他两位创始人共同创立了英伟达

1994 与 SGS-Thomson Microelectronics 达成了首个战略合作伙伴关系，为该公司制造单芯片图形用户界面加速器

1996 推出首款支持 Direct3D 的 Microsoft DirectX 驱动程序

1997 推出全球首款 128 位 3D 处理器 RIVA 128。

1998

- 与台积电签约建立多年战略合作伙伴关系
- 扩展RIVA处理器系列，RIVA 128ZX提供了业内超快的3D处理能力，RIVA TNT是第一款多纹理3D处理器。

1999

- NVIDIA 发明了图形处理器，由此走上了重塑行业的道路
- 宣布以每股 12 美元的价格首次公开募股

2000

- 收购显卡技术先驱 3dfx
- 发布全球首款笔记本电脑 GPU – GeForce2 Go
- 微软选择NVIDIA为其首款Xbox游戏机提供图形处理器

2001

- 成为发展最快的半导体公司，收入达到 10 亿美元，并被纳入标准普尔 500 指数
- 携 NFORCE 进军集成显卡市场
- 推出业内首款可编程GPU-NVIDIA GeForce3，使开发者能够创建定制视觉效果

2002

- 第 1 亿台处理器出货
- 推出NVIDIA 推出游戏之道，旨在鼓励游戏开发者充分利用 GPU 的强大功能

2003

- 收购无线领域图形和多媒体技术的领导者MEDIA Q
- 推出“Dusk、Ogre 和 Time Machine”的三重演示，展示了栩栩如生的几何形状的头发、动态模糊、基于时间的着色、高级皮肤着色和其他突破性的图形功能

2004

- 与暴雪娱乐合作，发布了采用3D图形技术游戏“魔兽世界”，这款大型多人在线游戏很快成为全球最热门的游戏
- 推出SLI技术，允许将多个GPU连接在一起，并显著提升了单台机器的图形处理能力
- 帮助美国国家航空航天局重建了火星地形；借助NVIDIA技术，漫游者号传输的数据在逼真的虚拟现实中实现渲染，让科学家们可以如同在火星表面上自由移动一样探索火星

2005

- 收购总部位于台湾的核心逻辑技术开发商Uli Electronics
- 为索尼 PLAYSTATION 3开发处理器

2006

- 第 5 亿台图形处理器出货
- 推出 CUDA，这是一种用于通用 GPU 计算的革命性架构。借助 CUDA，科学家和研究人员能够利用 GPU 的并行处理能力来应对最为复杂的计算挑战
- 收购 Hybrid Graphics，这是一家为手持设备开发嵌入式 2D 和 3D 图形软件的开发商

2007

- 创下第一季度 10 亿美元的营收业绩
- 推出 Tesla GPU，让此前在超级计算机中提供的计算能力广泛用于药物研发、医学成像和天气建模等领域研究人员的工作
- 收购 PortalPlayer，这是一家为个人媒体播放器提供半导体、固件和软件的供应商

2008

- 推出 Tegra 移动处理器，其功耗比普通 PC 笔记本电脑低 30 倍，并可提供酷炫的性能
- 收购 Mental images，这家公司是视觉渲染软件领域的领导者，其 Iray 软件与 Quadro GPU 相结合，通过逼真的设计渲染效果为创意专业人士提供即时反馈
- Apple 为其突破性产品 MacBook、MacBook Pro 和 MacBook Air 笔记本电脑采用 GeForce 9400M GPU
- 收购 AGEIA，这家公司是游戏物理技术开发商，其 PhysX 软件在游戏中用于再现物理性质对物理世界中的物体的影响效果

2009

- 与 Google 合作，在其 Tegra 处理器上运行 Android 系统
- 与 Siemens Healthcare 携手创造出全球首个 3D 超声波
- 推出了代号为“Fermi”的新一代 CUDA GPU 架构

2010

- Tesla GPU 为全球超快的超级计算机，即中国的 Tianhe-1A，提供动力支持
- 推出 Optimus 技术，这是笔记本电脑的一项突破，可实现自动管理 GPU 以平衡电池寿命和性能
- 奥迪选择 NVIDIA GPU 为全球所有奥迪汽车的导航和娱乐系统提供支持

2011

- 推出全球首款双核移动处理器 Tegra 2，在此基础上打造出首款 Android 平板电脑
- 第10亿台图形处理器出货
- 与英特尔达成为期六年的交叉授权协议，赢得15亿美元的授权费
- 在CES大会上推出“Project Denver”,这是一款基于超高效ARM架构的定制CPU

2012

- NVIDIA GRID 将图形引入云端，首款虚拟化 GPU 问世
- 推出功能强大的基于 Tegra 3 的平板电脑和智能手机
- 推出基于 Kepler 的 GeForce GTX 600 系列，可提供世界上超快的游戏性能

2013

- 令人震撼的游戏和娱乐便携设备 NVIDIA SHIELD
- 面向游戏玩家推出 GeForce GTX TITAN，采用与世界顶级超级计算机相同的“DNA”
- 收购 Portland Group，进而推动 NVIDIA 为加速计算革命创建开发者工具的进程
- NVIDIA GRID 视觉计算设备问世，可为小型企业网络上的几乎任何设备提供超快的 GPU 性能
- 发布全球超快的四核移动处理器Tegra 4和首款完全集成的4G LTE 移动处理器 Tegra 4i

2014

- 推出 Maxwell（第 10 代架构），助力 GeForce GTX GPU 在性能、图形和效率方面取得突破性进展
- NVIDIA Tegra K1 正式发布，这是一款192核超级芯片，将全球速度超快的GPU的DNA注入了移动设备

2015

- NVIDIAGeForce GTX TITAN X问世，这是有史以来功能极其强大的处理器，专为训练深度神经网络而打造
- NVIDIA Tegra X1 是一款256核移动超级芯片，可为深度学习和计算机视觉应用程序带来1 Teraflops 的处理能力
- Jetson TX1 是模块化超级计算机，支持新一代智能自主机器

2016

- 推出 NVIDIA Iray® VR，通过模拟光线和材质来创建逼真的交互式虚拟环境
- NVIDIA DRIVE PX2 支持功能强大的车载人工智能，使汽车行业走上自动驾驶汽车的道路
- 推出第11代GPU架构NVIDIA Pascal，为更为先进的NVIDIA Tesla 加速器和GeForce GTX 显卡提供支持
- 推出NVIDIA DGX-1,这是全球首款一体化深度学习超级计算机，可强力支持人工智能应用程序

2017

- 借助NVIDIA Isaac 机器人模拟器，可以更轻松地训练和部署智能机器人
- NVIDIA推出NVIDIA Volta GPU架构，借助NVIDIA Tesla V100 GPU加速器为DGX系列AI超级计算机提供支持
- NVIDIA SHIELD 借助Google Assistant 和SmartThings Hub技术将AI带入家庭
- 模块化NVIDIA Jetson TX2 AI 超级计算机为AI城市的功能强大的智能机器人、无人机和智能摄像头打开了大门

2018

- 推出NVIDIA DGX-2，这是首款能够提供2Petaflops(1petaflop等于每秒钟进行1千万亿次的数学运算)计算能力的单一服务器，由NVIDIA V100 GPU和革命性的GPU互联结构NVIDIA NVSwitch提供支持
- NVIDIA Clara 平台问世，提升了数百万种传统医疗仪器的功能，并未采用AI的医疗设备开创了未来
- **推出NVIDIA Turing GPU架构，为全球首款具备实时光线追踪功能的GPU提供支持，而实时光线追踪长期以来被视为计算机图形技术的终极目标**
- 推出NVIDIA DRIVE constellation，这是一款模拟系统，可在VR种模拟自动驾驶汽车安全行驶数十亿英里
- 推出NVIDIA Jetson AGX Xavier,支持轻松创建和部署用于制造、配送、零售、智慧城市等领域的AI机器人应用程序
- NVIDIA 推出RAPIDS，这是一个开源GPU加速平台，可加速数据科学和机器学习发展

2019

- 推出NVIDIA Studio,这个平台可以大幅提升全球4000万在线和工作室创意工作的性能和可靠性
- NVIDIA DRIVE AGX Orin 是一款适用于自动驾驶汽车和机器人的高度先进的软件定义平台
- 推出NVIDIA参考设计平台，帮助公司为日益增多的HPC应用程序快速构建基于GPU加速Arm的服务器
- NVIDIA Jetson Nano 和 Jetson Xavier NX 加入NVIDIA 适用于嵌入式物联网应用程序的高性能、低功耗处理器系列
- 丰田和沃尔沃集团使用NVIDIA DRIVE 端到端平台，开发和训练安全的自动驾驶汽车，并部署到全球的交通运输行业
- 推出NVIDIA EGX边缘计算平台，可将加速AI的强大功能引入企业边缘

2020

- 隆重推出NVIDIA Ampere GPU架构，助力实现一种新的功能强大且灵活的数据中心
- 收购了高性能互联技术领域的领头羊Mellanox，HPC领域的两家卓越公司合二为一
- 成为全球市值最高的半导体公司

2023

- Open AI推出了ChatGPT，成为了历史上用户增速最快的应用，一时间火遍全球；但是其需要超过1万枚英伟达A100 GPU芯片来提供算力支持

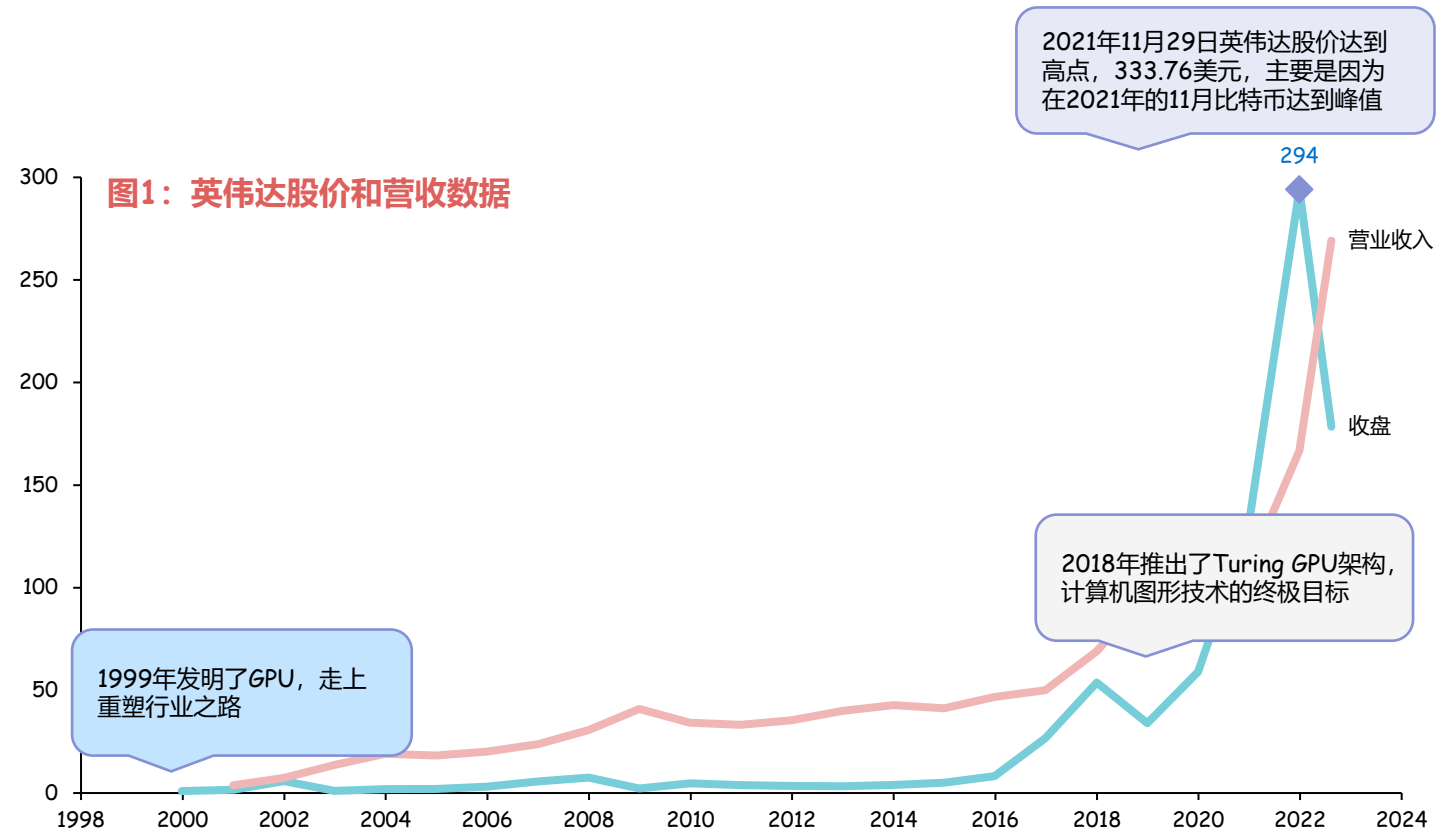


图1：英伟达股价和营收数据

图2：英伟达股价和比特币指数走势

数据来源：investing.com,wind

英伟达产品和解决方案分类

我将英伟达的股价、营业收入以及比特币指数绘制在了上一页，可以看到在2016年股价开始起飞，而且2021年英伟达股价达到了创纪录的333.67美元，这是因为比特币的价格在2021年的11月达到了创纪录的新高，而市场对于英伟达股价的预期与比特币走势较为密切，因为比特币挖矿需要英伟达的显卡，因而在比特币开始下跌时，英伟达的股价也开始了下跌。从图二中也可以看到，英伟达股价和比特币指数的走势除了红框标注的部分不一样外，其他部分，尤其是红框之后的走势高度一致，我也计算了两只股票的日数据之间的相关系数，为89.72%。显示两者的价格高度相关。

解决方案	AI 和数据科学	数据分析	机器学习	深度学习训练	
		深度学习推理	对话式AI		
	数据中心和云计算	面向企业IT的加速计算	NVIDIA launchPad	云计算	
		托管	边缘计算	网络	
		本地	虚拟化	ML Ops	
	设计和虚拟化	增强现实和虚拟现实	多显示器	渲染	
		3D协作	图像虚拟化	工程模拟	直播
	边缘计算	AI-ON-5G	智能视频分析	工业	
		机器人	边缘部署管理	边缘解决方案	
	高性能计算	高性能计算和AI	仿真与建模	科学可视化	
	自动驾驶汽车	Chauffeur	Concierge	训练	
		高精地图	仿真	自驾计程车	
		货车运输业	ADAS		

英伟达—人类未来科技的心脏

硬件	游戏和娱乐	GeForce 显卡	游戏笔记本电脑	G-SYNC显示器	
	笔记本电脑和工作站	游戏笔记本电脑	NVIDIA RTX 桌面工作站	专业笔记本电脑中的NVIDIA RTX	
		DGX station	NVIDIA RTX 数据科学工作站	Studio 设计本	
	云和数据中心	Grace CPU	DGX系统	EGX 平台	
		NVIDIA OVX	HGX 平台	DRIVE constellation	
	网络	DPU	以太网	InfiniBand	
	GPU	GeForce	NIVIDA RTX/Quadro	数据中心	Titan RTX
	嵌入式系统	Jetson	DRIVE AGX	Clara AGX	
软件	应用框架	3D协作-Omniverse	汽车-DRIVE	云端AI视频流-Maxine	语音AI-Riva
		数据分析-RAPIDS	医疗健康-Clara	高性能计算	智能视频分析-metropolis
		推荐系统-Merlin	机器人-Issac	电信-Aerial	
	应用和工具	NGC目录	NVIDIA NGC	3D-Omniverse	
		数据中心	GPU监控	NVIDIA RTX experience	
		NVIDIA RTX 桌面管理器	RTX加速的创意应用程序	视频会议	NVIDIA 工作台
	游戏和娱乐	GeForce experience	INVIDIA broadcast APP	(Omniverse) Machinima	
	基础架构	AI enterprise 套件	云原生支持	集群管理	
		边缘部署管理	推理服务	IO加速	
		软件	虚拟GPU		

图3：英伟达行业词云图

我将英伟达目前所涉及到的行业绘制了词云图，如上所示，让大家有个直观的印象，下面的篇幅将主要更加深入地研究英伟达的各种产品、解决方案以及行业等等。我们同样以前面的产品分类进行分析：

1 硬件　2 软件　3 解决方案

显卡

GeForce RTX30系列

GeForce RTX20系列

配备专用的光线追踪技术和 AI 核心，提供强大的性能和前沿的功能；

GeForce RTX16系列

NVIDIA Turing™ 架构,:高帧率游戏体验;

笔记本电脑

GeForce RTX30 系列笔记本电脑

NVIDIA 第二代 RTX 架构 – Ampere；尤为逼真的光线追踪效果和 NVIDIA DLSS 等先进的 AI 功能；由 AI 赋能的全新 Max-Q 技术，可令轻薄高性能笔记本电脑的表现远超以往；

G-SYNC显示器

G-SYNC显示器

- **获得与人眼能感知的亮度相似的真实亮度，以及比传统显示器更宽的色域范围；**
- **G-SYNC 可更大限度地降低从输入键盘指令到屏幕上显示相应动作之间的延迟，更好地满足高端玩家的需求；**
- **NVIDIA 与 LCD 显示器生产商合作，面向市场推出更宽范围的高刷新率显示器 – 从 75 Hz 一直到 360 Hz；**
- **先进的 G-SYNC ULTIMATE 显示器支持 DCI-P3 色域，能重现级别过渡更平滑、更为真实的颜色；**
- **脉冲式显示屏的响应速度超快，物体在移动时如丝般顺滑，如水晶般清晰；**
- **玩 3D 游戏时，无论是桌面窗口模式还是全屏模式，G-SYNC 都能提供同样顺畅的无撕裂体验；**

硬件-游戏与娱乐

硬件-笔记本电脑和工作站👇

NVIDIA RTX 桌面工作站	这是一种桌面级显卡，即台式工作站显卡，也就是用于台式机的显卡。NVIDIA RTX 专业桌面产品是数百万创意和技术专业人士的首选。借助世界上可视化领域功能超强的 GPU 获得非凡的桌面体验，此 GPU 具有大容量内存、先进的企业特性、经过优化的驱动以及逾 100 个专业应用认证
用于 NVIDIA RTX 专业笔记本电脑	NVIDIA RTX™ 专业笔记本电脑 GPU 集速度、便携、大容量显存、企业级可靠性和新的 **RTX 技术**于一身，采用实时光线追踪、高级图形和加速 AI 技术，能够随时随地处理要求十分严苛的创意、设计和工程任务
NVIDIA DGX station	NVIDIA DGX station A100： ①适用于数据科学团队的人工智能超级运算；②服务器级 AI 系统，但无需数据中心的电力和冷却系统；③专为在公司办公室、实验室、研究机构甚至在家中工作的敏捷数据科学团队设计，不需要复杂的安装或大量 IT 投资；④使用经 GPU 全面优化的软件堆栈和高达 320 GB 的 GPU 内存来训练大型模型，响应更迅速
NVIDIA RTX 数据科学工作站	硬件采用更为先进的GPU（NVIDIA RTX 和NVIDIA Quadro RTX 专业级GPU）；软件采用 NVIDIA CUDA-X AI 打造且经过测试和优化的综合性堆栈。此堆栈采用 RAPIDS 数据处理和机器学习库、NVIDIA 优化 XGBoost、TensorFlow、PyTorch 及其他领先的数据科学软件，可为企业提供加速工作流程，以便提高数据准备、模型训练和数据可视化的速度
Studio 设计本	NVIDIA Studio 设计本和台式电脑专为加速创作而打造，具有色彩绚丽的显示器以及高速的内存和存储；通过 Dynamic Boost 2.0 或控制声音的 Whisper Mode 2.0 等高级 NVIDIA 功能，AI 应用加速、DLSS 2.0 高速渲染以及 NVIDIA Omniverse 和 NVIDIA Broadcast 等专属应用相结合，以提供更高性能

硬件-云和数据中心👇

NVIDIA Grace CPU	专为解决全球最富挑战的计算难题：分为两款产品①NVIDIA Grace Hopper 超级芯片②NVIDIA Grace CPU 超级芯片 NVIDIA Grace Hopper 超级芯片将Grace CPU 与 Hopper GPU 相结合，专为解决巨型 AI 和 HPC 挑战；NVIDIA Grace CPU 超级芯片通过 NVLink-C2C 技术带来 144 个 Arm® v9 核心以及 1 TB/s 内存带宽；
NVIDIA DGX 系统	NVIDIA DGX™ 系统针对企业 AI 开发和规模提供出色的解、方案。 英伟达的AI系统产品组合：NVIDIA DGX Station™ A100（人工智能工作组设备）、NVIDIA DGX™ A100（AI训练、推理和分析）、NVIDIA DGX H100（完善的AI平台）
NVIDIA EGX 平台	NVIDIA EGX是一个加速计算平台，它无需将数据传到云端或数据中心，在数据产生的地方就能够进行实时感知、理解以及处理；而以往只能在云数据中心的强大机器上处理，这意味着需要传递大量的数据，会造成时延
NVIDIA OVX	专为满足 Omniverse 数字孪生的需求而打造：数字孪生彻底改变了企业设计、测试和优化复杂的系统和流程的方式，这需要多个自主系统在同一时空中进行交互。 NVIDIA ® OVX™ 专用于为通过数据中心进行大规模工业数字孪生提供技术支持，以实时创建和运行非常复杂的模型和逼真的仿真环境。
NVIDIA HGX 平台	功能强大的端到端AI超级计算平台： 专为模拟仿真、数据分析和 AI 的融合而构建
DRIVE constellation	自动驾驶需要在各种情况下进行大规模开发和测试才能部署，但是如果在现实世界里测试和模拟成本会非常高，因此，需要一个虚拟的仿真模拟平台，drive constellation 正是这样一个平台

硬件-网络

DPU

NVIDIA BlueField 数据处理器:

通过对各种高级网络、存储和安全业务进行卸载、加速和隔离，BlueField DPU 可为云、数据中心或边缘计算等环境中的各种工作负载提供安全加速的基础设施。BlueField DPU 将强大的计算能力、完整的片上基础设施可编程性及高性能网络相结合，鼎力支撑要求严苛的工作负载。

以太网

再强大的算力也需要稳定可靠的网络支持，以太网产品包括：

NVIDIA BlueField® 数据处理器 (DPU)： BlueField 将业界出色的 NVIDIA ConnectX® 网卡与 Arm 核心阵列相结合，可提供具有数据中心基础架构级芯片 (DOCA) 可编程性的专用硬件加速引擎。

ConnectX 智能网卡： 业界出色的 ConnectX 系列智能网卡可提供非常广泛、先进的硬件加速引擎。可以借助智能网卡在数据中心内实现快速、高效的以太网

NVIDIA Spectrum® 以太网交换机： Spectrum 交换机可为 AI、云和企业提供优越的性能（高达 400GbE）和规模，并有多种网络操作系统可供选择，包括 NVIDIA Cumulus Linux™、SONiC、Onyx 和 DENT/SwitchDev。

NVIDIA LinkX® 线缆和收发器： 旨在充分提高高性能计算网络的性能，满足在以太网元素之间建立高带宽、低延迟和高度可靠连接的要求。

Infini Band

InfiniBand 网卡： InfiniBand 网卡 (HCA) 可提供超低延迟、超高吞吐量和创新的 NVIDIA 网络计算引擎。

DPU： NVIDIA® BlueField® DPU 集强大的计算能力、高速网络和广泛的可编程性于一体，能为要求严苛的工作负载提供软件定义、硬件加速的解决方案。

InfiniBand 交换机： InfiniBand 交换机系统提供超高的性能和端口密度。

路由器和网关系统： 通过使用 InfiniBand 路由器、InfiniBand 长距离连接 (NVIDIA MetroX®-2) 和 InfiniBand to Ethernet 网关系统 (NVIDIA Skyway™)，InfiniBand 系统能够提供超强可扩展性和子网隔离能力。

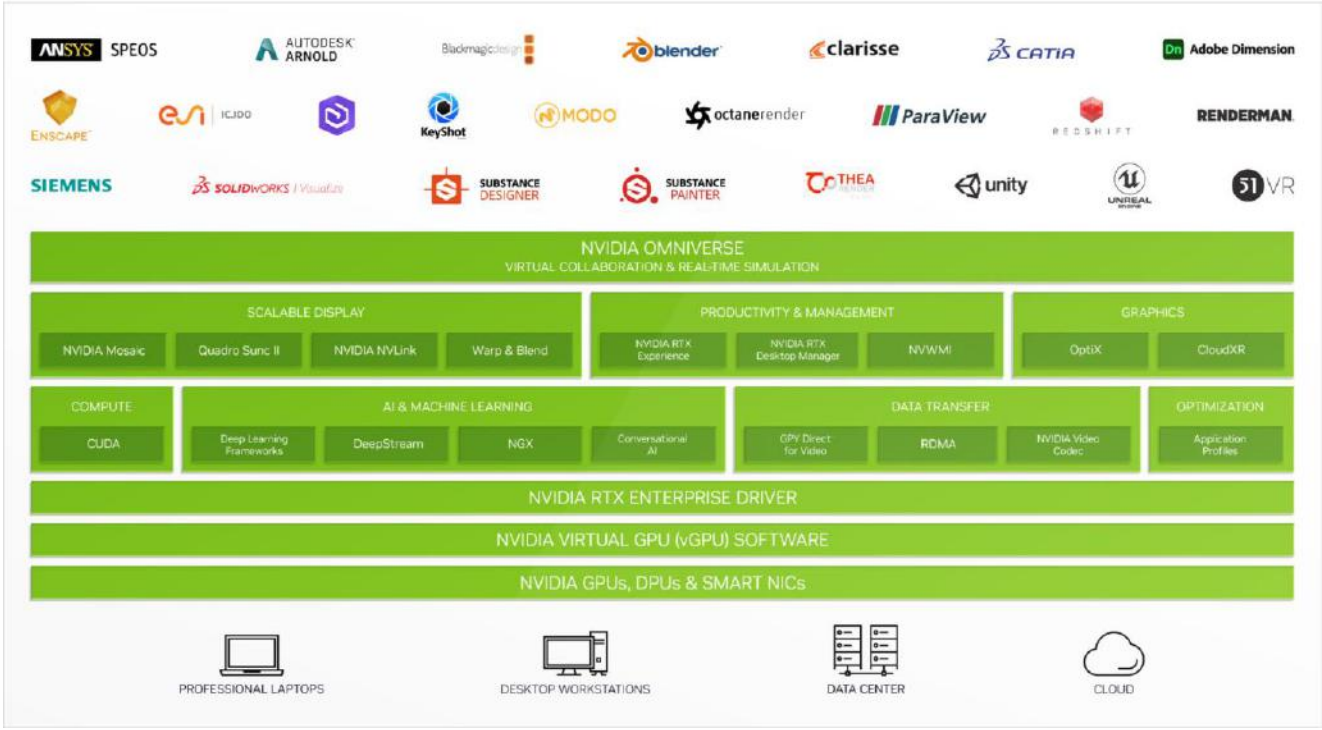
LinkX InfiniBand 线缆和收发器： NVIDIA LinkX® 线缆和收发器旨在更大限度地提高 HPC 网络的性能，满足这类网络在 InfiniBand 元素之间建立高带宽、低延迟和高度可靠连接的要求。

硬件-GPU

NVIDIA RTX/Quadro

全球出色的专业视觉平台

NVIDIA RTX 和 NVIDIA Quadro 专业解决方案助力设计师、艺术家、科学家和研究人员以更快的速度探索他们的大胆想法。在过去 20 年里，NVIDIA 开发了一个完整的生态系统，为专业人士提供完成最佳作品所需的一切，包括强大的硬件、高级软件和工具、跨行业平台以及庞大的第三方应用程序网络。



从桌面到云端的 NVIDIA RTX 和 Quadro 解决方案

专业笔记本电脑

NVIDIA RTX 和 NVIDIA Quadro RTX GPU 可在功能强大、轻薄的外形中提供一流的性能

桌面版工作站

NVIDIA Quadro 和新的基于 NVIDIA RTX Ampere 的专业 GPU 为新一代桌面工作站带来了实时光线追踪、AI 和高级图形的新技术

适用于专业可视化的EGX平台

将基于 NVIDIA Ampere 架构的 GPU 与 NVIDIA Mellanox® 智能网卡和NVIDIA 虚拟 GPU 软件相结合，借助支持多个工作负载的解决方案扩展数据中心基础架构

虚拟工作空间

NVIDIA RTX 虚拟工作站 (vWS) 软件可在任何地方使用任何设备加速要求非常苛刻的应用程序，同时确保 NVIDIA RTX 和 NVIDIA Quadro 解决方案的可靠性能。

云

NVIDIA RTX 虚拟工作站可从 NVIDIA 云服务提供商合作伙伴处访问，在数分钟内即可启动 GPU 加速虚拟工作站，只在需要时为所需买单

专业解决方案

从平板电脑到迷你工作站，可获得全尺寸工作站的强大功能和性能，以及由 NVIDIA Quadro GPU 提供支持的专用解决方案中的应用程序兼容性

嵌入式解决方案

搭载 NVIDIA Quadro 的嵌入式解决方案可供想要设计可提供高级图形、计算、深度学习和 AI 功能的自定义解决方案的系统集成商使用

硬件-GPU

NVIDIA RTX/Quadro

推动各行业的创新

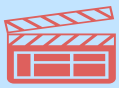
建筑、工程以及施工

世界各地的 AECO（建筑、工程、施工和运营） 公司在合作开展建筑和基础设施设计时，依靠 NVIDIA 先进的视觉计算技术来加快工作流程。建筑师、工程师和设计师使用 NVIDIA RTX™ 助力的工作站支持实时光线追踪、虚拟现实、工程模拟和采用 AI 技术的应用。NVIDIA RTX™ 虚拟工作站 (vWS) 软件为远程处理大型复杂 BIM 模型的设计团队提供桌面级图形性能。利用面向 AEC 的 NVIDIA Omniverse™，项目团队可以改变概念设计流程。

制造和产品开发

改进产品开发，推动制造业发展

NVIDIA RTX™ 虚拟计算平台可通过装有 NVIDIA 虚拟 GPU (vGPU) 软件的物理或虚拟工作站进行访问，使得当今领先的设计和工程团队能够利用逼真的可视化、实时模拟、AI 和增材制造等功能，在竞争中保持领先地位，以远超以往的速度将令人惊叹的新产品推向市场。



媒体和娱乐

通过虚拟制作、渲染、人工智能来加速影视制作流程。

虚拟制作：通过使用 NVIDIA 认证系统、NVIDIA 网络解决方案和 NVIDIA Omniverse Enterprise 平台，将虚拟制作集直接连接到艺术家，实时创建、迭代和协作。

渲染： NVIDIA RTX™ 具有专门用于光线追踪的 RT Core 和用于 AI 降噪的 Tensor Core、超级采样等多种功能，能够实时打造精美、照明精确的渲染。

人工智能：使用 AI 助手来减少重复性任务并启用新的创意功能。

能源

从数据中提取价值：利用 NVIDIA AI 工具，将常规上游业务、管道和炼油厂传感器以及维护流程中的大量数据转变为实际可行的深入见解

强力支持计算：无论是在数据中心本地还是在云中，都可借助高性能计算加速地球物理和工程应用程序

保护健康和环境：确保遵守适当的个人防护装备 (PPE) 协议，并使用 AI 技术观察设备、预测和检测故障，从而识别安全隐患，拯救生命

AI 和数据科学

AI 正在推动全球各行各业的变革。随着公司日益依靠数据来推动运营，对 AI 技术的需求也在不断增长。从语音识别和推荐系统到医疗成像和改进的供应链管理，AI 技术能为企业提供其团队完成毕生工作所需的计算能力、工具和算法。

医疗健康

AI 正在医疗健康领域创造新的可能性。计算生物学的进步正在加速药物研发的每个阶段，新一代软件定义的医疗设备支持实时感知，智慧医院正在改善临床体验，并且加速计算正在解锁人类基因组以实现更精准的医疗。

零售

领先的零售商正利用 AI 来减少损耗、改善预测、实现仓库物流自动化、决定店内促销活动和实时定价、为客户提供个性化服务和推荐，以及在实体店和网店提供更出色的购物体验

电信

通过针对高带宽和低延迟进行优化的 5G 网络提供加速服务，例如借助 NVIDIA CloudXR 和 AI 推理为更智能的城市提供增强现实和虚拟现实

金融服务

利用AI加速金融服务：
大型数据集、永久性市场波动、远程办公。智能技术可以攻克现代金融服务业所面临的关键挑战。借助 NVIDIA 的 AI 技术（包括深度学习、机器学习和自然语言处理 [NLP]），金融机构可以加强风险管理、改善数据支持的决策和安全性，并提升客户体验。

数据中心

NVIDIA Hopper 架构

英伟达在2022年3月发布了NVIDIA Hopper 架构，主要用于数据中心。Hopper 采用先进的台积电 4N 工艺制造，拥有超过 800 亿个晶体管，采用五项突破性创新技术为 NVIDIA H100 Tensor Core GPU 提供动力支持。与上一代 NVIDIA Megatron 530B 聊天机器人的 AI 推理速度相比，实现了令人难以置信的30 倍提速，是世界上最大的生成语言模型。

Transformer 引擎	在当前的计算平台上一般大型AI模型可能需要数月来训练，这对于企业而言太慢了；而 Transformer引擎是全新Hopper架构的一部分，将显著提升 AI 性能和功能，并助力在几天或几小时内训练大型模型
NVLink Switch 系统	AI 和高性能计算 (HPC)（包括新兴的万亿参数模型）领域的计算需求不断增长，在这一趋势的推动下，对于能够在每个 GPU 之间实现无缝高速通信的多节点、多 GPU 系统的需求也在与日俱增；**NVLink 是一种 GPU 之间的直接互连**，可扩展服务器内的多 GPU 输入/输出 (IO)。**NVSwitch 可连接多个 NVLink，在单节点内和节点间实现以 NVLink 能够达到的最高速度进行多对多 GPU 通信**
NVIDIA 机密计算	虽然数据在存储中和在网络传输时处于加密状态，但在数据处理期间并不受保护。NVIDIA 机密计算通过保护使用中的数据和应用来弥合这一差距。NVIDIA Hopper 架构引入了具有机密计算功能的加速计算平台
第二代 MIG	若不使用 MIG，则同一 GPU 上运行的不同工作（例如不同的 AI 推理请求）会争用相同的资源（例如显存带宽）。显存带宽更大的工作会占用其他工作的资源，导致多项工作无法达成延迟目标。借助 MIG，工作可同时在不同的实例上运行，每个实例都有专用的计算、显存和显存带宽资源，从而实现可预测的性能，同时符合服务质量并尽可能提升 GPU 利用率
DPX 指令	动态编程是一种算法技术，通过将复杂递归问题分解为更简单的子问题来解决。通过存储子问题的结果，之后也不必重新计算它们，从而减少了指数级问题解决的时间和复杂性。Hopper 架构引入了 DPX 指令，与 CPU 相比将动态编程算法速度提高了 40 倍，与 NVIDIA 前一代 Ampere 架构 GPU 相比，则提高了 7 倍。这大幅加快了疾病诊断、实时路由优化甚至图形分析的速度。

硬件-GPU👇

TITAN RTX

NVIDIA® TITAN RTX™ 是专为研究人员、开发者和创作者度身设计，并由 Turing™ 架构提供动力支持。

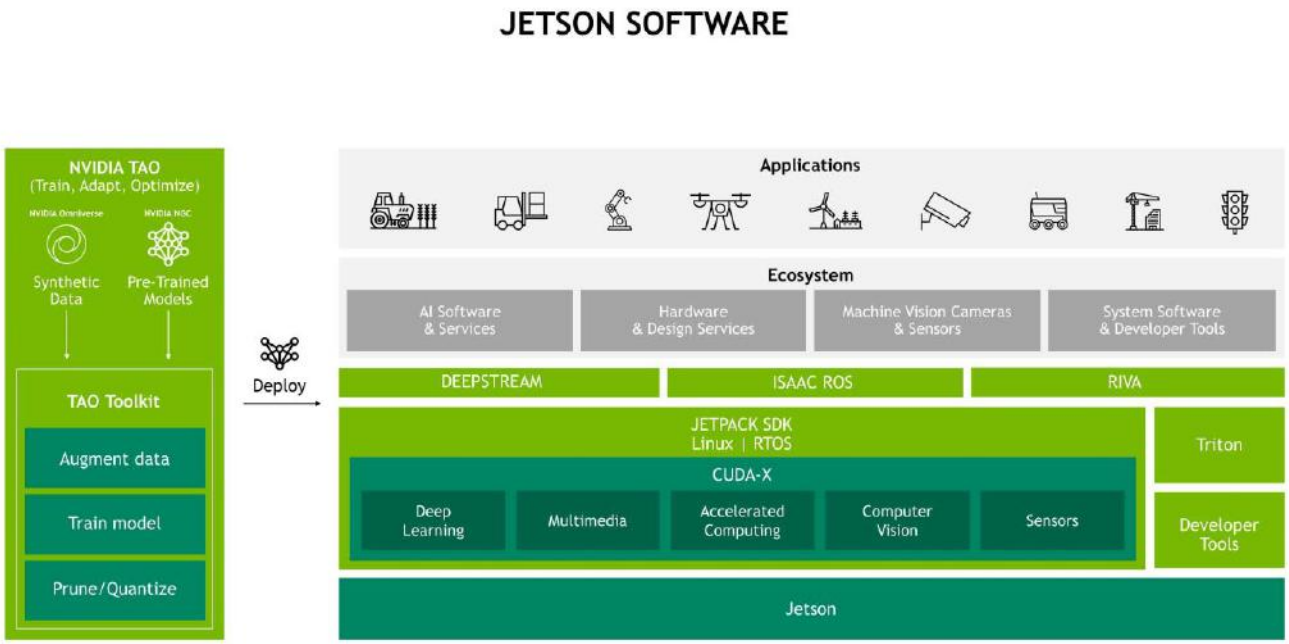
TITAN RTX 可以将诸如 ResNet-50 和 GNMT 之类的高级模型训练速度比 Titan XP 快 4 倍。 RAPIDS 训练模型的速度比 CPU 快 3 倍。 TITAN RTX 内置多精度 Turing Tensor 内核，可提供 FP32，FP16，INT8 和 INT4 的突破性性能，从而可以更快地训练和推理神经网络。 TITAN RTX 配备了前一代TITAN GPU 和 NVIDIA NVLink™ 的两倍的存储容量，使研究人员和数据科学家能够在 GPU 内存上进行比以往更大的神经网络和数据集实验。

硬件-嵌入式系统👇

Jetson

边缘计算的人工智能平台

单一且统一的嵌入式软件堆栈



NVIDIA Jetson™ 是世界领先的平台，适用于自主机器和其他嵌入式应用程序。该平台包括 Jetson 模组（外形小巧的高性能计算机）、用于加速软件的 NVIDIA JetPack™ SDK，以及包含传感器、SDK、服务和产品的生态系统，从而加快开发速度。

所有的 Jetson 模组均由同一软件堆栈提供支持，因此只需一次开发，即可在任意地方部署。Jetson 平台由强大的 Jetson 软件堆栈提供支持，旨在为 AI 应用程序提供端到端加速，并加快上市速度。

硬件-嵌入式系统

Drive AGX

适用于自动驾驶汽车的硬件

NVIDIA DRIVE® 嵌入式超级计算平台处理来自摄像头、普通雷达和激光雷达传感器的数据，以感知周围环境、在地图上确定汽车的位置，然后规划并执行安全的行车路线。这款 AI 平台外形紧凑、节能高效，支持**自动驾驶**、**座舱功能**和**驾驶员监控**，以及其他安全功能。

NVIDIA DRIVE Hyperion

用于量产自动驾驶汽车的平台

此自动驾驶汽车参考架构通过将基于 DRIVE Orin™ 的 AI 计算与完整传感器套件（包含 12 个外部摄像头、3 个内部摄像头、9 个雷达、12 个超声波、1 个前置激光雷达和 1 个用于真值数据收集的激光雷达）相集成，能够加速开发、测试和验证。DRIVE Hyperion 具有适用于自动驾驶 (**DRIVE AV**) 的完整软件栈，以及驾驶员监控和可视化 (**DRIVE IX**)，能够无线更新，在车辆的整个生命周期中添加新的特性和功能。它还可以跨代兼容，因此合作伙伴可以利用当前使用的 DRIVE Orin 平台，无缝迁移到 NVIDIA DRIVE Atlan™ 及后续平台。

NVIDIA DRIVE Orin

NVIDIA DRIVE Orin™ SoC（系统级芯片）可提供每秒 254 TOPS（万亿次运算），是智能车辆的中央计算机。它是理想的解决方案，为自动驾驶功能、置信视图、数字集群以及 AI 驾驶舱提供动力支持。借助可扩展的 DRIVE Orin 产品系列，开发者只需在整个车队中构建、扩展和利用一次开发投资，便可从 L2+ 级系统一路升级至 L5 级全自动驾驶汽车系统。

NVIDIA DRIVE AGX Pegasus

NVIDIA DRIVE AGX Pegasus 利用两块 NVIDIA Xavier™ SoC 和两块 Turing™ GPU 的强大功能，实现了 320 TOPS 的超级计算能力。该平台专为各种类型的自主系统（包括机器人出租车）而设计和打造。

NVIDIA DRIVE AGX Xavier

NVIDIA DRIVE AGX Xavier 可为 L2+ 级和 L3 级自动驾驶提供每秒 30 TOPS 的运算。其核心是 NVIDIA 首次生产的车规级 Xavier 系统级芯片，该芯片采用了六种不同类型的处理器，包括 CPU、GPU、深度学习加速器 (DLA)、可编程视觉加速器 (PVA)、图像信号处理器 (ISP) 和立体/光流加速器。

硬件-嵌入式系统

Clara AGX

使用 NVIDIA Clara™ 开发者套件构建新一代智能医疗设备。Clara 开发者套件能够高速处理输入/输出 (IO)、加速图像重建、进行可扩展的 AI 推理和 3D 可视化，为开发 AI 驱动型设备提供完整的医疗健康专用系统。

用于 AI 驱动型医疗设备的强大计算功能

NVIDIA Clara AGX 开发者套件

NVIDIA Clara AGX™ 开发者套件为医疗设备提供实时流式连接和 AI 推理。结合内置 NVIDIA® Jetson AGX Xavier™ 的 ARM® 片上系统 (SoC) 的灵活性、NVIDIA RTX™ 6000 GPU 的强大性能，以及 NVIDIA ConnectX®-6 智能网卡的 100GbE 连接，Clara AGX 为开发软件定义、支持 AI、实时、即时可用的医疗设备提供了一个易于使用的平台。

NVIDIA Clara Holoscan 开发者套件

NVIDIA Clara Holoscan 开发者套件配备高性能 NVIDIA AGX Orin™ 模组、强大的 RTX A6000 GPU，并具有 ConnectX-7 智能网卡的稳健连接性能。以上组合可提供的性能是上一代的三倍，这使NVIDIA Holoscan 开发者套件成为开发新一代软件定义医疗设备的理想解决方案。

利用 NVIDIA Clara Holoscan，构建新的 AI 功能，提高研发团队的生产力。Clara Holoscan 是一个医疗设备混合计算平台，集硬件系统、优化的库、SDK 以及核心微服务于一体，这些都是在任意位置（从嵌入到边缘到云）开发和运行端到端流式和成像应用所需的要件。

软件-应用框架

元宇宙应用- Omniverse

Omniverse 是一个实时协作引擎和仿真模拟平台，专为虚拟协作和物理级准确的实时模拟打造；该平台集成了英伟达过去二十多年在AI、HPC和图形各方面的技术、算法、标准，是英伟达创建元宇宙虚拟平行宇宙的技术平台底座。Omniverse 分为两种版本，一种是为个人客户打造的，目前是免费的；另外是企业版本的，目前是收费的。我们从两种不同的版本入手，探索Omniverse平台的功能：

FEATURES	个人版本 （最适合创作者和开发者）	企业版本 （最适合团队）
连接到行业领先的 3D 设计工具	√	√
无限制多应用协作	√	√
最多 2 个用户的多用户协作	√	√
2个用户以上的多用户协作		√
可扩展的实时RTX渲染	√	√
使用 PhysX 5.0、Blast、Flow 进行物理级准确模拟	√	√
检查、修改或使用 300 多个预构建的扩展来开发和构建增强的解决方案	√	√
针对英伟达认证系统进行测试和优化		√
将 Nucleus 协作扩展到数据中心或私有云		√
企业安全管理（SSO、SSL）		√
企业部署工具		√
英伟达企业支持		√

软件-应用框架

元宇宙应用- Omniverse

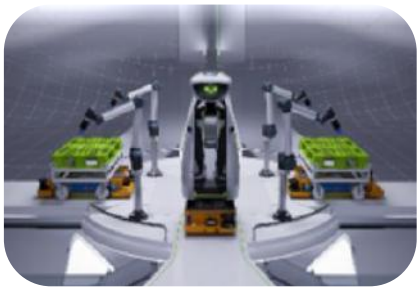
Omniverse平台的五大关键要素

Nucleus

平台的数据库和协作引擎

Connect

扩展和附加软件层



Kit

构建本地Omniverse应用
程序和微服务的工具包

Simulation

Omniverse kit的仿真插件或微服务

RTX renderer

渲染器
(软件从模型生成图像的过程)

软件-应用框架

元宇宙应用- Omniverse

本节从平台的整体架构介绍一下Omniverse平台：

应用层

Audio2Face
借助 NVIDIA 的深度学习AI技术，仅从音频源即可快速轻松地生成富有表现力的面部动画

Code
供开发人员和高级用户轻松构建自己的Omniverse 扩展、应用程序或微服务

Create
利用强大的 USD 格式来加速高级场景合成以及实时组装、照明、模拟和渲染 3D 场景

Create XR
使用沉浸式 VR 和 AR 浏览用户的Omniverse 场景并与之交互

DRIVE Sim
在虚拟世界中加速自动驾驶汽车的开发

Isaac Sim
一种可扩展的机器人仿真应用程序和合成数据生成工具

Kaolin
一套用于加速 3D 深度学习研究的工具

Machinima
使用强大的工具和逼真的渲染重新混合、重新创建和重新定义动画视频游戏的故事情节

RTX Remix
使用 RTX Remix 改装平台创建令人惊叹的经典游戏 RTX 重制版

View
强大的可视化APP

云层

Omniverse Cloud
使个人和团队能够从任何地方设计、发布和体验丰富的交互式 3D 互联网应用程序

核心技术层

MDL
可以自由地在支持的应用程序之间共享基于物理的材质和灯光

PhysX
可扩展的多平台物理模拟解决方案，支持从智能手机到高端多核 CPU 和 GPU 的各种设备

RTX
开创了以前所未有的速度模拟物理世界的新一代应用程序

USD
一个开放且可扩展的生态系统，用于在 3D 世界中进行描述、组合、模拟和协作，最初由皮克斯动画工作室发明

硬件层

GeForce Graphics Cards
GeForce RTX 30 Series GPU

Nvidia Certified Systems
用于合作伙伴的加速计算

Nvidia Studio Products
随时随地进行创作的最佳笔记本电脑和 PC 台式机

Nvidia Ovx
专为在数据中心规模创建和操作Omniverse 应用程序而构建

RTX Graphics Cards
NVIDIA RTX Series GPU

软件-应用框架

汽车- Drive

专为自动驾驶开发所设计的NVIDIA DRIVE平台覆盖了从汽车到数据中心的全部要素。**DRIVE Hyperion**是一种车载解决方案，一种汽车架构，包含了传感器、用于计算的DRIVE AGX，以及强大的自动驾驶和智能座舱功能所必须的软件。在数据中心，英伟达提供用户进行自动驾驶开发所需的硬件和软件，包括用于训练DNN进行感知的**NVIDIA DGX™**，以及用于生产数据集和验证整个自动驾驶全堆栈的**DRIVE Sim**。

Nvidia DRIVE 平台主要包括：

自动驾驶开发平台

DRIVE Hyperion
开发level-2和level3高速公路自动驾驶解决方案的开发平台和参考架构

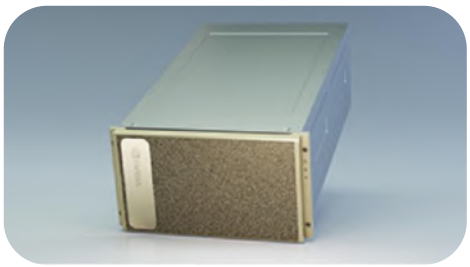
模块化软件栈

DRIVE SDK
包括驾驶操作系统、软件开发工具包、以及高级应用程序

模拟平台

DRIVE Sim
利用英伟达的核心技术，包括NVIDIA RTX™，Omniverse™以及AI来提供一个强大的云基础的计算平台，能够产生大范围的用于自动驾驶开发和验证的现实世界场景。

深度神经网络训练平台



DRIVE DGX
DGX 以及深度学习软件开发工具包

软件-应用框架

云端AI视频流- Maxine

NVIDIA Maxine为实时音频和视频沟通做好了充分的准备。无论是一个视频会议，呼叫客服中心，或者是直播，Maxine 都能实现清晰的通信以增强虚拟交互。Maxine的优势有以下几点：

最先进的NVIDIA AI功能

为开发人员提供世界级的预训练模型部署优质的音频和视频质量功能

实时AI性能

包含了加速以及优化人工智能功能，可以在GPU上进行实时推理，产生低延迟的音频、视频和增强网络弹性的增强现实效果。

完整的AI管道

提供视频解码、转码、编码、对话式人工智能，计算机视觉，视频流，以及分析来完成AI管道。

多云、可定制的部署

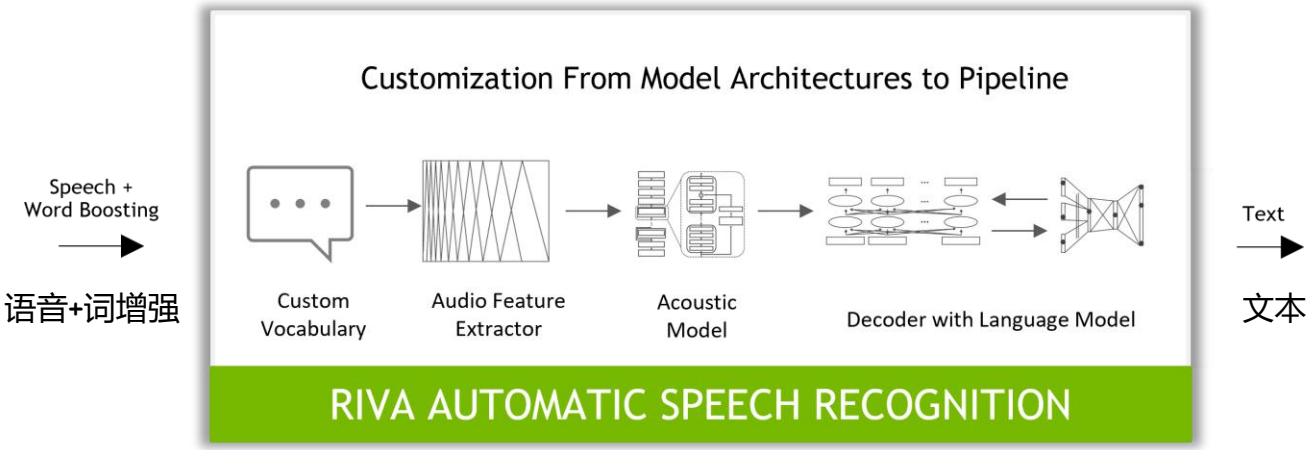
Maxine 的云原生微服务可以灵活、快速的部署和更新

软件-应用框架

语音AI- Riva

NVIDIA Riva 是一款用 GPU 加速的 SDK，可针对用户的使用案例构建定制的实时语音 AI 应用。由于基于语音的应用在全球广泛使用，解决方案需要跨多种语言与人类进行交互。语音 AI 应用需要识别行业特定术语，并作出自然的实时响应。Riva 包含先进的**自动语音识别 (ASR)** 和**文字转语音 (TTS)** 功能，且实时运行。

自动语音识别



自定义词汇　音频特征提取器　声学模型　　用语言模型进行解码

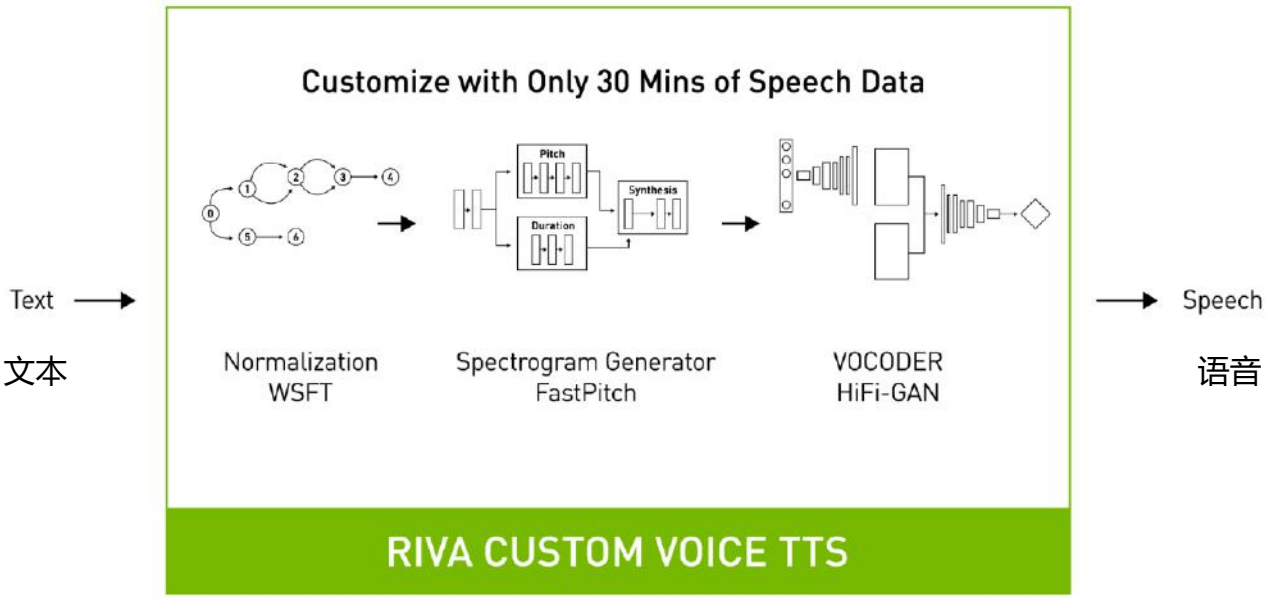
- Riva 提供开箱即用的卓越自动语音识别 (ASR)，可针对任何领域或部署平台进行定制。
- 该服务可处理成百上千个音频流输入，并以尽可能低的延迟返回流式转录文稿。
- Riva 制作流程基于各种特定于领域的数据进行训练，并且可以针对不同的语言、口音、区域、词汇和上下文进行进一步调整。
- 端到端流程经过 GPU 优化，包含可定制的特征提取、解码、标点符号、声音和语言模型。

软件-应用框架

语音AI- Riva

NVIDIA Riva 是一款用 GPU 加速的 SDK，可针对用户的使用案例构建定制的实时语音 AI 应用。由于基于语音的应用在全球广泛使用，解决方案需要跨多种语言与人类进行交互。语音 AI 应用需要识别行业特定术语，并作出自然的实时响应。Riva 包含先进的**自动语音识别 (ASR)** 和**文字转语音 (TTS)** 功能，且实时运行。

文字转语音



标准化WSFT　　频谱图生成器　　声码器

- Riva 提供模仿人类的文字转语音 (TTS) 神经声音，这些声音使用先进的频谱图生成和声码器模型。Riva 制作流程经过自定义和优化，可在 GPU 上高效实时运行。
- Riva TTS 将原始文本作为输入内容，在流式传输模式下或在批量模式下的整个序列末尾生成后，即可返回音频区块。
- Riva 定制语音功能使任何企业只需提供 30 分钟的数据，即可为其品牌、虚拟助理或呼叫中心创建独特的语音。
- 使用 Riva 创建新语音需要在 A100 GPU 上进行不到一天的训练，而使用替代技术则需要数周时间。

软件-应用框架

数据分析-RAPIDS

RAPIDS是一款面向Machine Learning（机器学习）、大数据处理市场的开源产品；是一个加速平台，利用 GPU 的强大功能轻松加速数据科学、机器学习和 AI 工作流程；并行开展数据加载、数据处理和机器学习，将端到端数据科学流程的速度提高 50 倍。RAPIDS 以 NVIDIA® CUDA-X AI™ 为基础，融合了显卡、机器学习、深度学习、高性能计算 (HPC) 等领域多年来的发展成果。

构建高性能生态系统

RAPIDS 由一系列开源软件库和 API 组成，用于完全在 GPU 上执行数据科学流程，从而可将训练时间从几天缩短到几分钟；**RAPIDS 以 NVIDIA® CUDA-X AI™ 为基础**，融合了显卡、机器学习、深度学习、高性能计算 (HPC) 等领域多年来的发展成果。

执行速度更快

RAPIDS 会在底层利用 NVIDIA CUDA®，通过在 GPU 上运行整个数据科学训练流程，帮助加速工作流程。这可以将模型训练时间从几天缩短到几分钟。

使用相同工具

通过隐藏 GPU 的工作复杂性，甚至隐藏数据中心架构内的后台通信协议，RAPIDS 提供了完成数据科学的简单方法。随着越来越多的数据科学家使用 Python 等高级语言，必须要在实现加速的同时避免代码变更，才能迅速缩短开发时间。

在任何位置大规模运行

RAPIDS 的运行位置不受限制，在云端或本地均可。您可轻松将其从工作站扩展到多 GPU 服务器，再到节点集群，并可在生产环境中与 Dask、Spark、MLFlow 和 Kubernetes 搭配部署。

针对大数据的超速性能

针对小型及大规模的大数据分析问题，GPU 可以节省大量成本和时间。RAPIDS 使用 10TB 大小的常见 API（如 Pandas 和 Dask），相较于最高的 CPU 基准，其在 GPU 上的运行速度要快 20 倍。NVIDIA 解决方案仅使用 16 台 NVIDIA DGX A100 即可达到 350 台基于 CPU 的服务器的性能，而且在提供 HPC 级性能的同时，其成本效益提高了 7 倍以上。

软件-应用框架

数据分析-RAPIDS

数据处理演进逻辑：

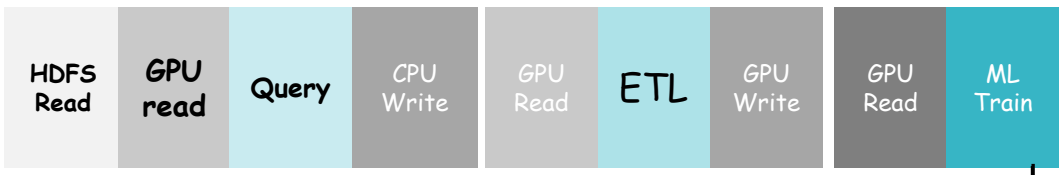
Hadoop Processing，Reading from Disk

25-100提升
少量代码
语言灵活
优先内存

Spark In-Memory Processing

5-10x提升
较多代码
语言僵化
主要基于GPU

Traditional GPU Processing



50-100x提升
相同量代码
语言灵活
主要基于GPU

RAPADIS

常见的数据处理任务有多个步骤，而 Hadoop 无法高效处理这些步骤。Apache Spark 通过在系统内存中保存所有数据解决了这个问题，这让数据流程变得更加灵活复杂，但也引入了新的瓶颈。在拥有数百个 CPU 节点的 Spark 集群上，即使是分析几百 GB 的数据也可能要花费数小时，甚至数天时间。为发挥数据科学的真正潜力，GPU 必须位于数据中心设计的中心，它包含以下五个要素：计算、网络、存储、部署和软件。**一般来说，相较于 CPU，GPU 上的端到端数据科学工作流程要快 10 倍。**

注：HDFS是一种分布式文件系统，可以存储非常大的文件，但不适用于低延时、多方读写的数据访问；read一般表示读取，Query表示的是请求查询，write一般表示存储；ETL是一种数据技术，表示的是数据的抽取、清洗和加载；ML train表示的是机器学习训练

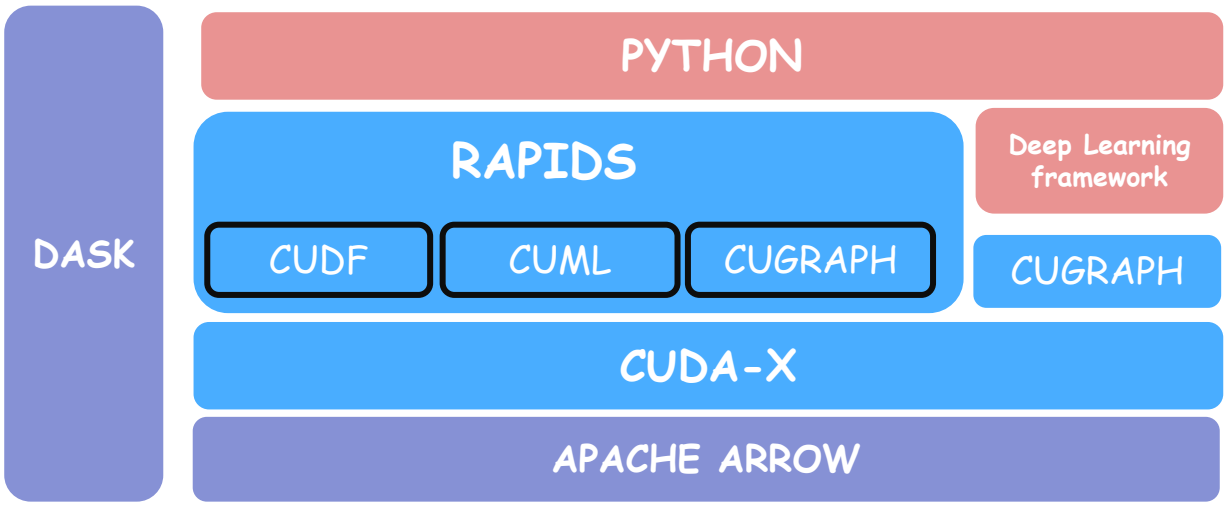
软件-应用框架

数据分析-RAPIDS

核心技术：

RAPIDS 依靠 CUDA 基元进行低级别计算优化，但通过用户友好型 Python 接口实现了 GPU 并行化和高显存带宽。RAPIDS 支持从数据加载和预处理到机器学习、图形分析和可视化的端到端数据科学工作流程。它是功能完备的 Python 堆栈，可扩展到企业大数据用例。

Machine learning to Deep learning : ALL on GPU



数据加载和预处理

RAPIDS 的数据加载、预处理和 ETL 功能基于 Apache Arrow 构建，用于加载、连接、聚合、过滤及以其他方式处理数据

机器学习

RAPIDS 的机器学习算法和数学基元遵循熟悉的类似于 scikit-learn 的 API。单块 GPU 和大型数据中心部署均支持 XGBoost、随机森林等主流工具。

核心技术

图形分析

RAPIDS 的图形算法（如 PageRank）和功能（如 NetworkX）高效利用了 GPU 的大规模并行计算能力，可将较大图形的分析速度提高 1000 倍以上。

可视化

RAPIDS 的可视化功能支持 GPU 加速的交叉过滤。受原始版本的 JavaScript 启发，它可以对超过 1 亿行表格数据集进行超快速的交互式多维过滤。

软件-应用框架

NVIDIA CLARA

加速计算和 AI 正在强效助力新一代医疗设备和生物医学研究。NVIDIA Clara™ 提供单一平台，用于医学影像、基因组学、患者监控和药物研发，并可部署在嵌入式系统、边缘、每个云等任何地方，助力医疗健康行业进行创新并加快实现精准医疗的目标。

NVIDIA Clara™ 提供四大产品：

NVIDIA Clara™

NVIDIA CLARA DISCOVERY

NVIDIA Clara Guardian

NVIDIA Clara Holoscan

NVIDIA Clara Parabricks

NVIDIA Clara Discovery 集 GPU 加速及优化的框架、工具、应用和预训练模型于一体，用于计算药物研发。Clara Discovery 专为支持跨学科工作流而构建，可帮助科学家和研究人员更快地将药物投放市场，并为疾病机制研究提供新的可能性。

GPU 助力的深度学习算法和 Transformer 模型将加速药物研发的每个阶段：

训练能够理解化学空间的大型语言模型 (LLM)

分子动力学模拟

蛋白质结构预测

生成药物设计

软件-应用框架

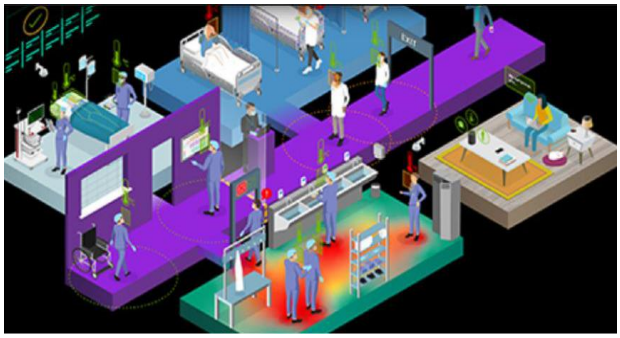
NVIDIA CLARA

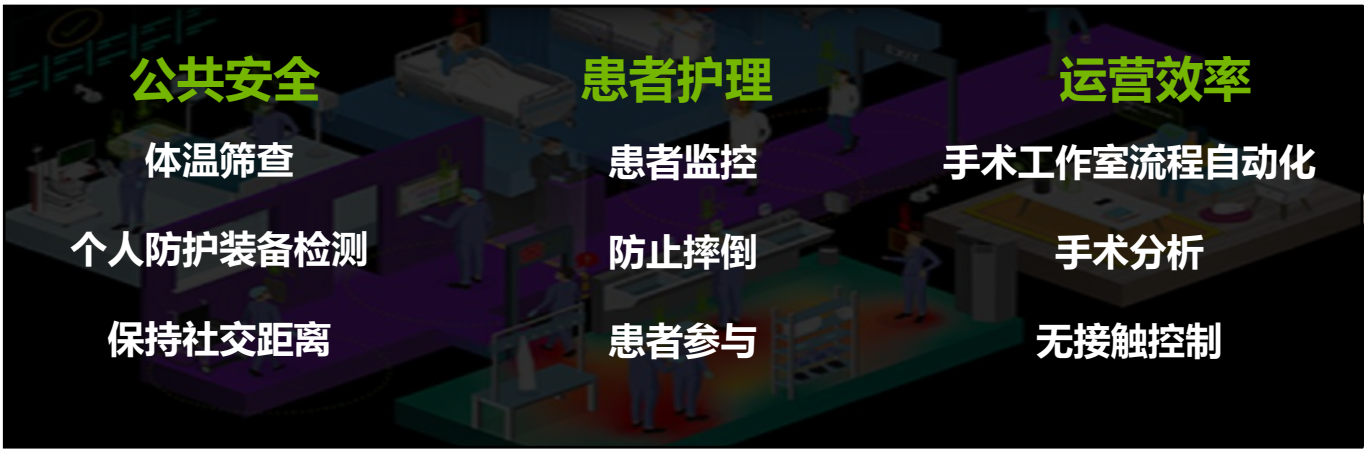
NVIDIA Clara™ 提供四大产品：

NVIDIA Clara™

NVIDIA CLARA DISCOVERY | NVIDIA Clara Guardian | NVIDIA Clara Holoscan | NVIDIA Clara Parabricks

当下，对智能医院的需求非常迫切。智能传感器可以充当“眼睛和耳朵”，在关键方面（从体温筛查和防护装备检测到保持安全社交距离）确保安全和出色运营。 NVIDIA Clara Guardian 既是一种应用程序框架，也是一个合作伙伴生态系统，它将智能传感器和多模态 AI 相结合，以在医疗健康机构中改善对患者的治疗





软件-应用框架

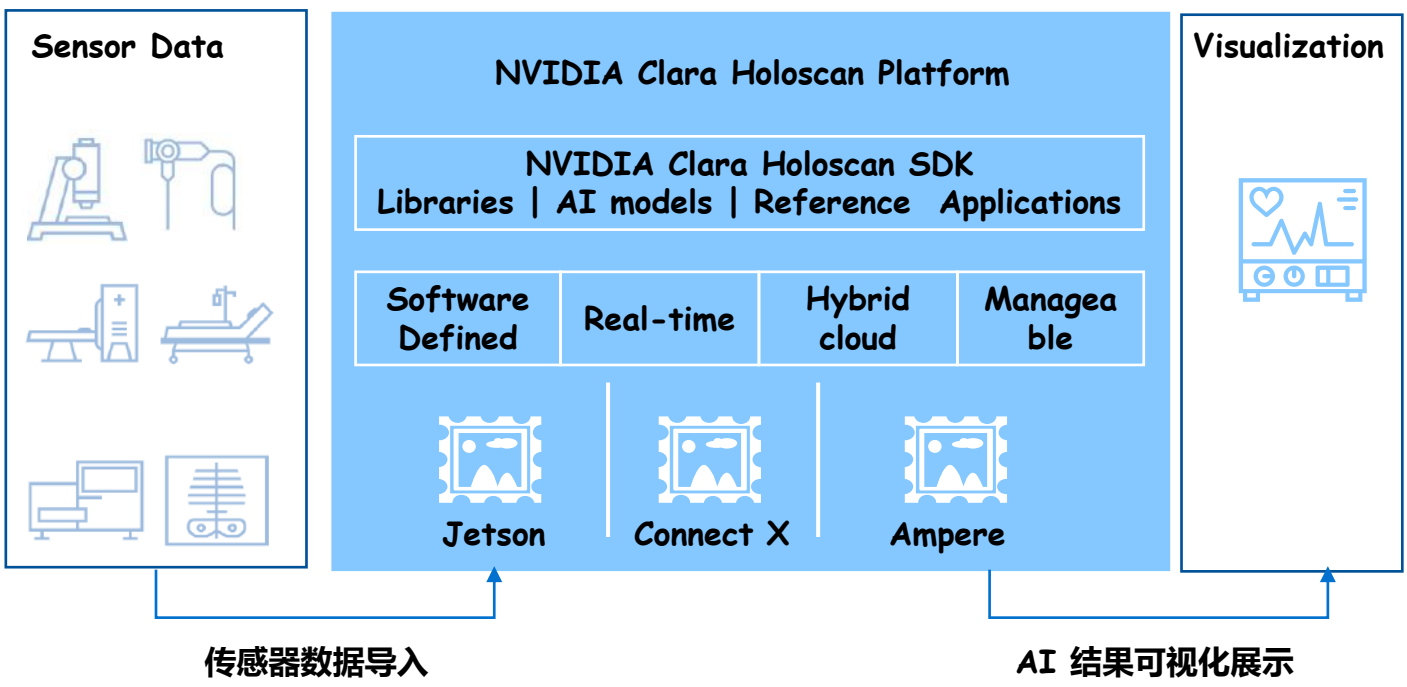
NVIDIA CLARA

NVIDIA Clara™ 提供四大产品：

NVIDIA Clara™

NVIDIA CLARA DISCOVERY | NVIDIA Clara Guardian | NVIDIA Clara Holoscan | NVIDIA Clara Parabricks

NVIDIA Clara Holoscan 是医疗设备的人工智能计算平台，它结合了用于低延迟传感器和网络连接的硬件系统、用于数据处理和人工智能的优化库，以及运行流媒体、成像和其他应用程序的核心微服务，从嵌入式到边缘到云。Clara Holoscan 使医疗设备开发人员能够创建下一代支持 AI 的医疗设备并将其推向市场。



软件-应用框架

NVIDIA CLARA

NVIDIA Clara™ 提供四大产品：

NVIDIA Clara™

NVIDIA CLARA DISCOVERY | NVIDIA Clara Guardian | NVIDIA Clara Holoscan | NVIDIA Clara Parabricks

NVIDIA Clara Parabricks 是一款 GPU 加速的计算基因组学工具包，可为测序中心、临床团队、基因组学研究人员以及新一代测序仪器开发者提供快速准确的分析，进而助力更快速、更准确的基因组学分析。

使用特点：

使用一流的工具加速应用

使用加速工具进行黄金标准的种系、体细胞和 RNA 的快速分析。

获得高达 80 倍的性能提升

与仅使用CPU的解决方案相比，速度提升高达 80 倍，计算成本降低高达 50%。

提高测序分析的准确性

借助 Clara Parabricks 和 GPU，将深度学习的强大功能应用到基因组分析。

利用灵活的工作流

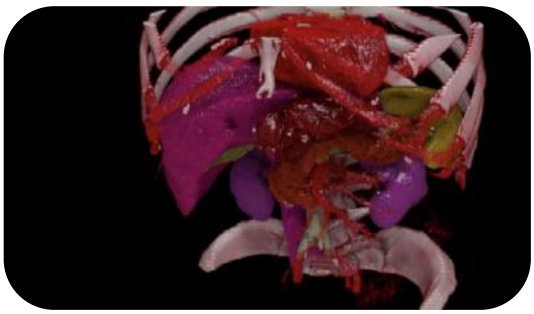
通过使用工作流描述语言 (WDL) 和 NextFlow 配置工具来创建自定义工作流。

软件-应用框架

NVIDIA CLARA

加速计算和 AI 正在强效助力新一代医疗设备和生物医学研究。NVIDIA Clara™ 提供单一平台，用于医学影像、基因组学、患者监控和药物研发，并可部署在嵌入式系统、边缘、每个云等任何地方，助力医疗健康行业进行创新并加快实现精准医疗的目标。

医疗影像和医疗设备

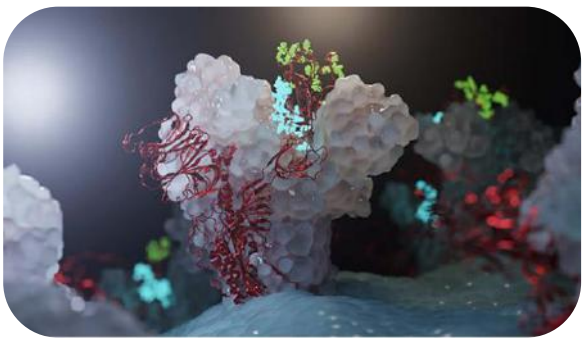


Clara Holoscan 是一个特定于领域的 AI 计算平台，可提供加速的全栈基础设施，以满足在临床边缘实时对串流数据进行软件定义的可扩展处理的需求。借助 NVIDIA Clara Holoscan，医疗设备开发者可以加速新一代 AI 设备的开发，从而构建能够将 AI 应用直接用于手术室的设备。

基因组学

Clara Parabricks® 可为开发者提供企业级的一站式 GPU 加速测序软件和技术栈，以构建用于基因组学中的高性能计算、深度学习和数据分析的应用。

生物制药



Clara Discovery 包含多种框架、应用和 AI 模型，可通过这三者的能力共同加速药物研发，为基因组学、显微镜学、虚拟筛选、计算化学、可视化、临床成像等领域的研究提供支持。

图片来源：英伟达官网

智慧医院

Clara Guardian 是一种应用框架，可为医院提供视频分析和对话式 AI 功能，简化智能传感器的开发和部署，从而优化临床体验。

Clara Imaging 利用开源框架、AI 辅助数据标记、AI 推理和预训练模型加速医疗影像 AI 工作流。

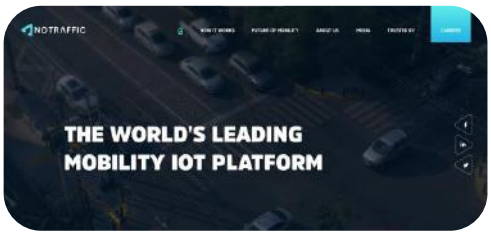
软件-应用框架

智能视频分析- Metropolis

NVIDIA® Metropolis运行在 NVIDIA 的 EGX 平台上，是一个应用框架、一组开发者工具，同时也是一个合作伙伴生态系统，可将可视化数据和 AI 整合起来，提高众多行业的运营效率和安全性。它将帮助处理和理解数万亿传感器生成的海量数据，这些数据来自于无人零售、简化库存管理、智能城市的交通工程、工厂车间的光学检查、医疗健康机构的患者护理等不同领域。通过这个前沿技术，以及广泛的 Metropolis 开发者生态系统，企业可以创建、部署和扩展从边缘到云端的 AI 和物联网 (IoT) 应用。

大约有 10 亿个摄像头 -物联网 (IoT) 终端传感器，已经部署在全球的城市和空间中，因此基于 AI 的视频分析优化对于无人零售、简化库存管理、智慧城市的交通工程、工厂车间的光学检测、医疗设施的患者护理等至关重要。下面介绍一下Metropolis不同行业的**合作伙伴**。

交通管理



NOTRAFFIC

Assaia

医院运营

ARTISIGHT

石油钻井等安全性要求较高的场所

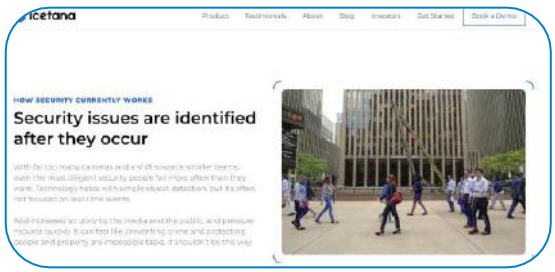
Herlin

注：以上均为视频，大家可以点击查看

软件-应用框架

智能视频分析- Metropolis

校园管理

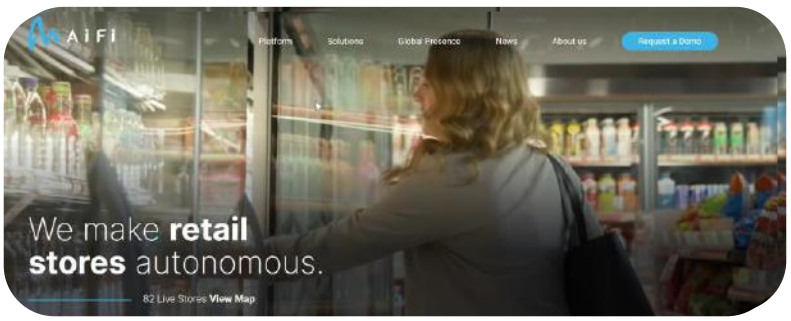


ICETANA

机场安全

IPSOTEK

零售运营



AIFI

质量控制

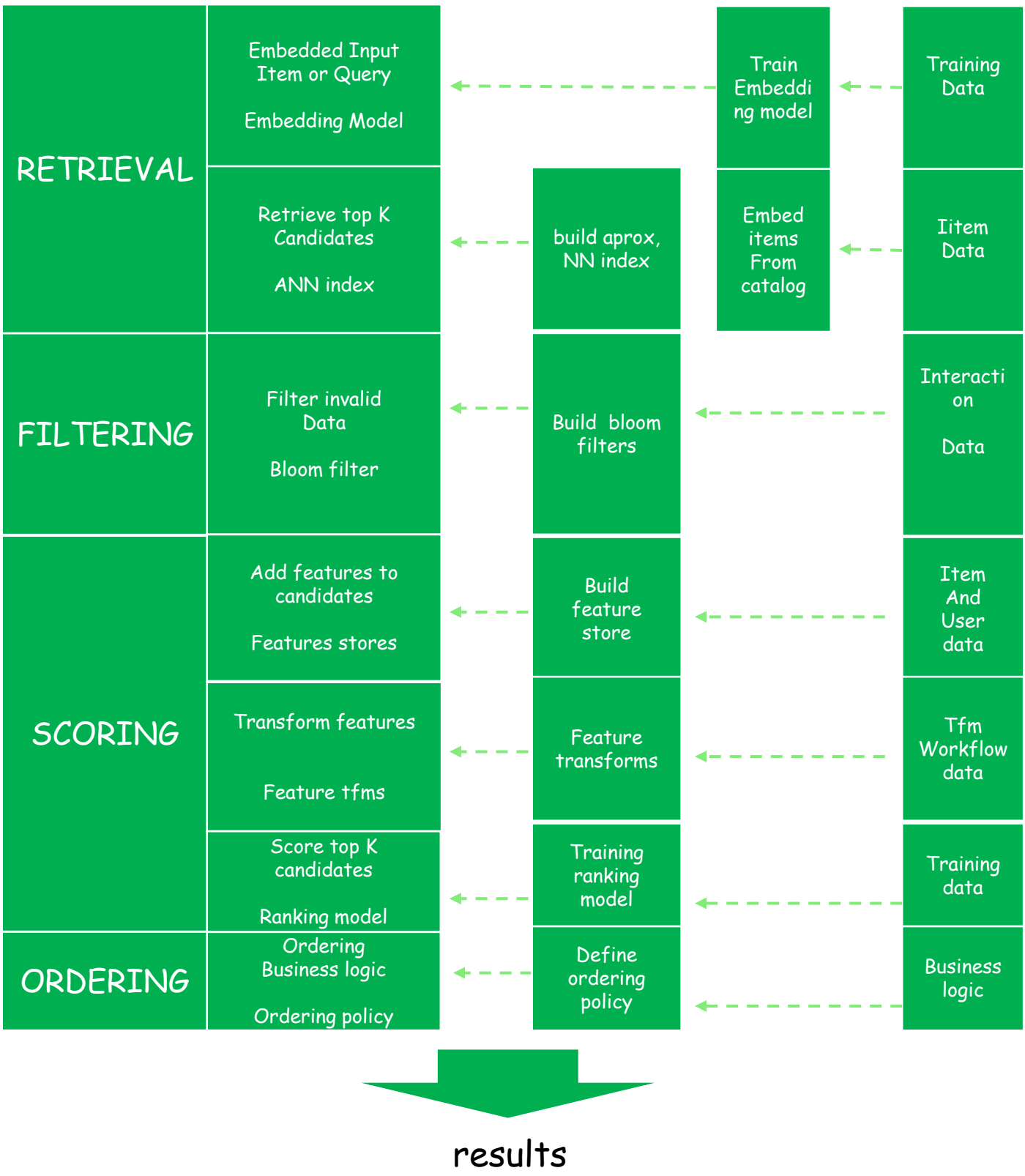
DATA MONSTERS

废料处理

RECYCLEYE

软件-应用框架

推荐系统- Merlin



软件-应用框架

推荐系统- Merlin

英伟达 Merlin 支持数据科学家，机器学习工程师，以及研究人员来大规模地建立高性能推荐系统。Melin包括库、方法、工具，通过解决通用的预处理程序、特征工程、训练、推理和部署到生产挑战上来精简推荐系统的构建。Merlin的组件和功能被优化来支持数百TB数据的检索、过滤、评分和排序，都可以通过易于使用的API进行访问。有了Merlin，更好的预测、更多的点击率，以及更快的生产部署触手可及。

Interoperable End-to-End Solution (可交互的端到端解决方案)

Merlin Models

Merlin模型是一个库，可以为推荐系统和在GPU和CPU上进行机器学习和高级的深度学习模型的高质量实施提供标准化的模型。训练用于检索和排序的模型在10行代码之内

Merlin NVTabular

Merlin NVTabular是一个特征工程和预处理库，旨在高效地操作TB级的推荐系统数据集以及显著缩减数据准备时间

Merlin HugeCTR

Merlin HugeCTR 是一个深度神经网络框架，专为GPU上的推荐系统而设计。它通过分层内存提供模型并行训练和推理以实现性能最大化和可扩展性。

Merlin Transformers4Rec

MerlinTransformers4Rec是一个库，可以精简基于会话式推荐管道的构建。该库使构建推荐器时更容易探索和应用流行的转换器架构。

Merlin Distributed Training

Merlin支持跨多个GPU的分布式训练。组件包括MerlinSOK以及Merlin 分布式嵌入。TensorFlow (TF) 用户有权使用 SOK (TF 1.x) 和 DE (TF 2.x) 来利用模型并行性来扩展训练。

Merlin Systems

Merlin Systems 是一个简化新模型和工作流部署到生产环境的库。它可以使深度学习工程师和操作人员使用50行代码部署端到端的推荐系统。

软件-应用框架

推荐系统- Merlin

Built on NVIDIA AI
(在英伟达AI平台上搭建推荐系统)

英伟达AI 使数百万的从业者和公司可以用英伟达AI平台来加速他们的工作流。英伟达Merlin，是英伟达AI平台的一部分。英伟达Merlin建立在平台内并且使用了额外的AI软件。

RAPIDS	cuDF	NVIDIA Triton Inference Server
RAPIDS是一组开源软件库以及API，可完全在GPU上实现端到端的数据科学和分析管道	cuDF是一个python GPU DataFrame库，可以用于加载、连接、聚合以及操作数据	利用 NVIDIA Triton™ 推理服务器通过延迟和 GPU 利用率的正确组合以最大限度地提高吞吐量，在 GPU 上高效地运行推理

软件-应用框架

机器人- Issac

NVIDIA Isaac是NVIDIA于2018年推出的一个针对机器人打造的自主机器平台，平台包括软件、硬件、一个虚拟世界的机器人模拟器。英伟达提供Isaac平台软件开发工具以用于部署商业级、AI驱动的机器人

ISAAC Sim (机器人仿真平台)

Domain Randomization	Scenario management	Sensor Models	Robot Models

ISAAC APPs

CARTER (indoor robot)	Kaya (Getting started)	Leonardo (manipulation)	Custom Aplications

ISAAC GEMs(机器人的算法功能模块)

Muti-class segmentation DNN	3D object pose estimation (with depth))DNN	Object detection DNN	Stereo depth DNN
Stereo visuallnertial odometry	Superpixels	Global planner with cost maps	GPU accelerated AprilTags
Muti Lidar localization &collsion avoidance	Motion planning for navigation &Manipu;ation	Reinforcement learning	Deepstreaming Robotics
2D skeleton pose Estimation DNN	Audio DNNs	Sensor and robot integrations	……and more

ISAAC Engine（记录数据、可视化、监控模块之间通信情况）

ComputionalGraph &CUDA Messging	Visualization Tools	Python API &ROS Bridge

NVIDIA HW Acceleration(硬件加速)

TensorRT	cuDNN	CUDA-X	RTX	PhysX

NVIDIA AGX

NVIDIA DGX

软件-应用和工具

NVIDIA NGC

面向 AI、机器学习和高性能计算的服务、软件和支持的门户；NVIDIA NGC™ 提供一系列完全托管的云服务，包括用于 NLU 和语音 AI 解决方案的 NeMo LLM、BioNemo 和 Riva Studio。AI 从业者可以利用 NVIDIA Base Command 进行模型训练，利用 NVIDIA Fleet Command进行模型管理，并利用NGC专用注册表安全共享专有 AI 软件。此外，**NGC还拥有一个 GPU 优化的 AI 软件、SDK 和 Jupyter Notebook 的目录**，可帮助加速AI工作流，并通过 NVIDIA AI Enterprise 提供支持。

探索NGC目录

NGC目录提供了什么?

NGC 目录为 AI、机器学习和 HPC 提供完整的 GPU 优化的容器集合，这些容器经过测试，可随时在本地、云端或边缘依托受支持的 NVIDIA GPU 运行。此外，NGC 目录还提供预训练模型、模型脚本和行业解决方案，可轻松集成到现有工作流中。

NGC 目录可解决哪些挑战?

编译和部署深度学习框架既耗时又容易出错。优化 AI 软件需要专业知识。构建模型需要专业知识、时间和计算资源。NGC 目录可提供 GPU 优化软件和工具，帮助解决这些挑战。借助这些软件和工具，数据科学家、开发者、IT 和用户可以专注于构建自己的解决方案。

NGC 目录中包含哪些内容?

每个容器都有一组预先集成的 GPU 加速软件。该堆栈包括选定的应用或框架、NVIDIA CUDA® 工具套件、加速库和其他必要的驱动，所有这一切均经过测试和调整，无需进一步执行任何设置即可立即协同运行。

NGC 目录提供了哪些AI软件

TensorFlow、PyTorch、MxNet、NVIDIA TensorRT、RAPIDS™ 等众多热门 AI 软件。

软件-应用和工具

NVIDIA NGC

适用于所有用例的平台

语言建模	推荐系统	图像分割	翻译
语言建模是一项自然语言处理 (NLP) 任务，可确定句子中的词汇在假设顺序下出现的概率	推荐系统是一种信息过滤系统，旨在预测用户对物品的“评分”或“偏好”	图像分割属于图像处理领域，可将图像划分为体现独特对象或子部分的多个小组或区域	机器翻译是将文本从一种语言翻译成另一种语言的任务
物体检测	**ASR**	**文本转语音**	**HPC**
物体检测不仅包括检测图像和视频中物体是否存在及其位置，还可根据其是否为日常物体进行分类	自动语音识别 (ASR) 系统包括向交互式虚拟助手发出语音命令、将音频转换为在线视频中的字幕等	语音合成或文本转语音是人为地根据原始转录内容生成人类语音的任务。当移动设备将网页上的文本转换为语音时，会使用文本转语音模型	高性能计算 (HPC) 是推动计算科学发展的关键工具之一，该科学计算领域已扩展到了各个方向

部分应用

DeepZen	Neurala	克莱姆森大学	亚利桑那大学
DeepZen是一家专注于研究逼真语音且让语音具备情感的 AI 公司	AI 初创公司Neurala 如何其 Brain Builder 平台的深度学习训练和推理速度提升 8 倍	克莱姆森大学的 HPC 管理员支持 GPU 优化的容器，帮助科学家加速研究	亚利桑那大学使用 NGC 目录中的容器直接在无人机上创建 3D 激光点云来加速科学研究

软件-应用和工具

NVIDIA 工具

NVIDIA RTX Experience	为企业用户提供必要的生产力工具：高达 8K 的内置桌面录制功能、新版驱动更新提醒、游戏和应用程序优化以及桌面管理工具访问权限
NVIDIA RTX Desktop Manager	借助 NVIDIA® RTX 桌面管理软件，用户可以轻松管理单显示器或多显示器工作空间，获得更大的灵活性和控制力，从而有效管理屏幕有效使用区域和桌面。
GPU 加速创意应用	包含3D&渲染应用（例如Adobe Substance 3D Designer）、视频编辑应用、摄影应用（PS）、图形设计应用、直播应用（OBS）、机构可视化应用
视频会议 NVIDIA Broadcast 应用	NVIDIA Broadcast 应用可将任何房间变为家庭办公室。无论是远程办公、协作、创作还是学习，AI 增强型语音和视频都能显著提高电话会议的质量。这让您可以随时随地发挥最佳工作表现。
NVIDIA 数据科学工作台	借助 NVIDIA数据科学工作台，用户可以轻松简化数据科学开发环境，并更大限度地提高工作效率。一键直达，从 Workbench 中获取关键资料，例如 NVDIA NGC catalog中的容器、软件自动更新以及 Docker 化 GitHub 内容。最后，用户可以快速方便地访问大批量的数据科学工具。轻松复制内容，访问容器化解决方案，利用多个 CLI，获取新闻和活动等。

软件-应用和工具

游戏和创作

GeForce Experience

过游戏中浮窗功能 NVIDIA ShadowPlay™ 技术，在保障游戏性能的前提下，轻松截取游戏视频和截图

经过开发者协作调优，并在数千种硬件配置中进行广泛测试，Game Ready 驱动程序能够实现可靠稳定、性能出众的游戏体验

借助功能强大的 NVIDIA Ansel 照相模式，可轻松截取专业级游戏图像，然后直接分享到微博或 Shot With GeForce 网站。在支持此功能的游戏中，可选择截取超清、360、HDR 或立体的图像

NVIDIA Freestyle 游戏滤镜允许用户在玩游戏时将后处理滤镜应用于游戏中。可以直接通过游戏内浮窗来调整颜色或饱和度，从而改变游戏观感和氛围，或者应用效果显著的后处理滤镜（例如 HDR）

软件-应用和工具

游戏和创作

Omniverse Machinima

Omniverse™ Machinima 测试版是一种参考应用，可助力用户进行实时协作，对虚拟世界中的角色及其环境进行操作处理并实现动画化。对于希望从这些虚拟世界内部利用高保真渲染器的技术艺术家、内容创作者和行业专业人士，Omniverse Machinima 能够提供可轻松制作游戏动画的工具。

NVIDIA RTX Remix

借助 RTX Remix MOD 平台，轻松为经典游戏打造出惊艳的 RTX 重制版。Mod 平台基于 NVIDIA Omniverse构建，可让 MOD 爱好者轻松截取游戏素材，使用功能强大的 AI 工具自动增强材质，以及借助光线追踪和 DLSS 快速创建令人惊艳的 RTX 重制版游戏。

软件-基础架构

AI Enterprise 软件套件

NVIDIA AI Enterprise 是 NVIDIA AI 平台的操作系统，对于使用广泛的框架库构建的应用至关重要：用于语音 AI 的 NVIDIA® Riva、用于推荐系统的 NVIDIA Merlin™、用于医疗影像的 NVIDIA Clara™ 等。它经过认证，可随时随地（从数据中心到公有云）进行部署，并包含全球企业支持，可保证AI 项目如期进行。以下是其基础架构：

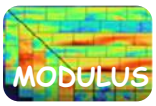
Application workflows

医学影像

语音人工智能

客户服务

推荐系统



物理机器学习

视频

视频分析

物流

对话式人工智能

机器人

自动驾驶

网络安全

NVIDIA AI Enterprise

AI and Data science development and deployment tools

加速数据准备	大规模训练AI模型		推理优化	大规模部署
NVIDIA RAPIDS	NVIDIA TAO Toolkit	PyTorch/ TensorFlow	NVIDIA TensorRT	NVIDIA Triton Inference Server

Cloud native management and orchestration

NVIDIA GPU operator

NVIDIA network operator

Certified partner integrations
MLOps I Container orchestration virtualization

Infrastructure optimization

NVIDIA vGPU

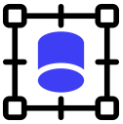
NVIDIA Magnum IO

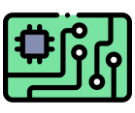
NVIDIA CUDA-X AI

Accelerated infrastructure


cloud


Data center


Edge


Embedded

解决方案-AI和数据科学

AI和数据科学概览

前面我们将英伟达的主要硬件和软件产品进行了梳理，那么接下来将进入其解决方案的部分，前面的软硬件包含了太多技术的内容，很多人可能并不理解，而解决方案，也就是英伟达能为客户解决怎样的问题，这样可能更容易理解他们到底在做什么事情，解决什么样的问题，创造什么样的价值。

AI和数据科学都包括哪些内容

数据分析	机器学习	深度学习训练	深度学习推理
对话式AI	预测与预报	语音AI	大型语言模型
实操实验			

解决方案-AI和数据科学

数据分析

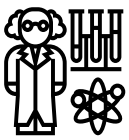
传统的数据分析工作流迟缓而繁琐，数据的准备、训练和部署都依赖于 CPU 计算。加速数据科学可显著提升端到端数据工作流的性能，在降低成本的同时加速价值产出。

行业面临的挑战

- 数据准备是一个复杂且耗时的过程，占据了数据科学家大部分的时间。
- 迭代会耗费大量时间，导致分析结果不够可靠。
- 对数据集进行下采样会导致结果欠佳。

尽管数据分析释放了巨大的潜力，但传统基于 CPU 的数据处理和分析却增加了业务运营的开销和复杂性，导致投资回报率降低。**加速数据科学开创了数据分析的新时代**，使得企业和从业者能够充分利用自身的数据和基础架构。

加速分析的优势



计算科学家

数据工程师

IT和DEVOPS专业人员

等待时间更少
减少等待流程结束的时间，从而让更多的时间用在迭代和测试解决方案上

处理速度更快
通过大规模的数据转换，更快地交付高质量的数据集，以支持从业者和整个企业的运营。

开销更低
借助 GPU 加速充分利用预算，无需增加购买、部署和管理更多 CPU 的成本。

结果更优
借助高性能处理分析数 TB 级的数据集，以提高结果的准确性和报告的速度。

互操作性更广
在众多热门分析库间轻松共享设备内存，从而避免昂贵且费时的数据复制操作。

决策更佳
利用拥有的所有数据做出更好的业务决策，提高企业绩效并更好地满足客户需求。

无需重构
不用学习新工具，仅需对代码进行少量修改，即可加速并扩展您现有的数据科学工具链。

无需重构
不用花费大量时间转换文件格式，只需要选择能够在企业中发挥出色效果的数据格式即可。

无缝扩展
借助统一且直观的架构，轻松从桌面扩展至多节点、多 GPU 集群。

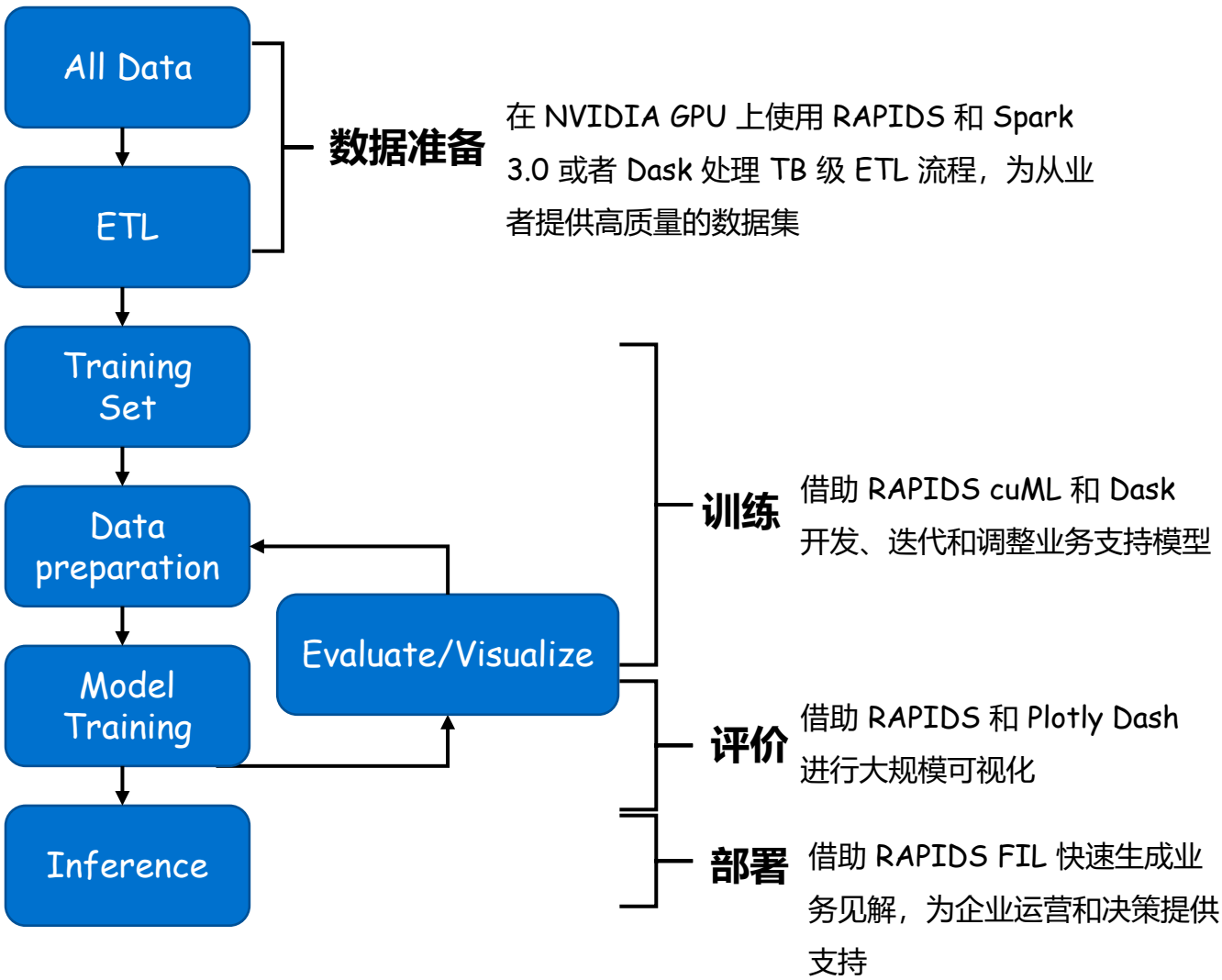
解决方案-AI和数据科学

数据分析

NVIDIA 端到端加速分析

NVIDIA 提供的端到端解决方案结合了针对高性能数据分析进行优化的硬件和软件，使企业能够轻松地从自身的数据中获得大量收益。借助 RAPIDS 和 NVIDIA CUDA，数据从业者可以加速 NVIDIA GPU 的分析流程，将数据分析操作所需时间（如数据加载、处理和训练等）从几天缩减至几分钟。借助熟悉的Python语言或基于Java的语言发掘CUDA的功能，从而简化加速分析流程。

从机器学习到深度学习，GPU包揽一切



注：ETL是指数据仓库技术，用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。

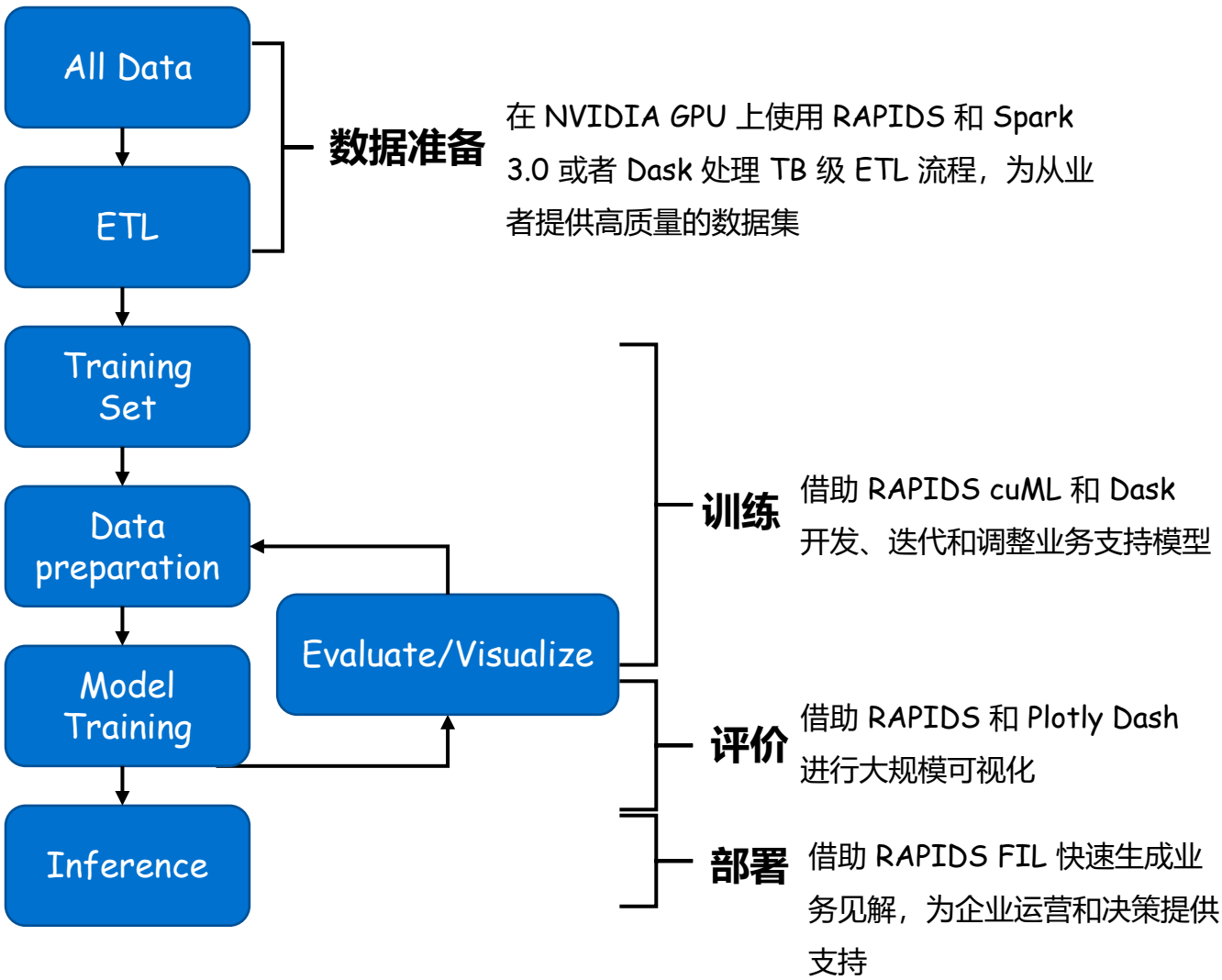
解决方案-AI和数据科学

机器学习

NVIDIA 端到端加速分析

NVIDIA 提供的端到端解决方案结合了针对高性能数据分析进行优化的硬件和软件，使企业能够轻松地从自身的数据中获得大量收益。借助 RAPIDS 和 NVIDIA CUDA，数据从业者可以加速 NVIDIA GPU 的分析流程，将数据分析操作所需时间（如数据加载、处理和训练等）从几天缩减至几分钟。借助熟悉的Python语言或基于Java的语言发掘CUDA的功能，从而简化加速分析流程。

从机器学习到深度学习，GPU包揽一切



注：ETL是指数据仓库技术，用来描述将数据从来源端经过抽取（extract）、转换（transform）、加载（load）至目的端的过程。

解决方案-AI和数据科学

机器学习

机器学习是人工智能的一个分支。机器学习研究和构建的是一种特殊算法（而非某一个特定的算法），能够让计算机自己在数据中学习从而进行预测。机器学习包含了很多种不同的算法，深度学习就是其中之一，其他方法包括决策树，聚类，贝叶斯等。

机器学习的挑战

模型迭代会增加额外开销

迭代意味着等待结果返回，并在计算能力方面产生更费用。尽管迭代可能带来更出色的结果，但为了更快提供解决方案，数据科学团队通常会限制迭代次数

缩减取样意味着降低模型的准确性

数据科学团队经常发现，由于算力的限制，他们不得不对数据集进行缩减取样，最后导致结果不准确，业务决策也不甚理想。

将模型投入生产是一项艰巨的任务

将模型投入生产非常耗时且繁琐，通常涉及大量代码重构，并会造成周期时间延长和价值生成延迟。

数据科学家

加速机器学习的优势

缩短等待时间

通过对比CPU的行业标准快18倍的解决方案，缩短等待流程完成时间，从而获得更多时间迭代和测试解决方案

改善结果

通过高性能处理来分析数TB的数据集，获得准确性更高的结果，并提高报告生成速度

无需重构

通过高性能处理来分析数TB的数据集，获得准确性更高的结果，并提高报告生成速度

IT基础架构专业人员

减少开销

与基于CPU的行业标准相比，该解决放啊的性价比高7倍，可通过GPU加速充分利用预算

更明智的决策

利用所有的数据，做出更明智的业务决策，改善企业表现，更好地满足客户需求

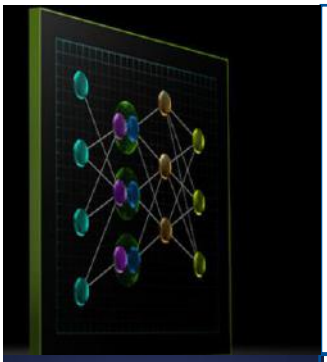
无缝扩展

从台式机轻松扩展为具有一致直观架构的多节点、多GPU集群

解决方案-AI和数据科学

深度学习推理平台

目前，先进的 AI 服务愈加受到市场青睐，其中包括图像和语音识别、自然语言处理、视觉搜索和个性化推荐。与此同时，数据集不断扩大，网络也变得越来越复杂，用户期望的延迟要求也愈发严格。NVIDIA 的推理平台可在云中、数据中心、网络边缘以及自主机器等平台上提供至关重要的性能、效率和响应速度，以支持新一代 AI 产品和服务。

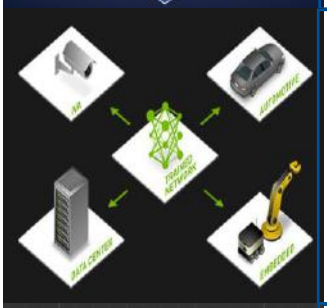


借助 NVIDIA TensorRT 发挥 NVIDIA GPU 的全部潜能

NVIDIA® TensorRT™ 是一款高性能推理平台，在充分发挥 NVIDIA Tensor Core GPU 的强大功能方面发挥着关键作用。与仅使用 CPU 的平台相比，TensorRT 可使吞吐量提升高达 40 倍，同时还可更大限度地降低延迟。使用 TensorRT，可以从任何框架入手，并在生产环境中快速优化、验证和部署经过训练的神经网络。TensorRT 还可在 NVIDIA NGC 目录上使用。

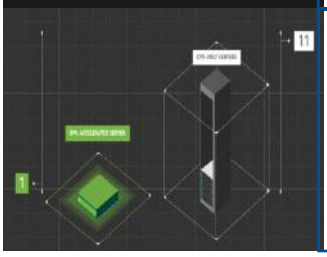
借助 NVIDIA Triton 推理服务器简化部署

NVIDIA Triton 推理服务器（以前称为 TensorRT 推理服务器）是一款开源软件，可简化深度学习模型在生产环境中的部署。借助 Triton 推理服务器，团队可以通过任何框架（TensorFlow、PyTorch、TensorRT Plan、Caffe、MXNet 或自定义框架），在任何基于 GPU 或 CPU 的基础设施上从本地存储、Google 云端平台或 AWS S3 部署经过训练的 AI 模型。它可在单个 GPU 上同时运行多个模型，以更大限度地提高利用率，并可与 Kubernetes 集成以用于编排、指标和自动扩展。



支持统一、可扩展的深度学习推理

通过搭载统一架构，各深度学习框架上的神经网络均可由 NVIDIA TensorRT 进行训练和优化，然后部署到边缘进行实时推理。借助 NVIDIA DGX™ 系统 、 NVIDIA Tensor 核心 GPU 、 NVIDIA Jetson™ 和 NVIDIA DRIVE™ ，NVIDIA 提供了一个端到端的，完全可扩展的深度学习平台，如 MLPerf 基准套件所示。



显著节省成本

要使服务器保持更高生产效率，数据中心管理者必须在性能与效率之间进行权衡。对于深度学习推理应用和服务而言，一台 NVIDIA T4 服务器可取代多台通用 CPU 服务器，从而降低能耗并节省购置和运营成本。

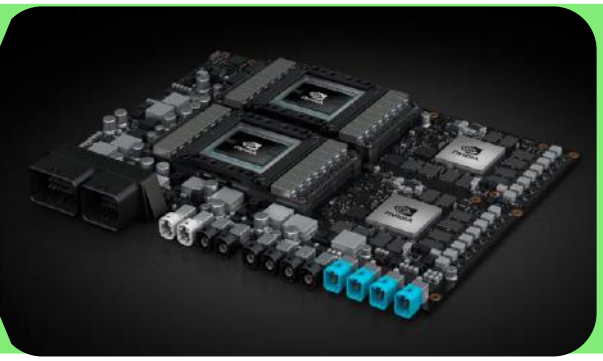
解决方案-AI和数据科学

深度学习推理平台

推理解决方案包括：数据中心、自动驾驶汽车、智能视频分析、嵌入式设备。主要介绍在自动驾驶方面的解决方案。

为安全自动驾驶解决方案提供支持，利用低延迟计算增强自动驾驶汽车性能。

安全是自动驾驶的第一需求，而**这需要高性能 AI 计算解决方案以极高的精确性处理传感器数据**。NVIDIA DRIVE™ Xavier 和 Pegasus 平台能够实时推理深度神经网络。基于深度学习的感知、定位和路线规划使车辆能够了解其周围环境并安全行驶。但技术总是在不断发展。随着自动驾驶车辆解决方案的发展，工程师继续利用各种深度学习框架在数据中心训练新的深度神经网络 (DNN) 模型。支持各大主要框架的 NVIDIA DRIVE 可通过无线更新变得更智能。即使在自动驾驶汽车投入生产后，该平台也可以适应新的框架和模型，从而实现更多的附加功能和更高级别的自动驾驶。

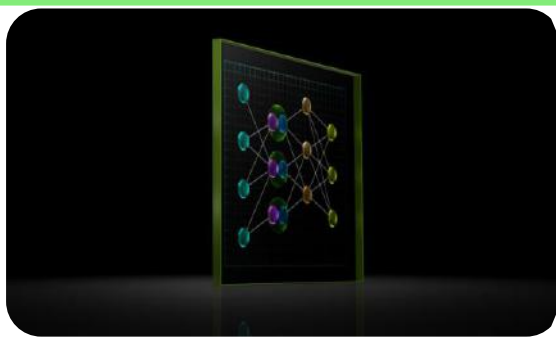


极具创新的架构

用于车内 AI 计算的 NVIDIA DRIVE 架构可以从 DRIVE Xavier（一款用于 L3/4 级自动驾驶的小型节能计算机）扩展至 DRIVE Pegasus（一款用于 L5 级无人驾驶汽车应用程序的功能强大的 AI 超级计算机）。DRIVE Xavier 基于世界超强性能的片上系统，而 DRIVE Pegasus 可提供无与伦比的每秒 320 万亿次运算 (TOPS)。

利用 NVIDIA TensorRT 进行优化

借助 NVIDIA 的统一架构，DNN 可以在数据中心的 NVIDIA® DGX™ 系统上进行训练，并且可以部署在自动驾驶汽车上，以便 DRIVE Xavier 或 DRIVE Pegasus 进行实时推理。NVIDIA TensorRT™ 可编程推理加速器更可进一步优化深度学习模型。它可以快速验证并将经过训练的 DNN 部署到汽车平台，以及加速推理深度学习应用程序的生产部署。



数据越多，响应越快

超过 370 家公司都在其自动驾驶汽车研发阶段应用 NVIDIA DRIVE。利用 TensorRT 优化后，NVIDIA 其中一家合作伙伴 TuSimple 已将推理性能提高 30%。TuSimple 成立于 2015 年，主要开发适用于自动驾驶长途货运方面的技术。TensorRT 带来的性能提升使得 TuSimple 能够在合理的响应时间内 分析更多的摄像头数据，并将新的 AI 算法添加到其自动驾驶卡车中。

解决方案-AI和数据科学

对话式AI

加速从语音识别到语言理解和语音合成的完整流程：

AI 驱动的语音和语言服务开创了进行个性化自然对话的革命性途径，但这些服务必须满足严格的准确性和延迟要求，以便实现实时互动。借助 NVIDIA 的对话式 AI SDK，开发者可以快速构建和部署先进的 AI 服务，在单个统一的架构中驱动多个应用程序，仅需少量的前期投资即可提供高精度、低延迟的系统。

NVIDIA 对话式AI应用程序解决方案

训练解决方案

轻松开发NVIDIA NeMo模型	借助NVIDIA TAO工具包进行更智能的训练	在NVIDIA DGX A100系统上进行训练
使用开源框架NVIDIA NeMo构建、训练和微调先进的语音和语言模型	使用产品级NVIDIA 预训练模型和NVIDIA TAO工具包将开发时间加快10倍	以超快的速度和超高的可扩展性学习包含数十亿个参数的强大语言模型，从而加快求解速度

部署解决方案

借助 NVIDIA Riva 简化部署	使用 NVIDIA EGX 平台在边缘进行部署
在云端、数据中心和边缘部署经过优化的对话式AI服务，以获得出色的性能	通过在边缘处理大量语音和语言数据实现实时对话，同时避免网络延迟

解决方案-AI和数据科学

对话式AI

对话式AI应用程序：

多人讲话转录

传统的语音到文本算法已经得到发展，它们现在能够转录会议、讲座和社交对话，同时识别演讲者并标记他们的贡献。NVIDIA Riva （对话式 AI 的应用框架）允许用户在呼叫中心和视频会议上创建准确的转录，还允许用户自动处理医患互动中的临床笔记。借助 Riva，用户还可以自定义模型和制作流程，以满足具体用例需求。

虚拟助理

虚拟助理可以通过类似人类的方式与客户互动，从而为联络中心、智能扬声器和车内智能助手的互动提供支持。由于缺少对话跟踪等关键组件，仅凭语音识别、语言理解、语音合成和 vocoding 等 AI 驱动的服务自身还无法支持此类系统。Riva 为这些主干服务补充了易于使用的组件，使它们可以扩展到任何应用程序。

解决方案-AI和数据科学

预测与预报

对话式AI应用程序：

预测与预报是帮助企业建模未来趋势的强大工具。借助 NVIDIA 加速数据科学，企业可以使用大规模数据集并获得高度准确的见解，以推动数据驱动型决策。

预测与预报面临的挑战

非常耗时

创建准确的预测需要大量数据。随着大数据用例的持续增长，CPU 性能已经成为主要瓶颈。这些限制增加了周期时间和成本。

非常昂贵

通过企业级基础设施缩短周期时间。大规模 CPU 基础设施会产生巨大的成本，而这会降低数据驱动型企业的投资回报。

非常繁琐

将大规模预测过程投入生产困难重重。通常需要进行重大的软件重构和团队之间的交接，因此可能会大大延迟洞察的产生。

NVIDIA全栈数据科学解决方案

无论是从头开始构建新模型，还是需要微调关键业务支持流程，NVIDIA 都能提供解决方案来加快企业的预测速度。通过全面开发软件和硬件，NVIDIA 提供了企业级解决方案，帮助企业轻松获得见解并部署模型，以改善运营或更好地为客户服务。**借助 RAPIDS 和 CUDA，数据科学家可以加速依托 NVIDIA GPU 运行的预测和预报流程，将数据加载、处理和训练等操作所需时间从几天减少到几分钟。**通过熟悉的 Python 或 Java 语言进行 NVIDIA 加速计算，从而轻松踏入加速数据科学领域。

预测与预报

客户应用

Walmart labs

沃尔玛是全球规模最大的零售企业之一。为了不断取得进步并满足客户需求，沃尔玛需要推动整体业务创新。沃尔玛实验室正是推动这种创新的智囊团。**沃尔玛实验室高度依赖数据科学来准确预测全球数千家门店的库存需求。**虽然他们的运营一直非常高效，但通过 RAPIDS 和 NVIDIA GPU，其预测性能提高了 1.3%，这给沃尔玛节省了数百万美元，并吸引了更多优质客户。

100倍 加速模型训练

20倍 降低计算成本

通过改进预测流程节省数百万美元

Capital one

100倍 加速模型训练

97% 降低计算成本

Capital One 一直是一家数据驱动型公司。为了更好地为客户服务，**Capital One 经历了一次重大转型，从一家使用科技的银行转变为一家经营银行业务的科技公司。**在这一转型过程中，Capital One 对数据科学产生了浓厚的兴趣。为了推动数据驱动型文化，Capital One 采用了 RAPIDS、Dask 和 NVIDIA GPU，大大提高了预测运营的绩效和投资回报。在 NVIDIA 加速数据科学的推动下，Capital One 获得了真正“彻底改变银行业务”所需的工具。

以更少的成本获得准确的结果，从而为客户提供更好的服务

解决方案-AI和数据科学

大型语言模型

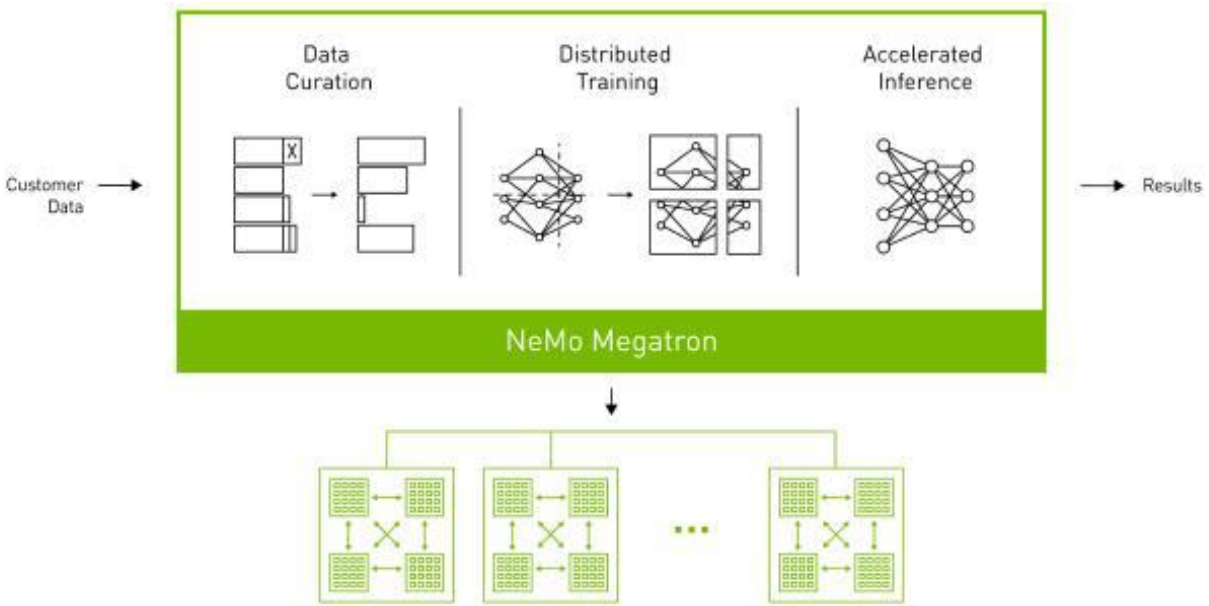
大语言模型 (LLM) 代表着 AI 领域的重大进步，并有望通过习得的知识改变该领域。在过去几年中，LLM 的规模每年增加 10 倍，而且随着这些模型的复杂程度和规模的增加，其性能也在不断发展。然而，LLM 的开发与维护并非易事，这使得大多数企业都无法使用 LLM。

NVIDIA NeMo LLM 服务

- 在 NVIDIA AI 平台上运行的 NeMo LLM 服务为企业提供了在私有云和公有云上自定义并部署 LLM 或通过 API 服务访问 LLM 的捷径
- NeMo Megatron 是一种端到端框架，用于训练和部署具有数十亿或数万亿参数的 LLM
- 通过 NVIDIA Triton 助力 LLM 推理

借助NVIDIA BioNeMo扩展药物发现研究

BioNeMo 是一款基于 NVIDIA NeMo Megatron 构建的 AI 赋能药物研发云服务和框架，用于在超级计算规模下训练和部署大型生物分子 Transformer AI 模型。服务包括预训练 LLM、对蛋白质、DNA、RNA 和化学的通用文件格式的原生支持，还提供可供 SMILES（用于分子结构）和 FASTA（用于氨基酸和核苷酸序列）使用的数据加载器。BioNeMo 框架也可供下载，以便用户可以在自己的基础架构上运行。

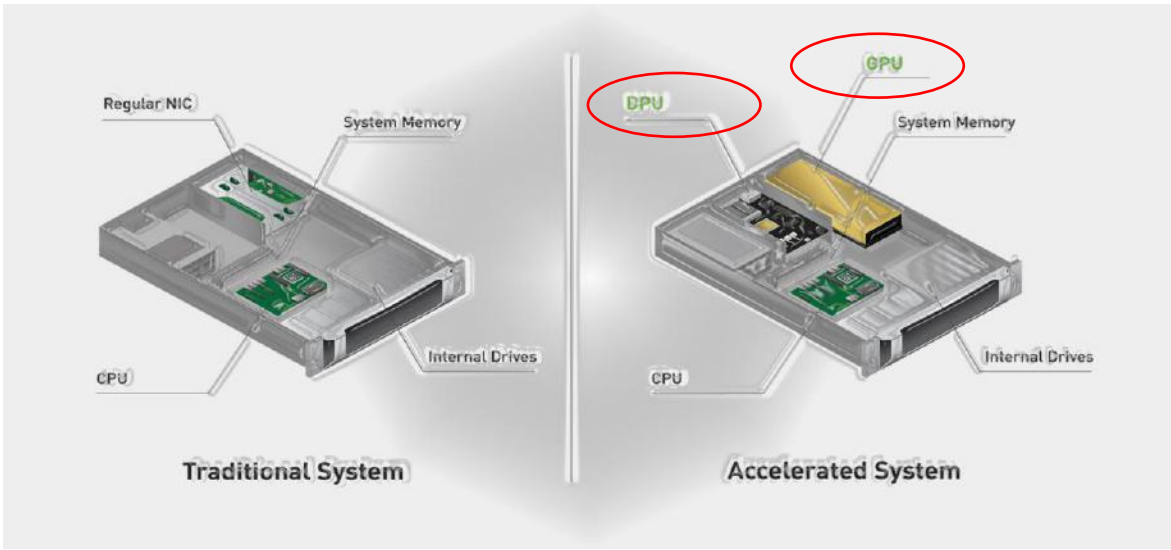


解决方案-数据中心和云计算

面向企业IT的加速计算

加速系统剖析

加速系统是计算机发展历程的下一个阶段。就像如今的智能手机都配备了图形处理器和 AI 处理器一样，每个服务器和工作站都将配备计算加速器，为当今使用的现代应用（包括 AI、可视化和自主机器）提供支持。其中许多系统还会配备数据处理器，可加速对云原生和云计算框架至关重要的网络、存储和安全服务。



NVIDIA 加速计算平台

随着加速计算越来越普及，对针对特定用途进行优化的加速系统的需求也日益增加。NVIDIA 定义了一系列加速平台，每个平台都由根据用例需求而设计的硬件系统，以及能够实现业务应用的运营和管理的软件栈组成。

NVIDIA EGX 适用于数据中心和边缘的主流多用途加速服务器	**NVIDIA HGX** 助力合作伙伴构建AI超级计算机	**NVIDIA DGX** 先进的AI超级计算机
NVIDIA OVX 适用于omniverse 工业数字孪生的超级计算机	**NVIDIA AGX** AI赋能的自主机器	**NVIDIA IGX** 用于边缘AI的高级功能安全性和可靠性

边缘计算

创造一个更快、更智能、联系更紧密的世界

零售商店、城市街道、仓储车间、医院的数十亿个物联网传感器正在生成大量数据。从这些数据中更快地挖掘见解，就可以改进服务和简化运作，甚至还可以拯救生命。但要做到这一点，企业需要实时做出决策，而这需要将他们的 AI 计算带到数据所在位置，即网络边缘。

边缘计算

在边缘，物联网和移动设备通过**嵌入式处理器收集数据**。边缘计算将 AI 的强大功能直接应用于这些设备，**在数据源（而不是在云端或数据中心）处处理获取的数据。**这加速了 AI 工作流，为实时决策制定和软件定义的自主机器提供动力支持。

降低延迟

边缘计算在行动点处理数据，可减少或消除数据传输过程，进而加速AI工作流

提高可靠性

在本地处理敏感数据，就不需要将其发送到云端，因此可以更好地保护敏感数据

降低成本

将数据发送到云端需要有足够的带宽和存储，而在本地处理数据可以减少这方面的成本

更广的覆盖范围

边缘计算可在本地进行，无需链接互联网，如此以来就扩展了AI的覆盖范围

边缘计算

借助NVIDIA解决方案使边缘触手可及

AI 和云原生应用、物联网及其数十亿的传感器，以及 5G 网络使得在边缘大规模部署 AI 成为可能。探索企业边缘、嵌入式边缘以及工业级边缘中的 NVIDIA 解决方案，这些方案将行动点的自动化智能和实时制定决策的可能性转化为现实成果。

企业边缘计算

NVIDIA EGX™ 平台履行在边缘计算领域的承诺，提供强大的分布式计算、安全的远程管理以及与行业领先技术的兼容性。此平台整合了 NVIDIA 认证系统™、嵌入式平台、软件和管理服务，能让企业充分运用边缘 AI 的强大功能

工业边缘AI

满足在受监管的工业环境中部署 AI 的非常具体的要求，以确保达到所需的生产力、安全性和合规性。NVIDIA IGX 是一个工业级边缘 AI 平台，使企业组织能够以安全的方式自信地提供 AI，以支持人类和机器协作

自主机器和嵌入式边缘

借助功能强大的AI 计算机，为节能高效的自主机器带来新一代边缘产品。NVIDIA Jetson™ 平台可为边缘带来出色的新功能，加速产品开发和大规模部署

解决方案-数据中心和云计算

虚拟化

NVIDIA 虚拟 GPU (vGPU) 软件为众多工作负载（从图形丰富的虚拟工作站到数据科学和 AI）提供强大的 GPU 性能，使 IT 能够利用虚拟化的管理和安全优势以及现代工作负载所需的 NVIDIA GPU 的性能。NVIDIA **vGPU 软件安装在云或企业数据中心服务器的物理 GPU 上，会创建虚拟 GPU**，这些 GPU 可以在多个虚拟机（可随时随地通过任意设备访问）之间共享。

虚拟GPU的优势

裸机性能：提供几乎与裸机环境无差别的性能

管理和监控：利用常见的数据中心管理工具，例如实施迁移

出色资源利用率：使用部分或多GPU虚拟机实现调配GPU资源

提高业务连续性：响应不断变化的业务需求和远程团队

增强现实和虚拟现实

NVIDIA 专业级扩展现实技术

扩展现实 (XR) 是多种沉浸式技术的统称，包括虚拟现实 (VR)、增强现实 (AR) 和混合现实 (MR)。扩展现实 (XR) 采用可实现规模化运作的技术，推动专业工作流程的变革。在手中持握虚拟模型、穿行于整个虚拟建筑或在虚拟环境中模拟复杂的外科手术过程 – XR 正在彻底改变人们在工作场所的经历。

NVIDIA 的 XR 平台包含 NVIDIA VR Ready GPU、创新型 XR 工具和开发软件，有助于提供出色的 VR 体验。

NVIDIA CloudXR

NVIDIA CloudXR 是基于NVIDIA RTX技术的突破性创新成果，可跨5G和Wi-Fi网络提供VR和AR体验

NVIDIA VRWorks

NVIDIA VRWorks 为虚拟现实带来物理属性逼真的视觉、听觉、触觉交互和模拟环境，将虚拟现实的临场感提升到了更高境界

NVIDIA VR Ready

NVIDIA VR Ready计划可确保用户的系统组件（GPU\CPU\HMD和驱动）能够呈现出色的沉浸式虚拟现实体验

NVIDIA Omniverse

NVIDIA Omniverse Create 应用可实现高级场景合成，让用户能与其制作出的场景进行交互

NVIDIA 虚拟现实捕捉与回放（VCR）

NVIDIA VCR 可让开发者和用户准确捕捉并回放VR会话，实现性能测试、场景故障排除等操作

渲染

高级渲染解决方案

渲染，就是将 3D 模型转换成 2D 图像，并最终呈现在屏幕上的过程。虽然这里只有一句话，但是这一句话里面包含了太多的数学、物理和计算机方面的知识，它描述了我们用计算机来虚拟化真实世界的基本逻辑。渲染过程是需要计算机进行运算且消耗时间的。

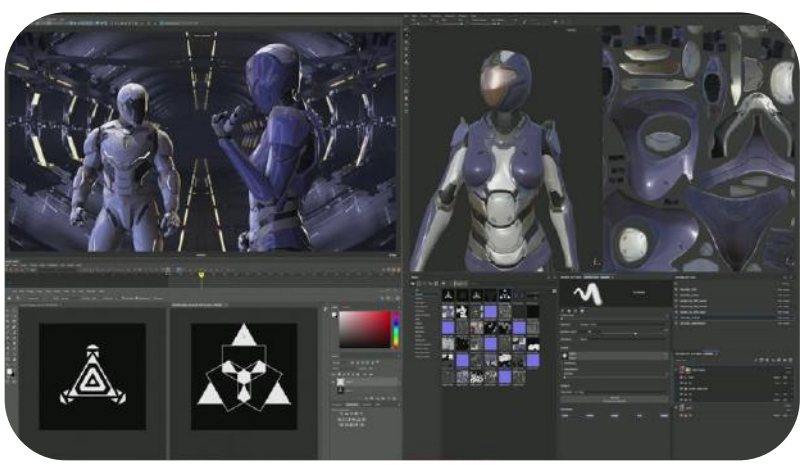
NVIDIA® RTX 平台提供现今超快速的 GPU 渲染解决方案。将 NVIDIA RTX™ 或 NVIDIA Quadro 显卡的强大功能与 NVIDIA RTX™ 加速应用相结合，各行业的设计师和艺术家可以将先进的渲染技术引入其专业工作流程中。

利用 RTX 技术实现光速渲染

如今，专业人士创造的内容数量远胜以往。在渲染单张图像时，传统的 CPU 渲染往往需耗时几分钟乃至几小时，而这种速度显然无法满足他们的需求。借助 NVIDIA RTX 加速的光线追踪和 AI 降噪功能，他们便可在应用程序视口中使用交互式光线追踪技术，进而实现创意设计流程转型。

适用于各种渲染工作流程的优化解决方案

使用 NVIDIA RTX 和 NVIDIA Quadro GPU 以及 NVIDIA NVLink ™ 技术，能够处理极其复杂的渲染工作负载，高达 96GB 的 GPU 显存适用于大型场景和多应用程序工作流程。 用户可以使用 NVIDIA OptiX™ AI 加速的降噪技术自动处理并加速复杂的渲染任务，并通过 NVIDIA EGX™ 平台在数据中心实现高性能的灵活渲染。



解决方案-设计和仿真

渲染

利用 NVIDIA vMaterials 打造出色的逼真效果

NVIDIA 的 vMaterials 目录汇集了使用 NVIDIA 材质定义语言 (MDL) 描述的各类真实材质。vMaterials 由 NVIDIA 材质专家设计和验证，具有准确、可控和一致的特点。利用此目录，专业人士可在他们的设计中加入逼真材质，并在支持的应用程序之间共享基于物理性质的材质和光线。

使用 NVIDIA Iray 模拟现实效果

NVIDIA Iray 是一项十分先进的渲染技术，让专业人士能够模拟光线和材质的物理变化，为交互式和批量渲染工作流程生成逼真的图像。RTX GPU 支持为领先的图形应用程序带来实时光线追踪和 AI 加速的降噪技术，为设计师和数字艺术家提供打造具有惊人视觉效果的渲染所需的工具，同时显著加速工作流程。

解决方案-设计和仿真

图像虚拟化

NVIDIA RTX 虚拟工作站

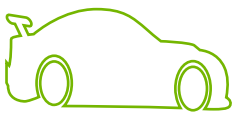
NVIDIA RTX 虚拟工作站 (vWS) 软件与领先的 GPU 相结合，可以为视觉计算提供强劲动力，从数据中心或云向任何设备提供极其强大的虚拟工作站。数百万创意专业人员和技术专业人员可以随时随地访问要求极高的应用程序，不仅可以获得堪比物理工作站的卓越性能，而且还可满足更高的安全性需求。

NVIDIA RTX vWS 是唯一支持 NVIDIA RTX 技术的虚拟工作站，将光线追踪和 AI 去噪等高级功能引入了虚拟环境。支持新一代 NVIDIA GPU，可以最大程度发挥出色性能，让设计师和工程师可以更快地创作出优秀作品。IT 可以将数据中心的任何应用程序虚拟化，使用任何设备实现工作站性能，获得与物理工作站别无二致的体验。

NVIDIA RTX vWS客户

设计更好的空间

建筑公司Gould Evans利用NVIDIA RTX vWS使美国各地100多名设计师实现协作



赢得比赛

SportPesa Racing Point F1使用NVIDIA RTX vWS 设计并制造具有加速CAD NVIDIA 应用的赛车

加速癌症研究

荷兰癌症研究所更新其IT基础设施以加快研究进程并简化患者护理

助力数字艺术家

DNEG助力数字艺术家随时随地访问性能堪比物理工作站的虚拟工作站

解决方案-设计和仿真

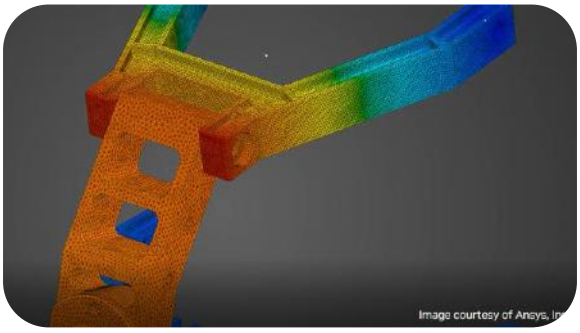
工程模拟

适用于工程模拟分析的先进 GPU 加速

NVIDIA适用于工程模拟和分析的业内领先解决方案能够提供高性能、可扩展性及企业级可靠性。借助 NVIDIA GPU 超大显存容量和超强性能，工程师和分析师能够利用所需的计算能力执行复杂模拟并解决极具挑战性的问题。

适用于计算流体动力学（CFD）的优化解决方案

利用 NVIDIA RTX 和 NVIDIA Quadro GPU 快速分析复杂的流体流动情况。通过能够提供高达 96 (GB) GPU 显存的 NVIDIA NVLink® 技术，加快超大型模型和模拟的处理速度。负责进行复杂 CFD 建模的分析师以及负责评估早期概念的流体效应的工程师都可以极大地缩短解决时间，几小时可以缩短到几秒，而几天可以缩短到几小时。



适用于结构分析的实时有限元分析（FEA）

探索概念构想并快速迭代，从而针对复杂结构工程问题制定出更好的决策。各种经验水平的工程师和设计师都可以使用 NVIDIA 解决方案快速模拟包含大型装配体、建筑模型和复杂材料的场景。

加快使用计算电磁学（CEM）的电子设计的速度

在设计高性能电子产品和组件时，能够提高创新速度并降低成本。借助 GPU 加速，模拟电磁性能，以准确预测电磁辐射、干扰和信号传输。

解决方案-设计和仿真

直播

聚焦媒体和娱乐行业的未来

影视行业正在不断创新和重塑其制作流程，满足观众对高质量内容日益增长的需求，同时在世界范围内的员工队伍中保持紧凑的工作节奏。此外，串流服务对持续发布和刷新的需求日益增长，以满足日渐扩大的订阅者群体。因此，各工作室纷纷开始转向 AI、虚拟制片、实时渲染和内容协作，无论是桌面端或数据中心抑或是云端，从而显著加快制作速度。

影视行业　直　播　广　告　游　戏

顶尖的影视工作室利用 NVIDIA 技术，打造更为先进、视觉效果丰富的影视与电视节目。利用前沿技术（如 AI、仿真、实时光线追踪和虚拟制片）助力影片打破票房纪录、打造超人气电视节目以及斩获奥斯卡金像奖。

影视行业解决方案

NVIDIA Omniverse Enterprise

革新工作室的制作流程。NVIDIA Omniverse 是专门用于加速创意制作的开放平台。在交互式仿真环境中，借助高端内容制作工具和无缝式协作之间的一键式互操作，团队能够以惊人的速度开展内容创作。

RTX 专业解决方案

业界广受好评的故事片需要最新技术进展的加持，从移动设备和台式机，到数据中心和云，皆包含在内。NVIDIA RTX 专业解决方案为您提供幕后支持，使您精准有力地讲述精彩绝伦的故事。

云端工作室

随着远程办公的工作室逐渐增加，RTX Virtual Workstation(vWS) 使艺术家能够随时随地访问功能强大的虚拟工作站，并享受与实体工作站水准无异的性能。

直播

应用案例

NVIDIA 是 14 年来奥斯卡最佳视觉效果奖的背后推手

14 年来，在获得奥斯卡最佳视觉效果奖提名的电影中，大量炫丽的视觉效果和电影艺术都离不开 NVIDIA 技术的支持。

AI"目之所及"的精彩：ILM 使用 Omniverse DeepSearch 打造绚丽的天空

知名的 Industrial Light Magic 工作室利用 Omniverse AI 助力的搜索工具筛选庞大的 3D 场景数据库。

影视行业　**直　播**　广　告　游　戏

NVIDIA 技术正帮助主播革新内容创作、传播和观看的方式。新款企业级 NVIDIA RTX 专业 GPU 与基于互联网协议 (IP) 的 NVIDIA 高速低延迟解决方案相结合，实现新颖的直播工作流。借助 AI 解决方案和新一代基础架构，在使用各种设备开辟全球市场的同时，掌握新的创新能力和深入的客户见解。

解决方案-设计和仿真

直播

影视行业　直　播　广　告　游　戏

随着高性能联网设备的普及，消费者的行为正在发生迅速改变。代理商和公司内部的创意团队必须对动态、体验性、品牌化的内容进行创新，以新方式吸引受众。在 NVIDIA RTX、实时技术和深度学习的加持下，广告商和品牌领导者将能够创造沉浸式实时体验，并在各个平台上获得更深入的消费者见解分析。

加速广告行业工作流程

视频制作
实时进行编辑和颜色分级，无需预先缓存或生成代理。在 GPU 加速应用程序（如 Blackmagic Design DaVinci Resolve 和 Adobe Premiere Pro）中通过惊艳的 360 度全景视频为观众打造沉浸式体验。

多渠道传输
使用 AI 将 3D 图像和视频动态传输到所有设备。有了 NVIDIA CloudXR™、智能手机、平板电脑、增强现实和虚拟现实技术，头显设备便化身为打开新世界的窗口。

人工智能
使用 DaVinci Resolve、Autodesk Flame 和 Adobe Creative Cloud 等 AI 加速应用程序更快地创建内容。通过深度学习利用精准的内容定位减少获取见解的时间并更快速地销售广告存货。

广告行业解决方案

解决方案-设计和仿真

直播

广告行业解决方案

桌面和移动工作站

长久以来，NVIDIA GPU 一直是艺术家和设计师的企业级桌面显卡的标准配备。新的 NVIDIA RTX 工作站由 NVIDIA Ampere 架构提供动力支持，并结合实时光线追踪和高级可视化功能，为专业的创意工作流程提供出色的功能和性能。

数据科学

借助 NVIDIA 提供动力支持的数据科学工作站，获取将海量数据转变为见解分析所需的强大性能，并打造令人惊叹的客户体验。

DGX

借助 NVIDIA DGX™ 解决方案应对 AI 训练及推理、数据分析、高级模拟和可视化等复杂的 AI 挑战。

案例

WPP 提供个性化的 3D 内容和体验

全球大型营销服务组织率先在 Omniverse Cloud 上推出汽车营销服务，为先进的汽车品牌提供定制的 3D 内容和体验，首先是 Mazda 和 McLaren.metaverse 应用，可在任何位置使用任何设备上进行。

解决方案-机器人开发和边缘计算👇

AI-on-5G

NVIDIA AI on-5G 是一个统一的平台，汇集了边缘 AI 和 5G 的发展成果，可加速企业和行业的数字化转型。**5G 可为数十亿台设备提供基础连接**，将 AI 算法和应用程序的范围扩展到边缘的所有连接对象，创造了新的用例和新的市场。

NVIDIA AI-on-5G 由超融合的 **NVIDIA EGX™ 计算平台**、用于软件定义的 5G 虚拟无线电局域网 (vRAN) 的 **NVIDIA Aerial™ SDK** 和**企业 AI 应用程序**组成，包括多种 SDK，例如 NVIDIA Isaac™ 和 NVIDIA Metropolis™。此解决方案可在本地部署，并由企业进行管理，也可以由像 Google Cloud 这样的超扩展器进行管理。

Metropolis 边缘 AI-on-5G 平台现已全面推出

NVIDIA 和 Mavenir 正在实现智能边缘网络构建方式的创新。企业和电信公司可以使用融合边缘服务器一起部署 AI 和 5G。

NVIDIA 与 Google Cloud 联合打造 AI-on-5G 开放创新实验室

NVIDIA 与 Google Cloud 合作，率先打造 AI-on-5G 实验室，加速 5G 网络运营商对于 AI 应用的开发

NVIDIA Aerial 5G 平台扩展对 Arm 的支持

NVIDIA 宣布在 NVIDIA Aerial A100 AI-on-5G 平台中支持基于 Arm 的 CPU，助力企业通过软件定义的云原生 5G vRAN 来部署智能服务。

解决方案-机器人开发和边缘计算

工业级AI

世界前沿的工业公司正在实施 NVIDIA 技术，以部署大规模 AI 项目。GPU 加速计算可使工业公司实现工业级 AI 应用，从而利用海量传感器和操作数据来优化运营，同时缩短分析时间并降低成本。

NVIDIA IGX 是一个工业级边缘 AI 平台，通过 NVIDIA IGX 加速智能机器的部署，可提供高性能、高级的功能安全性和可靠性

加快AI开发和部署速度

NVIDIA 与合作伙伴合力开发出 AI解决方案，旨在加快部署由 GPU加速的深度学习和机器学习模型

提升精确度

运用适度学习技术，提升工业检测和预测性维护算法的精确度

在工业规模上发挥AI优势

利用设备中的庞大数据加快训练AI算法，并大规模优化企业运营

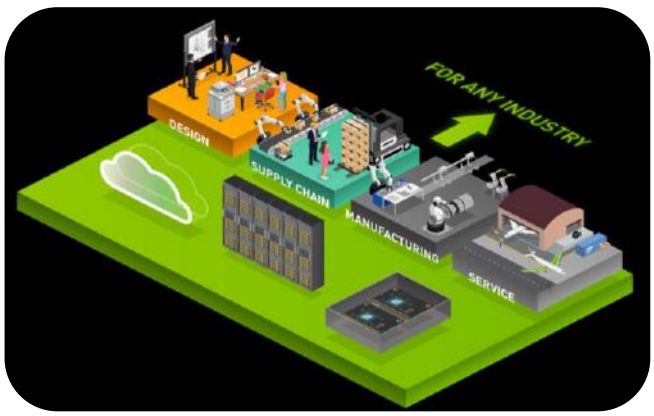
适用于AI和高性能计算的单个平台

设计
- 高性能计算
- 建模与仿真
- 可制造性与可服务性的设计

供应链
- 预测
- 供应链优化

制造
- 机器人
- CV检测
- 预测维护
- 进程控制

服务
- 预测维护
- 现场检测
- 物流优化
- 零件库存优化



解决方案-机器人开发和边缘计算👇

工业级AI

NVIDIA 工业解决方案

边缘加速计算

当今的工业边缘计算需要GPU驱动的计算能力，以支持工厂内的工业检测和机器人，以及现场设备和资产的预测性维护。NVIDIA Tesla 级边缘GPU和 Jetson 解决方案可加速此类及其他一些应用的最强大的边缘计算系统

为数据中心带来强大计算能力

NVIDIA Tesla GPU 加速的计算平台可显著加快深度学习和机器学习模型的训练以及 HPC 工作负载，从而提供前所未有的深入洞见。Tesla GPU 适用于所有主流计算机系统和服务器制造商，可提升 HPC 应用程序的性能，并在数据中心内训练 AI 模型。

利用云实现数据中心的大众化

云计算通过实现数据中心的大众化和彻底改变企业的运作方式引发了行业变革。全球各大云平台均可按需获取 NVIDIA GPU，NVIDIA GPU Cloud (NGC) 更可提供用于简化部署的 GPU 加速容器，其中包括 TensorFlow、PyTorch、MXNet 等深度学习框架。

加速软件部署

NVIDIA 的软件库和 SDK 共同造就了一种可扩展的解决方案，使客户能够在云端、服务器上或边缘位置部署推理和 AI 功能。这种软件投资旨在加快客户部署，同时降低总体开发成本。这些 SDK 投资包括用于嵌入式设备的 JetPack，用于 IVA 的 DeepStream用于机器人的 Isaac、用于推理的 TensorRT，用于调优 DNN 的 TAO 工具包 、用于容器和 AI 软件的 NVIDIA GPU Cloud 等等。

解决方案-机器人开发和边缘计算

边缘部署管理

随时随地部署AI

NVIDIA Fleet Command 是一项云服务，能够跨分布式边缘基础设施安全部署、管理和扩展 AI 应用程序。Fleet Command 专为 AI 打造，是适用于 AI 生命周期管理的一站式解决方案，可提供简化部署、分层安全保护和详尽监控功能，从而支持您仅用几分钟，即可零基础实现 AI。

简化的边缘AI

只需单击几下，即可完成从软件安装到边缘部署的操作。“无需IT人员介入”的经过测试和优化的解决方案可简化设置和管理，并可在边缘运行超密集的应用程序

简化的部署

轻松将AI部署到任何位置。简化的界面可提供集中式AI管理和一键式资源调配，因此可以用户可以将应用程序部署和扩展到多个位置，从而加快获得AI见解的速度

AI生命周期管理

借助采用易用性设计的功能来简化AI生命周期管理，无线更新应用程序、扩展应用程序，并监控AI运行状况，从而通过一个屏幕优化部署和维护

分层安全保护

确保应用程序数据从云到边缘始终收到保护。Fleet Command 遵循领先的安全协议，对传输中的数据和静态数据进行加密，并提供持续监控

解决方案-高性能计算

HPC和AI

利用 GPU 加速的 HPC 和 AI 提高准确度

加速计算正在助力研究人员更快取得重大科学突破。研究人员已经很快意识到，在 AI 的助力下，他们可在更短时间内获得高精度结果，且可与科学模拟结果相媲美。这一结果已推动 AI 在高性能计算 (HPC) 中的应用。

HPC 和 AI 的使用对象有哪些?

HPC和AI可应用于几乎各个领域。

研究人员

研究人员可以**利用AI增强HPC模拟技术**，为各类科学工作负载取得更快、更出色的结果

工程师

工程师们正在**利用AI评估医疗设备、制造机器人以及汽车零部件**等各种设计

分析师

金融机构的分析师正在利用AI**识别和预测市场趋势、标记欺诈交易，以及加快在线支付**

……

NVIDIA Modulus——解锁工业与科学模拟功能

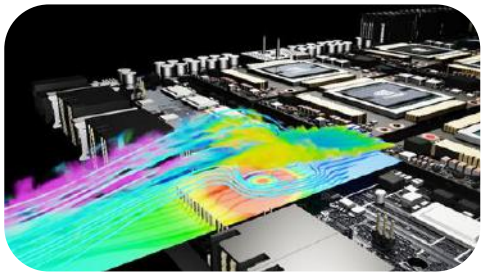
NVIDIA Modulus是一个AI加速模拟工具包，是一个基于物理信息的神经网络 (PINN) 工具包，旨在解决使用 AI 和物理技术过程中所面临的挑战。无论用户是希望着手于 AI 驱动的物理模拟，还是处理复杂的非线性物理问题，NVIDIA Modulus 都可助解决正向、反向或数据同化问题。

HPC和AI

HPC 和 AI 的实际应用

HPC 和 AI 应用非常广泛，例如帮助解决地球气候问题、加速科学发现，及模拟工作流程以更快完成各种任务。

模拟产品设计工作流程

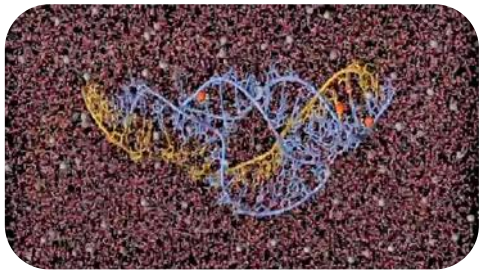


NVIDIA Modulus 是一种基于新型PINN架构的端到端AI驱动模拟框架。借助高精度数值求解器，Modulus 已经协助解决了多物理场问题，执行自动设计空间探索的速度更高达传统模拟器的1000倍。

模拟地球系统

HPC和AI用于众多地球科学领域，包括极端天气预测、物理仿真、即时预报、中期预报、不确定性量化、偏差修正、生成对抗网络、数据图像修复、网络-HPC耦合、PINN和地质工程等。

加速科学发现



从高能物理到生命科学和医疗健康领域，深度学习和人工智能与传统HPC的融合可加速各个领域的科学发现。

……

解决方案-高性能计算

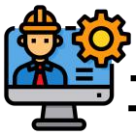
仿真与建模

谁会使用模拟和建模?

模拟和建模用于各行各业。使用模拟和建模，研究人员可以制造对抗疾病的新药，工程师可以模拟复杂的现实世界问题，分析师也可以来创建金融模型。

研究人员

研究人员正在使用GPU来更快地运行大规模模拟，更快地获取更深入见解，从而更快地公布新的发现结果



工程师

机械工程、地球科学和制造业领域的工程师正在基于GPU提供动力支持的系统进行复杂设计建模，分析其工作

分析师

金融机构正在使用NVIDIA GPU**从海量数据集提取见解**，从而做出实时决策

加快模拟工作负载处理速度

从流体模拟到分子动力学，应用程序帮助科学家、工程师和研究人员跨领域开展工作。如今，由 GPU 加速的成千上万的此类应用程序支持研究人员更高效地完成其毕生工作。可从 NVIDIA NGC™ 目录中获取使用主要的 HPC 应用程序。

GROMACS

GROMACS是一款**分子动力学应用程序**，旨在模拟包含数百到数百万个粒子的系统的牛顿运动方程

LAMMPS

大规模原子/分子并行模拟器（LAMMPS）是专为分子动力学模拟设计的软件应用程序

NAMD

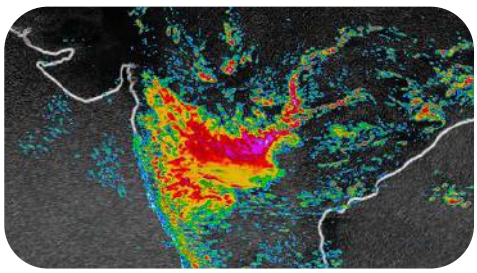
纳米尺度的分子动力学（NAMD）是一款并行分子动力学代码，专为实现大型生物分子系统的高性能模拟而设计

解决方案-高性能计算

仿真与建模

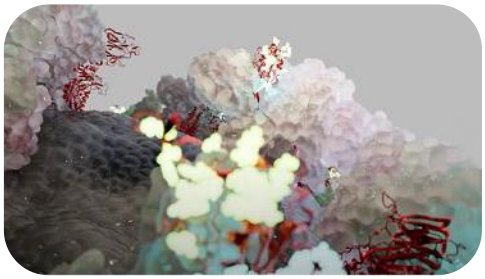
加速模拟和建模的实际应用

预测天气模式



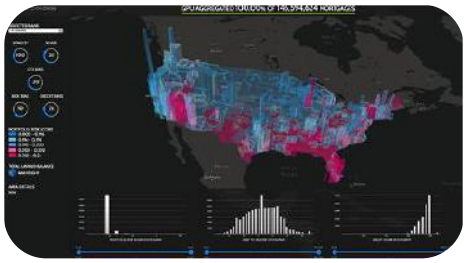
探索模拟如何应用于天气预报和气候建模中，其中包括自动化特征检测如何识别恶劣天气、太阳风暴和近地天体的威胁，以及加速模型和数据同化技术如何做出更准确的预测

新型冠状病毒模拟



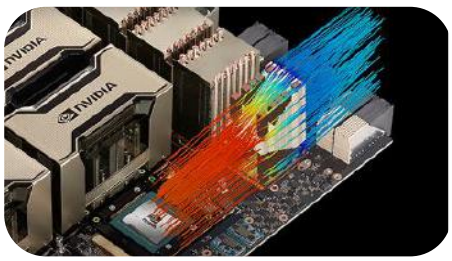
结合超过20万块NVIDIA GPU的计算能力和志愿者提供的其他计算资源，folding@Home项目对突刺蛋白进行了超大规模模拟

加速金融模型



随着金融模型的规模和复杂程度的不断提升，越来越多的数据科学家和开发者转向使用HPC来加速算法和模拟

加快工程模拟速度



帮助油藏工程师仅使用比基于CPU的解决方案更少的硬件资源，更快地开发更准确、可靠和高预测的模型

……

解决方案-高性能计算

科学可视化

加速模拟和建模的实际应用

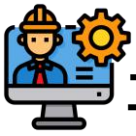
体验高性能计算 (HPC) 模拟的更佳方式就是通过可视化。NVIDIA 加速的科学可视化使研究人员能够以**交互式速度可视化其大型数据集**并更好地与全球各地的团队开展协作，从而加快数据分析和科学推广的速度。

谁会使用科学可视化?

科学可视化可应用于多个领域，包括实验室的研究人员、工作室的创意艺术家以及解决复杂技术难题的工程师等。

研究人员

研究人员正在使用科学可视化从大型HPC数据集收集见解，从而可视化蛋白质折叠、分析化学对接、了解超新星等



工程师

工程师正在使用科学可视化来分析机器人、制造系统和结构工程等各种使用案例的设计

创意艺术家

创意艺术家正在**将科学数据转化为逼真的视觉效果**，帮助研究人员和普通观众更好地理解其艺术背后的科学

加速科学可视化工作负载

NVIDIA 提供 NGC™ 目录中的各种可视化软件，使研究人员能够与同事远程协作，并以交互方式实时可视化其科学数据集，从而加快科学发现并更快地发布结果。

NVIDIA IndeX	NVIDIA IndeX® 是一个 3D 立体交互式可视化框架，支持科学家和研究人员可视化大量高性能计算数据集并与其进行交互

解决方案-高性能计算

科学可视化

NVIDIA Omniverse	通过使用 NVIDIA Omniverse™，研究人员和开发者可以构建自定义 3D 和仿真工作流，并可视化大规模 3D 数据集。Omniverse 基于开放标准构建而成，能够与领先的高性能计算工具和框架连接，例如 ParaView、NanoVDB、NeuralVDB、NVIDIA IndeX 和 NVIDIA Modulus。借助 Omniverse，高性能计算 (HPC) 团队可以统一数据集，跨地区开展协作。
VMD	VMD 专为生物分子系统（如蛋白质、核酸、脂质膜和碳水化合物结构）的建模、可视化和分析而设计。
NeuralVDB	NVIDIA NeuralVDB 能够实现由 AI 提供支持的大规模立体数据表示，也能够显著提升 OpenVDB 的效率。OpenVDB 是用于仿真和渲染稀疏体积数据（如水、火、烟雾和云）的行业标准库。
NVIDIA Modulus	NVIDIA Modulus 是一种神经网络框架，它以控制偏微分方程 (PDE) 形式将物理学的力量与数据相结合，以构建具有近乎实时延迟的高保真参数替代模型，并能够借助 Omniverse 扩展程序构建可视化。
NVIDIA HPC SDK	NVIDIA HPC SDK 包含经过验证的编译器、库和软件工具，对于更大限度地提高开发者工作效率以及 HPC 应用的性能和便携性而言至关重要。

解决方案-高性能计算

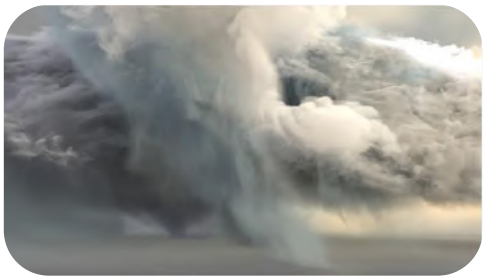
科学可视化

科学可视化具有可视化分子仿真、仿真大量数据以及数据提取和筛选等多种用例。

使用 Omniverse 提高研究人员的气候数据收集速度

美国国家海洋和大气管理局 (NOAA) 已委派 Lockheed Martin 和 NVIDIA 构建一个系统，利用 NVIDIA Omniverse 在 10 分钟或更短的时间内将最新气候数据中的复杂可视化结果输出给研究人员

可视化猛烈的龙卷风



借助 NVIDIA 渲染、仿真和 GPU 加速技术，威斯康星大学的气候研究人员通过协作式交互科学工作工作流，进一步了解这些不可预测的风暴的复杂性

使用 Omniverse 实现电影级气候可视化效果

气候仿真会产生大量的 3D 数据，但分析通常受限于 2D 预测。NVIDIA Omniverse 融合了大规模科学数据与电影级渲染功能，从而能够实现对复杂气候现象的交互式探索

星系风的交互式可视化效果

NVIDIA IndeX 是一种立体可视化工具，支持用户通过交互方式可视化整个数据集并加速收集更深入见解的过程。用户可以更改彩色地图来突出数据的细微属性，查看时间系列的横截面，并利用环境光遮蔽和阴影等功能来检查数据的关键组件

……

解决方案-自动驾驶汽车

Chauffeur

NVIDIA DRIVE® Chauffeur

NVIDIA DRIVE® Chauffeur 基于 NVIDIA DRIVE Orin™ 和 NVIDIA DRIVE SDK 构建。该平台采用**感知层、地图构建层和规划层**，以及基于高质量真实驾驶数据和合成数据训练的 DNN（深度神经网络），旨在处理高速公路和城市交通场景。这些感知输出可用于自动驾驶和地图构建，为日常驾驶提供私人司机。

丰富的感知输出

DRIVE Chauffeur 采用 NVIDIA DRIVE Perception 构建，**旨在检测和分类目标、可行驶空间、车道和道路标记，以及交通灯和标志**。它可以估计与每个检测到的目标的距离，并融合来自多个异构传感器模态的输入。它与 NVIDIA DRIVE DNN 结合，形成用于自动驾驶的端到端感知管线。

多样化神经网络

DRIVE Chauffeur 采用 NVIDIA DRIVE Perception 构建，**旨在检测和分类目标、可行驶空间、车道和道路标记，以及交通灯和标志**。它可以估计与每个检测到的目标的距离，并融合来自多个异构传感器模态的输入。它与 NVIDIA DRIVE DNN 结合，形成用于自动驾驶的端到端感知管线。

日常通勤实现舒适和便利

DRIVE Chauffeur 直接与 DRIVE Concierge 配合，**旨在消除日常驾驶的压力和麻烦**。DRIVE Concierge 支持时刻享受智能服务，使用 NVIDIA DRIVE IX 和 NVIDIA Omniverse ACE 实现实时对话式 AI。该服务与 DRIVE Chauffeur 紧密集成，**提供 3D 车内可视化、代客泊车和智能召唤功能**。

解决方案-自动驾驶汽车

Concierge

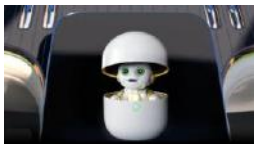
为每位乘客提供智能体验

NVIDIA DRIVE® Concierge 平台的应用延伸至了驾驶舱之外的领域，可提供**信息娱乐和游戏服务**，并充当每位乘客的数字助手。它建立在核心的 NVIDIA 技术之上，包括 DRIVE 平台、DRIVE IX 智能体验软件和 Omniverse™ ACE (Avatar Cloud Engine)，从而提供真正独特的用户体验。**利用对话式 AI、自然语言理解以及推荐引擎技术**，DRIVE Concierge 可满足每位驾驶员和乘客的需求，并与 DRIVE Chauffeur 紧密配合，提升驾乘体验。

集中式计算平台

DRIVE Concierge 将在跨域的 NVIDIA DRIVE 计算平台上运行。它能够在单芯片上虚拟化和托管多个虚拟机。借助这种集中式架构，DRIVE Concierge **可无缝编排驾驶员信息、驾驶舱和信息娱乐功能，并提供多区域舱内体验**。每位乘客均可通过控制单独的显示器和音频区域来定制车载体验，从而确保享受个性化体验。

AI 赋能的助手



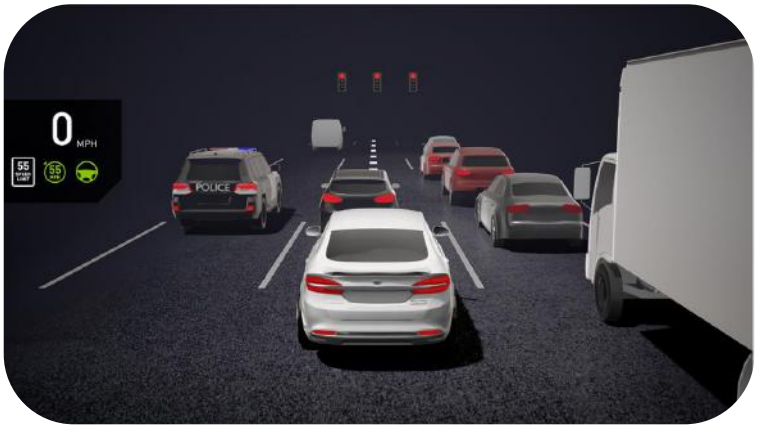
NVIDIA Omniverse ACE 是一个**基于云的AI模型和服务的集合**，供开发人员轻松构建、定制和部署交互式虚拟形象。该平台为创建 AI 助手打开了大门，**用户可通过语音 AI、计算机视觉、自然语言理解、推荐引擎以及仿真技术轻松定制 AI 助手。**借助 DRIVE Concierge，虚拟助手可提供建议、预订服务、调控车辆设置并发出提醒。

解决方案-自动驾驶汽车

Concierge

安心巡航

驾驶员和乘客可以借助精美的 3D 图形，随时了解车辆 AI 的想法。DRIVE Concierge 与 DRIVE Chauffeur 紧密集成，**可提供低延迟、高品质的 360 度 4D 可视化服务，因此驾驶员可以舒适地享受行程，并相信 AI 司机能带领自己安全抵达终点。**



全天候贴身泊车员

停车是一项复杂但又十分基础的任务。DRIVE Concierge 能显示检测到的交通标志和标记，以及边界和斜坡，从而执行垂直、侧方和倾斜停车（包括进出泊位）。**它甚至能帮助驾驶员寻找可用泊位，通过具有增强图形效果的 360 度 3D 环绕视图来指导自动停车系统。**DRIVE Concierge 还能够显示动态线条投影来指明路径，以及车辆与其他物体、相邻汽车、路缘和保险杠的距离，从而使其更好地与停车边界对齐。

解决方案-自动驾驶汽车

Concierge

随时随地，畅玩游戏

在充电时、等待时或在背座上体验高性能游戏。**受益于自动更新和无限的云存储，乘客可通过 NVIDIA® GeForce NOW™ 云游戏，在无需下载的情况下访问 1000 多款游戏。**

值得信赖的搭档

借助 DRIVE Concierge，驾驶员能够获得随时相伴的智能助手，畅享高级别便捷与安全体验。**DRIVE Concierge 可以监控驾驶员状态，以确保驾驶员将注意力集中在道路上，同时还可以监控乘客，以确保乘客的安全，以及保障没有贵重物品遗落在车厢内。**所有这些功能都可在 DRIVE 平台上运行，确保低延迟和高安全性的解决方案。

解决方案-自动驾驶汽车

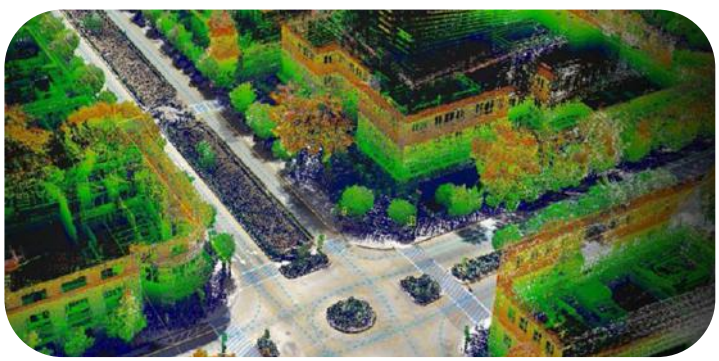
高精地图

适用于自动驾驶汽车的 DRIVE Map

NVIDIA DRIVE® Map 是一个**多模态地图构建平台**，旨在实现高级别的自主驾驶并同时提高安全性。它兼具真实数据建图的准确性，以及基于 AI 的车队源建图的及时性和规模。**DRIVE Map 具有四个定位层（分别是摄像头、激光雷达、雷达和 GNSS），可提供更先进的 AI 驱动所需的冗余和各种功能。**

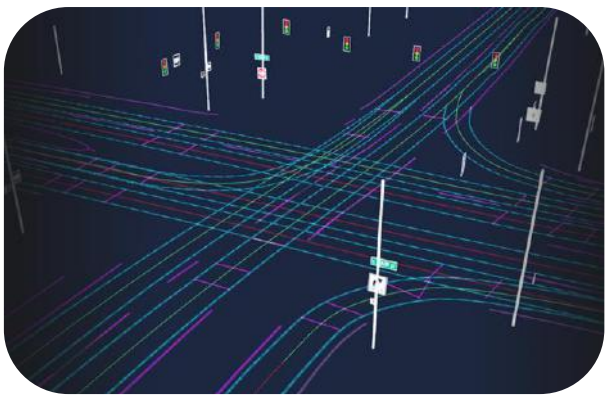
真值地图数据

DRIVE Map 真值地图引擎专为实现更高的准确性而设计，**配备 NVIDIA DRIVE Hyperion 数据采集车辆，使用摄像头、雷达、激光雷达和差分 GNSS/IMU 等丰富多样的传感器创建 DRIVE Map。**在高速公路和城市环境等选定环境中，它能够实现 5 厘米的精确度，从而达成更高级别的自主性 (L3/L4)。



车队源地图数据

DRIVE Map 专为实现近乎实时的操作和全球可扩展性而设计。它基于真实数据和车队源数据，代表了数百万车辆的集体记忆。DRIVE Map 利用 DRIVE Hyperion 传感器套件和其他合作伙伴传感器组中的数据流，包含所有必需的特征和语义信息（包括动态和行为信息），可提供安全舒适的驾驶体验。



解决方案-自动驾驶汽车

高精地图

全球覆盖

DRIVE Map 旨在为全球各地的辅助型汽车和自动型汽车提供支持。**NVIDIA 正在创建北美、欧洲和亚洲主要高速公路（总里程超过 50 万公里）的高精度地图，此大型高精度地图将由数百万辆客车不断扩展和更新。**

多功能数字孪生

DRIVE Map 对于自动驾驶开发和驾驶功能（例如感知、定位、预测、路线规划和控制等）至关重要。地图数据所示的细节和密度有助于各种应用模拟和重现特定位置的数字孪生。**Omniverse 上的 NVIDIA DRIVE Sim 能够进一步增强这些地图生成的数字孪生，适用于自动驾驶模拟、远程操作和车队管理模拟等不同类型的应用。**

解决方案-自动驾驶汽车

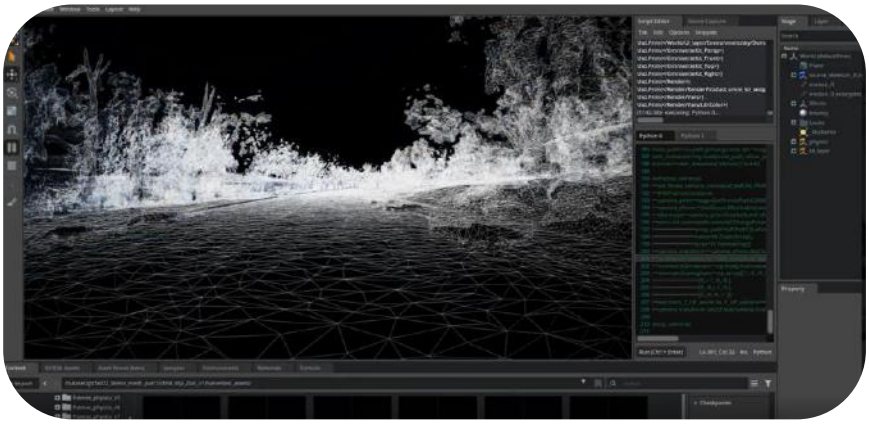
仿真

由 Omniverse 提供动力的 NVIDIA DRIVE Sim

自动驾驶汽车是我们这个时代最复杂的工程挑战之一，需要拥有由出色的开发团队带来的出色的工具。在现实世界中我们无法安全和彻底地对自动驾驶汽车进行测试，面对这一痛点，**NVIDIA DRIVE Sim™ 构建了一个物理精准的仿真平台，能够快速、高效地进行大规模的自动驾驶汽车测试和验证。**

神经重建引擎

NRE 是一套 AI 工具，可以将真实世界的数据直接带入仿真中，大大增加真实感并加快生产速度。**NRE 可将驾驶过程中收集的视频数据转换为交互式 3D 测试环境，**开发者可在此环境中修改场景、添加合成对象，并应用随机化技术，使初始场景更具挑战性。



用于闭环仿真的 DRIVE Sim

NVIDIA DRIVE Sim 通过将使用物理精准的仿真与高保真 3D 环境相结合，为自动驾驶汽车开发创建出虚拟的试验场。闭环测试可在单个软件组件或整个自动驾驶堆栈上完成。

解决方案-自动驾驶汽车

仿真

用于合成数据生成的 DRIVE Replicator

借助 NVIDIA DRIVE Replicator™，**开发者可以为罕见和复杂场景创建多样化的合成数据集**，包括基于物理性质的传感器数据和像素准确的真值标签。这些标签包括深度、速度、遮挡和其他难以标记的参数。

模拟用户体验

借助 NVIDIA DRIVE Sim，汽车制造商可以完全在虚拟世界中设计车辆，从而简化漫长的传统流程。工程师和产品经理还可使用它验证车辆设计，以确保它们符合当地安全标准。另外，汽车的潜在买家也可以从中受益，他们将能够坐在家中舒适地配置和体验汽车。

兼具开放性与扩展性

DRIVE Sim 是一个开放式模组化可扩展平台，可让用户根据自己的需求自定义仿真器。用户可以使用随附的 SDK，为传感器模型、车辆动力学、交通模型或自定义硬件的界面轻松构建扩展程序。DRIVE Sim 还拥有丰富的合作伙伴生态系统，这些合作伙伴可以提供兼容扩展程序。

DRIVE Sim 生态系统合作伙伴

MathWorks

借助 MathWorks RoadRunner 快速将环境导入 DRIVE Sim。

CARLA

利用兼容的仿真就绪内容的共享生态系统。

解决方案-自动驾驶汽车

仿真

AEVA

在 DRIVE Sim 中使用 Aeva 的传感器模型在仿真中访问 FMCW 激光雷达技术。

arbe

在 DRIVE Sim 中访问 Arbe 的感知成像雷达模型。

CEPTON

将 DRIVE Sim 与 Cepton 的高保真传感器模型结合使用。

Hum3D

从 Hum3D 市场将资产直接引入 DRIVE Sim。

dSPACE

使用 dSPACE HIL 和 DRIVE Sim 模拟整个车辆。

LUMINAR

在 DRIVE Sim 中使用 Luminar 的传感器模型实现细致的激光雷达仿真。

解决方案-自动驾驶汽车

自驾出租车

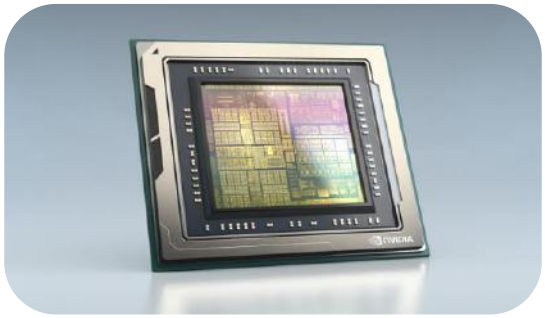
自驾出租车是带有轮子、可以自主驾驶的复杂超级计算机，需要高性能的计算能力以及独特的端到端程序才能开发、推出并持续改善。NVIDIA DRIVE® 提供适用于自驾出租车开发的全方位人工智能运算解决方案，确保全自驾车辆能处理大量数据，并执行冗余且多元的深度神经网络，以安全无虞地运作。

智能基础架构

数据中心开发对于大规模部署自驾出租车而言十分关键。若要能在全球数千种环境下运行，便需要使用大量数据进行密集的深度神经网络训练。 NVIDIA DRIVE 基础架构采用 GPU 技术的整合开放式高效能运算能力，是大规模开发、训练、测试和验证自驾 DNN 的唯一解决方案。

全天候进行人工智能计算

开发自驾出租车所需的高性能计算能力和车辆本身一样重要。 NVIDIA DRIVE AGX 人工智能计算平台能够同时执行所有需要的冗余且多元 DNN，进而取代人类驾驶。DRIVE AGX 系列以单一的可扩充架构打造，每秒最高可进行 2,000 兆次运算，确保能在无人驾驶的情况下全天候运作。



解决方案-自动驾驶汽车

自驾出租车

持续提升

自驾车公司可结合数据中心和车辆内解决方案，建立持续不断的端到端开发周期。当 DNN 在数据中心学习新功能时，经过验证的算法能以无线方式传送到车辆的运算平台，让车辆永远采用最新且最棒的技术。这种持续不断的开发周期能为驾驶提供令人惊艳的新体验，并转型为全新的经营模式。

丰富的自驾出租车生态系统

autoX　cruise　DiDi　oxbotica

pony.ai　WeRide 文远知行　ZOOX

更多

货车运输业

自驾卡车

卡车和物流在推动世界进步方面扮演的角色越来越重要。事实上，美国境内的卡车货运量占了所有货运量的 70% 以上，而随着电子商务崛起，高效率货运的需求也持续加速增长。 NVIDIA DRIVE® 提供唯一专为改善全天候作业的严苛环境而设计的可扩充平台，可为自驾卡车做好长途运输的准备。

◆ Level 2+ 驾驶辅助

即便有真人驾驶，人工智能也可利用 Level 2+ 的驾驶辅助来提升安全性与效率。有了 NVIDIA DRIVE，制造商可以整合环境侦测、环绕视觉、主动式安全功能和人工智能辅助驾驶功能，减轻驾驶长途卡车的疲劳。此系统包含驾驶员监控功能，可确保驾驶员将注意力保持在前方路况上。

◆ 转运中心到转运中心

一般而言，大多数的长途运输路线会在主要货运站或转运中心之间行驶。可扩充的 NVIDIA DRIVE 平台让这些经常行驶的路线能够执行 Level 4 自驾或无人驾驶作业，提高日常营运的安全性与效率。

解决方案-自动驾驶汽车

货车运输业

◆ 设施到设施

在装卸区内运输货物时，车辆通常需要在具有地理围栏的区域和产业道路上行驶。NVIDIA DRIVE 提供设施到设施的自动驾驶功能，可开发新型商用车辆，包括无人驾驶且无驾驶室的卡车。

◆ 长途运输的生态系统

einride　FAW　LOCOMATION

NAVISTAR　Plus　tusimple

ADAS

高级驾驶员辅助系统 (ADAS)

NVIDIA DRIVE® 平台是适用于高度自动化的监督式驾驶的全方位解决方案。此平台包含主动式安全、自动驾驶、停车，以及人工智能座舱功能，可将自动驾驶等级从 Level 2+ 提升至最高等级。

已验证硬件

NVIDIA DRIVE 支持一整套 ADAS 功能。 它采用 NVIDIA DRIVE Hyperion™ 8，其中包括三个 NVIDIA Orin™ 系统级芯片 (SoC) — 两个用于主动安全、自动驾驶和高级停车应用，一个用于智能座舱功能 — 来自异构、高保真传感器模式的 360° 感知场景解释。 NVIDIA DRIVE ADAS 解决方案由安全设计、可扩展架构支持，以确保系统在广泛的操作设计域中的稳健性和高性能。

NVIDIA DRIVE 高级驾驶员辅助平台

NVIDIA DRIVE 软件堆栈是一个完整解决方案，有助于构建和部署先进的 ADAS 应用程序，包括从感知、定位、计划和地图构建到计划和控制、驾驶员监控和自然语言处理在内的各种领域。该解决方案包含 NVIDIA DRIVE OS 安全操作系统，以及用于实现全面中间件功能的 NVIDIA DriveWorks SDK。NVIDIA DRIVE AV 和 DRIVE IX 堆栈还提供了 DNN，能够实现感知、地图构建、规划，以及智能驾驶舱功能。

解决方案-自动驾驶汽车

ADAS

端到端开发

通过利用 GPU 和 AI 的强大功能，开发者可以全面训练 DNN 以实现 ADAS 感知、规划、控制等。DRIVE Constellation™ 和 DRIVE Sim™ 模拟平台提供的虚拟试验场几乎覆盖各种驾驶条件，力求在与车辆相同的硬件上测试并验证 DNN。DRIVE 基础设施与 NVIDIA DRIVE 平台相结合，创建了一个持续的开发周期以便不断改进。

行业-建筑、工程、施工与运营的未来

设计建筑、工程、施工与运营（AECO）

本部分我们进入行业的部分，继续探索人工智能会如何改造每个行业。

将 AECO 从概念转变为施工

建筑与基础设施的设计、施工和运营带来了复杂的挑战。为了克服这些挑战，世界各地的建筑、工程、施工和运营 (AECO) 公司使用 NVIDIA 技术来优化设计、减少危险和提高协作成效，即使在远程工作时也是如此。随着 AI、3D 图形虚拟化、虚拟现实 (VR) 和 NVIDIA Omniverse™ 等协作解决方案取得突破，许多公司正在改变工作流，以重塑我们这个世界的未来。

1 加快设计工作流

先进的技术（如沉浸式 VR、实时光线追踪和 3D 图形虚拟化）极大地改进了 BIM 模型的设计和可视化，同时提高了工作效率和创新水准

2 实现数据驱动的决策

借助 GPU 助力的数据科学和 AI，AECO 团队可以利用大量数据（包括设计、环境、模拟和结构）深入了解从交通流量到天气条件的各种情况

3 确保安全，更大限度地节省成本

借助 GPU 助力的数据科学和 AI，AECO 团队可以利用大量数据（包括设计、环境、模拟和结构）深入了解从交通流量到天气条件的各种情况

利用 NVIDIA 技术优化 AECO 工作流

设计 | 构建 | 操作

实时建筑可视化

世界各地的 AECO 公司在合作开展建筑和基础设施设计时，依靠 NVIDIA 先进的视觉计算技术来加快工作流程。建筑师、工程师和设计师使用 NVIDIA RTX™ 助力的工作站支持实时光线追踪、虚拟现实、工程模拟和采用 AI 技术的应用。NVIDIA RTX™ 虚拟工作站 (vWS) 软件为远程处理大型复杂 BIM 模型的设计团队提供桌面级图形性能。利用面向 AEC 的 NVIDIA Omniverse™（一个图形和模拟平台，允许围绕着逼真的数字孪生开展实时协作），项目团队可以改变概念设计流程。

行业-建筑、工程、施工与运营的未来

设计建筑、工程、施工与运营（AECO）

利用 NVIDIA 技术优化 AECO 工作流

设计 | 构建 | 操作

更高效的施工工作流

世界各地的 AECO 公司在合作开展建筑和基础设施设计时，依靠 NVIDIA 先进的视觉计算技术来加快工作流程。建筑师、工程师和设计师使用 NVIDIA RTX™ 助力的工作站支持实时光线追踪、虚拟现实、工程模拟和采用 AI 技术的应用。NVIDIA RTX™ 虚拟工作站 (vWS) 软件为远程处理大型复杂 BIM 模型的设计团队提供桌面级图形性能。利用面向 AEC 的 NVIDIA Omniverse™（一个图形和模拟平台，允许围绕着逼真的数字孪生开展实时协作），项目团队可以改变概念设计流程。

设计 | 构建 | 操作

智慧城市

AI 彻底改变了智能解决方案，能打造出可持续发展程度更高的城市、维护基础设施并改进面向居民和社区的公共服务。而这一切，都始于能够从数万亿个传感器和其他物联网 (IoT) 设备收集数据，并将复杂的视频数据转化为可行见解。NVIDIA Metropolis™ 平台提供先进的技术和范围广泛的开发者生态系统，帮助创建、部署和管理基于 AI 的视频分析应用。

行业-建筑、工程、施工与运营的未来

设计建筑、工程、施工与运营（AECO）

AECO 解决方案

Omniverse

借助面向 AECO 的 NVIDIA Omniverse，建筑师和设计师可以**更有效地迭代概念**，以更快的速度呈现惊艳的模型可视化效果。通过在搭载了 RTX 的机器（从笔记本电脑到数据中心的机器）上运行 Omniverse，可以**从根本上转变协作和沟通方式**。

工作站

NVIDIA RTX 助力的台式机和移动工作站可提供 AECO 项目团队所需的高性能，进而满足当今严苛工作流的超高要求。

虚拟 GPU

AECO 项目团队可以体验虚拟桌面基础设施 (VDI) 的所有优势（如移动性和安全性），同时能够在图形密集型软件工具中实现流畅处理。

行业-建筑、工程、施工与运营的未来

设计建筑、工程、施工与运营（AECO）

AECO 解决方案

GPU 加速的基于物理性质的实时渲染**可提供非常逼真且准确的 3D 模型**，以加快设计决策、提高创新能力并帮助顺利推进 AECO 项目。

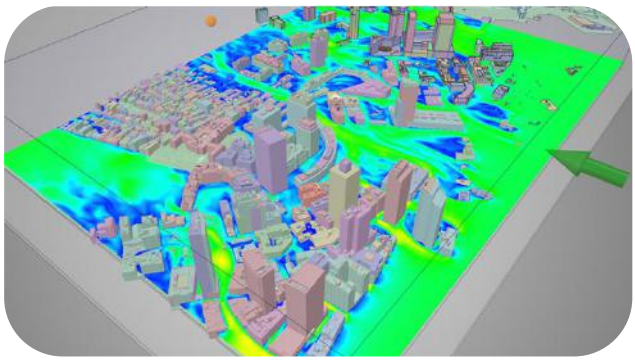
GPU 渲染

扩展现实

NVIDIA CloudXR 等技术创新正在强化高保真增强现实工作流，并使用户摆脱传统的连接束缚，实现虚拟现实体验。**沉浸式 VR 可改善整个建筑设计流程中的 3D 设计和可视化工作流，甚至可拓展到培训和营销**。

模拟

如今，AECO 团队在设计流程早期采用模拟技术。NVIDIA 企业用 GPU 提供强大助力和性能，可以加快非常复杂的模拟工作流。



行业-消费互联网

定义下一代消费互联网

流式传输，搜索，购买，送货

如今，消费者只需动动指尖，全球好物便近在眼前。**在线用户的迅猛增长使得消费互联网行业成为机器学习、深度学习和数据科学的庞大用户之一。从推荐程序到智能聊天机器人，再到增强型AI的视频会议，这些功能都依赖于快速，智能的工具和基础架构。**因此，领先的公司纷纷使用 NVIDIA 解决方案来充分利用其大量的数据、确定发展趋势，并率先开发塑造在线体验的新产品。

简化AI模型

为应对不同 AI 模型的激增及随之而来的网络复杂性，NVIDIA 的基于 GPU 的 SDK 汇集了深度学习和机器学习建模、训练和推理

处理海量数据集

数据集不仅在高速增加，它们还会以不同的质量演变为不同的格式。这使得 GPU 加速数据科学在处理现代运算工作负载方面显得至关重要

提高部署速度

调整基础设施监控、管理和软件部署可能颇具挑战性。借助 NVIDIA GPU，工程师可以构建用于支持数据的技术基础设施和管线

适用于定制消费应用程序的解决方案

◆ NVIDIA Riva 的会话式 AI

NVIDIA Riva 是一种 GPU 加速应用程序框架，可用于构建提供实时性能的多模态会话式 AI 服务。**Riva 包含经过预训练的会话式 AI 模型、NVIDIA AI 工具包中的工具，以及经过优化的端到端服务（如邮件应用程序、基于语音的助理和聊天机器人），可进行自动通信并大规模创建个性化的客户体验。**它将视觉、听觉同其他感官功能相结合，可为呼叫中心助理和其他虚拟助理提供助力。

行业-消费互联网

定义下一代消费互联网

◆ NVIDIA Merlin 的推荐系统

NVIDIA Merlin 是一种用于构建基于深度学习的高性能推荐系统的框架。前面已经介绍过了哈~

◆ 采用 RAPIDS 的加速数据科学项目

在消费者行为和预测分析等领域中，数据科学有助于在更短时间内获得业务见解。**NVIDIA 加速数据科学基于 NVIDIA® CUDA-X AI™ 构建，配备 NVIDIA RAPIDS™ 数据处理和机器学习库，可为数据科学工作流程提供 GPU 加速软件，从而更大限度提高生产效率、性能和 ROI。**

◆ NVIDIA Maxine 的AI视频会议

NVIDIA Maxine™ AI平台SDK使视频会议提供商可以通过超分辨率，凝视校正，实时字幕等功能大大提高云中的流传输质量。 除了减少视频带宽外，Maxine的完全加速平台还包括创新功能，例如面部对齐，噪声消除和虚拟助手。

◆ NGC：经 GPU 优化的软件

NGC™ 是 NVIDIA 的 GPU 优化软件的中心，用于简化和加速端到端工作流程。**数据科学家、开发者和研究人员可以快速部署带有 BERT、推理和推荐系统容器的 AI 框架。**用户服务需要不断的创新，而 NGC 可以确保更快的 AI 实施，从而更快地提供解决方案。

定义下一代消费互联网

◆ 采用 NVIDIA DGX 的通用 AI 工作负载

在消费互联网行业，AI 辅助服务的增长受到复杂 AI 模型的激增、不断增加的数据集以及繁琐的部署和管理工作流程的阻碍。这些挑战导致计算架构运行缓慢且成本高昂。NVIDIA DGX™ A100 为 IT 总监、数据科学家和数据工程师提供了一个能够统一所有 AI 工作负载、简化基础设施并加速投资回报的平台。

◆ GPU 云计算

消费互联网服务可在全球和多个平台上使用。NVIDIA 的 GPU 加速解决方案可通过所有高级云平台提供，使公司能够根据需要轻松扩展和访问巨大的计算能力。**NVIDIA T4 Tensor Core GPU 能够加速各类云工作负载，包括高性能计算、深度学习训练与推理、机器学习、数据分析以及图形处理。**它使企业能够打造新的客户体验，以帮助提高服务的可访问性和可扩展性。

行业-网络安全

利用 AI 实现网络安全现代化

随着数字世界的不断扩展，网络威胁的数量和复杂程度也与日俱增。公司等各种组织正通过 AI 做出响应，在海量数据中发现威胁，并构建基于零信任、全方位安全架构的系统，从数据中心边界到每台服务器边缘皆安全无虞。

英伟达如何构建更强大、更智能的网络安全

简化AI

利用 NVIDIA AI 框架中的行业标准 API，快速定制和部署网络安全解决方案

加速AI性能

利用 GPU 加速的性能（比仅使用 CPU 的服务器快 600 倍），支持对整个网络中的每台服务器、每个数据包、每个用户和每台机器进行 AI 推理和实时监控

高性能网络

借助 NVIDIA® BlueField® DPU 的卸载、加速和隔离功能，实现网络、存储和安全服务。

放心部署

使用 **NVIDIA 认证系统**™，可以部署硬件解决方案，以安全、优化的方式运行现代加速工作负载

行业-网络安全👇

利用 AI 实现网络安全现代化

网络安全举例：

👉行为分析

企业需要保护极其庞大的数据网络。NVIDIA Morpheus **通过监控企业数据中心内的每个用户、服务、帐号和机器来启用数字指纹识别，以确定何时发生了可疑交互。**当在 NVIDIA 认证服务器中整合了 NVIDIA GPU、NVIDIA DPU 加速器和 NVIDIA DOCA 遥测时，为数据中心带来了更高的安全水平。

👉泄露敏感数据

检测泄露敏感数据的传统方法依赖于基于规则的静态模型，这些模型受限于训练数据的质量。相反，**NVIDIA Morpheus 会检查生成的原始数据包信息是否存在潜在泄露。**驻留在 NVIDIA BlueField-2 DPU 上的 **DOCA 遥测代理将原始数据包直接传送至 Morpheus。自然语言处理（NLP）模型可确定敏感信息（例如密码和私钥）是否在数据包中泄露。**数据包立即被标记，并将建议操作发送回 DOCA。这些实时警报会传递给管理员，以便可以立即对受损数据进行补救。

👉欺诈检测

检测欺诈通常需要大量已标记数据，并需要领域专家来处理和标记这些数据。这将此类欺诈检测限制在可以进行如此投资的组织中。通过使用图神经网络（GNN），Morpheus 解锁了以前在没有大量标记数据的情况下无法实现的功能，**即现在可以使用以前需要标记数据的一小部分来获得高精度结果。**这使企业能够以较低的成本实现欺诈检测，从而节省数亿美元。

行业-网络安全

利用 AI 实现网络安全现代化

网络安全举例：

网络钓鱼检测

网络钓鱼电子邮件仍然是勒索软件和恶意软件的三大初始感染媒介之一。**远程工作和教学的趋势加剧了这种情况，导致攻击面扩大。NVIDIA Morpheus 网络安全 AI 框架提供了一个 NLP 模型，可以分析电子邮件来识别网络钓鱼企图。**Morpheus 能够使用内置于网络日志加速器（CLX，Morpheus 的构建块之一）中的自定义序列分类器来分析电子邮件的整个原始正文和/或电子邮件中的 URL。

勒索软件

勒索软件是成本最高的网络安全漏洞类型之一。随着威胁行为者的攻击变得更加复杂，全球范围内的攻击事件有所增加，如果能带来经济回报，攻击可能会变得更加频繁。**NVIDIA Morpheus 为现代企业数据中心提供基于 AI 和机器学习的优化网络安全管道，并结合作为高速、高保真数据源的 NVIDIA BlueField 和在入口点识别、阻止勒索软件的 NVIDIA AppShield。**

利用 AI 实现网络安全现代化

加速向可持续的未来转变

先进的能源公司使用 NVIDIA 技术来变革行业并提高全球人民的生活品质。该公司通过开发可再生能源，构建更智能、更具弹性的电网运营，加快能源勘探和生产，以及确保员工和社区的安全条件，帮助我们迈向更光明、更可持续的未来

从数据中提取价值

利用 NVIDIA AI 工具，将常规**上游业务、管道和炼油厂传感器以及维护流程中的大量数据**转变为实际可行的深入见解

强力支持计算

无论是在数据中心本地还是在云中，**都可借助高性能计算加速地球物理和工程应用程序**

保护健康和环境

确保遵守适当的个人防护装备 (PPE) 协议，**并使用 AI 技术观察设备、预测和检测故障，从而识别安全隐患，拯救生命**

简化行业细分流程

上游 | 中游 | 下游 | 电网

勘探和生产

地震数据处理

无论是使用逆时偏移 (RTM)、Kirchhoff，还是全波形反演 (FWI) 算法，**与仅采用 CPU 相比，基于 NVIDIA GPU 的地震波处理技术可将地震数据处理速度提升高达 5 倍，从而缩短石油开采时间。**借助 NVIDIA GPU，处理地震勘测的地质学家可以在最复杂的数据集上应用高级过滤器并解释结果。

行业-能源

利用 AI 实现网络安全现代化

上游 | 中游 | 下游 | 电网

勘探和生产

地球科学可视化

无论是使用本地工作站还是虚拟桌面，**NVIDIA 专业解决方案均能够提高可视化和重度计算的吞吐量。**高性能计算（HPC）和 AI 可在解释器桌面改善三维地震道属性的计算和区域盆地的可视化分析。

油藏模拟

借助先进的建模和模拟技术，更大限度地提高油藏分析性能。在CUDA® 软件上运行的 NVIDIA GPU 能够加快并减少模型处理周期，使研究人员能够以最少的时间获得最大的价值。

健康、安全和环境

保护员工、承包商和环境是当今能源公司极其重要的工作。借助 NVIDIA Metropolis，**企业可以通过将深度学习应用于视频流以实现员工安全、交通管理和资源优化等应用，从而使井场更安全、更智能。**

上游 | 中游 | 下游 | 电网

使用深度学习和机器学习算法，石油和天然气公司可以确定在条件变化时优化其运营的最佳方式。

管道优化

利用大数据和 AI 系统优化管道填充方式，检测腐蚀以识别潜在泄漏，并自动化超声波流量计以提高吞吐量。这些技术还可用于监控运输位置，验证其安全性。包括优化商品交易的需求预测以及优化运输和管道容量的其他领域。

行业-能源

利用 AI 实现网络安全现代化

上游 中游 下游 电网

预测性维护

实时识别机器中的差异并预测设备的剩余使用寿命，避免断电、停机和不必要的维护成本。借助 NVIDIA® DGX™ A100等基于 GPU 的深度学习服务器，**油井运营商可以可视化和分析大量生产和传感器数据，例如泵压、流速和温度。**借助 NVIDIA Omniverse™ 和 NVIDIA Modulus，全球能源巨头西门子能源公司正在构建数字孪生，每年可帮助在热回收蒸汽发生器的预测性维护方面节省 17 亿美元。其他重点领域包括容量优化、经济预测和站点监控。

性能优化

通过确定性能欠佳、虚拟测试运营或资产变更的根本原因，并更大限度地减小所提议变更的意外风险，提高提炼资产的可靠性和性能。

上游 中游 下游 电网

发电和配电：**借助可以预测需求、发电和管理能源资源的智能弹性电网**，向更具可持续性的未来加速前行。

碳氢化合物

利用地震处理技术勘探新的能源矿藏，发现可用的碳氢化合物储量。借助 NVIDIA GPU 以更少的时间构建准确的地下模型。

可再生能源

借助 **AI 优化可再生能源生产**并降低运营成本，例如**风力发电机的检测和维护。**

行业-能源

利用 AI 实现网络安全现代化

上游 **中游** **下游** **电网**

电力零售

通过在 NVIDIA GPU 上运行的机器学习模型中**使用高级计算基础设施数据**（AMI）预测未来的负载需求。

电网

提高基于 NVIDIA GPU 运行的**复杂电网运算优化和建模技术**的性能。

行业-金融服务

实现更智能、更安全的金融服务

利用 AI 加速金融服务

大型数据集、永久性市场波动、远程办公。智能技术可以攻克现代金融服务业所面临的关键挑战。借助 NVIDIA 的 AI 技术（包括深度学习、机器学习和自然语言处理 [NLP]），金融机构可以加强风险管理、改善数据支持的决策和安全性，并提升客户体验。

金融行业如何使用AI——银行

AI 驱动的银行

AI 能够改变银行的运营方式。以前所未有的速度和规模帮助银行改善业务和功能的各个方面，包括**在海量数据中识别关键洞察、计算风险，以及自动化日常任务**。在此助力下，AI 驱动的银行能够提高工作效率、扩大服务范围、减少风险，并大大改善客户服务。

欺诈检测

一些极为成功的 AI 案例与打击交易欺诈有关，此类问题可造成数十亿美元的损失。检测出真正的欺诈对于实现这一目的至关重要，但从历史数据来看，传统系统所产生的误报信号远远多于真正的欺诈信号。现今，**在精尖机器学习和深度学习技术的助力下（例如 NVIDIA Triton™ 推理服务器），欺诈检测能力不仅得到了提高，误报率也随之大幅下降。**AI 正在彻底变革像金融服务业这样价值数万亿美元的产业，同时为世界各国的发展提供助力。许多公司都在利用 AI 改善客户成果、降低成本以及打击欺诈，其中就包括 PayPal、美国运通和中国平安。

实现更智能、更安全的金融服务

金融行业如何使用AI——银行

改进客户服务

对话式 AI 使消费者能够管理所有类型的金融交易，包括账单支付、汇款以及开设新账户。通过提供这些自助服务互动方式，银行可节省客户服务人员的时间，使其专注于具有更高价值的互动和交易。**对话式 AI 的核心是深度学习模型，这些模型需要强大的计算能力，以便训练聊天机器人使用特定领域的术语和金融服务语言进行交流。**这些模型经过训练后，机器人需要能够与客户实时开展逼真对话。这种低延迟性能，以及训练深度学习模型所需的计算能力，均可由 NVIDIA GPU 实现。

个性化银行优惠

在某些大型商业平台上，推荐服务所产生的收入占比高达 30%，而这等同于数十亿美元的销售额。**因此，银行和保险公司都在使用推荐系统推动客户采取各式行动，而这些行动就包括推荐客户访问某一网页，以及推荐客户优先考虑偿还某项债务。**推荐系统还通过向消费者提供个性化消息，提高客户忠诚度和对银行的满意度来提升转化率。NVIDIA Merlin™ 是一种基于 GPU 的端到端推荐系统框架，可提供快速的特征工程和较高的训练吞吐量，进而为深度学习推荐模型的快速实验和生产再训练提供支持。Merlin 也可实现低延迟、高吞吐量，可用于生产的推理，从而提供快速准确的个性化客户互动方式。

行业-金融服务

实现更智能、更安全的金融服务

金融行业如何使用AI——保险

保险公司摒弃了传统的索赔管理方法，转而采用积极的数字化转型，以及完全由分析驱动的方法。其中包括：**利用 AI 实现对简单案件的自动化索赔处理，通过实施 AI 赋能的服务加速复杂案件的处理**，以及通过打造新的数字化服务提高客户满意度。

金融行业如何使用AI——交易

加速交易计算

更快的处理速度能造就更明智的交易策略、更成功的交易完成和更高的收入。GPU 助力的硬件加速可加快获取见解的速度，让业务运营能够保持竞争优势。借助 NVIDIA 技术，金融机构能够利用 AI 和高性能计算 (HPC) 的强大功能，从大量数据中进行学习，并对市场波动做出快速响应。

金融行业如何使用AI——交易

用于数字支付的欺诈预防策略

向亲朋好友转账。在线支付账单。在零售商店用手机付款。在线支付、移动支付、店内支付、企业对消费者 (B2C) 和企业对企业 (B2B) 等支付方式推动着全球经济的发展。 **AI 可以帮助银行更好地检测和预防支付欺诈，并改进反洗钱 (AML) 和了解客户 (KYC) 系统的流程。**通过在支付领域应用 AI，使资金流动更加安全、透明，并为公司和客户带来更好的支付体验。

行业-金融服务

实现更智能、更安全的金融服务

金融行业如何使用AI——金融科技

现代金融科技领域中的可解释的 AI

金融科技正在推动全球创新，改变企业、消费者和资金在各行各业（从金融服务、零售到交通运输等行业）的交互方式。 **利用 AI，推荐引擎使金融科技交互更加个性化，对话式 AI 增强了自助服务，并且深度学习欺诈检测模型的使用提高了交易的安全性。**而这仅仅是开始。

成功案例

防止欺诈并挫败网络犯罪

美国运通已通过采用此**增强型实时欺诈检测系统**提高准确性，并更好地保护客户和商家。该系统在 2 毫秒的紧迫延迟要求下运行，与基于 CPU 的配置（无法实现目标）相比，其性能提升 50 倍。通过结合 GPU 加速的 LSTM 深度神经网络模型和他们长期应用的梯度提升机模型（GBM），最高可将特定细分市场上的欺诈检测准确率提高 6%。

对话式智能提高盈利能力并带来其他优势

Talkmap 是一款革命性的对话式智能和自动化平台，**它使用具有对话理解能力的 AI 驱动的机器学习，从与客户和潜在客户的对话中获取见解，**使手动、缓慢且昂贵的工作实现自动化。过去需要 8 到 24 个月才能完成的工作现在只需几个小时即可完成，而且其结果更为准确、可靠。

行业-金融服务

实现更智能、更安全的金融服务

部署渐进式 AI 以简化复杂的业务挑战（Applica）

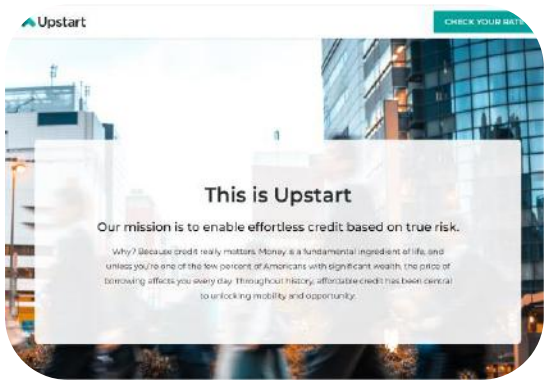
“ V100 GPU 使我们能够运行独有的 2D 布局感知语言模型，不管在我们的机器人文本自动化平台上使用哪种语言或布局，都能轻松处理文档。NVIDIA GPU 的速度比 CPU 快 50 倍，这使我们能够真正实现智能的文本自动化。结合 Applica 平台的易用性，用户无需具备数据科学或深度学习背景，即可轻松利用该技术。这样一来，我们就能携手为所有人提供 AI。”

高效的 AI 客服（Voca.ai）助力改善客户体验

Voca.ai 使用 NVIDIA GPU 来运行基于 AI 的 Voca 客服。这些客服能够解读对话，从而引入人工客服，以便其处理高价值和复杂的对话。在 NVIDIA 的支持下，Voca 的 AI 技术提高了与潜在客户（呼入和呼出）立即互动的能力。基于 AI 的 Voca 客服可以轻松处理通话量峰值，并且不会增加成本。

基于真实风险，轻松获取贷款

利用 Upstart 的 AI 平台，由 Upstart 提供支持的银行通常获得更高的批准率和更低的损失率，并能提供其客户所需的出色“数字优先”贷款体验。超过三分之二的 Upstart贷款能够在完全自动化的条件下实现即时批准。



行业-医疗健康和生命科学

借助 AI 推动医学研究发展

医疗健康行业要求实施新的计算模式，以满足个性化医疗、新一代诊所、更高护理品质的需求，另外还要求生物医学研究取得突破，以治疗疾病。借助 NVIDIA，医疗健康机构可以利用人工智能和高性能计算 (HPC) 的力量来定义医学的未来。

通过加速计算为医疗健康解决方案提供动力支持

药物研发

借助加速计算，研究人员能够以虚拟方式同时建立数百万个分子的模型，筛查数百种潜在药物，从而降低成本并缩短解决问题的时间

基因组学

通过使用HPC加速人口研究和癌症基因组学研究中的基因组分析，可以帮助识别罕见疾病并更快地将定制地疗法推向市场，从而进一步实现精准医疗目标

医疗影像

AI驱动的工具可以成为一双额外的“眼睛”，帮助临床医生快速阅读图像、计算测量结果、监控变化和确定紧急检查结果，以优化工作流程和改善患者护理

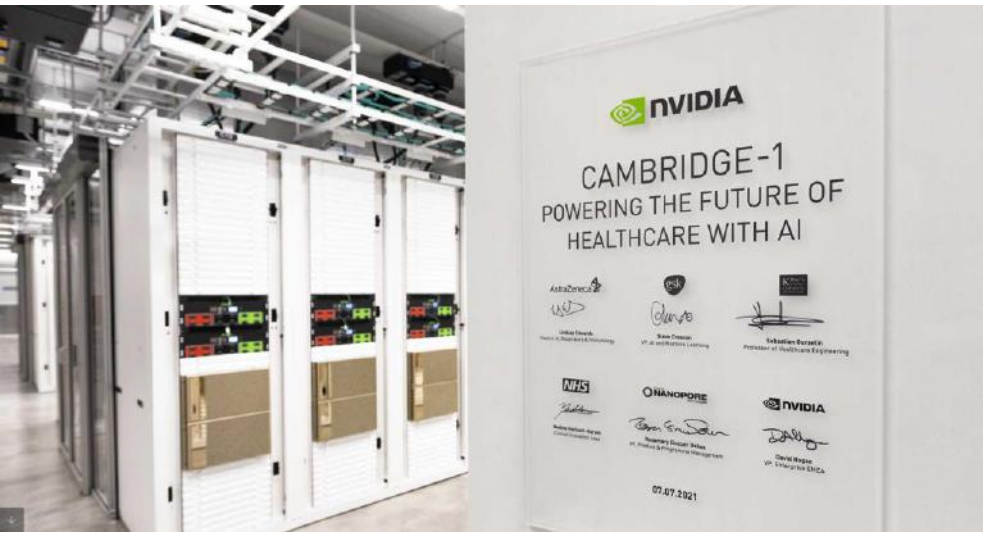
智慧医院和医疗仪器

从智能传感器到支持实时高级图像处理的医疗仪器，边缘AI可以提供即时见解、优化患者护理和实现智慧医院的承诺

行业-医疗健康和生命科学

借助 AI 推动医学研究发展

Cambridge-1



Cambridge-1是英伟达为推动英国的数字生物学、基因组学、量子计算和 AI 研究而打造的第一款超级计算机，这款超级计算机将 AI 和仿真相结合，使顶尖科学家和医疗健康专家能够利用其强大功能来加速数字生物学革命，并推动该国领先的生命科学行业的发展。

使用案例

- AstraZeneca 正通过 Cambridge-1 **训练模型学习化学语法**，并处理全切片数字病理学图像，推动现代机器学习 (ML) 和 AI 发展，从而加速药物研发。
- AI 对于发挥人类遗传学的价值至关重要。借助 Cambridge-1 的计算资源，GSK 正使用先进的 ML **来利用大型数据库中的遗传案例**，并研发变革性药物。
- 伦敦国王学院正在构建能够合成人脑的 3D MRI 图像的深度学习模型。Cambridge-1 能够加速生成这些合成数据，以便研究人员了解不同因素影响大脑、解剖学和病理学的方式。
- Peptone 旨在改变使用无监督式学习和强化学习技术设计蛋白质药物的方式。该公司计划使用 Cambridge-1 **设计有助于治疗多种炎症性疾病的抗体**。

行业-医疗健康和生命科学

借助 AI 推动医学研究发展

医疗行业解决方案

为数据中心带来强大计算能力

NVIDIA DGX™ A100 由 NVIDIA A100 Tensor Core GPU 构建，可实现出色的计算性能，助力新药研发、揭示基因突变以更好地对抗疾病，并推动医疗健康创新。**NVIDIA DGX A100 集专家支持和易于部署的基础架构于一身，是用于构建 AI 数据中心的基础模组。**

实时边缘 AI

NVIDIA 的边缘解决方案旨在收集并计算网络边缘的连续数据流。借助先进的图像、视频和信号处理技术，内嵌 AI 的医疗仪器可以帮助外科医生进行创伤性更低和更具针对性的手术，帮助放射科医生确定诊断结果，以及帮助超声检查医生进行快速准确的超声心动图检查。利用 NVIDIA EGX 平台，引领医疗健康走向前沿。

AI 就绪型医院数据中心

NVIDIA 和 VMware 的 AI 企业就绪平台推出了一款适用于当今解决方案融合的全堆栈架构。它专为实现弹性扩展而构建，为核心医院应用程序和 AI 工具提供单一平台，帮助临床医生及其服务的患者获取更好的体验。

按需云计算

新款 NVIDIA GPU 在全球所有主要的云平台上都可用。通过简化的 IT 管理、可以按需提高或降低的计算能力，以及 NGC™（针对 GPU 优化的深度学习、机器学习和 HPC 软件中心）的访问权，组织可以专注于构建解决方案、收集见解并实现业务价值。

行业-教育

高等教育和研究的计算新时代

建立由 AI 提供支持的大学

大学培养学生，为应对不断变化的世界做好准备。人工智能和数据科学等新兴技术即将成为高等教育和科研机构的核心结构，以及实现科学突破的关键。先进的大学正在利用 NVIDIA GPU 加速的超级计算机、实验室和计划，为学生、教师和研究人员提供改变世界所需的工具。

探索佛罗里达大学的AI之旅

人工智能领导力的基本组成部分

解决社会最大问题的力量

超级计算机对于人工智能的发展至关重要，佛罗里达大学(UF)正在基于英伟达的前沿技术构建学术界最强大的人工智能系统。 无论是寻找模拟气候变化的新方法还是加速药物发现，研究人员都将能够比以往更快地解决最复杂的挑战

为社区创建独特解决方案的数据

每个社区都有独特的挑战和可用于创建强大的 AI 解决方案的数据。 UF 计划探索从农业到医疗保健的各种问题，寻找可以改善佛罗里达人和所有美国人生活的答案。

行业-教育

高等教育和研究的计算新时代

对顶尖人才的投资

超级计算机和新颖的挑战更容易吸引研究人员和教育工作者。UF 正在招聘 100 名专注于 AI 的教师，此外还有 500 名将 AI 整合到他们的教学和研究中。NVIDIA 将提供他们世界一流的 AI 培训。

未来熟练使用培训

与人工智能相关的技能和素养变得越来越重要。 UF 的新课程是在 NVIDIA 的 AI 专家的帮助下创建的，将把它们传授给整个大学的专业——从商业到音乐再到艺术。

与政府、学术界和工业界的合作伙伴关系

UF 与 NVIDIA 以及地方、州和联邦官员密切合作，以创建他们最先进的系统。 他们的工作将与其他学术和研究机构保持一致，推动人工智能发展并建设国内经济。

行业-教育

高等教育和研究的计算新时代

高等教育机构如何使用 AI

生命科学

借助 AI 推进医学和研究，斯坦福大学的研究人员与 NVIDIA、牛津纳米孔技术公司、谷歌、贝勒医学院和加州大学圣克鲁兹分校的合作者共同创造了一项吉尼斯世界纪录，以实现最快的 DNA 测序。

物理

向 AI 教授物理，NVIDIA Modulus 是一种神经网络框架，它将偏微分方程 (PDE) 形式的物理力量与数据相结合，以构建具有近乎实时延迟的高保真参数化代理模型。 无论是想着手解决 AI 驱动的物理问题，还是想为复杂的非线性、多物理系统设计数字孪生模型，NVIDIA Modulus 都提供支持。

气候

利用 AI 和高性能计算 (HPC) 创造变革，NVIDIA AI 正在改变我们识别、评估和根据气候科学知识采取行动的方式。 地球数字孪生、超级计算机模拟和 GPU 加速天气预报的创新为气候变化及其影响打开了全新的视角。 我们现在能够实时探索人类行为和气候变化缓解策略的后果。

网络安全

借助 AI 实现网络安全现代化，日益复杂的网络安全攻击的威胁越来越大，导致美国政府采取零信任方法。 借助 NVIDIA Morpheus，开发人员可以利用 AI 框架和工具以更少的代码行构建高性能、零信任的网络安全解决方案。

机器人与仿真

加速机器人解决方案的开发用于机器人技术的端到端 NVIDIA Isaac™ 平台可以通过增强的开发、模拟和部署加速开发过程。

数据科学

为高性能数据科学提供动力，通过介绍 NVIDIA RAPIDS™ 生态系统的八个不同教程和备忘单，可以更好地了解如何大幅加速 Python 数据科学工作流程。

行业-餐厅

提升快捷餐厅的服务速度和效率

借助 AI 在客户下单量方面保持领先地位

AI 技术能够帮助快捷餐厅 (QSR) 解决劳动力短缺问题，并助力其实现更敏捷高效的运作。借助计算机视觉和语音识别功能，餐厅运营者可以通过支持 AI 的食品亭提供自动点餐服务，预测订单准备时间，优化人员配置，并加快送餐上门的速度，从而提高客户满意度和品牌忠诚度。

语音食品亭

Violet 是一个由 AI 驱动的客户服务助理，随时准备接受用户的订单。 这个 NVIDIA Tokkio 参考应用程序利用 NVIDIA Omniverse™ Avatar Cloud Engine (ACE) 创建交互式化身，可以看到、感知、智能交谈并提供建议，以增强餐厅等场所的客户服务。

顾客订单分析

利用 NVIDIA Metropolis 在商店和餐厅中部署智能视频分析 (IVA)，提供丰富的数据和见解，帮助管理者提高效率，增强顾客体验，并且更快地做出重要决策。 先进的 QSR 正在使用 IVA 来改进队列和候时管理、路边提货服务、生产质量和食品安全、公共卫生管理、外部和内部安全，以及消费者参与。

预测和库存优化

NVIDIA 认证系统™ 和 NVIDIA RAPIDS™ 数据科学库可使科学家减少提取、转换和加载 (ETL) 操作，加快模型训练速度，并改善预测准确性。QSR 能够更频繁地运行预测，并可将预测准确性提高 20%，以确保其拥有满足顾客订单需求的正确产品。

行业-零售业

开创智慧零售新时代

改进预测，增加收入，减少紧缩

领先的零售商正利用 AI 来减少损耗、改善预测、实现仓库物流自动化、决定店内促销活动和实时定价、为客户提供个性化服务和推荐，以及在实体店和网店提供更出色的购物体验。

智能商店

通过使用来自摄像头和传感器的数据，零售商正利用 AI 来减少损耗、消除缺货问题，以及清楚查看客户行为。同样的基础设施也能加快结账速度。**零售商使用 AI 创建智能商店的多种方式：**商店仿真、资产保护、自主购物、缺货和库存管理、商店分析、商店运营、个性化促销和增强现实体验。

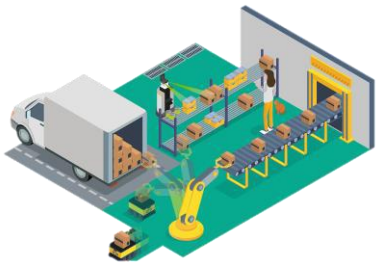
需求预测

AI 还能改进需求预测和库存管理。需求预测使用来自各种来源的数据，以确保在合适的时间和合适的商店发售合适的产品。使用机器学习提高预测准确度能对供应链优化产生重大影响。有效的预测不仅仅要考虑人口统计数据和地点。许多外部事件（例如天气或当地体育赛事）也可能会影响供给和需求。通过基于 NVIDIA GPU 运行 NVIDIA RAPIDS™ 软件库，零售商可以将机器学习算法的训练速度提升高达 20 倍。这意味着他们可以使用更多的数据，并以更高的准确度更快地处理数据。

行业-零售业

开创智慧零售新时代

仓库物流



AI 在仓储物流行业的应用

仓储物流是优化、集成、自动化和管理运营或配送中心的产品流的艺术。结合适用于智能视频分析 (IVA)、机器人技术、自动化和供应链管理的 NVIDIA AI 解决方案，实现更高的运营效率和流程吞吐量。负责处理订单的仓库机器人能够在做出决策之前评估所有变量并适应变化的情况。通过自动报告，它们可以优化路线，提供端到端的可见性，并提高订单拣选、包装和投递的准确性。

推荐系统和搜索

推荐和视觉搜索

了解顾客行为对于希望推动业务发展的零售商而言变得更为重要。通过视频分析技术助力的 AI 应用，可以使零售商像一直在线一样，了解店内的顾客行为。通过深入了解客流量大的通道、停留时间和人口统计数据，零售商可以改善推销效果和在店内举行实时促销活动，以增加收入并提供更好的体验。对于电子商务，零售商正使用 GPU 驱动的机器学习和深度学习算法来打造更快、更准确的推荐引擎，从而将收入提升 60%。

对话式AI

零售商正在个性化客户体验，将消费者数据转化为切实可行的见解，**并借助实时对话式 AI 改善客户服务。**NVIDIA 借助 NVIDIA Riva SDK 使这些想法成为可能。NVIDIA Riva SDK 是用于语音识别和语音合成、语言理解等的端到端工作流。Riva 包含先进的自动语音识别 (ASR) 和文字转语音 (TTS) 功能，且实时运行。Riva 定制语音功能使任何企业只需提供 30 分钟的数据，即可为其品牌、虚拟助理或呼叫中心创建独特的语音。它可以针对不同的语言、口音、领域、词汇和上下文进行进一步调整。

行业-零售业👇

开创智慧零售新时代

零售软件合作伙伴生态系统

这些是能力卓越的工程师，他们创造出多项能使零售业受益的新工具。一些是成熟的公司，一些是 NVIDIA 初创加速器 NVIDIA 初创加速计划成员，他们已经为零售业开发出基于 GPU 的颠覆性 AI 工具。探索处于第四次工业革命前沿的企业和组织，为智能商店、仓库物流和全渠道管理提供新功能。👇

Deep North — **利用 AI 见解改善店内体验**

利用 Deep North 的技术，实体零售商店和运输枢纽可通过更好的客户体验从数字世界吸引更多消费者。

Clarifai — **加速 AI 辅助的数据标注**

Clarifai 的新 Labeler 工具可缩短标注图像所需的时间，为数据科学家提供帮助。

AiFi — **助力实现迷你型无人销售店和智能货柜**

AiFi 正在助力全球商店实现自动化，并为每个人提供采用 AI 技术的自动结账解决方案。

Dematic — **管理和扩展 AI 部署**

Dematic 正在利用边缘 AI 随时随地管理其应用程序、服务器和软件部署。

行业-超级计算

利用超级计算应对全球范围内的巨大挑战

预测天气， 研发新药， 开发新能源。 凭借可达到（甚至超过）千万亿次量级的性能，超级计算可为研究人员提供模拟和预测世界所需的强大功能。 提高工作效率和增加科研模拟数量对研究人员科技突破成果的数量和质量有着深远的影响。

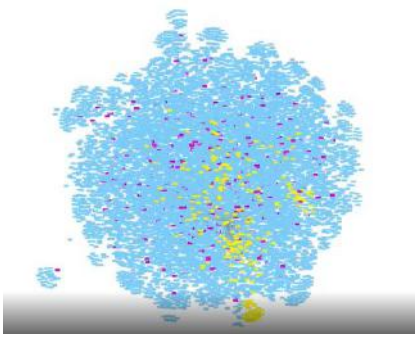
助力世界上运行速度超快的超级计算机

NVIDIA GPU 正为美国和欧洲运行速度超快的超级计算机提供助力。在美国，橡树岭国家实验室的 Summit 是非常先进的超级计算机，它将高性能计算 (HPC) 和人工智能 (AI) 融为一体，可提供超过 200 petaFLOPS 的双精度计算，满足 HPC 的需要，并提供 3 exaFLOPS 的混合精度计算，加速科学探索。此外，全球多个超级计算机中心都在采用 NVIDIA Ampere 架构。**他们将利用该架构引领科学进入百万兆级时代，同时模拟更大的模型、训练并部署更深层次的网络，以及开辟 AI 支持模拟的新兴混合领域。**

英伟达技术的实际应用场景

破解基因谜题

从可持续生物能源到更深入地理解疾病，橡树岭国家实验室的生物科学部依靠由 NVIDIA V100 Tensor Core GPU 提供支持的 Summit 超级计算机，破解之前无法解答的基础性生物学难题。



行业-超级计算

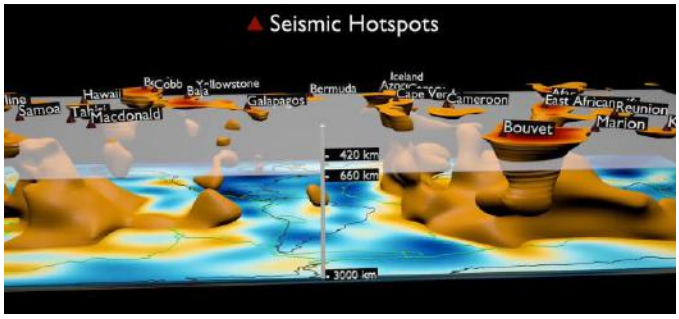
利用超级计算应对全球范围内的巨大挑战

探索超新星爆发

橡树岭国家实验室的科学计算小组正使用性能和内存皆出色的 NVIDIA Volta GPU 探索超新星爆发，及研究核科学的新领域。

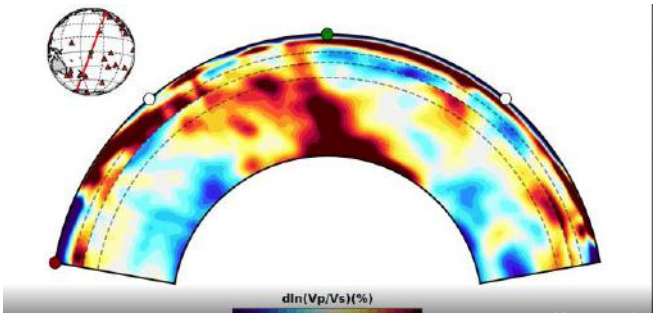
实现科学创新

从地震、清洁能源到超新星，能源部 (DOE) 帮助科学家利用橡树岭领导计算设施中的 GPU 加速 Summit 超级计算机，探索其理论。



绘制地球的内部构造

普林斯顿大学地球科学系使用由 NVIDIA Volta GPU 助力的 Summit 观察和模拟地震数据，并对地球的内部构造进行全局成像

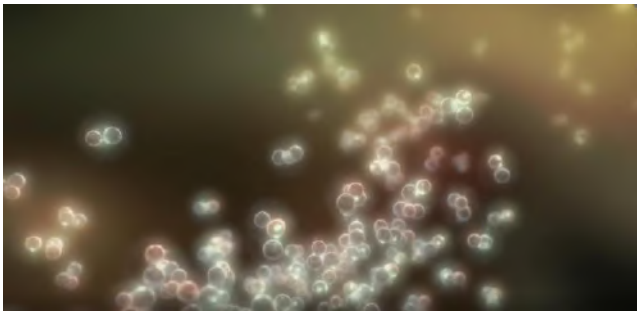


行业-超级计算

利用超级计算应对全球范围内的巨大挑战

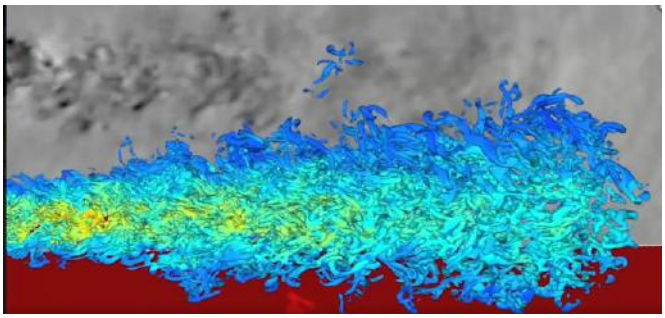
创造取之不尽的能源

普林斯顿等离子体物理学实验室正在使用橡树岭国家实验室由 NVIDIA Volta GPU 助力的 Summit 超级计算机，模拟并预测下一个聚变反应堆的等离子体活动。



推动更高效的燃烧性能

佐治亚理工学院正在寻求在不稳定性最小的情况下实现最高的燃烧性能。Summit 强劲的计算性能可以驱动更可靠且可预测的流模型，从而实现多种规模和多重物理模拟。



行业-电信

借助 AI 推动 5G 连接新时代

高带宽，低延迟，全覆盖：

5G、物联网 (IoT) 和边缘计算的融合正在大幅提升网络性能。因此，全球知名的电信公司纷纷开始使用 NVIDIA 技术来构建软件定义的基础设施，以便满足扩大内存和处理海量数据的需求，并在边缘实现智能服务。

在电信领域开发新功能

AI赋能的电信行业

网络运营变得越来越复杂，而且会生成大量有价值的数据。通过充分利用 AI 来及时获取富有实用价值的分析洞见，电信公司可以优化网络运营、提升客户体验，并发掘新的收入来源。

T-Mobile 在其客户体验中心部署了一个基于对话式 AI 的聊天机器人、一个自助服务中心，以及坐席辅助功能和转录服务。现在，他们能将客户与坐席间的对话转录为实时文本，更好地协助客户体验中心坐席。T-Mobile 称，在前 18 个月中，他们实现了相当可观的投资回报率。——案例

AI-on-5G

NVIDIA AI-on-5G 是一个统一的平台，可将 AI 和 5G 的发展成果汇聚于边缘，加速各行业企业的数字化转型。5G 为数十亿台设备提供基础连接，将 AI 算法和应用的覆盖范围扩展到边缘的所有已连接对象，创造了新的用例和市场。

行业-电信

借助 AI 推动 5G 连接新时代

在电信领域开发新功能

电信边缘服务

各行各业的企业都在采用计算机视觉、扩展现实和机器人等前沿 AI 技术来增加业务价值。电信公司凭借在连接和边缘基础设施方面的独特优势来提供这些服务。

可以确定的是，5G 将改变运营商的服务模式（从每月的语音、数据和短信套餐转变为针对特定客户需求进行调整且具有更高价值的服务），同时也势必会越来越多地用到边缘计算。——案例

VNF 和 NFVI 加速

加速内容、服务和应用交付

英伟达如何加速5G发展?

NVIDIA GPU 使基于云的虚拟现实 (VR)、智慧城市、云游戏、360 度沉浸式视频、互联无人机和自动驾驶汽车等应用成为现实。随着电信行业需求的增长，上述应用无需更改软件功能即可充分利用 GPU 计算能力。