

GPU 引领行动装置的创新与发展



袁永清

NVIDIA 中国区教育科研医疗行业总监

13911712713

dayuan@nvidia.com





游戏



专业可视化



数据中心



行动装置

视觉计算的世界领导者

GTX 980Ti

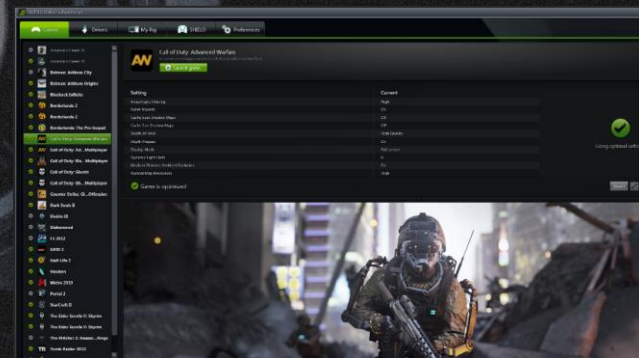
NVIDIA

GEFORCE: 全球第一的游戏平台

全球2亿游戏玩家

最前沿的科技

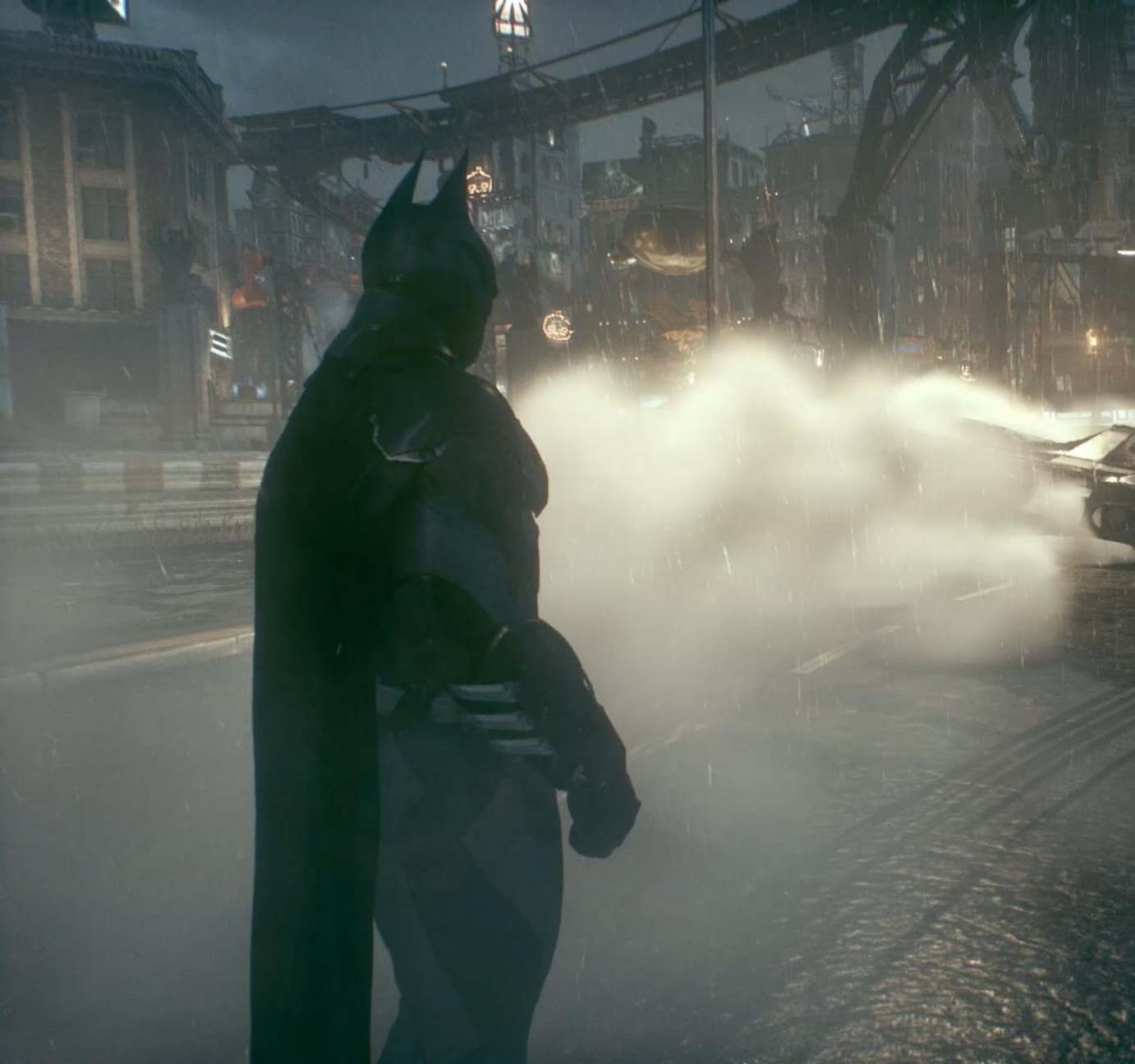
极为震撼的体验









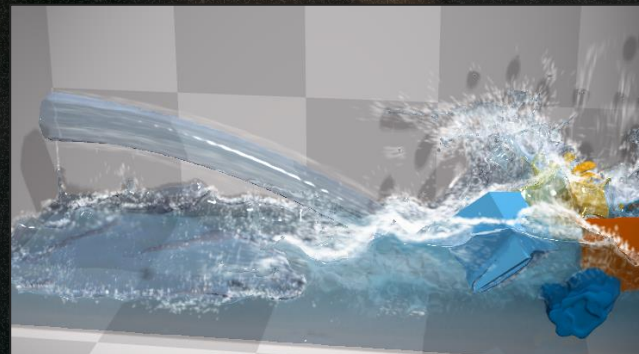


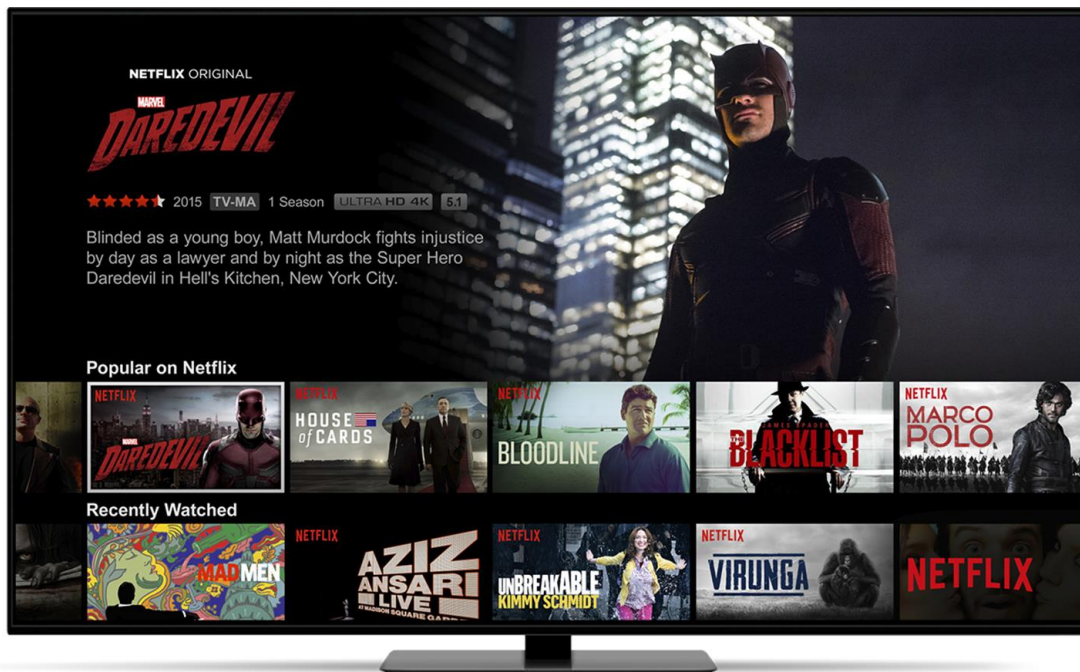
GAMEWORKS: 让游戏开发者发挥极致

300 NVIDIA visual effects engineers

Libraries, algorithms, tools for game developers

Featured in 12 blockbuster titles in 2015



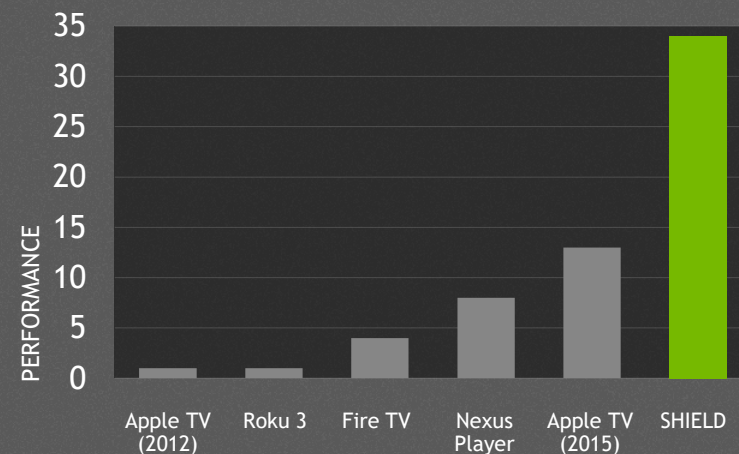


SHIELD: 最佳的 Android 游戏机

Most powerful smart TV device

A world of apps and content in 4K

Amazing games



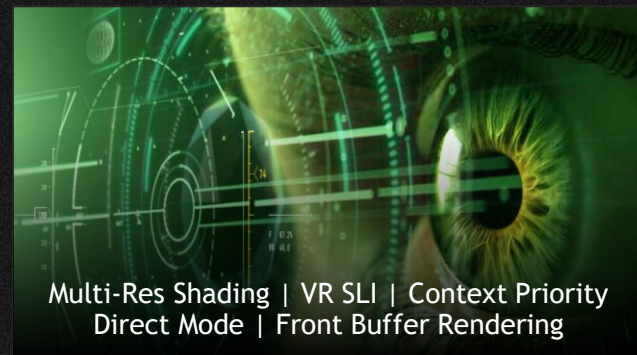


VR: 视觉游戏的革新体验

Transforming gaming, design,
film, news

Requires powerful GPUs:
Oculus min-spec = enthusiast GPU

VRWorks VR: Software for VR
headset and game developers



Multi-Res Shading | VR SLI | Context Priority
Direct Mode | Front Buffer Rendering



QUADRO: 全球第一专业 工作站

Enables professional designers to
bring their visions to life

90% workstation market share

Since 2009, every film nominated
for “Best Visual Effects” made
with NVIDIA





物理基础的渲染

Simulating light and materials to mirror real-world conditions

Potential to eliminate the need for physical prototypes

NVIDIA **Iray** and **DesignWorks** software bringing PBR to millions





医疗影像

GPUs are essential in medical imaging

Power the GE Revolution CT scanner

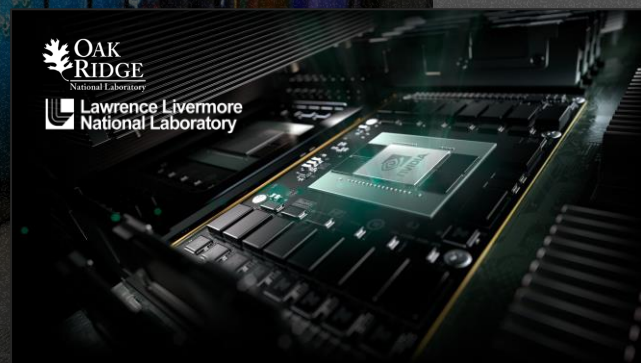
High-quality imagery while reducing radiation by up to 82%

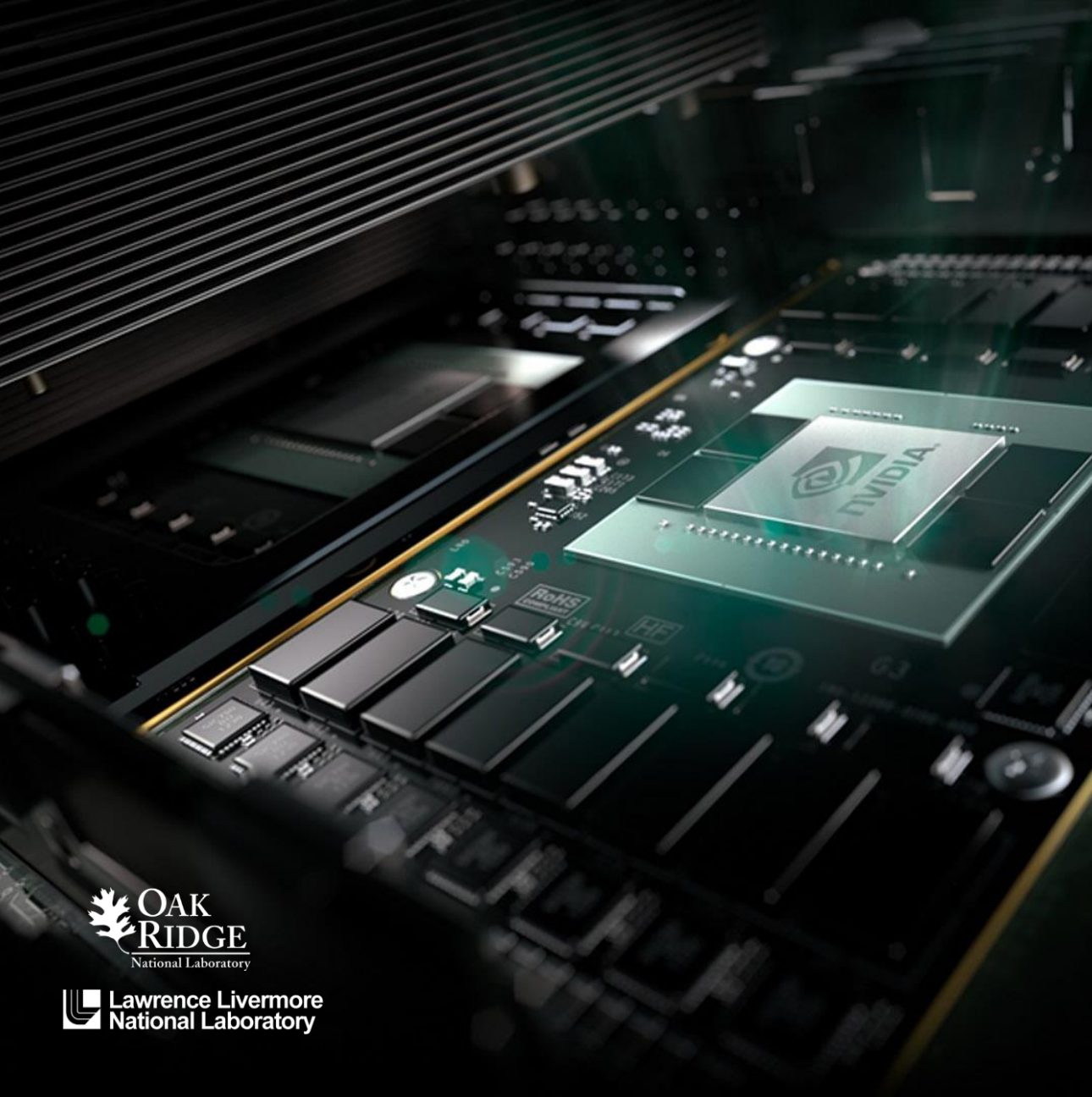


TESLA : 全球第一的 GPU 加速平台

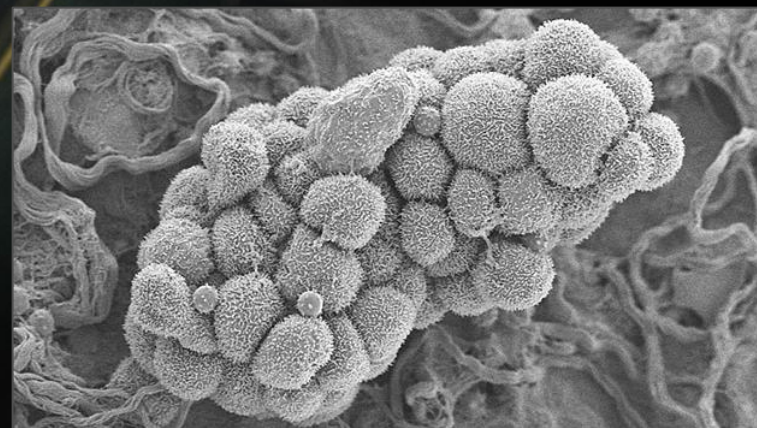
Tesla platform powers top supercomputers in U.S., Europe, Japan; Next-gen supercomputers for U.S. Dept. of Energy

CUDA programming model is accessible, easy to program



A close-up, high-angle photograph of an NVIDIA Tesla GPU mounted on a green printed circuit board. The GPU is a square chip with the NVIDIA logo and 'TESLA' branding. Various electronic components, capacitors, and connectors are visible on the board. The lighting is dramatic, highlighting the textures of the components.

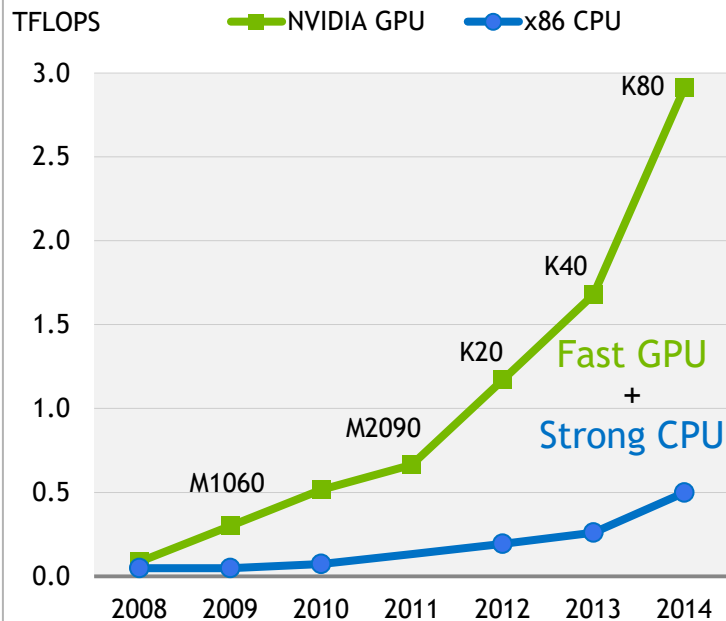
科学家和研究人员利用 Tesla® GPU 的并行处理功能在诸多领域 (例如地震研究和癌症疫苗接种) 开展创新性工作。近 800 所院校在教授 GPU 计算。最近，美国能源部选中 NVIDIA 为世界上最快的超级计算机提供支持，这些计算机将在橡树岭和劳伦斯利福摩尔国家实验室投入使用。



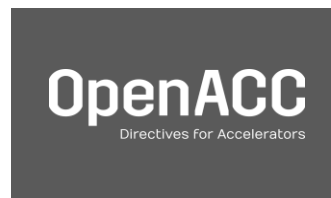
TESLA 加速计算平台

Focused on Co-Design from Top to Bottom

Fast GPU Engineered for High Throughput



Productive Programming Model & Tools



Expert Co-Design

APPLICATION

MIDDLEWARE

SYS SW

LARGE SYSTEMS

PROCESSOR

Accessibility



400 GPU加速应用软件



www.nvidia.com/appscatalog



创新的驾驶体验

NVIDIA processors power digital cockpits and infotainment systems

超过一千万辆汽车配备了 NVIDIA GPU



HONDA



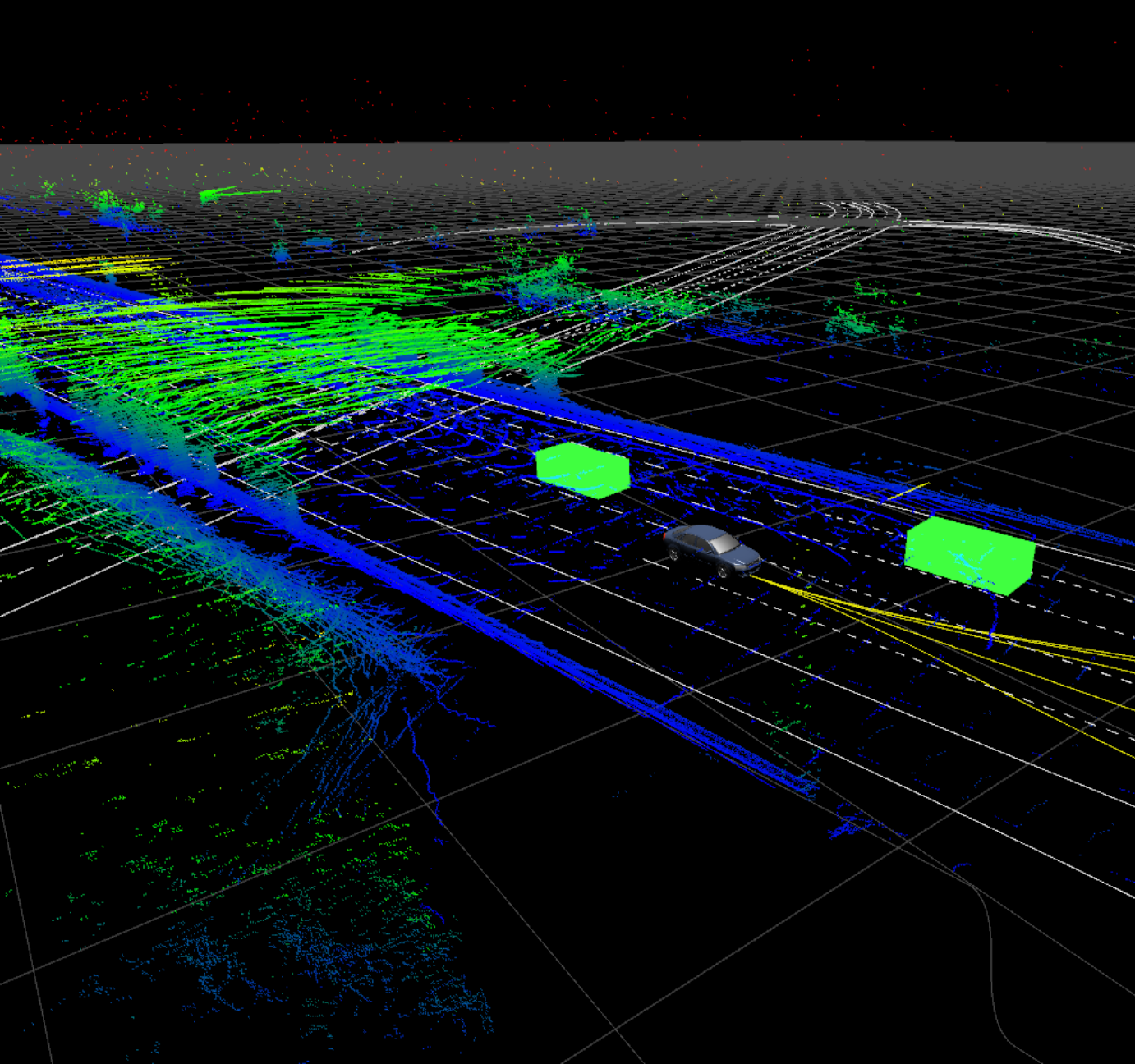
TESLA





未来，许多汽车显示器的分辨率将达到数百万像素。世界上一些最具创新性的汽车(包括奥迪、BMW、本田、兰博基尼、特斯拉和大众的车型)的数字驾驶舱和信息娱乐系统都由 NVIDIA 处理器提供支持。目前已有逾一千万辆配备 NVIDIA 处理器的汽车上路，而且未来几年还将增加 2500 万辆。





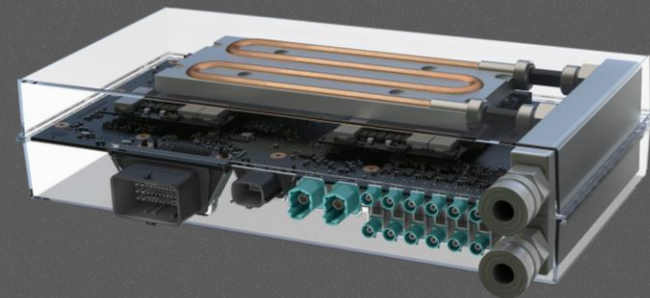
无人驾驶的大脑

Automakers: Increasing driver assistance capabilities

New entrants: Building driverless vehicles

Requires mobile supercomputer that “sees and understands”

DRIVE PX: World’s first AI supercomputer for self-driving cars



移动的超级电脑

Bringing AI and machine learning to a world of robots and drones

Jetson TX1 is the first embedded computer designed to process deep neural networks

1 TeraFLOPS in a credit-card sized module



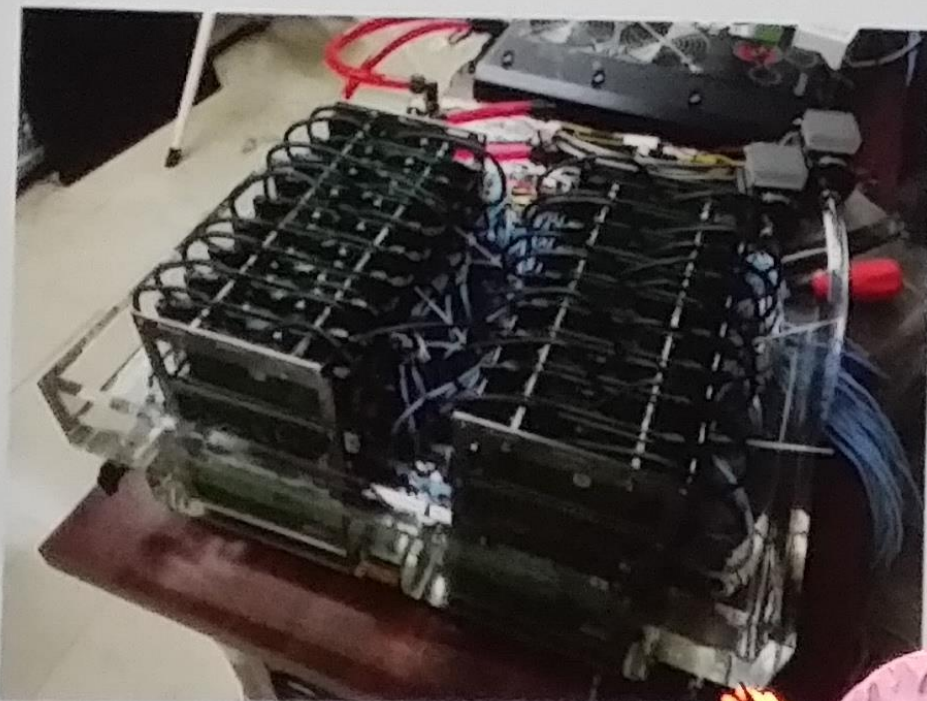
JETSON TX1

Supercomputer on a module



	KEY SPECS
GPU	1 TFLOP/s 256-core Maxwell
CPU	64-bit ARM A57 CPUs
Memory	4 GB LPDDR4 25.6 GB/s
Storage	16 GB eMMC
Wifi/BT	802.11 2x2 ac / BT Ready
Networking	1 Gigabit Ethernet
Size	50mm x 87mm
Interface	400 pin board-to-board connector
Power	Under 10W

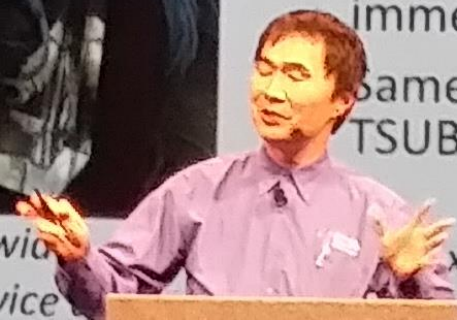
GoldenBox "Proto1" (NVIDIA K1-based) Demo at SC14 Tokyo Tech. Booth #1857



- 36 Node Tegra K1, ~11TFlops SFP
- ~700GB/s BW
- 100-700Watts
- Integrated mSata SSD, ~7GB/s I/O
- Ultra dense, Oil immersive cooling
- Same SW stack as TSUBAME2

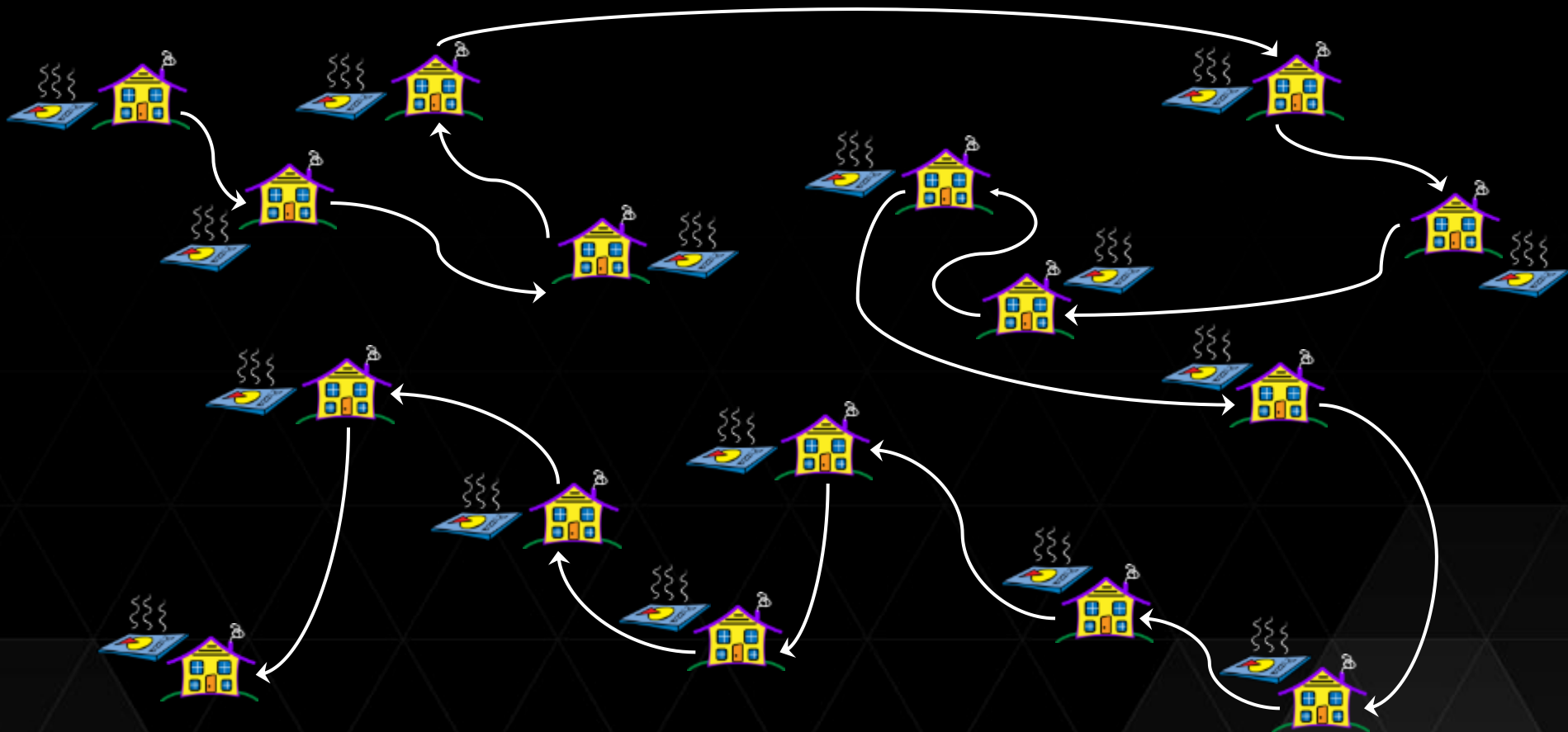


2022: x10 Flops, x10 Mem Bandwidth, x10 node density, with new device

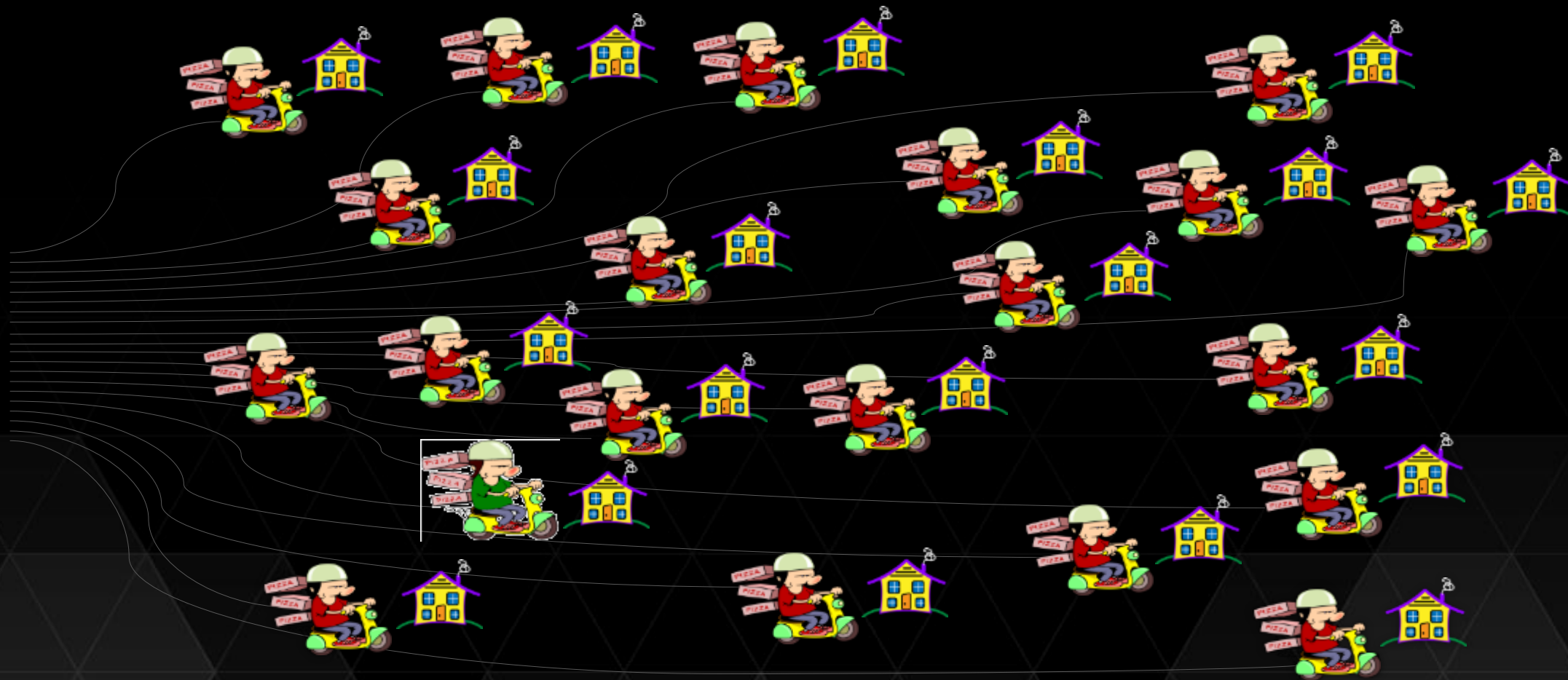


x10 NVM, x10 I/O

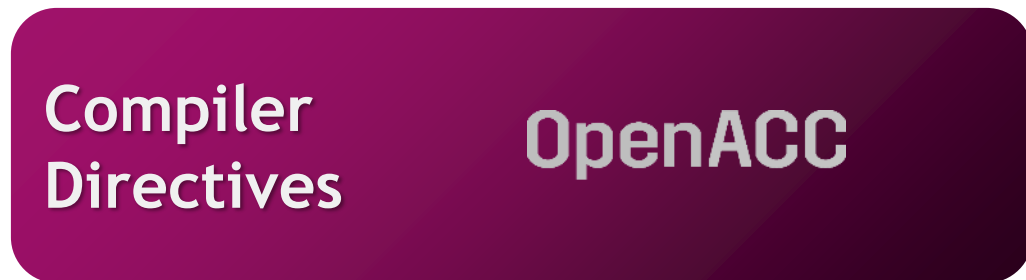
CPU – 送披萨



NVIDIA GPU – 送披萨



跨CPU的通用编程模型



GPU 完善的生态环境

Numerical Packages

MATLAB
Mathematica
NI LabView
pyCUDA

Debuggers & Profilers

cuda-gdb
NV Visual Profiler
Parallel Nsight
Visual Studio
Allinea
TotalView

GPU Compilers

C
C++
Fortran
Java
Python

Auto-parallelizing & Cluster Tools

OpenACC
mCUDA
OpenMP
Ocelot

Libraries

BLAS
FFT
LAPACK
NPP
Video
Imaging
GPULib

Consultants & Training



OEM Solution Providers

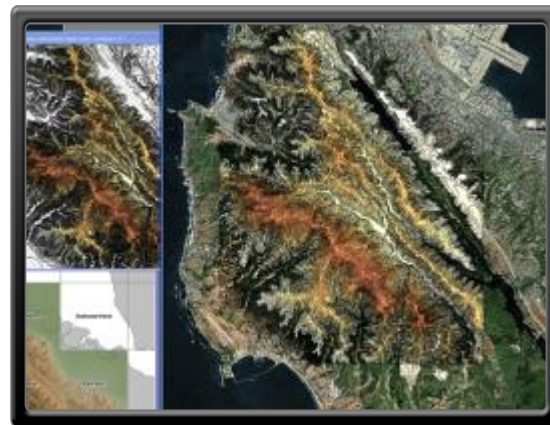




计算机视觉

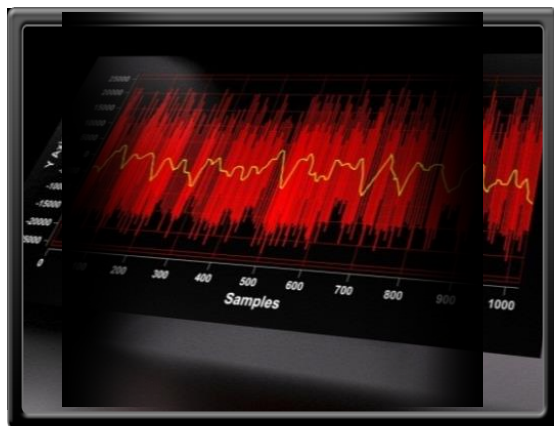


视频增强

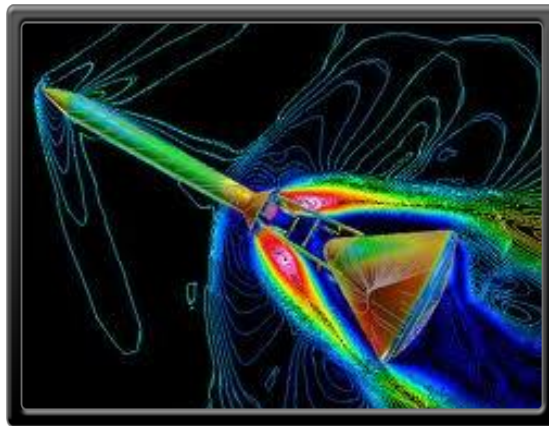


卫星成像

GPU加速10x-100x



信号处理



空气动力学



隐形与天线

生物医药研究 - H1N1

世界上第一个完整的甲
型H1N1流感病毒仿真

更准确和完整的模型

加深了对药物相互作用的
理解



中国科学院
CHINESE ACADEMY OF SCIENCES

GPU的应用领域：国防及公共安全

视频分析

视频增强

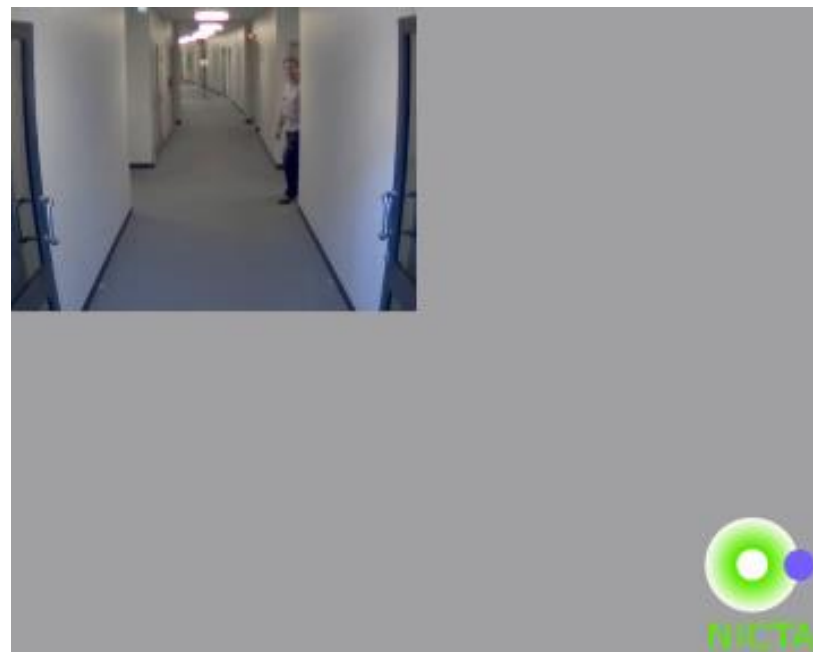
视频稳定

人像识别

指纹识别

雷达成像

卫星图形处理



视频和影像搜索和分析



- 人工审查数小时、数日、甚至数年的视频和图像，耗尽了人力分析资源
- Nervve允许自动搜索目标，这些目标以多种视频/图像格式和类型，包括EO，红外和激光雷达。Nervve需要最少的培训。
- GPU加速20x，寻找视频对象的速度超过20,000 FPS
- 有了Nervve, 分析人员能够取证搜索或实时搜索、获取排名结果、实时报警。

几秒钟就可以在一小时
视频中找到可疑对象



一小时视频 5 秒钟搜完

平安城市应用： IV007视频侦查系统

iCare 捷尚®

Your security, i Care.
智者领先 诚者共赢

- 刑事案件发生之后，刑侦人员需要调阅大量监控视频，查找关键线索。
- 视频多、查找难、清晰度差，为案件迅速侦破带来了困难。

已有审看方式，耗费大量人力、时间

如何有效快速审看视频？

▶ 看



面对海量视频，无法有效查找目标

如何在海量视频中查找关键目标？

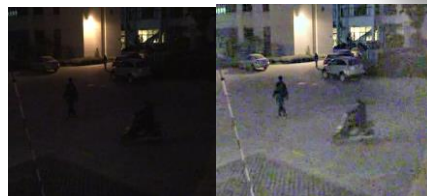
▶ 搜



视频模糊，容易丢失重要线索

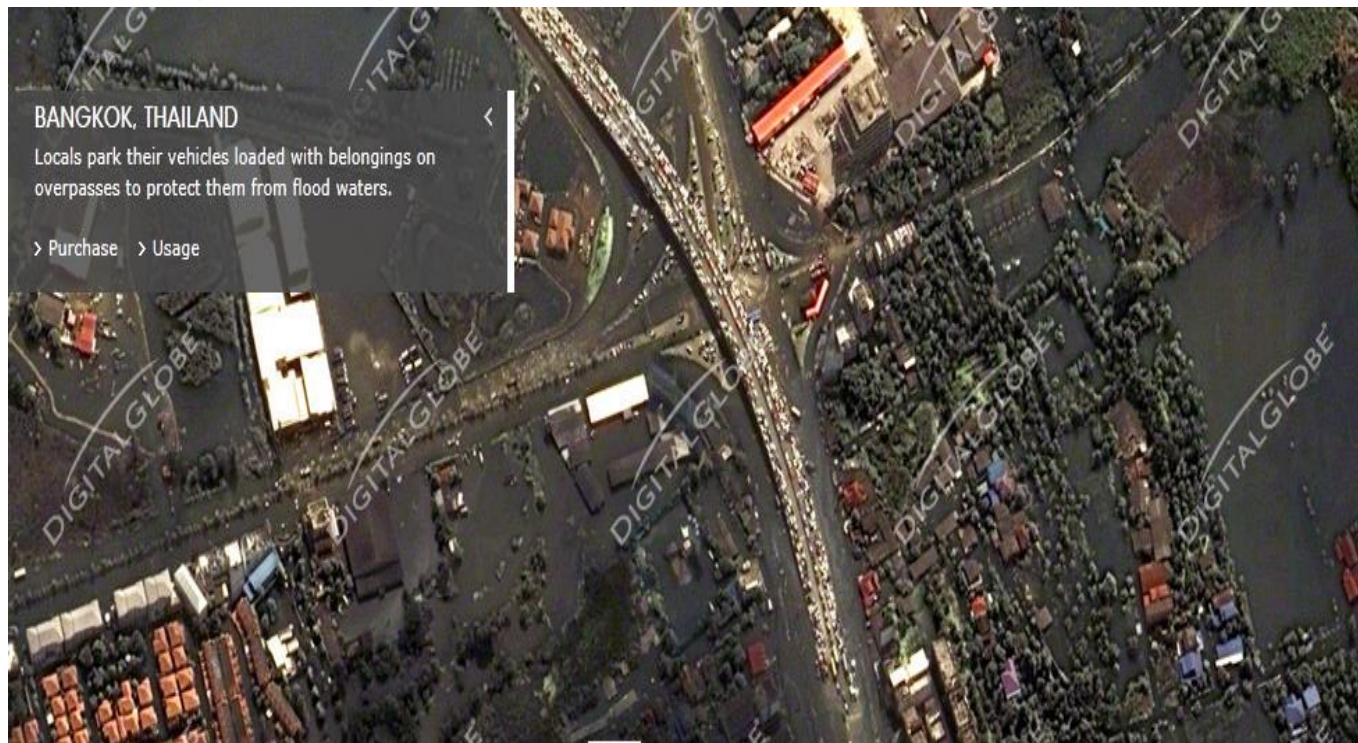
如何从模糊影像中恢复关键信息？

▶ 辩



数字城市 – 实时遥感图像处理救灾

- 高分辨率地球成像卫星每天收集超过200万平方公里的图像
- 由于地表变化造成明显的图像失真，几乎不可能做图像分析
- “正射校正”和“笔锐化”算法可以较校和恢复图像
- 使用GPU正射校正，加速12x
- 救援人员可以立即评估情况，并快速部署资源



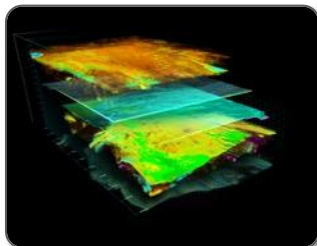
加速计算的革新

科学计算



AMBER · GROMACS · NAMD

石油天然气



RTM · FWI · Elastic

人工智能

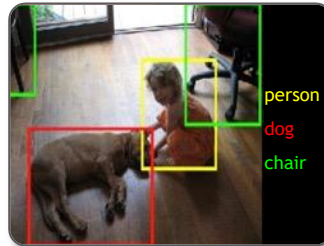
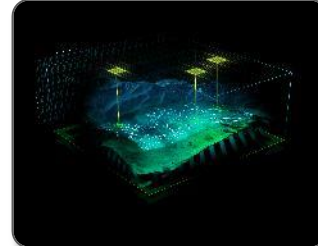


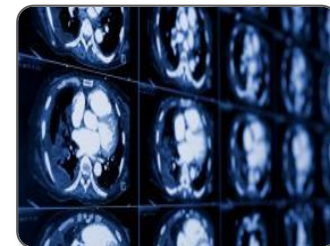
Image & Voice Recognition

国防安保



Signal · Image · Video

医疗影像



MRI · Tomography · Ultrasound

Optimized Applications in the Data Center

Accelerated Data Center

Accelerated Algorithms

Servers & Interconnects

Developer Tools

GPU Accelerators

CUDA

Tesla Accelerated Computing Platform

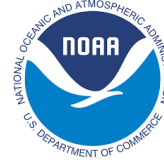
加速气象与气候预测的发展



MeteoSwiss Deploys World's First Accelerated Weather Supercomputer

2x higher resolution for daily forecasts

14x more simulation with ensemble approach for medium-range forecasts



NOAA Chooses Tesla To Improve Weather Forecast Research

Develop global model with 3km resolution, five-fold increase from today's resolution

Improved resolution requires 40x computational complexity



GOOGLE DEEPMIND ALPHAGO CHALLENGE



深度学习—开启了人工智能新篇章



“如果你问我未来20年最重要的技术是什么，
我会告诉你人工智能。
人工智能会像200年前的电力一样重要。”
——Kevin Kelly

“人工智能可能在接下来的100年之内
就将人类取而代之。”
——Stephen William Hawking

“未来机器人的数量和智能化程度
决定GDP全球排名”
——孙正义

人工智能上升为国家战略



奥巴马政府推动“大脑活动图谱计划”



欧盟“人类大脑项目”



我国正在积极启动“中国脑计划”

DEEP LEARNING IN GOOGLE



Google大脑——建立更加智能化的计算机系统，去感知世界

目标

- 语音、视觉识别
- 自然语言理解
- 用户行为预测
- 识别环境及互动
- ...

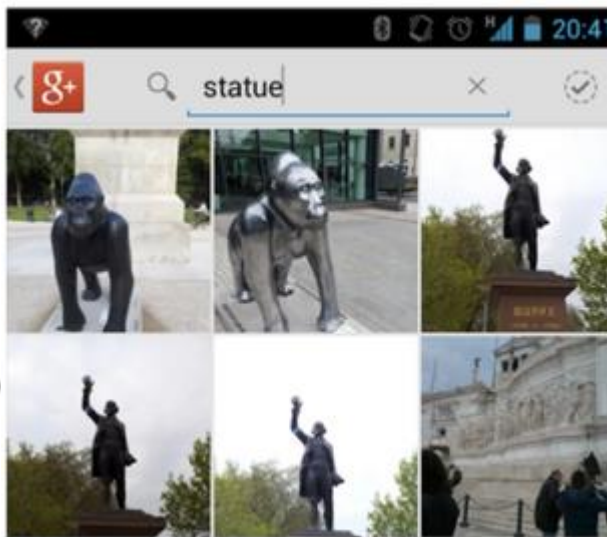
技术突破

- 避免每个应用每个算法
- 通用算法让机器从目标数据中学习

实现

- 从数据中学习
- 解决复杂问题、改善产品
- 最小化软件编程工作
- 最先进的科技水平

在过去2年里，通过深度学习的应用，发布了47款产品，例如：Photo search, Android speech recognition, StreetView, Ads placement...



还有...

图片自动标题

视频说明

Learn to action.....

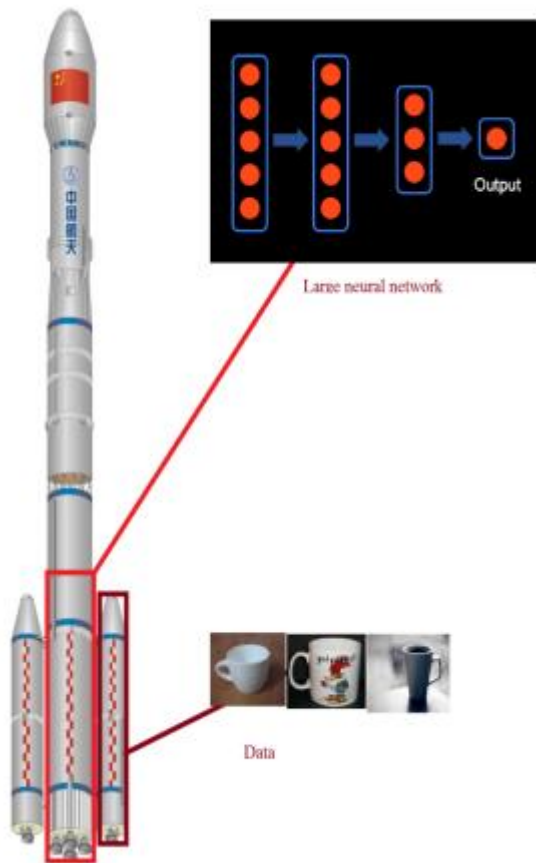


DEEP LEARNING IN 百度

Opportunities for Deep Learning

- 语音识别
- 图像识别
- 文字识别 (OCR)
- 语言翻译
- Web 搜索
- 精确推荐 (CTR)
- ...

今天, 用户在百度平台上的几乎每个服务请求, 都被深度学习系统所处理。



Images



Speech



Behavior



Baidu, Inc. (BIDU) - NasdaqGS ★ Follow

216.10 +0.02(0.01%) 3:01PM EDT - Nasdaq Real Time Price

Enter name(s) or symbol(s) GET CHART COMPARE EVENTS TECHNICAL INDICATORS CHART SETTINGS RESET

Jul 24, 2014: BIDU 204.27



"We deepened our investment in advanced technologies like Deep Learning, which is already yielding near term enhancements in user experience and customer ROI and is expected to drive transformational change over the longer term."

阿里巴巴深度学习

Image classification at Alibaba

- Product Display Classification
- Fashion Style Classification
- Buy-by-photo mobile app, search for visually similar products by images
- Leverage Caffe framework



- 商品分类
- 拍照购
- 商品风格预测
- 反黄
- 商品质量控制
- 人脸识别
- 图像搜索
- OCR
- 标签识别
- 语音输入
- 微信语音开放平台
- 消息转文字



DEEP LEARNING IN 腾讯



- 语音输入
- 微信语音开放平台
- 消息转文字



选择文件	选择文件	选择文件
tag识别	相似图搜索	商品检索
run: 11m:30s.042ms		
result	prob	
Labrador retriever 拉布拉多寻回犬 (犬种)	0.244519548018	1158
beagle (一种猎兔用) 小猎犬	0.20225943625	1171
Walker hound, Walker foxhound 沃克猎犬, 沃克猎狐犬	0.158994840418	2004
Rhodanian ridgeback 罗得西亚脊背犬	0.0822668597102	1174
dog, domestic dog, Canis familiaris	0.0403209328851	211



- QQ
- QQ空间
- 微信



语音识别 — 讯飞超脑

语音呼叫导航系统应用于中国移动、中国联通，全面提供行业语义理解解决方案

发布新一代讯飞语音云和灵犀语音助手，语义理解覆盖衣食住行等多个方面

机器口语翻译系统获得IWSLT国际评测冠军

机器口语翻译系统获得NIST国际评测冠军



定义万物互联时代人机语音交互的标准



远场降噪



方言识别



全双工



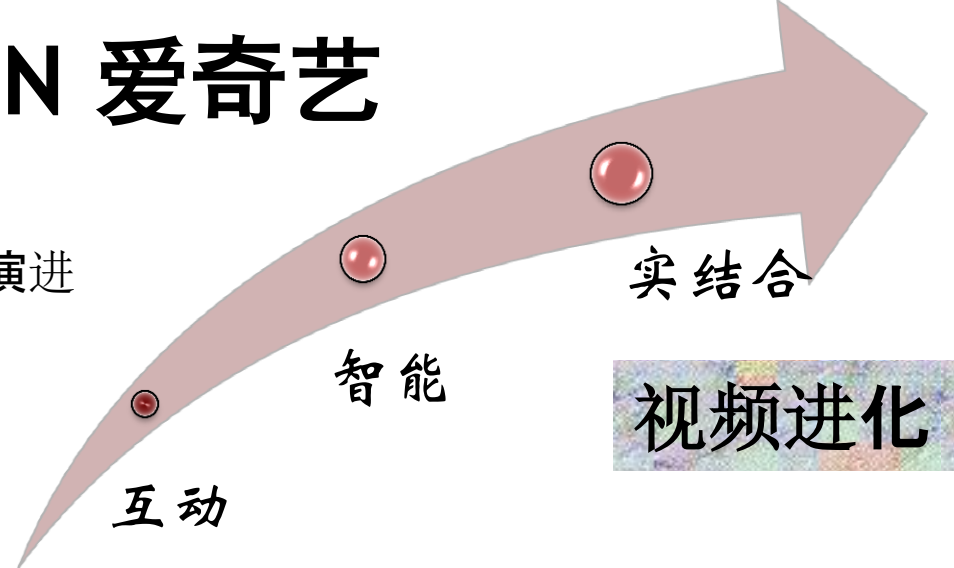
纠错



多轮对话

DEEP LEARNING IN 爱奇艺

- 视频**互动**——“人和视频”向“人到人”、“人和服务”演进
- 智能化
 - 用户分析——个性化推荐、通知、广告
 - 视频内容分析
 - 流量预测
- 虚实结合



智能化 - 视频内容分析



虚实结合 - Video In

- 真实世界中植入虚拟的物体



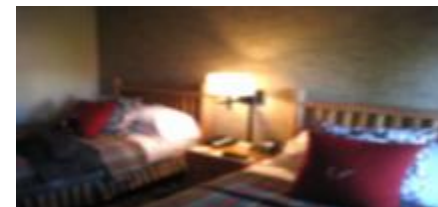
虚实结合 - Video out



DEEP LEARNING IN SENSETIME 商汤科技

Vision

- 文强
- 图片识别
- 图片分类
- 图片修复服务

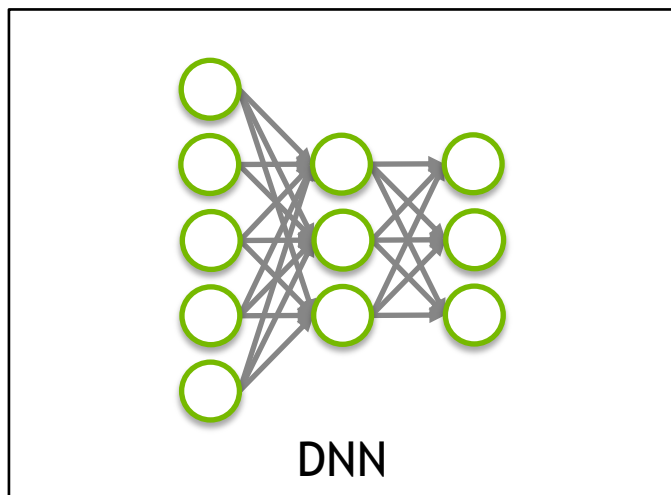


识别

- 人脸比对检测
- 目标检测
- 姿态识别
- 行人检测
- 群体突发事件识别
- 动态人脸识别



深度学习的三大要素



“Google’s AI engine also reflects how the world of computer hardware is changing. (It) depends on machines equipped with GPUs... And it depends on these chips more than the larger tech universe realizes.”

WIRED

Deep learning with COTS HPC systems

A. Coates, B. Huval, T. Wang, D. Wu,
A. Ng, B. Catanzaro

ICML 2013

*“Now You Can Build Google’s
\$1M Artificial Brain on the Cheap”*

WIRED

GOOGLE DATACENTER



1,000 CPU Servers
2,000 CPUs • 16,000 cores

600 kWatts
\$5,000,000

STANFORD AI LAB



3 GPU-Accelerated Servers
12 GPUs • 18,432 cores

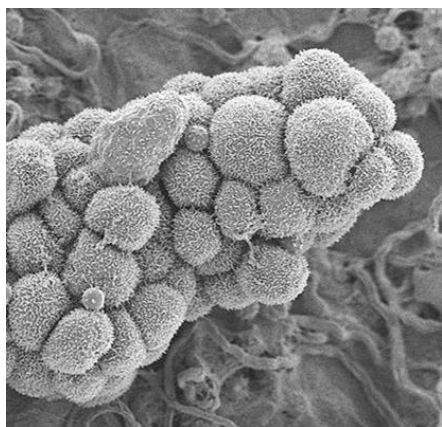
4 kWatts
\$33,000

无处不在的深度学习



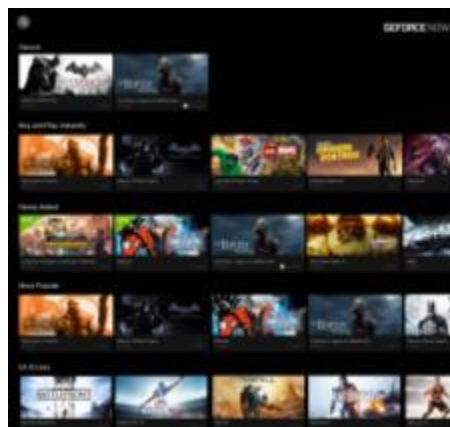
互联网

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation



医药生物

Cancer Cell Detection
Diabetic Grading
Drug Discovery



媒体娱乐

Video Captioning
Video Search
Real Time Translation



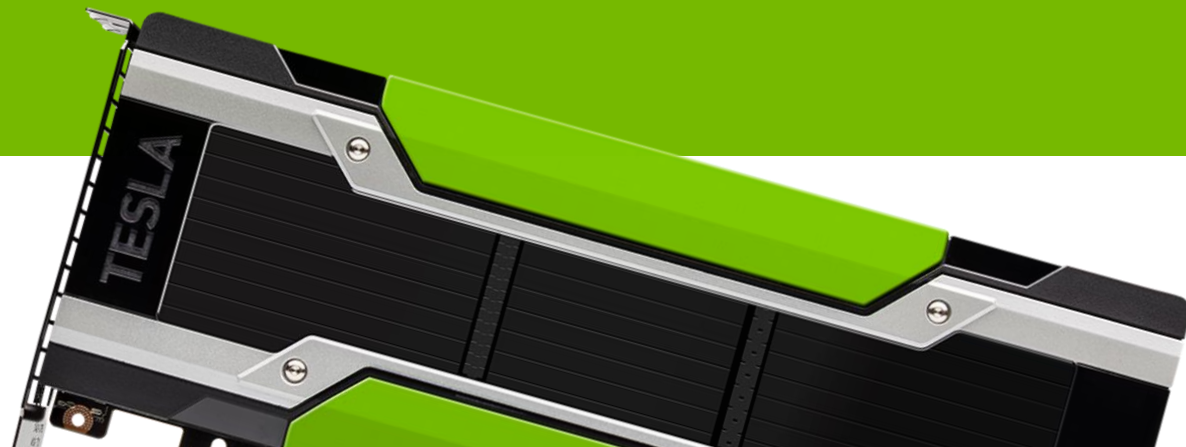
安保国防

Face Detection
Video Surveillance
Satellite Imagery



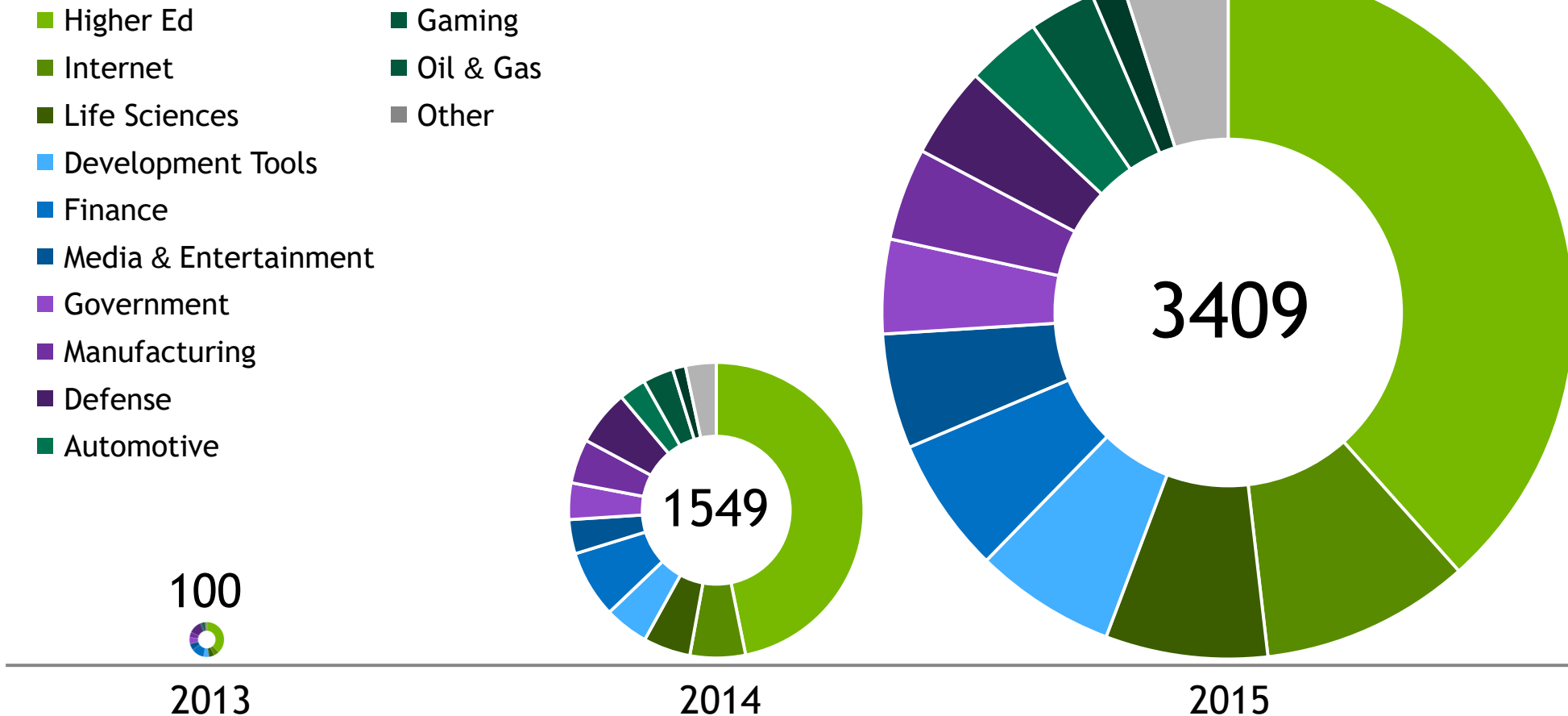
汽车，无人机，机器人

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign



EVERY INDUSTRY WANTS INTELLIGENCE

Organizations engaged with NVIDIA on deep learning

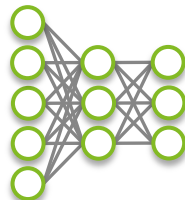


CUDA FOR DEEP LEARNING

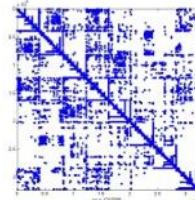
DEEP LEARNING SDK



DIGITS



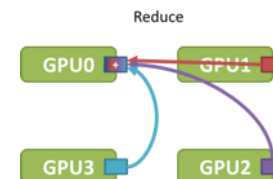
cuDNN



cuSPARSE

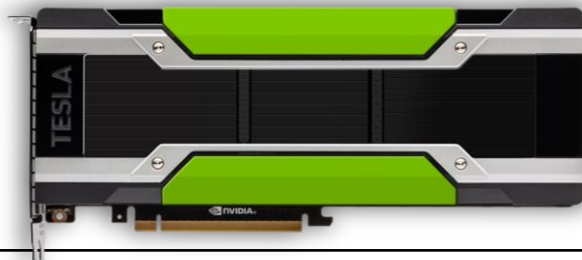


cuBLAS

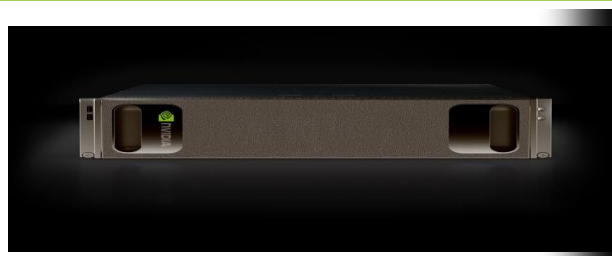


NCCL

TESLA M40



DGX-1



GPU CLOUD



NVIDIA DIGITS

Interactive Deep Learning GPU Training System

Process Data

The 'Process Data' interface shows the dataset 'voc_cropped@256x256'. It includes fields for Job Information (Directory, Image Type, Dimensions, Resize Mode) and Parse Folder (Folder, Number of categories, Training images, Validation images). A 'Create DB (train)' section shows the input file 'train.txt' and the number of DB entries (26759). A bar chart at the bottom displays the distribution of image counts across different categories.

Configure DNN

The 'Configure DNN' interface allows users to select a dataset (PASCAL VOC, ILSVRC 2012, MNIST Dataset) and configure solver options (Training epochs, Validation interval, Batch size, Base Learning Rate). It also features a 'Custom Network' section with a JSON configuration for a convolutional layer and a 'Pretrained model' field. A 'Model Name' field is set to 'ImageNet'.

Monitor Progress

The 'Monitor Progress' interface displays a graph showing Loss (train), Loss (val), and Accuracy (val) over time. The graph shows a decreasing trend in loss and an increasing trend in accuracy. A 'Dataset' section on the right provides details about the dataset (voc_cropped@256x256) and the training process (Done Wed Mar 11, 05:16:57 PM).

Visualize Layers

The 'Visualize Layers' interface shows a test image of the digit '8' and its predictions (8, 3, 0, 6, 4). It also displays the activations for the 'conv1' and 'pool1' layers, showing the feature maps extracted from the input image.

developer.nvidia.com/digits

NVIDIA SDK

The Essential Resource for GPU Developers

NVIDIA SDK

DEEP LEARNING

Deep Learning SDK

High-performance tools and libraries for deep learning



SELF-DRIVING CARS

NVIDIA DriveWorks™

Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous



VIRTUAL REALITY

NVIDIA VRWorks™

A comprehensive SDK for VR headsets, games and professional applications



GAME DEVELOPMENT

NVIDIA GameWorks™

Advanced simulation and rendering technology for game development



ACCELERATED COMPUTING

NVIDIA ComputeWorks™

Everything scientists and engineers need to build GPU-accelerated applications



DESIGN & VISUALIZATION

NVIDIA DesignWorks™

Tools and technologies to create professional graphics and advanced rendering applications



AUTONOMOUS MACHINES

NVIDIA JetPack™

Powering breakthroughs in autonomous machines, robotics and embedded computing



ADDITIONAL RESOURCES

More resources for GPU Developers



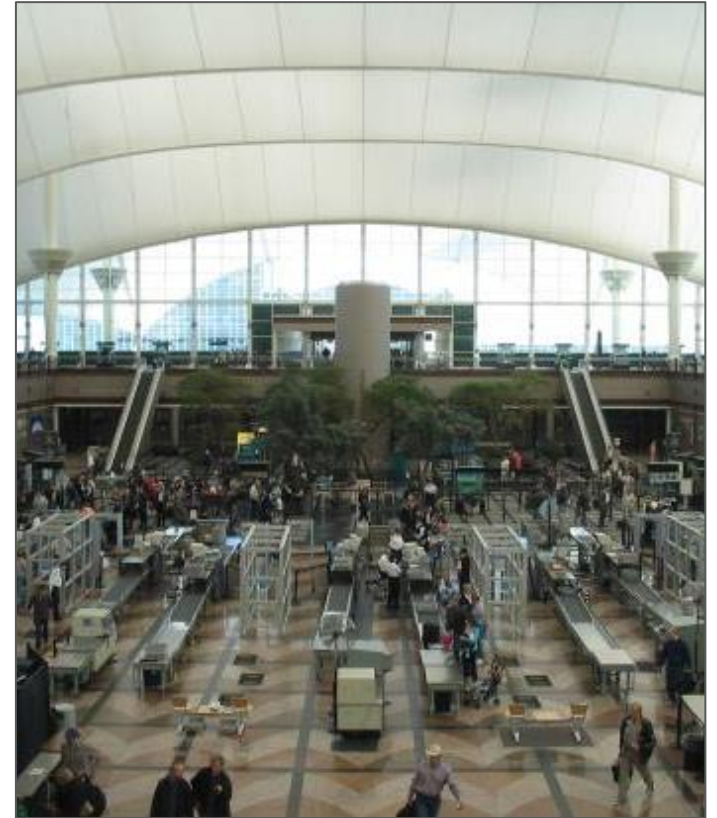
ASSISTIVE MACHINES ARE IMPROVING OUR LIVES



DELIVERY ASSISTANCE



ELDER CARE



PUBLIC SAFETY

© 2015 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, GeForce, GTX, Quadro, Tesla, GeForce Experience, Iray, NVIDIA GRID, NVIDIA GRID vGPU, SHIELD, DRIVE PX, GameWorks and DesignWorks are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

