《计算机系统设计》课程设计报告

湖南大學

HUNAN UNIVERSITY

选题名称：Tesla GPU 架构分析

姓　　名：王倩

学　　号：201708010630

专业班级：物联 1702

Tesla GPU 的 20 系列产品家族基于代号为“Fermi”的下一代 CUDA 架构，支持技术与企业计算所“必备”的诸多特性，其中包括 C++支持、可实现极高精度与可扩展性的 ECC 存储器以及 7 倍于 Tesla 10 系列 GPU 的双精度性能。

介绍

Tesla GPU 的 20 系列产品家族基于代号为“Fermi”的下一代 CUDA 架构，支持技术与企业计算所“必备”的诸多特性，其中包括 C++支持、可实现极高精度与可扩展性的 ECC 存储器以及 7 倍于 Tesla 10 系列 GPU 的双精度性能。Tesla? C2050 与 C2070 GPU 旨在重新定义高性能计算并实现超级计算的平民化。

与最新的四核 CPU 相比，Tesla C2050 与 C2070 计算处理器以十分之一的成本和二十分之一的功耗即可实现同等超级计算性能。

特性

基于新一代Fermi CUDA 架构的 GPU	与基于最新四核 CPU 的纯 CPU 系统相比，该 GPU 以十分之一的成本和二十分之一的功耗即可实现同等的集群性能。
448 个 CUDA 核心	每颗 GPU 最高可实现 515 Gigaflop 双精度峰值性能，从而让一台工作站即可实现 Teraflop 级甚至更高的性能。每颗 GPU 的单精度峰值性能超过 1 Teraflop。
ECC 存储器	能够满足工作站计算精度与可靠性方面的关键需求。能够为存储器中的数据提供保护功能，从而为应用程序增强数据完整性和可靠性。寄存器文件、L1/L2 高速缓存、共享存储器以及 DRAM 均受 ECC 的保护。
台式机上的集群性能	与一个小型服务器集群相比，配备多颗 GPU 的单台工作站能够更快地解决大型难题。
每颗 GPU 最多配备 6GB GDDR5 存储器	更大的数据集能够保存在直接附属于 GPU 的本地存储器上，从而实现了性能的最大化并减少了数据传输的情况。
NVIDIA®（英伟达？）并行DataCache?	能够为物理效果解算器、光线追踪以及稀疏矩阵乘法等诸多算法加速，在这些算法中，数据地址事先都是未知的。每个流式多处理器模块均包含一个可配置的 L1 高速缓存，所有处理器核心使用统一的 L2 高速缓存。
NVIDIA®（英伟达？）GigaThread?引擎	通过多项技术实现了吞吐量的最大化，其中包括 10 倍于上一代架构的高速上下文切换、并发内核执行以及改良的线程块调度。
异步传输	计算核心在 PCIe 总线上传输数据的同时还能够处理其它数据，因而增强了系统性能。即便是地震处理这类需要大量数据传输的应用程序，也能够通过事先将数据传输至本地存储

	器的方法来最大限度提升计算效率。
CUDA 编程环境受到各种编程语言与 API 的广泛支持	开发人员无论选择 C 语言、C++、OpenCL、DirectCompute 还是选择 Fortran 语言，都能够实现应用程序的并行机制，进而利用“Fermi”GPU 的创新架构。Microsoft Visual Studio 开发人员可以使用 NVIDIA®（英伟达？）Parallel Nsight 工具。
高速 PCIe Gen 2.0 数据传输率	实现了主系统与 Tesla 处理器之间带宽的最大化。让 Tesla 系统能够应用于几乎所有具备一条开放式 PCIe x16 插槽且符合 PCIe 规范的主系统。

规格

尺寸规格	9.75 英寸 PCIe x16 规格
Tesla GPU 的数量	1
CUDA 核心数量	448
CUDA 核心频率	1.15 GHz
双精度浮点性能（峰值）	515 Gflops
单精度浮点性能（峰值）	1.03 Tflops
专用存储器总容量* Tesla C2050 Tesla C2070	3GB GDDR5 6GB GDDR5
存储器频率	1.5 GHz
存储器接口	384 位
存储器带宽	144 GB/秒
功耗	247W 热设计功耗
系统接口	PCIe x16 Gen2
散热解决方案	主动式风扇散热器
软件开发工具	CUDA C/C++/Fortran、OpenCL 以及 DirectCompute 工具包。针对 Visu

	al Studio 的 NVIDIA ® （英伟达？）Parallel Nsight?

TESLA GPU

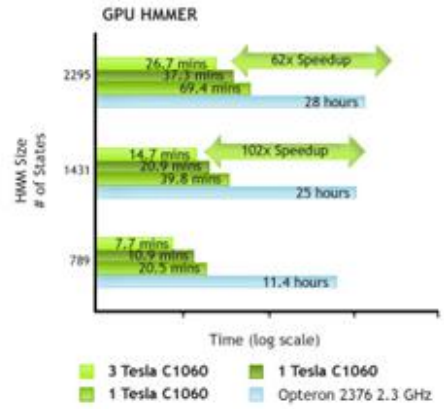
特性	Tesla K80	Tesla K40
GPU	2 颗 Kepler GK210	1 Kepler GK110B
峰值双精度浮点性能	2.91 Tflops (GPU 动态提速频率) 1.87 Tflops (基础频率)	1.66 Tflops (GPU 动态提速频率) 1.43 Tflops (基础频率)
峰值单精度浮点性能	8.74 Tflops (GPU 动态提速频率) 5.6 Tflops (基础频率)	5 Tflops (GPU 动态提速频率) 4.29 Tflops (基础频率)
存储器带宽 (ECC关闭)2	480 GB/s (每颗 GPU 240 GB/s)	288 GB/sec
存储器容量 (GDDR5)	24 GB (每颗 GPU 12GB)	12 GB
CUDA 核心数量	4992 个 (每颗 GPU 2496个)	2880

Tesla GPU 主要应用领域

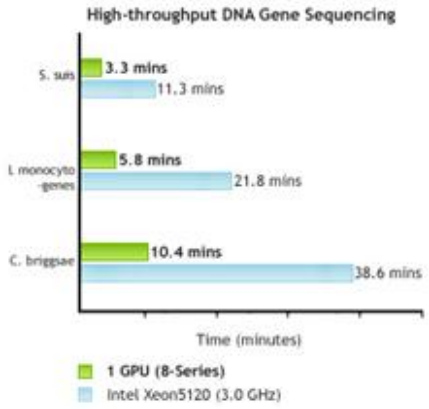
Tesla GPU 主要应用领域

生物信息学

排序以及蛋白质对接等极其密集型计算任务能够在支持 CUDA 的 GPU 上实现巨大性能提升。 当前，利用 GPU（图形处理器）来处理各种生物信息学以及生命科学代码的工作正如火如荼地进行着。



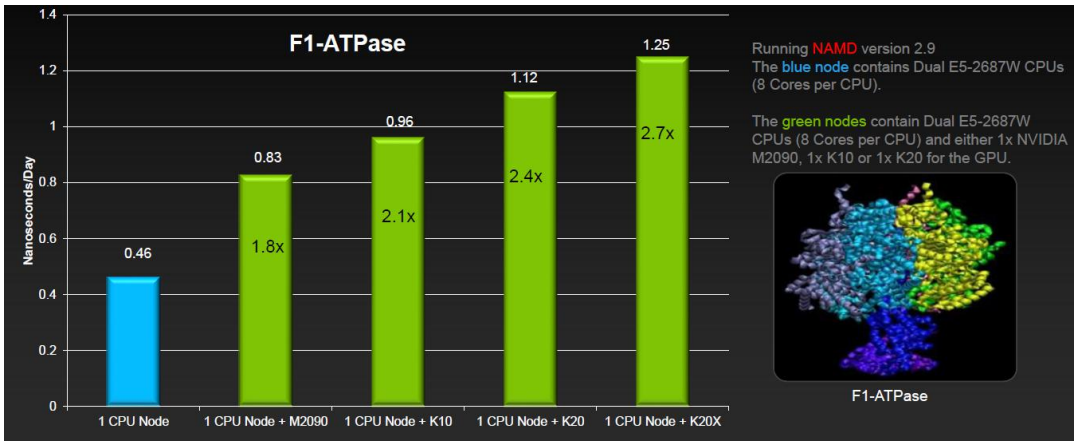
利用 GPU 加速 HMMER



使用 GPU 进行高吞吐量 DNA 序列对比

计算化学

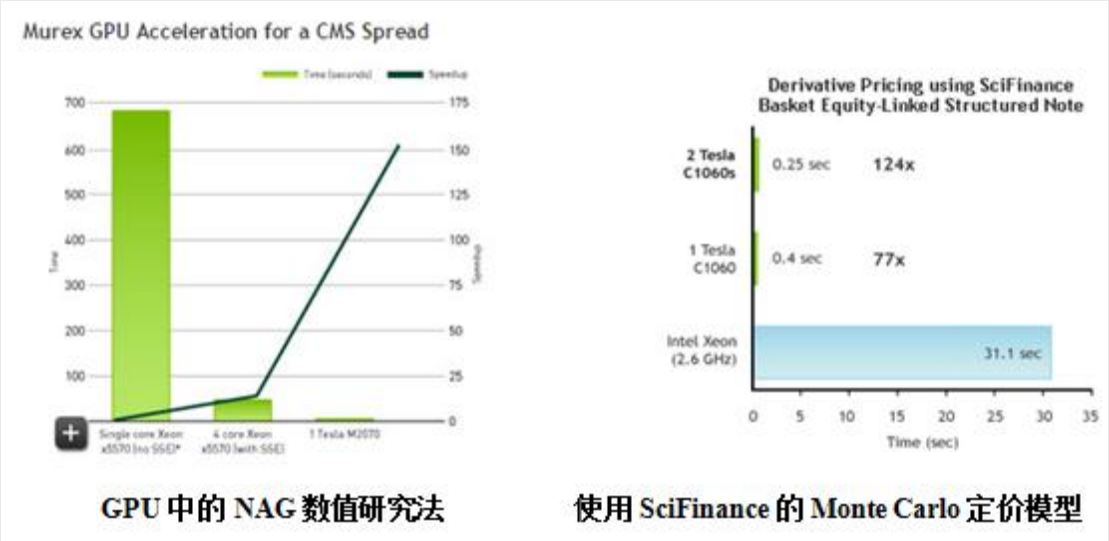
NVIDIA® Tesla® GPU 加速器让计算化学和生物学研究人员能够突破探索的极限。凭借 NVIDIA Tesla GPU，科学家可以把标准 PC 变成一个 “计算实验室”，它能够以五倍的速度运行一般分子动力学、量子化学、可视化和用于蛋白质折叠的对接应用、生物分子互动建模以及虚拟筛选。



NAMD F1-ATPase 测试对比

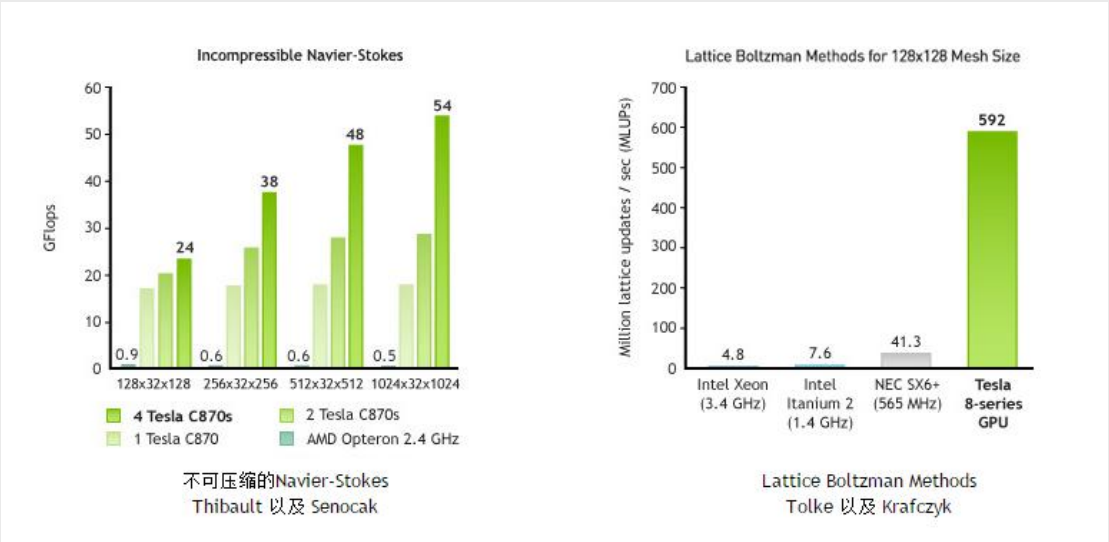
计算金融

目前，使用 CUDA GPU 平台进行的期权定价、风险分析、算法交易等工作正在进行之中。这项工作以及一些随机数字生成器以及蒙特卡洛模拟中具有代表性的图表如下：



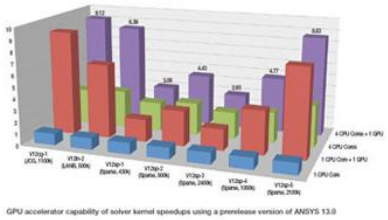
计算流体动力学

目前正在进行的针对纳维—斯托克斯（Navier-Stokes）模型以及 Lattice Boltzman 方法的几个项目已经表明，利用支持 CUDA 的 GPU 能够实现大幅的速度提升。这一工作在下面内容中进行了说明，开始部分为图表，接下来是技术报告的链接。利用 GPU 进行的天气建模以及海洋建模工作也在进行之中。

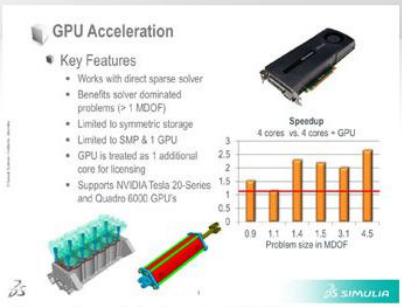


计算结构力学

目前在计算结构力学领域，我们所熟知的 ANSYS、Abaqus、MSC Nastran、IMPETUS Afea 等软件，都逐步开发了对 GPU 加速模块的支持，使用户方便使用到 GPU 强大的计算能力



Source: ANSYS Advantage, Vol. 4, Issue 2, 2010 [learn more]



Source: Abaqus 6.11 Overview, Eric Weybrant, Product Manager, SIMULIA [learn more]

数据挖掘、分析学及数据库

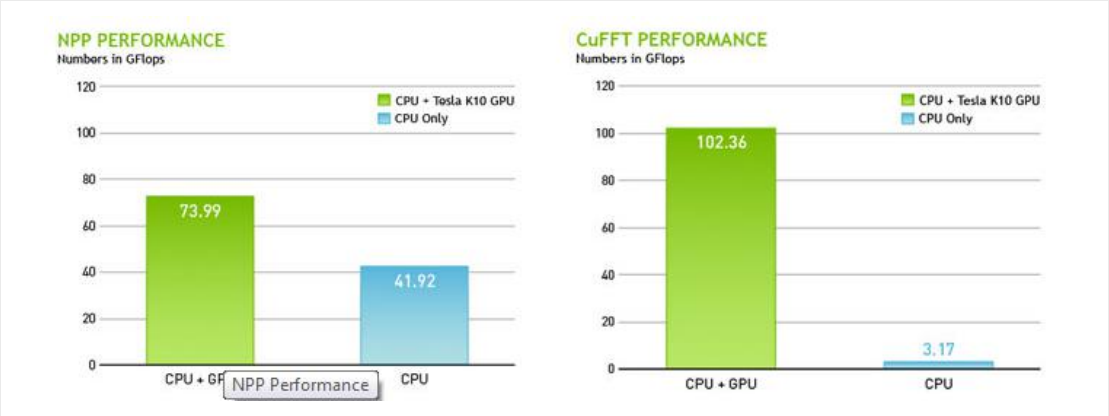
数据库是当今企业的命脉。搜索数据库并找到有用信息已经成为一个巨大的计算难题。学术界以及微软、Oracle、SAP 等公司的研究人员将依赖于支持 CUDA 的 GPU 来找到一款可扩展的解决方案。

采用 Nvidia GPU 技术的数据挖掘、分析学及数据库技术供应商

独立软件供应商（ISV）	描述	GPU优势
MathWorks MATLAB	数据并行数学 (MATLAB PCT、MDCS)	可大幅提升速度，以便为学生、科学家以及工程师提升生产率
Jedox Palo	利用联机分析处理技术可扩展Excel以便用于规划和分析	20-40倍速度提升：将决策时间从数天缩短至短短几小时
ParStream	利用GPU加速数据库和数据分析	10倍或更高速度提升：在区区几秒内即可搜索数以十亿计的纪录
Fuzzy Logix公司的Tanay Data Analytics	数据库内分析引擎	利用GPU加速金融模拟、数据挖掘以及统计方法
GPU-Quicksort	高度优化的Quicksort算法	10倍速度提升：在不到½秒的时间里即可对1600万以上的浮点数字进行排序
Arrayfire	针对 C、C++、FORTRAN 的 GPU 函数库	特定应用最多可实现100倍速度提升
Wolfram Mathematica	符号数学分析 (Mathematica)	在各种应用程序上均可实现大幅性能提升，其中包括线性代数、图像处理、金融模拟以及傅里叶变换
GPU-LIBSVM	libSVM的不同实现方式 / 支持向量	与libSVM相比，加速幅度从10倍至100倍不等
cuSVM	利用GPU分类和回归	

国防情报

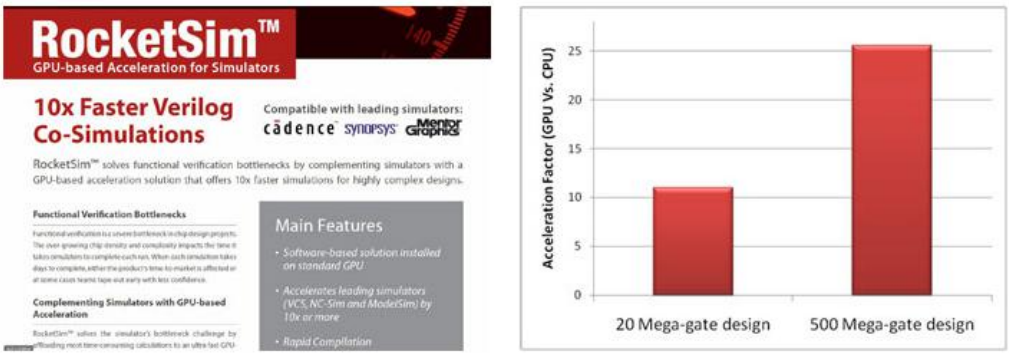
国防和情报机构每天需要对战略、情报信息进行及时准确的处理，通过对情报信息的收集分析来评估这些活动。这些海量信息来源于不同地方，如卫星、无人机、监控摄像机和雷达等，快速的对这些海量原始数据进行分析处理，这需要一个强大的计算平台及对应的基础设施。GPU 在该领域的应用，以其强大的计算能力改变了原有工作模式，在极大地提高了工作效率的同时又降低了功耗及客户总体拥有成本。



电子设计自动化（EDA）

EDA 涉及各种各样的软件算法和应用，它们都是设计复杂的下一代半导体和电子产品所需要的工具。 超大规模集成电路设计变得更加复杂，这对 EDA 构成了巨大挑战。应用程序的性能无法有效提升，因为伴随着扩展而出现的功耗增长与工艺性等问题阻碍了微处理器的性能提升。 数字系统的验证一般通过把逻辑模拟任务分配到多个大型计算中心来完成，每次耗时长达数周之久。 然而，模拟的性能通常滞后，从而导致验证不完整，漏掉了一些功能上的 Bug。 因此，半导体行业始终在寻求更快的模拟解决方案便不足为奇了。

高性能计算（HPC）领域中近来的趋势是不断发掘群核 GPU 极具竞争力的优势，方法是通过将这种 GPU 用作大规模并行的 CPU 协处理器，从而在计算量繁重的诸多 EDA 模拟上实现速度提升，这些模拟包括 Verilog 模拟、信号完整性与电磁学、计算光刻以及 SPICE 电路模拟等等。



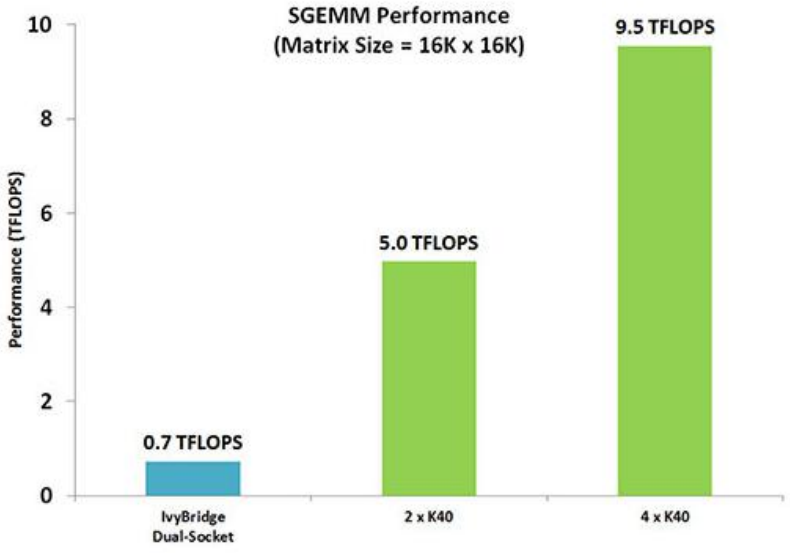
使用 RocketSim 在 GPU 上运行 Verilog 模拟

机器学习

工业与学术界的数据科学家已将 GPU 用于机器学习以便在各种应用上实现开创性的改进，这些应用包括图像分类、视频分析、语音识别以及自然语言处理等等。尤其是深度学习，人们在这一领域中一直进行大力投资和研究。深度学习是利用复杂的多级「深度」神经网络来打造一些系统，这些系统能够从海量的未标记训练数据中进行特征检测。

虽然机器学习已经有数十年的历史，但是两个较为新近的趋势促进了机器学习的广泛应用：海量训练数据的出现以及 GPU 计算所提供的强大而高效的并行计算。 人们利用 GPU 来训练这些深度神经网络，所使用的训练集大得多，所耗费的时间大幅缩短，占用的数据中心基础设施也少得多。 GPU 还被用于运行

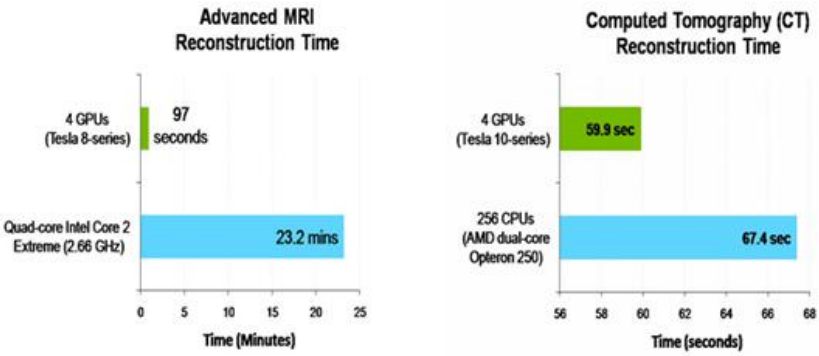
这些机器学习训练模型，以便在云端进行分类和预测，从而在耗费功率更低、占用基础设施更少的情况下能够支持远比从前更大的数据量和吞吐量。
将 GPU 加速器用于机器学习的早期用户包括诸多顶级规模的网络和社交媒体公司，另外还有数据科学和机器学习领域中一流的研究机构。 与单纯使用 CPU 的做法相比，GPU 具有数以千计的计算核心、可实现 10-100 倍应用吞吐量，因此 GPU 已经成为数据科学家处理大数据的首选处理器。



针对机器学习应用的基准测试程序

医疗成像

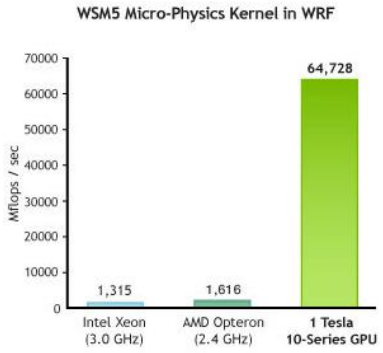
医疗成像是最早利用 GPU（图形处理器）计算加快性能的应用之一。GPU 在这一领域的应用日趋成熟，当前有多款医疗设备均配备了 NVIDIA Tesla GPU。



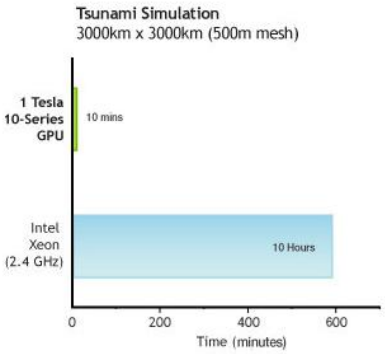
使用 GPU 加快高级 MRI 重建速度　　计算体层重建

天气、大气、海洋建模与空间科学

不少天气和海洋建模应用程序，如 WRF（天气研究与预测模型）和海啸模拟应用程序获得显著的性能提升，缩短了计算时间的同时亦改进了计算精度



利用CUDA实现的WRF加速
WRF总体加速幅度为1.25倍
Michalakes 和 Vachharajani



海啸模拟
Dr. Takayuki Aoki