——国防科大2020年高性能评测与优化课程小组讨论

# Fermi GF100 GPU architecture

**汇报：黄显栋　刘俊奇　詹俊伟**

**指导：龚春叶、甘新标、杨博**

# 目录

CONTENTS

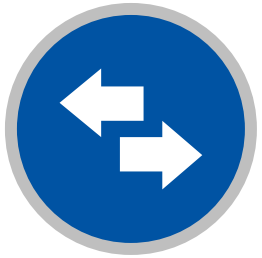
PART ONE

# 开发背景

# 开发背景

“采用Fermi架构的GF100,是一款图形处理同样出色的并行处理器”

GPU不再仅仅局限于进行图形处理，通用计算同样重要

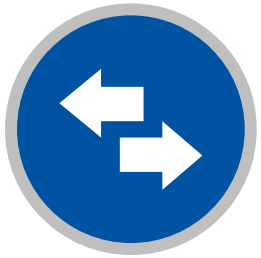


DirectX 11降低了使用GPU可编程单元进行通用计算的编程难度

CUDA编程模型采用统一处理架构，引入片内共享存储，降低了GPU并行计算开发难度

# 开发背景

40nm级的半导体工艺的支持，使更大规模的运算，控制和缓存资源的加入成为可能



NVIDIA早期在图形芯片设计上积累了足够成熟的设计经验

PART TWO

# 技术方案

# 技术方案

## 第三代的 SM（多流处理器）

- 每个SM包含32个CUDA Core，是GT200的4倍
- 8倍于GT200的双精度浮点操作的峰值计算能力

## 第二代的线程并行计算ISA

- 统一的地址空间，内存访问指令支持64位寻址，支持C++编程
- 针对OpenGL和DirectCompute做了优化设计
- 完全支持IEEE 754-2008标准的单双精度浮点计算

# 技术方案

## 增强的内存操作性能

- 支持GPU内存的ECC(错误检查与纠正技术)
- 增强了内存原子操作性能

## 千兆线程调度管理引擎

- 10倍的上下文切换能力
- 并发的kernel执行
- 支持block乱序执行

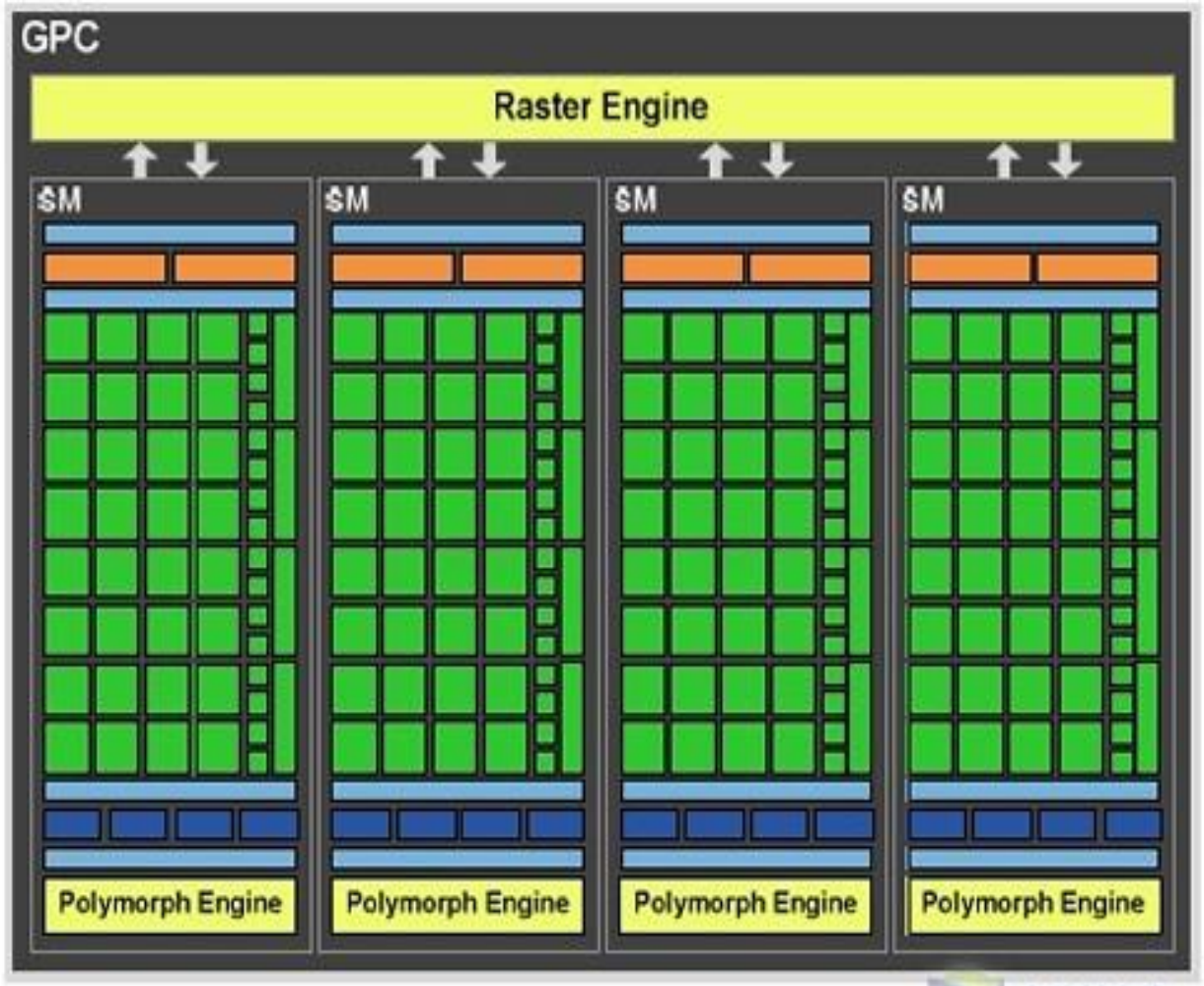
# 技术方案

- Fermi GF100 GPU基于图形处理团簇（GPC），可扩展流阵列多处理器（SM）和内存控制器（MC）。
- 一个完整GF100包括4个GPC，16个SM和6个内存控制器。


Fermi架构示意图

# 技术方案

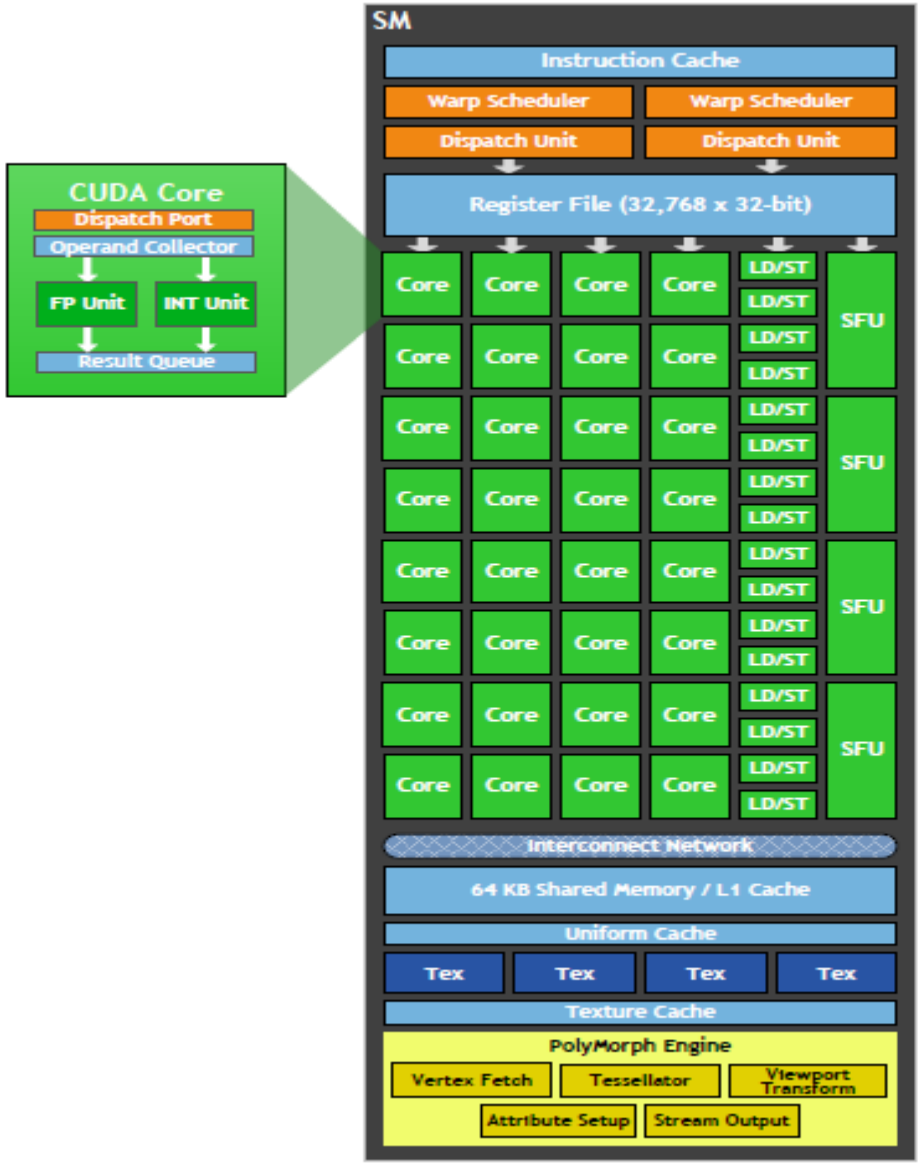
- GPC是GF100占主导地位的高层次的硬件模块。
- 每个GPC包含1个光栅引擎和4个SM单元。
- 除了计算单元它还包括两个重要特点——分别是1个可升级的光栅引擎（Raster Engine）、Z-cull和1个带有属性提取和细分曲面的多边形引擎（Polymorph Engine）




*Graphics Processing Cluster (GPC)*

# 技术方案

- GF100拥有512个CUDA Core，它们属于16个SM单元，每个SM单元包括32个CUDA内核。
- 每个SM是一个高度平行处理器，最多支持在任何规定时间完成对48个warp的处理 。每个CUDA Core是一个统一的处理器核心，执行顶点，像素，几何和kernel函数。
- 每组SM里4个纹理单元，共享使用12KB的L1纹理缓存，并和整个芯片共享768KB的L2缓存。




Fermi架构第三代流处理器群（SM）示意图

# 技术方案

- 光栅引擎以流水线的方式执行：边缘/三角形设定(Edge/Triangle Setup)、光栅化(Rasterization)、Z轴压缩(Z-Culling)等操作；每个时钟周期可处理8个像素。


光栅引擎（Raster Engine）

# 技术方案

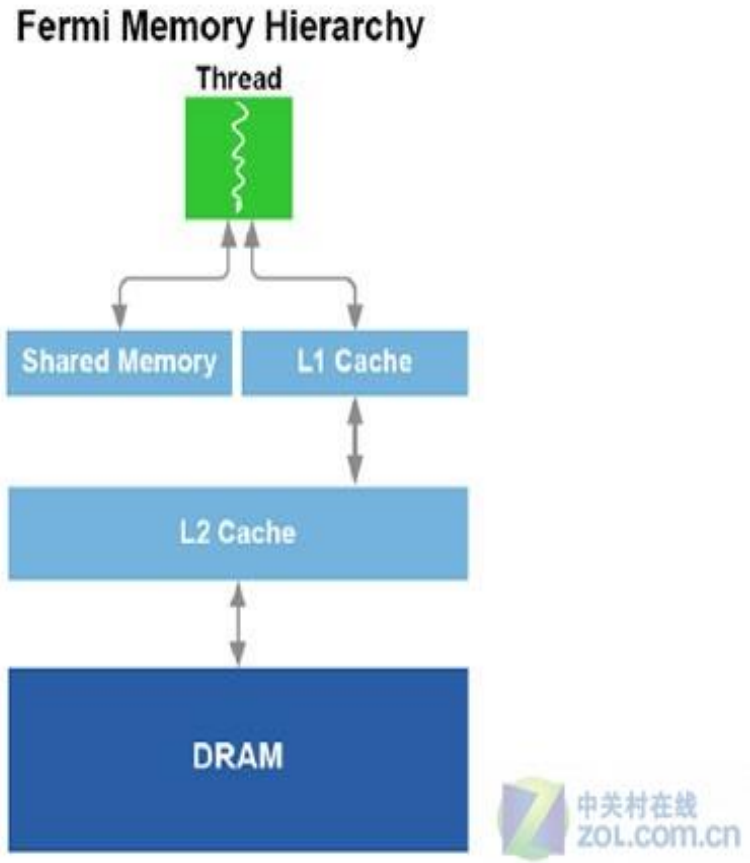
- 多边形引擎则要负责顶点拾取（Vertex Fetch）、细分曲面（Tessellation）、视口转换（Viewport Transform）、属性设定（Attribute Setup）、流输出（Stream Output）等五个方面的处理工作。


多边形引擎（PolyMorph Engine）

# 技术方案

- GF100的每一个SM中拥有64KB的可配置片上缓存，可以设置为：48KB共享缓存+16KB L1缓存；或者16KB共享缓存+48KB L1缓存。

- L1缓存可以用于处理寄存器溢出、堆栈操作和全局LD/ST；



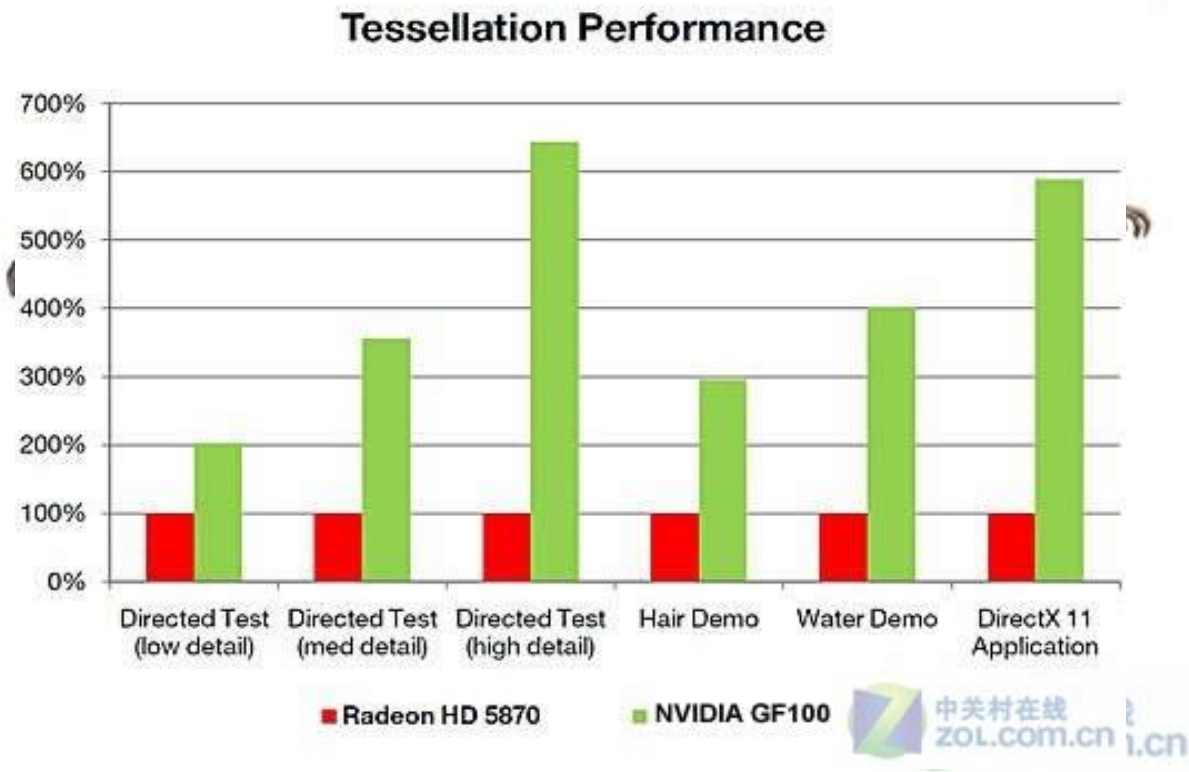

Fermi架构可配置缓存结构

PART THREE

# 效果

# 细分曲面技术带来的变革

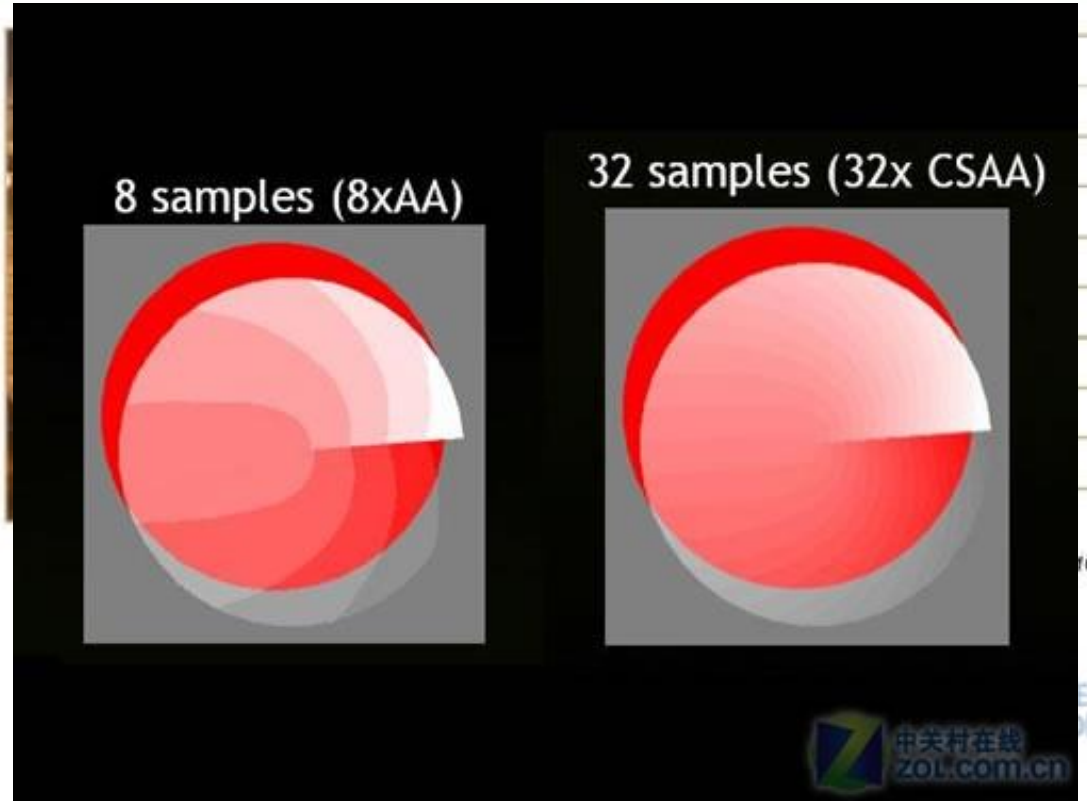
- 细分曲面技术是DirectX 11为我们带来的最重要的革新， 是创造具有更多纹理细节以及平滑边缘几何图形的最佳途径之一。
- 相比于其他产品，GF100上具有更多的多边形处理引擎，能更好的支持细分曲面技术。




Tessellation细分曲面效果性能对比

# 游戏画质进一步改善

- GF100可以给玩家更好的游戏画质体验，包括更先进的抖动采样技术和性能衰减更小的抗锯齿加速。 Fermi执行CSAA性能下降非常低，32X CSAA性能下降只有传统8X的0.7%左右。




Gather4抖动采样指令细节图解

8X和32X CSAA的效果对比

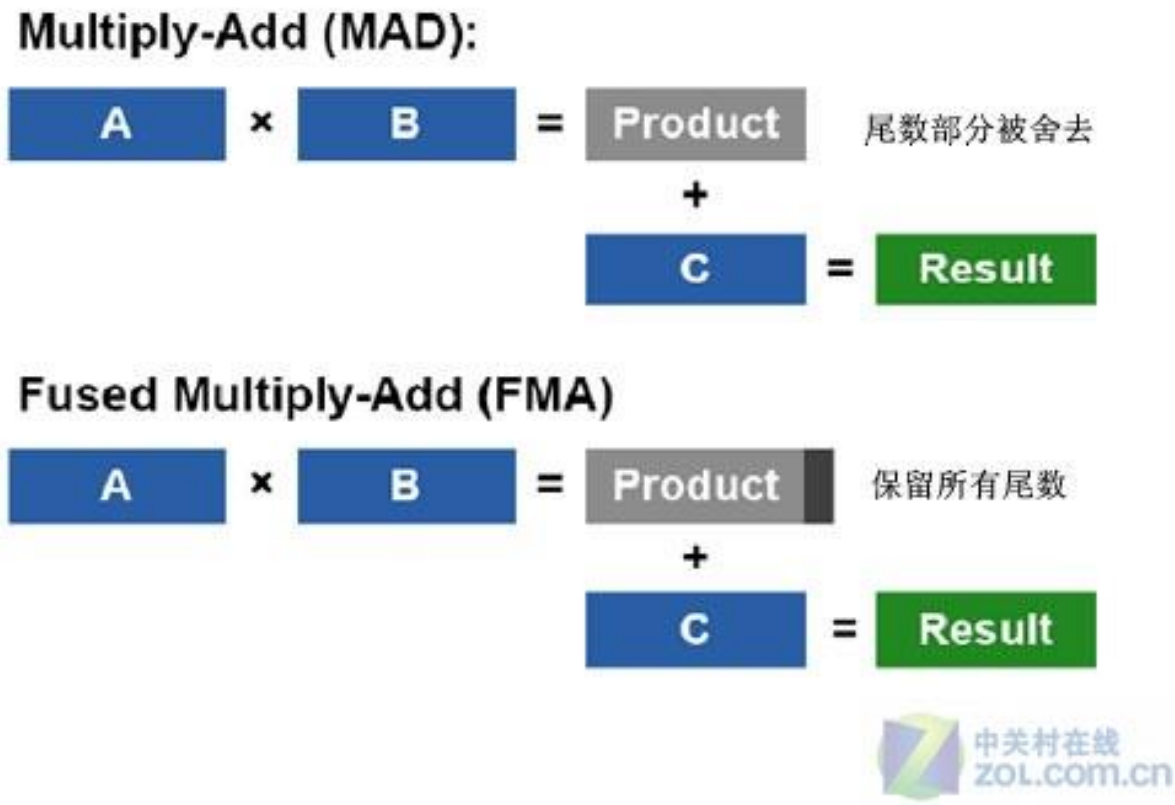
# 游戏计算性能提高

- NVIDIA宣称，GF100的游戏计算性能相比GT200有了大幅提高，比如PhysX流体DEMO演示程序3.0倍、《Dark Void》游戏物理2.1倍、光线追踪3.5倍、人工智能3.4倍。
- CUDA还可以用于游戏的AI寻路计算，可以高效的计算最短路径，并可以做冲突预测，GF100在寻路方面以提供3倍于GT200的性能。


GPU游戏计算架构模拟出非常真实的景深效果

# 底层计算单元的改进

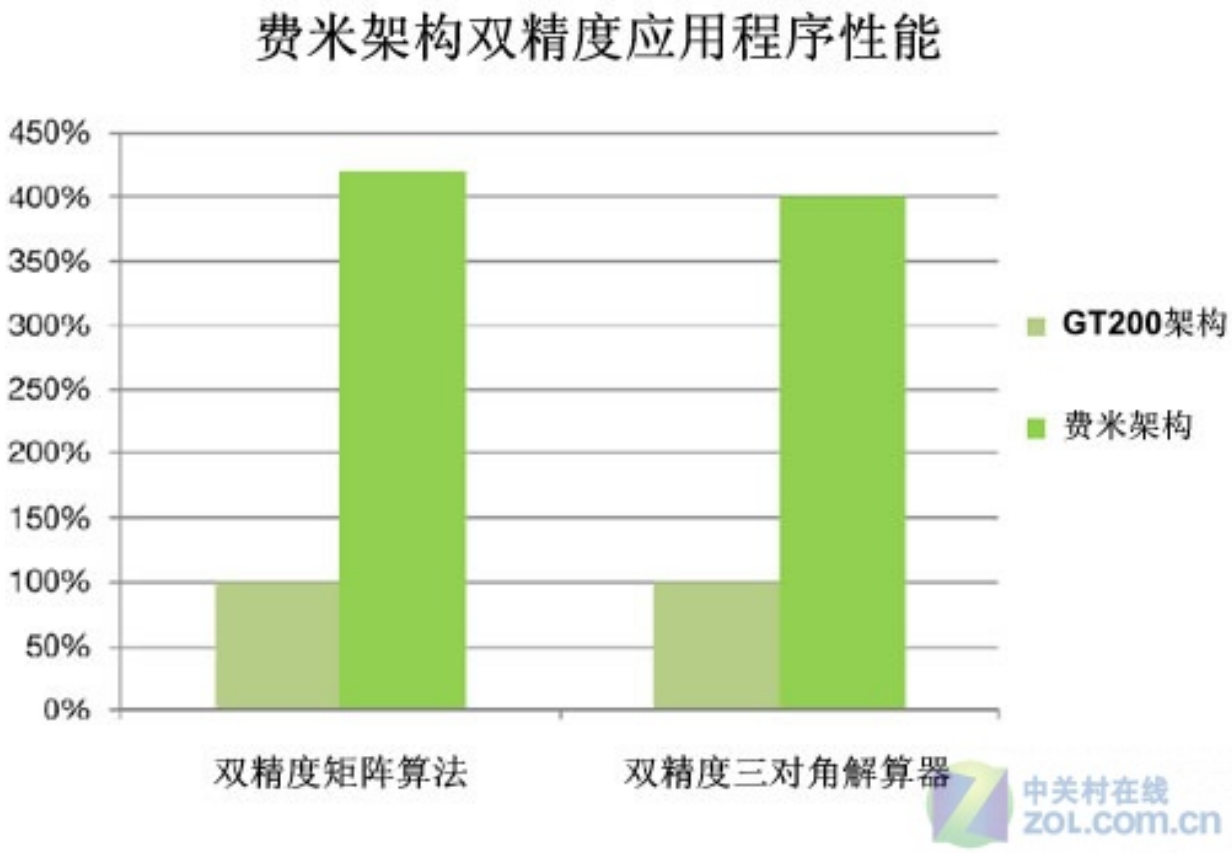
- 和MAD（multiply-add）指令相比，Fermi所支持的FMA指令在做乘运算和加运算的时候只在最后运算的时候作一次舍入，不会在执行加法的时候就出现精度损失，精度比把操作分离执行更高。




FMA指令在精度上的优势

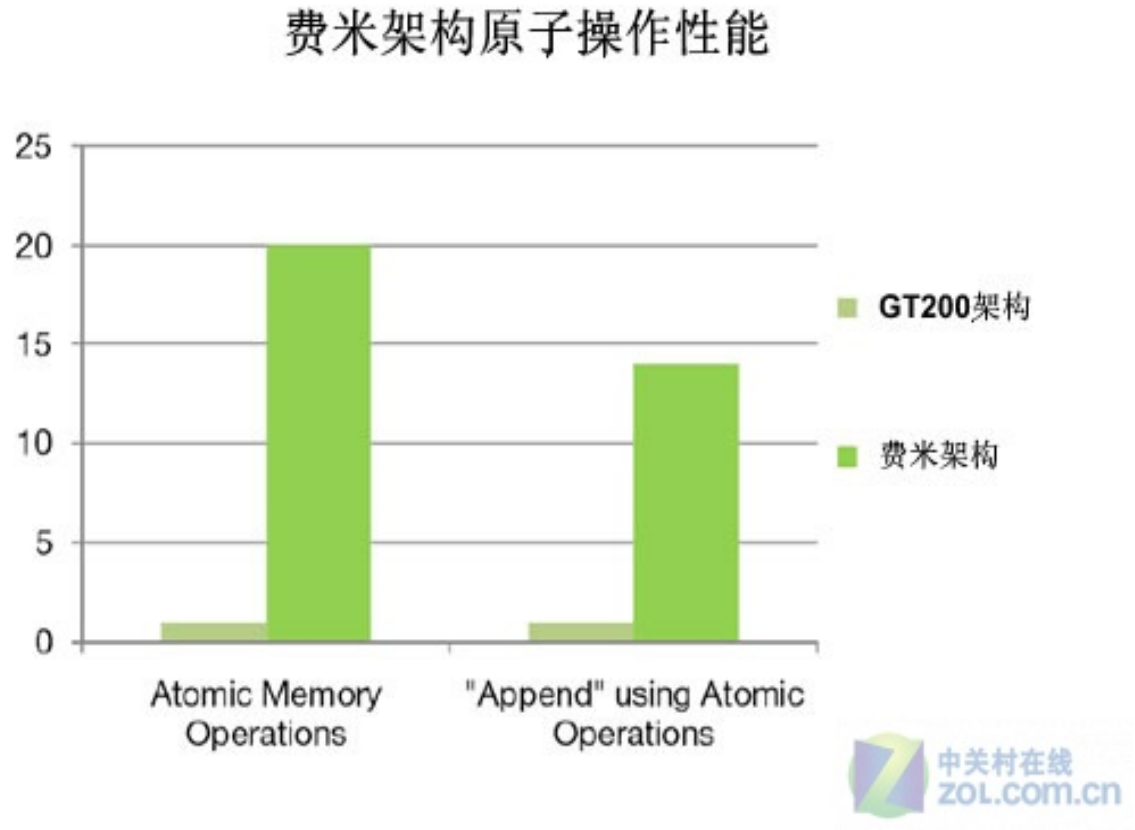
# 双精度浮点性能提升

- Fermi的双精度浮点（FP64）性能也大幅度提升，峰值执行率可以达到单精度浮点（FP32）的1/2，而过去只有1/8，AMD现在也不过1/5，比如Radeon HD 5870分别为单精度2.72TFlops、双精度544GFlops。



# 原子操作性能增强

- 大量原子操作单元和L2缓存的使用，大大增强了FermiGPU架构中的原子操作能力。在相同位置的原子操作，Fermi的速度比GT200快20倍，连续的内存操作是GT200的7.5倍。



PART FOUR

# 总结

# 总结

01 40nm工艺，30亿个晶体管的“大芯片”
基于图形处理但超越图形处理的设计

02 拥有更多的通用计算所需的内存，控制以及缓存等周边资源

03 图形处理的设计更偏向应用，几何形能大幅提升；同时也更偏向CPU

# THANKS