GP GPU

GPUコンピューティン No.2

GPUアーキテクチャの変遷と各種メモリバンド幅

東京工業大学　学術国際情報センター

青木　尊之

GP GPU

GeForce 8800 (2006)

GPU computing 世代の発の GPU

「シェーダ言語でアクセスできる高性能グラフィックプロセッサ」

◎Unified-Shader型アーキテクチャ
◎DirectX 10(Direct3D 10) Shader Model 4.0準拠
◎128ストリームプロセッサ
◎96 ROP
◎384bitメモリインターフェイス

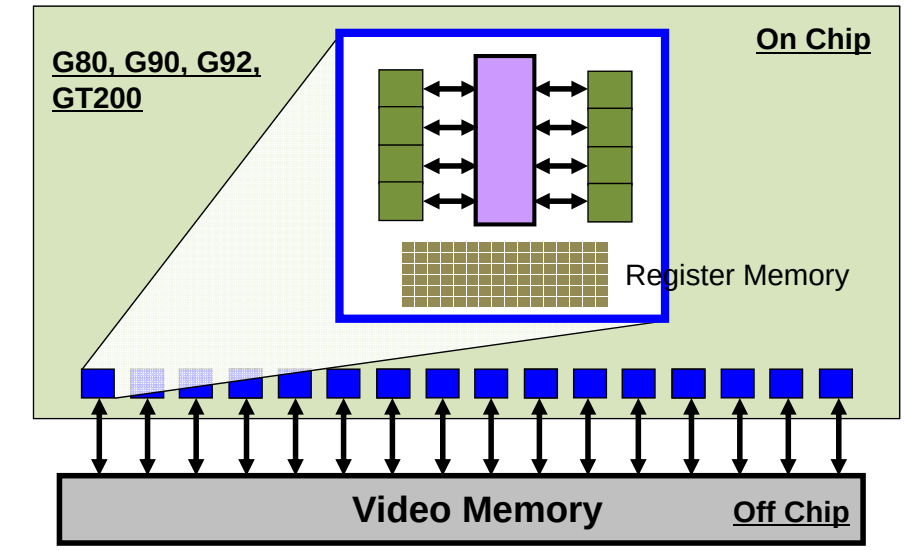
Jen-Hsun Huang

David B. Kirk

「c言語で利用できる膨大な浮動小数点並列プロセッサ」

◎極めて粒度の小さなマルチスレッディング
◎ライトバック制御が可能なキャッシュ
◎CUDA でプログラミング可能
◎スカラ型のIEEE 754“準拠”ストリームプロセッサ

NVIDIA G80 Architecture



Global memory	~4GB (VRAM
Streaming Multiprocessor	~30 (GTX280(GT200))
Shared memory	16 Kbyte
Streaming Processor	8 per SM, total 240

NVIDIA 8000 Series

	GeForce 8800 GT	GeForce 8800 Ultra	GeForce 8800 GTX	GeForce 8600 GTS	GeForce 8500 GT
コードネーム	G92	G8x	G80	G84	G86
API(HWで実装)	DirectX 10 Shader 4.0	DirectX 10 Shader 4.0	DirectX 10 Shader 4.0	DirectX 10 Shader 4.0	DirectX 10 Shader 4.0
コアクロック	600MHz	612MHz	575MHz	675MHz	450MHz
シェーダクロック	1,500MHz	1,500MHz	1,350MHz	1,450MHz	900MHz
メモリクロック	900MHz	1,080MHz	900MHz	1,000MHz	400MHz
メモリ転送レート	1,800MHz	2,160MHz	1,800MHz	2,000MHz	800MHz
メモリインターフェイス幅	256bit	384bit	384bit	128bit	128bit
メモリ帯域(GB/Sec)	57.6GB/sec	103.7GB/sec	86.4GB/sec	32.0GB/sec	12.8GB/sec
標準搭載メモリ量	512MB	768MB	768MB	256MB	
Stream Processor数	112	128	128	32	16
GFLOPS(SP)	336.0 GFLOPS	384.0 GFLOPS	345.6 GFLOPS	92.8 GFLOPS	28.8 GFLOPS
PCI Express	Gen2	Gen1	Gen1	Gen1	Gen1
製造プロセス技術	65nm	90nm	90nm	80nm	80nm
トランジスタ数	754M	681M	681M	289M	210M
ダイサイズ		470平方mm	470平方mm	160平方mm	122平方mm

GeForce GTX280, Tesla C1060 S1070

(2008)

GP GPU

1TFLOPSの浮動小数点演算(単精度)パフォーマンスを達成

GT200 core

◎倍精度浮動小数点演算のサポート
（倍精度浮動小数点演算パフォーマンスは
90GFLOPSと、単精度の12分の1の性能）

◎SP(Streaming Processor)数を
倍増 128 → 240

◎CUDA 2.0

	Tesla C1060	GeForce 8800 GTX
コードネーム	GT200	G80
コアクロック	602MHz	575MHz
シェーダクロック	1,300MHz	1,350MHz
メモリクロック	800MHz	900MHz
メモリ転送レート	1,600MHz	1,800MHz
メモリインターフェイス幅	512bit	384bit
メモリ帯域(GB/Sec)	102GB/sec	86.4GB/sec
Shared メモリ	16kB	16kB
標準搭載メモリ量	4096MB	768MB
Stream Processor数	240	128
GFLOPS(SP)	936 GFLOPS	518 GFLOPS
PCI Express	Gen2	Gen1
製造プロセス技術	65nm	90nm
トランジスタ数	14億	6億
ダイサイズ	550平方mm	470平方mm

TSUBAME 1.2 (2008.10)

Voltaire ISR9288 Infiniband x8
10Gbps x2 ~1310+50 Ports
~13.5Terabits/s
(3Tbits bisection)

NEC SX-8i

500GB
48disks

Storage
1.5 Petabyte (Sun x4500 x 60)
0.1Petabyte (NEC iStore)
Lustre FS, NFS, CIF, WebDAV (over IP)
60GB/s aggregate I/O BW

10Gbps+External NW

Unified Infiniband network

10,000 CPU Cores
300,000 SIMD Cores
~900TFlops-SFP,
~170TFlops-DFP
80TB/s Mem BW (x2 ES)

Sun x4600 (16 Opteron Cores)
32~128 GBytes/Node
10480core/655Nodes
21.4TeraBytes
50.4TeraFlops
OS Linux (SuSE 9, 10)
NAREGI Grid MW

GCOE TSUBASA
Harpertown-Xeon
90Node 720CPU
8.2TeraFlops

NEW: co-TSUBAME
72Node 586CPU (Low Power)
~5TeraFlops

PCI-e

ClearSpeed CSX600
SIMD accelerator
360 648 boards,
35 52.2TeraFlops

NVIDIA

Nvidia Tesla S1070: 170, total 680 GPU cards
High Performance in Many BW-Intensive Apps
10% power increase over TSUBAME 1.0 (130TF SFP / 80TF DFP)

NVIDIA Fermi GPU (2010)

GP GPU

◎「Streaming Multiprocessor(SM)」の改革

- プロセッサコア「CUDA Core(SP)」の数をSM当たり8個から32個に
- 倍精度浮動小数点演算の性能を単精度の1/2に拡張
  (従来は1/8)
- SM内部のライタブルメモリを16KBから64KBに拡張しコンフィギュラブルに(L1キャッシュ/Shard Memory)

◎メモリサブシステムとメモリ階層の拡張

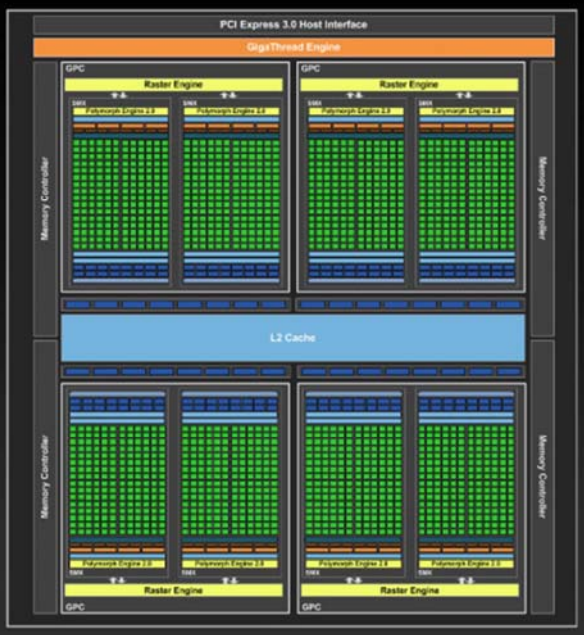
- L1/L2キャッシュ階層をリードオンリーからライタブルに
- オンチップも含めて全てのメモリでECCをサポート

- IEEE 754-2008スタンダードに単精度/倍精度とも対応
- 複数カーネルプログラムの同時実行が可能

Kepler GPU

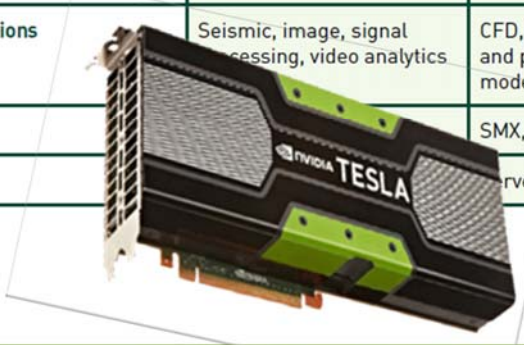
Kepler Block Diagram

- 8 SMX
- 1536 CUDA Cores
- 8 Geometry Units
- 4 Raster Units
- 128 Texture Units
- 32 ROP units
- 256-bit GDDR5



Kepler GPUs

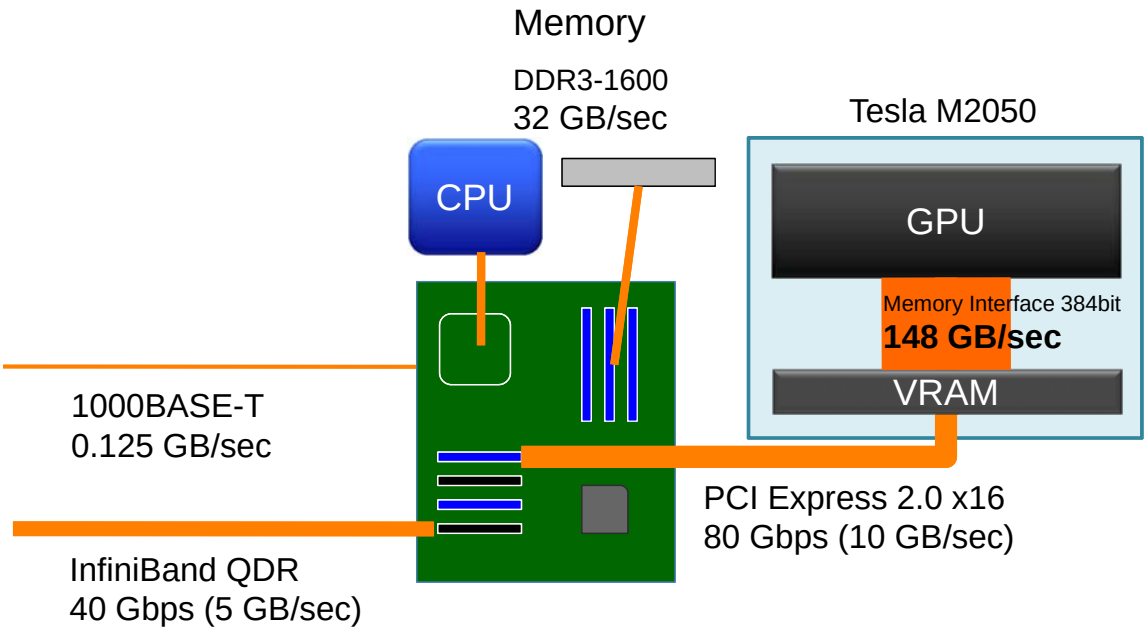
TECHNICAL SPECIFICATIONS	TESLA K10[a]	TESLA K20	TESLA K20X
Peak double precision floating point performance (board)	0.19 teraflops	1.17 teraflops	1.31 teraflops
Peak single precision floating point performance (board)	4.58 teraflops	3.52 teraflops	3.95 teraflops
Number of GPUs	2 x GK104s	1 x GK110	
Number of CUDA cores	2 x 1536	2496	2688
Memory size per board (GDDR5)	8 GB	5 GB	6 GB
Memory bandwidth for board (ECC off)[b]	320 GBytes/sec	208 GBytes/sec	250 GBytes/sec
GPU computing applications	Seismic, image, signal ...cessing, video analytics	CFD, CAE, financial computing, computational chemistry and physics, data analytics, satellite imaging, weather modeling	
Architecture features		SMX, Dynamic Parallelism, Hyper-Q	
System		...rvers and Workstations	Servers only



Tesla K20/K20x

Heterogeneous Computer

■ Several Bandwidth Bottle Necks



TSUBAME 2.0 : Thin Nodes

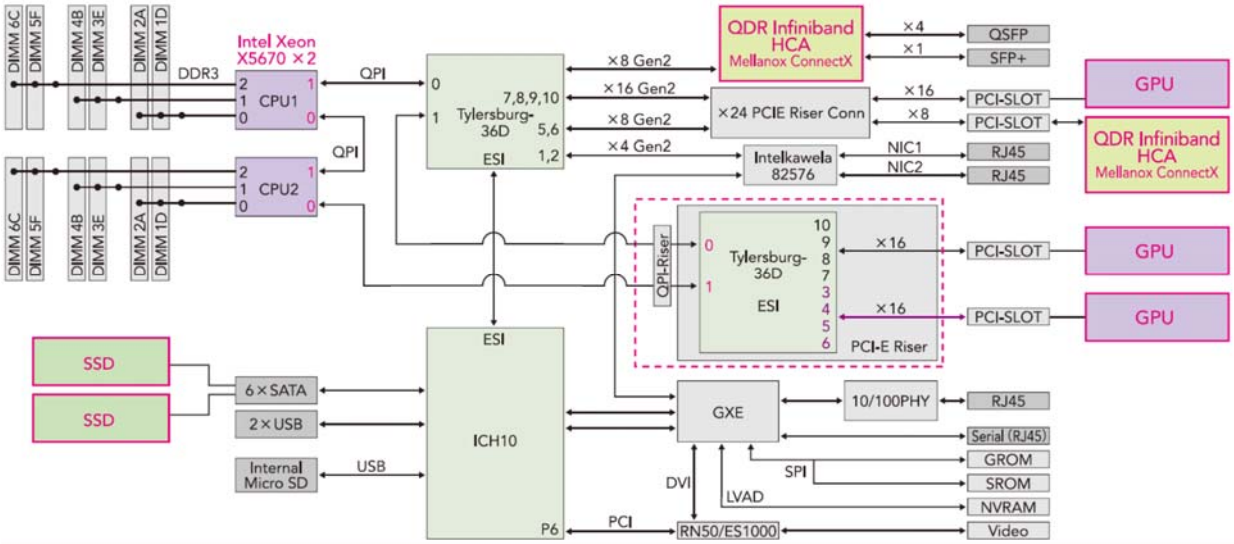
HP ProLiant SL390s

GPU : NVIDIA Tesla M2050 (Fermi Core) ×3 515GFLOPS VRAM 3GB/GPU

CPU : Intel Xeon X5670 2.93Ghz ×2
6 core/socket 76.7 GFLOPS (12cores/node) ※ Turbo boost: 3.196GHz

Memory : 58GB DDR3 1333MHz 一部 103GB

SSD : 60GB ×2 (120GB/node) 一部 120GB ×2 (240GB/node)



TSUBAME2.0: Bandwidthtest

TSUBAME2.0 に login:

```
$ cd /opt/cuda/5.0/samples/1_Utilities/bandwidthTest>
$ ./bandwidthTest
```

TSUBAME2.0: Bandwidthtest

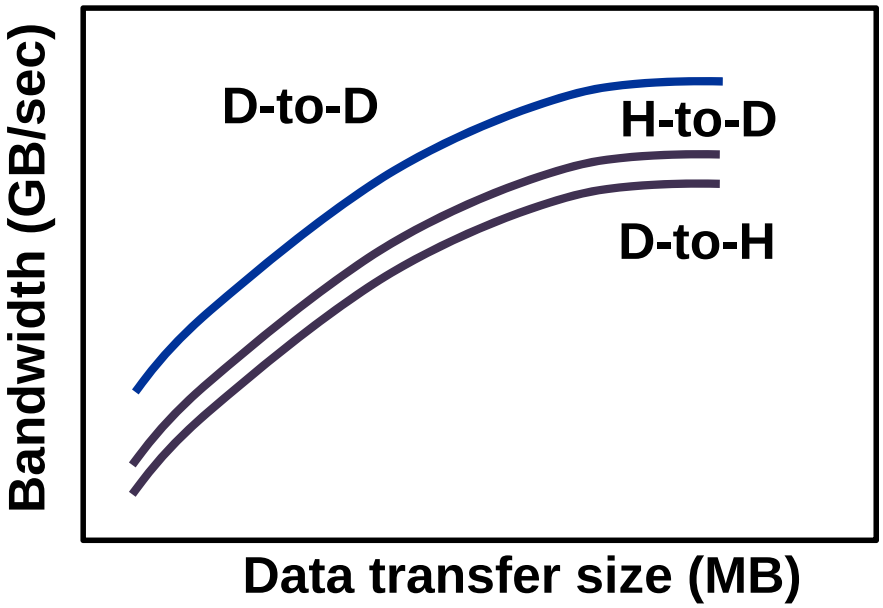

```
[CUDA Bandwidth Test] - Starting...
Running on...

 Device 0: Tesla M2050
 Quick Mode

 Host to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(MB/s)
   33554432                     5546.5

 Device to Host Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(MB/s)
   33554432                     4325.6

 Device to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(MB/s)
   33554432                     108451.0
```


TSUBAME2.0: Bandwidthtest


```
$ ./bandwidthTest --help

[CUDA Bandwidth Test] - Starting...
Usage:  bandwidthTest [OPTION]...
Test the bandwidth for device to host, host to device, and device to device transfers

Example:  measure the bandwidth of device to host pinned memory copies in the range 1024 Bytes
to 102400 Bytes in 1024 Byte increments
./bandwidthTest --memory=pinned --mode=range --start=1024 --end=102400 --increment=1024 --dtoh

Options:
--help  Display this help menu
--csv   Print results as a CSV
--device=[deviceno]     Specify the device device to be used
  all - compute cumulative bandwidth on all the devices
  0,1,2,...,n - Specify any particular device to be used
--memory=[MEMMODE]      Specify which memory mode to use
  pageable - pageable memory
  pinned   - non-pageable system memory
--mode=[MODE]   Specify the mode to use
  quick - performs a quick measurement
  range - measures a user-specified range of values
  shmoo - performs an intense shmoo of a large range of values
--htod  Measure host to device transfers
--dtoh  Measure device to host transfers
--dtod  Measure device to device transfers
--wc    Allocate pinned memory as write-combined
--cputiming     Force CPU-based timing always
Range mode options
--start=[SIZE]  Starting transfer size in bytes
--end=[SIZE]    Ending transfer size in bytes
--increment=[SIZE]      Increment size in bytes
```

TSUBAME2.0: Bandwidthtest

課題：転送データサイズ横軸、実行バンド幅を縦軸にしたグラフを作成せよ。（横軸log-縦軸linear）



TSUBAME2.0: Bandwidthtest

Host 側のメモリについては、pageable と pinned の両方を試すこと。

TSUBAME の GPU (M2050) で無くてもよい。その場合はGPUの詳細とHostのハードウェア構成を示すこと。

期限：2013年 5月9日（17:00）
場所：学術国際情報センター・国際棟 1F ポスト I7－3

氏名、学籍番号、日付とタイトル (Bandwidthtest) を書き、紙に印刷したものを提出すること。

または、Subject: Bandwidthtest とし、
メールで gpu_report2013@sim.gsic.titech.ac.jp に上記の内容を pdf ファイルとして提出すること。

