Code Optimization on Titan Black and Xeon Phi

Yuree Chung, Yongchull Jang, Hwancheol Jeong,Jangho Kim, Weonjong Lee, Jeonghwan Pak
SWME collaboration, BNL, SNU, UW
June, 23th ∼ June, 28th. 2014

1. GTX Titan Black

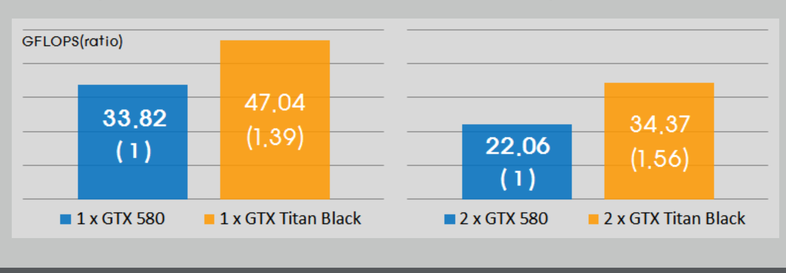
GTX Titan Black

- NVIDIA GTX Titan & Titan black show similar theoretical performance with NVIDIA Tesla K20X.
- However, GTX Titan & Titan black are about 3 times cheaper than Tesla K20X.

	Fermi	Kepler			
	GTX 580	Tesla K20X	GTX TITAN	GTX 780Ti	GTX TITAN black
SP TFLOPs	1.58	3.95	4.50	5.04	5.1
DP TFLOPs	0.20	1.31	1.27	0.21	1.3
Memory Size (GB)	1.5	6	6.1	3	6.1
Memory Bandwidth (GB/sec)	192.4	250	288.4	336	336

CG performance with old code

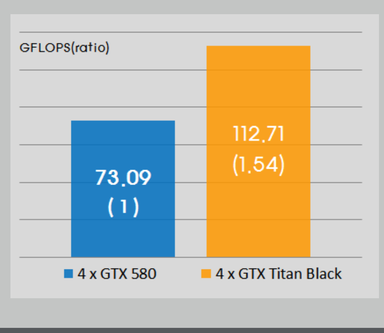
- CG performance for $20^3 \times 64$ MILC asqtad ensemble.



NPR code performance with old code

- Dominant part is one-color four fermion operator calculation.(Jangho's talk on Tue)

$$O^{f_1 f_2 f_3 f_4}_{i;l}(z) = \overline{\chi}^{f_1}_{i;c_1}(z_A)\overline{(\gamma_{S_1} \otimes \xi_{F_1})}_{AB}\chi^{f_2}_{i;c_2}(z_B) \times \overline{\chi}^{f_3}_{i;c_3}(z_C)\overline{(\gamma_{S_2} \otimes \xi_{F_2})}_{CD}\chi^{f_4}_{i;c_4}(z_D) \times [U_{i;AD}]_{c_1 c_4}(z)[U_{i;CB}]_{c_3 c_2}(z)$$



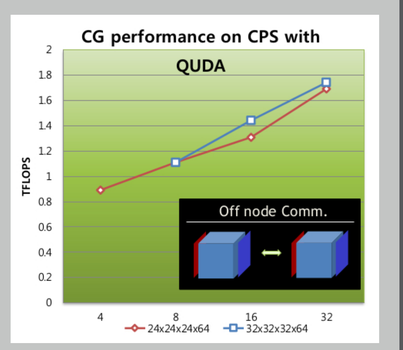
Tuning Schemes for Kepler

	Fermi	Kepler
Simultaneous blocks / SM(X)	8	16
Warp schedulers / SM(X)	2	4
Registers / Thread	63	255
Bandwidth (GB/s)	192 (GTX 580)	336 (Titan black)

- Adjust thread and block numbers
- Adjust shared memory + L1 cache ratio : 16 + 48, 48 + 16, or 32 + 32 (only for Kepler) possible
- Use 48 KB read only data cache
- Warp shuffle : exchange data between threads of a warp without using shared memory, which does not consume shared memory, and provides low latency than shared memory access

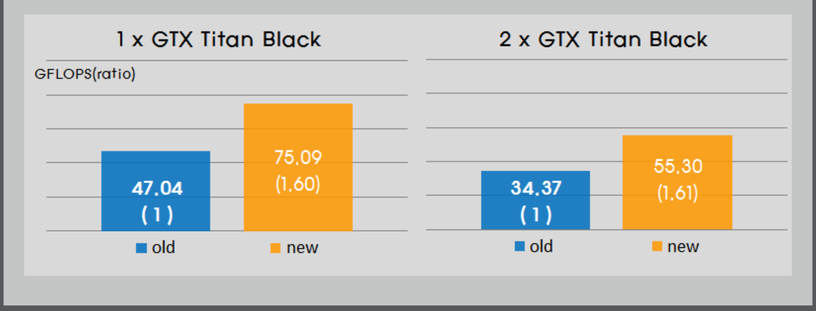
CPS & QUDA for staggered fermion

- QUDA - library for lattice QCD based on CUDA
- High performance CG and BiCG inverters of mixed precision
- We adopted QUDA staggered fermion inverter to CPS



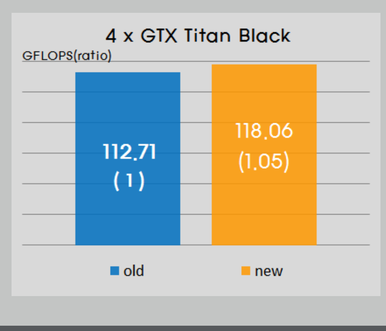
CG Performance with New CPS & QUDA

- CG performance for $20^3 \times 64$ MILC asqtad ensemble



NPR Code performance with new CPS

- Using our CUDA code (not using QUDA)



Optimization Strategy

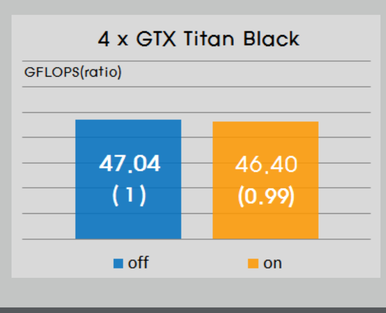
- CUDA 6.0 & CUDA compatibility 3.5
- Tune thread and block numbers and other tuning parameters (e.g. warp scheduler, number of shared memory variable)
- Dynamic Parallelism
- GPU Direct
- Controlling double precision performance mode

Double Precision Performance Mode

- GTX Titan & Titan black provides double precision performance mode
- However, single precision performance decreases a little bit when this mode is turned on
- Mixed precision code can not gain the full performance

Double Precision Performance Mode

- For mixed precision CG,



2. Xeon Phi coprocessor

Xeon Phi

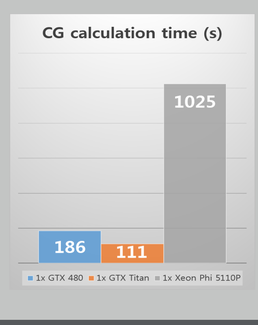
- Many CPU cores are put on a PCIe card as GPU's CUDA cores
- Easier to import a usual C code than GPU, which requires CUDA
  - OpenMP, OpenCL, MPI are available
- Provides 512 bit SIMD register
  - Simultaneous 8 double precision operations
  - Vectorization is very important.

	Intel Xeon Phi		NVIDIA GPU	
	7110X	5110P	Tesla K20X	GTX Titan Black
SP TFLOPS	2.44	2.02	3.95	5.1
DP TFLOPS	1.22	1.01	1.31	1.3
Memory Size (GB)	16	8	6	6.1
Mem. Bandwidth (GB/s)	352	320	250	336
Price (USD)	4130	2650	3800	1100

CG Performance with Old Code

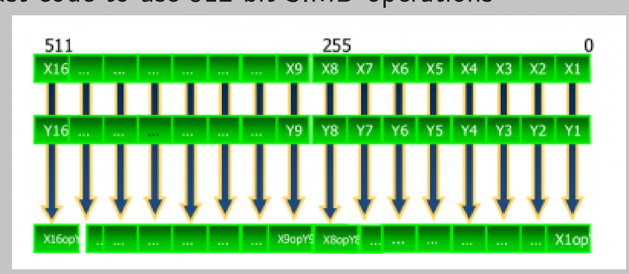
- CG calculation time for $20^3 \times 64$ MILC asqtad ensemble (only MPI is used with 200 Xeon Phi processors)

Device	CG time (ratio)
GTX 480	186 s (1)
GTX Titan	111 s (0.60)
Xeon Phi 5110P	1025 s (5.5)



Optimization Strategy

- Adjust code to use 512 bit SIMD operations



- Use OpenMP instead of MPI
  - Different from MPI, OpenMP provides memory sharing between processes
- New CPS is not compiled well on Xeon Phi.
  - Not support Intel compiler option
  - Not support 512bit SIMD vectorization

Conclusion

1. GTX Titan Black improves the performance of the LQCD GPU code by 40% ∼ 60% from that of GTX 580.

2. We achieved 60% better performance by adapting the latest CPS library and QUDA for staggered fermion.

3. We need to optimize our code by using 512 bit SIMD vectorization and OpenMP to obtain a proper performance from Xeon Phi.

jjome0@gmail.com