

Supplementary Material for Performance Optimization Using Partitioned SpMV on GPUs and Multicore CPUs

Wangdong Yang, Kenli Li, Zeyao Mo, and Keqin Li, *Senior Member, IEEE*

1 COMPUTING POWER

The computing power of CPU and GPU are shown in Table 1.

2 HETEROGENEOUS COMPUTING SYSTEMS WITH CPUs AND GPUS

GPUs are widely used in parallel computing, because there are more cores and co-processors to speed the computing in more and more computing systems. How to improve the computing performance of tasks in heterogeneous computing systems with two different types of processors? The keys are algorithm design and task allocation according to the characteristics of the different types of processors. GPUs are controlled on a host as speed devices, so data and instructions are transported into the device memory from the host memory. For GPUs provided by NVIDIA, NVIDIA provided CUDA (Compute Unified Device Architecture) to improve the development efficiency of parallel program [1]. The CUDA execution model shown in Fig. 1 is a collection of threads running in parallel. The number of threads is decided by the programmer to be executed. A collection of threads (called a block) runs on a multiprocessor at a given time. Multiple blocks can be assigned to a single multiprocessor and their execution is time-shared. For heterogeneous computing systems with CPUs and GPUs, each core of CPU can independently perform its instructions and data, which is called the MIMD model. If there is not dependency, it does not need synchronization between threads on various cores of CPU. For multicore CPUs, each core can be scheduled

TABLE 1
The Computing Power of CPU and GPU

Sparse Matrix	Dimensions	NNZ	GPU : A Core of CPU
ukerbe1	5981*5981	7852	69.71 : 1
PGPgiantcompo	10680*10680	24316	71.75 : 1
whitaker3_dual	19190*19190	28581	70.73 : 1
wb-cs-stanford	9914*9914	36854	72.83 : 1
circuit_2	4510*4510	21199	74.85 : 1
sts4098	4098*4098	38227	75.88 : 1
bayer03	6747*6747	59196	74.91 : 1
circuit_3	12127*12127	48137	72.93 : 1
big	13209*13209	91465	75.96 : 1
CurlCurl_0	11083*11083	62213	73.94 : 1
thermal	3456*3456	66528	72.97 : 1
lpl1	32460*32460	180248	76.98 : 1
ted_B	10605*10605	77592	73.95 : 1
cz10228	10228*10228	102876	74.98 : 1
bayer04	20545*20545	159082	75.01 : 1
raefsky5	6316*6316	168658	76.02 : 1
vibrobox	12328*12328	177578	75.04 : 1
coater2	9540*9540	207308	77.07 : 1
Kuu	7102*7102	173651	78.05 : 1
pcrystk02	13965*13965	491274	76.08 : 1
crplat2	18010*18010	489478	76.96 : 1
epb2	25228*25228	17027	74.91 : 1
goodwin	7320*7320	324784	78.09 : 1
pkustk01	22044*22044	500712	77.02 : 1
G_n_pin_pout	100000*100000	501198	77.03 : 1
Andrews	60000*60000	410077	75.04 : 1
ky2010	161672*161672	393889	74.02 : 1
circuit_4	80209*80209	307604	75.98 : 1
olesnik0	88263*88263	402623	76.03 : 1
bayer01	57739*57739	277774	77.98 : 1
Average Proportion			75.09 : 1
Standard Deviation			2.06

- Wangdong Yang, Kenli Li, and Keqin Li are with the College of Information Science and Engineering, Hunan University, Changsha, Hunan 410008, China; and the National Supercomputing Center in Changsha, Changsha, Hunan 410008, China.
E-mail: yangwangdong@163.com; lkl@hnu.edu.cn.
- Zeyao Mo is with the Institute of Applied Physics and Computational Mathematics, Beijing 100094, China.
E-mail: zeyao_mo@iapcm.ac.cn.
- Keqin Li is also with the Department of Computer Science, State University of New York, New Paltz, New York 12561, USA.
E-mail: lik@newpaltz.edu.

independently to perform the threads. So the partitioning methods for multicore CPU do not consider the uniformity of NNZ of rows in the block for SpMV. But the basic computing unit of a GPU is called streaming multiprocessor (SM). As a component at the bottom of the independent hardware structure, SM can be seen as an SIMD processing unit. Each SM contains some scalar processors (SP) and special function units (SFU). It needs synchronization between SPs of the same SM on GPU.

TABLE 2
Parameters of the Test Computer

Parameter	Description	Value
S_i	the size of an integer	4 bytes
S_s	the size of single-precision	4 bytes
S_d	the size of double-precision	8 bytes
C	number of stream processors	336
F	the clock rate of sp	1.5 Gflop/s
BW	bus width of global memory	256 bits
CR	clock rate of global memory	1.9 GHz
W	number of threads per warp	32
nc	number of cores per CPU	4
fc	the frequency of CPU	2.13 GHz

So the inequality of the load of different threads will have a larger impact on performance.

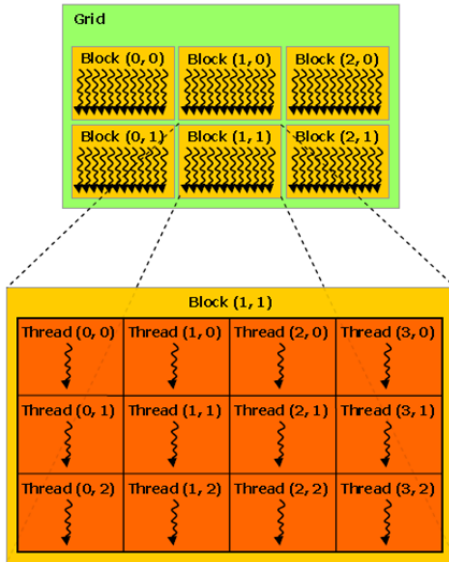


Fig. 1. The CUDA execution model.

3 ADDITIONAL EXPERIMENTAL RESULTS

3.1 Experiment Settings

The hardware parameters of the testing computer are shown in Table 2.

3.2 General Information of the Sparse Matrices Used in the Evaluation

We chose 50 sparse matrices to test for Algorithms 1, 2, 3, and 4 from the UF Sparse Matrix Collection [2]. The main features of all tested sparse matrices are shown in Table 3. The probability distributions of 10 representative tested sparse matrices are shown in Table 4, where the X axis is the NNZ of rows and the Y axis is the number of rows with x NNZ . The main features of 10 representative tested sparse matrices are shown in Table 5. Most of these matrices are derived from scientific computing and real engineering applications.

TABLE 3
General Information of All Sparse Matrices Used in the Evaluation

No.	Sparse Matrix	Dimensions	NNZ	Structure
1	PR02R	161070*161070	8185136	unsymmetric
2	pwtk	217918*217918	5926171	symmetric
3	TSOPF_RS_b300_c3	42138*42138	4413449	symmetric
4	af_shell10	1508065*1508065	27090195	symmetric
5	rajat31	4690002*4690002	20316253	unsymmetric
6	ohne2	181343*181343	11063545	unsymmetric
7	nlpkt120	3542400*3542400	50194096	symmetric
8	nlpkt160	8345600*8345600	118931856	symmetric
9	nlpkt200	16240000*16240000	232232816	symmetric
10	cage15	5154859*5154859	99199551	unsymmetric
11	atmosmod1	1489752*1489752	10319760	unsymmetric
12	M6	3501776*3501776	10501936	symmetric
13	hugetric-00020	7122792*7122792	10680777	symmetric
14	333SP	3712815*3712815	11108633	symmetric
15	AS365	3799275*3799275	11368076	symmetric
16	human_gene1	22283*22283	12345963	symmetric
17	NLR	4163763*4163763	12487976	symmetric
18	delaunay_n22	4194304*4194304	12582869	symmetric
19	asia_osm	11950757*11950757	12711603	symmetric
20	adaptive	6815744*6815744	13624320	symmetric
21	road_central	14081816*14081816	16933413	symmetric
22	Freescape1	3428755*3428755	18920347	unsymmetric
23	wikipedia-20051105	1634989*1634989	19753078	unsymmetric
24	CoupCons3D	416800*416800	22322336	unsymmetric
25	hugetrace-00020	16002413*16002413	23998813	symmetric
26	delaunay_n23	8388608*8388608	25165784	symmetric
27	road_usa	23947347*23947347	28854312	symmetric
28	Serena	1391349*1391349	32961525	symmetric
29	bone010	986703*986703	36326514	symmetric
30	channel-500x100 x100-b050	4802000*4802000	42681372	symmetric
31	delaunay_n24	16777216*16777216	50331601	symmetric
32	europe_osm	50912018*50912018	54054660	symmetric
33	wb-edu	9845725*9845725	57156537	unsymmetric
34	Flan_1565	1564794*1564794	59485419	symmetric
35	circuit5M	5558326*5558326	59524291	unsymmetric
36	Cube_Coup_dt0	2164760*2164760	64685452	symmetric
37	kron_g500-logn21	2097152*2097152	91042010	symmetric
38	ML_Geer	1504002*1504002	110879972	unsymmetric
39	torso1	116158*116158	8516500	unsymmetric
40	Si41Ge41H72	185639*185639	7598452	symmetric
41	pkustk14	151926*151926	7494215	symmetric
42	CurlCurl_3	1219574*1219574	7382096	symmetric
43	rgg_n_2_20_s0	1048576*1048576	6891620	symmetric
44	hugetrace-00000	4588484*4588484	6879133	symmetric
45	BenElechi1	245874*245874	6698185	symmetric
46	halfb	224617*224617	6306219	symmetric
47	troll	213453*213453	6099282	symmetric
48	thermal2	1228045*1228045	4904179	symmetric
49	t2em	921632*921632	4590832	unsymmetric
50	tmt_unsym	917825*917825	4584801	unsymmetric

3.3 The Relative Difference (%) of Partitioning in Test 1 and Test 2 Using Algorithms 1, 2, and 3

The relative difference of a partition over computing powers is calculated by

$$\sum_{i=1}^K (NNZ(A_i) - CP_i \times NNZ) / (CP_i \times NNZ) \times 100\%,$$

where CP_i is the ratio which the i th RVS should take in a partition. For all tested sparse matrices, the average relative difference of Test 1 using Algorithms 1, 2,

TABLE 4
Probability Distributions of 10 Representative Sparse Matrices

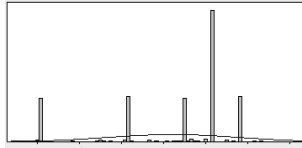
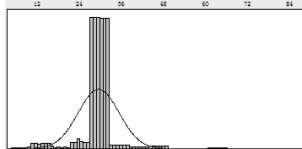
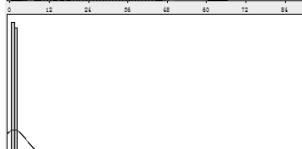
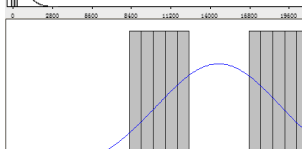
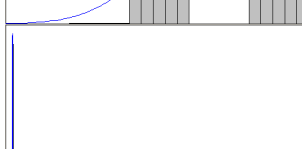


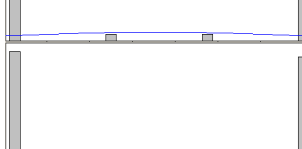
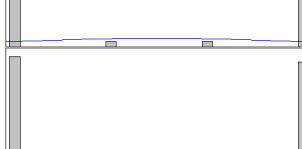
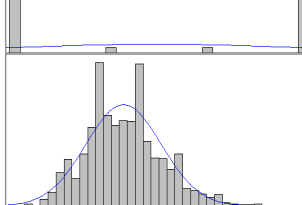
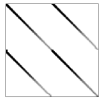
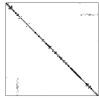



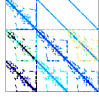
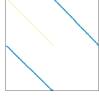
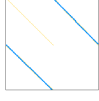

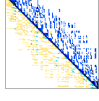
Sparse Matrix	Probability Distribution
Fluorem/PR02R	
Boeing/pwtk	
TSOPF/TSOPF_RS_b300_c3	
Schenk_AFE/af_shell10	
Rajat/rajat31	
Schenk_ISEI/ohne2	
Schenk/nlpkkt120	
Schenk/nlpkkt160	
Schenk/nlpkkt200	
vanHeukelum/cage15	

TABLE 5
General Information of 10 Representative Sparse Matrices Used in the Evaluation

No.	Sparse Matrix	Dimensions	NNZ	$E(X)$	$\max(i p_i > 0)$	Variance	Characteristics
1	Fluorem/PR02R	161070*161070	8185136	50.817	88	349.122	
2	Boeing/pwtk	217918*217918	5926171	27.194	90	39.039	
3	TSOPF/TSOPF_RS_b300_c3	42138*42138	4413449	104.7	20702	700463.1	
4	Schenk_AFE/af_shell10	1508065*1508065	27090195	17.964	25	27.294	
5	Rajat/rajat31	4690002*4690002	20316253	4.33	1252	1.22	
6	Schenk_ISEI/ohne2	181343*181343	11063545	61.01	3441	444.59	
7	Schenk/nlpkkt120	3542400*3542400	50194096	14.17	28	174.326	
8	Schenk/nlpkkt160	8345600*8345600	118931856	14.251	28	176.274	
9	Schenk/nlpkkt200	16240000*16240000	232232816	14.3	28	178.876	
10	vanHeukelum/cage15	5154859*5154859	99199551	19.244	47	32.910	

and 3 are 17.3600%, 0.0348%, and 0.0119% respectively, and that of Test 2 are 25.5616%, 0.0326%, and 0.0054% respectively, as shown in Table 6.

3.4 The Density of Algorithms 1, 2, and 3 in Test 2

The densities of all tested cases using Algorithms 1, 2, and 3 are shown in Table 7. The average densities of all tested cases using Algorithms 1, 2, and 3 are approximately 11%, 40%, and 61% respectively, as shown in Table 7.

3.5 The Time and Flop-rate of SpMV by Partitioning Using Algorithms 1, 2, 3, and 4

We define *flop-rate* as the ratio of computing scale to computing time. The scale of the computation for SpMV is $2 \times NNZ$, because each non-zero element should perform a multiplication and an addition operations. The *flop-rate* is calculated by $2 \times NNZ / T \times 10^{-9}$, where T is the computing time of SpMV (the unit is second). The unit of *flop-rate* is Gops, which is giga operations per second.

In Tables 8 and 9, A.1, A.2, A.3, and A.4 are Algorithm 1, Algorithm 2, Algorithm 3, and Algorithm 4 respectively.

It is observed from Tables 8 and 9 that (1) For all tested sparse matrices, the average *flop-rate* of SpMV im-

TABLE 6
The Relative Difference (%) of Partitioning in Test 1 and Test 2

Sparse Matrix No.	Test 1			Test 2		
	Algori. 1	Algori. 2	Algori. 3	Algori. 1	Algori. 2	Algori. 3
1	0.015137	0.000050	0.000075	0.022046	0.000094	0.000090
2	0.042082	0.000044	0.000038	0.071559	0.000058	0.000113
3	0.117656	0.001280	0.000845	0.930066	0.000621	0.000631
4	0.000210	0.000001	0.000001	0.003845	0.000015	0.000009
5	0.000058	0.000000	0.000000	0.001071	0.000007	0.000002
6	0.005187	0.000007	0.000006	0.035364	0.000093	0.000070
7	0.112423	0.000000	0.000000	0.851116	0.000008	0.000005
8	0.113389	0.000000	0.000000	0.853570	0.000002	0.000002
9	0.113976	0.000000	0.000000	0.855064	0.000002	0.000003
10	0.015706	0.000000	0.000000	0.187229	0.000005	0.000003
11	0.011358	0.000004	0.000005	0.017409	0.000013	0.000004
12	0.435861	0.000004	0.000002	0.595223	0.000008	0.000001
13	0.211904	0.000001	0.000001	0.311920	0.000002	0.000000
14	0.250463	0.000002	0.000002	0.122823	0.000007	0.000005
15	0.434957	0.000003	0.000001	0.587905	0.000003	0.000003
16	0.419225	0.001259	0.000675	0.623684	0.001186	0.000535
17	0.431119	0.000002	0.000002	0.575119	0.000005	0.000002
18	0.375083	0.000002	0.000002	0.269297	0.000003	0.000001
19	0.039969	0.000000	0.000000	0.066765	0.000003	0.000001
20	0.249868	0.000000	0.000000	0.249038	0.000002	0.000000
21	0.172779	0.000001	0.000000	0.259518	0.000001	0.000000
22	0.119557	0.000003	0.000002	0.174062	0.000009	0.000003
23	0.923503	0.000326	0.000341	0.788777	0.003118	0.000016
24	0.119583	0.000019	0.000011	0.138967	0.000028	0.000034
25	0.109162	0.000001	0.000000	0.175707	0.000001	0.000000
26	0.375055	0.000001	0.000001	0.268978	0.000004	0.000001
27	0.082774	0.000000	0.000000	0.241015	0.000001	0.000000
28	0.013059	0.000007	0.000006	0.017596	0.000015	0.000009
29	0.019110	0.000005	0.000007	0.025199	0.000016	0.000020
30	0.002806	0.000001	0.000001	0.006479	0.000003	0.000001
31	0.374930	0.000001	0.000001	0.269039	0.000001	0.000000
32	0.062636	0.000000	0.000000	0.086178	0.000000	0.000000
33	0.078374	0.000005	0.000024	0.119071	0.000226	0.000002
34	0.010041	0.000003	0.000003	0.016943	0.000014	0.000011
35	0.932942	0.013049	0.002542	0.573102	0.006282	0.000004
36	0.010478	0.000003	0.000002	0.017083	0.000011	0.000007
37	0.504595	0.000091	0.000213	0.676008	0.000269	0.000073
38	0.001208	0.000004	0.000007	0.001953	0.000010	0.000005
39	0.230542	0.000924	0.000716	0.201886	0.003434	0.000619
40	0.310538	0.000114	0.000231	0.387905	0.000231	0.000089
41	0.285593	0.000071	0.000053	0.301795	0.000159	0.000100
42	0.012601	0.000005	0.000009	0.020143	0.000020	0.000013
43	0.002879	0.000009	0.000004	0.006292	0.000013	0.000013
44	0.249393	0.000003	0.000001	0.278421	0.000002	0.000000
45	0.003877	0.000027	0.000019	0.009020	0.000060	0.000035
46	0.187733	0.000031	0.000048	0.322340	0.000087	0.000066
47	0.018385	0.000031	0.000041	0.020001	0.000092	0.000091
48	0.067612	0.000007	0.000005	0.143178	0.000014	0.000004
49	0.002026	0.000005	0.000003	0.003428	0.000025	0.000002
50	0.000598	0.000005	0.000000	0.000618	0.000017	0.000011
Average relative difference	0.173600	0.000348	0.000119	0.255616	0.000326	0.000054

proves by 42.51%, 2.99%, and 62.03% for single precision, 68.83%, 0.38%, and 72.51% for double precision by using Algorithm 3 compared with Algorithms 1, 2, and 4 in Test 1. (2) For all tested sparse matrices, the average *flop-rate* of SpMV improves by 65.22%, 36.17%, and 76.15% for single precision, 97.27%, 22.14%, and 93.80% for double precision by using Algorithm 3 compared with Algorithms 1, 2, and 4 in Test 2.

[2] T. A. Davis and Y. Hu University of Florida sparse matrix collection[J]. 2009.

REFERENCES

[1] NVIDIA CUDA C Programming Guide, Version 5.0, May 2012.

TABLE 7
The Density of Algorithms 1, 2, and 3 in Test 2

No.	Algori. 1	Algori. 2	Algori. 3
1	0.016467	0.577469	0.836821
2	0.003767	0.400875	0.568791
3	0.009656	0.076387	0.095698
4	0.028555	0.718542	0.899287
5	0.010111	0.010044	0.010319
6	0.050920	0.035942	0.059114
7	0.530805	0.601302	0.754727
8	0.097234	0.605333	0.747683
9	0.111251	0.607776	0.751522
10	0.178458	0.513652	0.730733
11	0.080523	0.989595	0.995940
12	0.005663	0.323769	0.709427
13	0.005557	0.499840	0.964949
14	0.057949	0.204244	0.359890
15	0.018906	0.297600	0.622746
16	0.012717	0.115207	0.524212
17	0.007771	0.235636	0.497643
18	0.005866	0.212730	0.538478
19	0.006747	0.138745	0.345847
20	0.010004	0.613229	0.813425
21	0.040600	0.162810	0.589617
22	0.027951	0.340085	0.481992
23	0.021558	0.029766	0.038766
24	0.016558	0.729891	0.885831
25	0.021923	0.499900	0.959493
26	0.051752	0.193704	0.501773
27	0.017207	0.159477	0.554522
28	0.047216	0.128357	0.319518
29	0.017472	0.876573	0.958250
30	0.030930	0.745519	0.882022
31	0.248478	0.205885	0.446343
32	0.068775	0.101551	0.290844
33	0.017841	0.045250	0.439920
34	0.039358	0.552926	0.797870
35	0.070571	0.028605	0.151519
36	0.071640	0.607658	0.810405
37	0.062469	0.042829	0.291227
38	0.920242	0.996261	0.996261
39	0.004731	0.059927	0.229719
40	0.103300	0.094806	0.339725
41	0.005774	0.247019	0.595847
42	0.034795	0.672557	0.914443
43	0.012863	0.296662	0.627111
44	0.425903	0.499739	0.949576
45	0.908115	0.908078	0.972066
46	0.339841	0.297948	0.606061
47	0.091446	0.109474	0.263385
48	0.376127	0.385026	0.701162
49	0.005902	0.996240	0.996242
50	0.003543	0.999058	0.999058
Average density	0.107076	0.395830	0.608356

