Parallel Processing Project: Max Flops

Aman Hogan-Bailey

University of Texas at Arlington

CSE-5351: Parallel Processing

Contents

Theoretical Maximum Flops for a Single Core (Part 1)

To determine the maximum GFLOPS of a single core of the Intel Xeon E5-2680 v4, the processor's clock speeds, and floating-point capabilities are needed. The `lscpu` command displays the base clock of 2.40 GHz and a maximum turbo frequency of 3.30 GHz. The Xeon E5-2680 v4 supports AVX2, allowing for 8 single-precision floating point operations per cycle. So, the formula for max FLOPS is as follows:

$$Max\ Flops = MaxFrequency * \#DoublesAVX * FMA * \#AVXUnits$$

For the base clock speed of 2.40 GHz, the calculation is:

$$38.4\ GFLOPS = 2.40 * 4 * 2 * 2$$

For the max clock speed using turbo frequency of 3.30 GHz, the calculation is:

$$52.8\ GFLOPS = 3.30 * 4 * 2 * 2$$

Max Flops Testing

The experiments use level 3 optimization. I ran the experiments on the UTA bell cluster and my PC using matrices of 3072x3072.

Max Flops using Loop Orders (Part 2)

In the experiments, the `ikj` loop ordering achieved the fastest execution times (1.02 seconds, 7.06 GFLOPS) and the highest GFLOPS compared to other loop permutations like `ijk` (Table 1). These results are expected, given how the CPU cache hierarchy functions and how the compiler optimizes. The `ikj` ordering accesses rows of matrix A and C sequentially, leveraging continuous row-major order memory. This structure enhances spatial locality and ensures contiguous memory accesses likely to remain in the cache. Additionally, it accesses elements of matrix B in a manner that maximizes temporal locality by reusing cached data across multiple iterations of the innermost loop. This effective cache utilization minimizes cache misses and reduces the reliance on slower main memory accesses, leading to improved data reuse and faster computation.

Max Flops with Cache Tiling (Part 3)

The analysis of the GEBP algorithm using various block sizes for kc/mc, m_r, and n_r reveals several insights into how these parameters influence performance. The most significant factor affecting performance is the size of kc/mc, which determines how much of the matrices can fit into the cache. As displayed in Table 3, the top three kc/mc sizes were 128x128, corresponding to an A block size of 128 KB, half of the L2 cache size. The top three best performing also had an m_r size of 32, and all had higher than 30 GFLOPS and 60% utilization. In Figure 1, we get a better visualization of this data. From this data, we can extrapolate:

- Any kc/mc above 150x150 had poor performance no matter the m_r or n_r.

- m_r between 20-100 had a high predictor of performance but dropped off after 128.
- n_r can increase performance at low m_rs, but very minimally.
- n_r has no effect at high m_rs.
- The best performance happens at half L2 cache utilization and m_r values between 32-80
- The results are roughly the same when performed on the PC (Figure 2)
- The highest GFLOPS achieved was 33.74 GFLOPS at 64% utilization.

The reason kc/mc is optimal at half the L2 cache is likely because this ensures that the data required for the algorithm is kept in the cache and reused efficiently. Fitting the data neatly in this manner reduces the number of cache misses and traversals to RAM or L3, which would increase the time required.

The reason m_r increases performance is likely due to compiler optimizations and modern CPUS being able to handle several instructions at once. The numbers in the Goto and Van de Geijn paper for m_r were significantly smaller, so these might be reasonable explanations for differing m_r values apart from unintended side effects like register spilling.

Conclusions

In conclusion, the experiments reveal insights into the performance characteristics of matrix multiplication when using optimized algorithms like GEBP. Theoretical maximum FLOPS calculations based on the processor's clock speeds show the potential for high computational throughput. In practice, performance peaked when kc/mc sizes were optimized to utilize half of the L2 cache, which minimized cache misses and maximized data reuse, leading to improved memory access efficiency. The best results occurred with m_r values between 32 and 80, indicating that modern CPUs, aided by compiler optimizations, can handle larger register blocking sizes than traditionally expected, possibly due to increased instruction-level parallelism and advanced caching mechanisms. The experiments demonstrate that careful tuning of blocking parameters, especially in relation to cache size and utilization, is crucial for achieving maximum performance, with the highest GFLOPS achieved at 33.74 and 64% utilization.

References

Goto, K., & Geijn, R. A. V. D. (2008). Anatomy of high-performance matrix multiplication. ACM Transactions on Mathematical Software (TOMS), 34(3), 1-25.

Tables

Table 1

Ordering of loops and respective time and Gflops on Bell Cluster

ijk	**6.27223**	**1.155531**
ikj	1.026673	7.059463
kij	1.465431	4.945821
kji	31.64932	0.229002
jki	31.36748	0.23106
jik	6.000089	1.207942
ijk	6.27223	1.155531

Table 2

Ordering of loops and respective time and Gflops on PC

Order	**Seconds**	**GFLOPS**
ijk	20.05508	0.361393
ikj	2.116447	3.424492
kij	3.049967	2.37634
kji	66.02297	0.109776
jki	66.40759	0.109141
jik	19.17403	0.377999

Table 3

Gflops, time, and utilization given kc, nr, and mr sizes on bell cluster sorted by GFLOPS

kc/mc	nr	mr	gflops	time (seconds)	util	A block (KB)	B Sliver (KB)
128	4	32	33.74	1.72	0.64	128/256.0	8/32.0
128	8	32	33.30	1.74	0.63	128/256.0	16/32.0
128	16	32	33.16	1.75	0.63	128/256.0	32/32.0
96	8	96	29.56	1.96	0.56	72/256.0	12/32.0
96	16	96	28.64	2.02	0.54	72/256.0	24/32.0
96	8	32	28.17	2.06	0.53	72/256.0	12/32.0
96	4	96	27.67	2.10	0.52	72/256.0	6/32.0
96	16	32	27.35	2.12	0.52	72/256.0	24/32.0
96	4	32	27.21	2.13	0.52	72/256.0	6/32.0
128	4	16	22.86	2.54	0.43	128/256.0	8/32.0
128	8	16	22.82	2.54	0.43	128/256.0	16/32.0
128	16	16	22.62	2.56	0.43	128/256.0	32/32.0
102	8	96	22.19	2.61	0.42	81/256.0	12/32.0
102	4	96	21.97	2.64	0.42	81/256.0	6/32.0
102	16	96	21.93	2.64	0.42	81/256.0	25/32.0
109	8	96	20.74	2.80	0.39	92/256.0	13/32.0
96	8	16	20.47	2.83	0.39	72/256.0	12/32.0
128	4	96	19.99	2.90	0.38	128/256.0	8/32.0
96	4	16	19.54	2.97	0.37	72/256.0	6/32.0
96	16	16	19.47	2.98	0.37	72/256.0	24/32.0
109	4	96	19.45	2.98	0.37	92/256.0	6/32.0
128	8	96	19.43	2.98	0.37	128/256.0	16/32.0
128	16	96	19.37	2.99	0.37	128/256.0	32/32.0
102	4	32	18.54	3.13	0.35	81/256.0	6/32.0
102	8	32	18.44	3.14	0.35	81/256.0	12/32.0
109	16	96	18.21	3.18	0.34	92/256.0	27/32.0
102	16	32	17.98	3.22	0.34	81/256.0	25/32.0
109	8	32	16.09	3.60	0.30	92/256.0	13/32.0
153	4	32	15.97	3.63	0.30	182/256.0	9/32.0
109	4	32	15.97	3.63	0.30	92/256.0	6/32.0
128	4	8	14.90	3.89	0.28	128/256.0	8/32.0
109	16	32	14.56	3.98	0.28	92/256.0	27/32.0
153	16	32	14.54	3.99	0.28	182/256.0	38/32.0
153	8	32	14.47	4.01	0.27	182/256.0	19/32.0
128	8	8	14.39	4.03	0.27	128/256.0	16/32.0
96	4	8	14.11	4.11	0.27	72/256.0	6/32.0
128	16	8	14.04	4.13	0.27	128/256.0	32/32.0

96	8	8	13.89	4.17	0.26	72/256.0	12/32.0
102	4	16	13.12	4.42	0.25	81/256.0	6/32.0
96	16	8	13.11	4.42	0.25	72/256.0	24/32.0
102	8	16	13.00	4.46	0.25	81/256.0	12/32.0
153	16	96	12.73	4.55	0.24	182/256.0	38/32.0
102	16	16	12.71	4.56	0.24	81/256.0	25/32.0
153	8	96	12.24	4.74	0.23	182/256.0	19/32.0
109	4	16	12.04	4.82	0.23	92/256.0	6/32.0
109	8	16	12.04	4.82	0.23	92/256.0	13/32.0
153	16	16	11.67	4.97	0.22	182/256.0	38/32.0
153	4	16	11.56	5.02	0.22	182/256.0	9/32.0
153	8	16	11.50	5.04	0.22	182/256.0	19/32.0
153	4	96	11.43	5.07	0.22	182/256.0	9/32.0
109	16	16	11.20	5.18	0.21	92/256.0	27/32.0
102	4	8	10.21	5.68	0.19	81/256.0	6/32.0
102	16	8	9.95	5.83	0.19	81/256.0	25/32.0
109	4	8	9.77	5.94	0.18	92/256.0	6/32.0
109	8	8	9.45	6.13	0.18	92/256.0	13/32.0
102	8	8	9.21	6.29	0.17	81/256.0	12/32.0
96	8	4	9.18	6.31	0.17	72/256.0	12/32.0
96	4	4	9.12	6.36	0.17	72/256.0	6/32.0
128	8	4	9.11	6.36	0.17	128/256.0	16/32.0
128	16	4	9.03	6.42	0.17	128/256.0	32/32.0
96	16	4	8.86	6.54	0.17	72/256.0	24/32.0
153	16	8	8.85	6.55	0.17	182/256.0	38/32.0
153	8	8	8.76	6.62	0.17	182/256.0	19/32.0
109	16	8	8.45	6.86	0.16	92/256.0	27/32.0
153	4	8	8.09	7.17	0.15	182/256.0	9/32.0
102	16	4	6.95	8.35	0.13	81/256.0	25/32.0
102	8	4	6.81	8.51	0.13	81/256.0	12/32.0
102	4	4	6.49	8.93	0.12	81/256.0	6/32.0
128	4	4	6.45	8.99	0.12	128/256.0	8/32.0
109	16	4	6.39	9.07	0.12	92/256.0	27/32.0
109	4	4	6.14	9.44	0.12	92/256.0	6/32.0
109	8	4	6.14	9.45	0.12	92/256.0	13/32.0
153	4	4	6.08	9.53	0.12	182/256.0	9/32.0
153	16	4	6.03	9.62	0.11	182/256.0	38/32.0
153	8	4	6.01	9.64	0.11	182/256.0	19/32.0

Table 4

Gflops, time, and utilization given kc, nr, and mr sizes on PC sorted by GFLOPS

kc/mc	nr	mr	gflops	time (seconds)	util	A block (KB)	B Sliver (KB)
128	4	32	32.47	1.79	0.51	128/256.0	8/32.0
128	16	32	32.18	1.80	0.51	128/256.0	32/32.0
96	4	96	32.14	1.80	0.51	72/256.0	6/32.0
128	8	32	31.50	1.84	0.50	128/256.0	16/32.0
96	16	96	30.92	1.88	0.49	72/256.0	24/32.0
96	8	96	30.85	1.88	0.49	72/256.0	12/32.0
96	4	32	30.50	1.90	0.48	72/256.0	6/32.0
96	16	32	26.52	2.19	0.42	72/256.0	24/32.0
96	8	16	24.59	2.36	0.39	72/256.0	12/32.0
96	4	16	24.50	2.37	0.39	72/256.0	6/32.0
128	4	16	24.33	2.38	0.38	128/256.0	8/32.0
128	8	16	24.25	2.39	0.38	128/256.0	16/32.0
96	16	16	24.25	2.39	0.38	72/256.0	24/32.0
128	16	16	23.79	2.44	0.37	128/256.0	32/32.0
128	16	96	20.19	2.87	0.32	128/256.0	32/32.0
96	8	32	19.89	2.91	0.31	72/256.0	12/32.0
128	4	96	19.84	2.92	0.31	128/256.0	8/32.0
128	8	96	19.81	2.93	0.31	128/256.0	16/32.0
109	4	96	19.74	2.94	0.31	92/256.0	6/32.0
109	8	96	19.53	2.97	0.31	92/256.0	13/32.0
109	8	32	18.99	3.05	0.30	92/256.0	13/32.0
102	4	32	18.88	3.07	0.30	81/256.0	6/32.0
109	4	32	18.22	3.18	0.29	92/256.0	6/32.0
102	8	32	17.90	3.24	0.28	81/256.0	12/32.0
153	4	32	17.78	3.26	0.28	182/256.0	9/32.0
153	8	32	17.49	3.32	0.28	182/256.0	19/32.0
153	16	32	16.77	3.46	0.26	182/256.0	38/32.0
96	4	8	16.66	3.48	0.26	72/256.0	6/32.0
109	16	32	16.42	3.53	0.26	92/256.0	27/32.0
96	16	8	16.40	3.54	0.26	72/256.0	24/32.0
128	16	8	16.29	3.56	0.26	128/256.0	32/32.0

109	16	96	15.95	3.63	0.25	92/256.0	27/32.0
96	8	8	15.93	3.64	0.25	72/256.0	12/32.0
153	4	16	15.91	3.64	0.25	182/256.0	9/32.0
128	8	8	15.91	3.64	0.25	128/256.0	16/32.0
102	4	16	15.63	3.71	0.25	81/256.0	6/32.0
128	4	8	15.55	3.73	0.24	128/256.0	8/32.0
102	16	16	15.22	3.81	0.24	81/256.0	25/32.0
153	16	16	15.13	3.83	0.24	182/256.0	38/32.0
153	8	16	14.70	3.94	0.23	182/256.0	19/32.0
102	16	32	14.36	4.04	0.23	81/256.0	25/32.0
153	16	96	14.18	4.09	0.22	182/256.0	38/32.0
109	16	16	14.10	4.11	0.22	92/256.0	27/32.0
153	8	96	13.96	4.15	0.22	182/256.0	19/32.0
153	4	96	13.90	4.17	0.22	182/256.0	9/32.0
102	8	16	13.86	4.18	0.22	81/256.0	12/32.0
109	4	16	13.47	4.30	0.21	92/256.0	6/32.0
109	8	16	13.27	4.37	0.21	92/256.0	13/32.0
102	8	96	12.55	4.62	0.20	81/256.0	12/32.0
102	4	96	11.55	5.02	0.18	81/256.0	6/32.0
102	16	8	11.24	5.16	0.18	81/256.0	25/32.0
153	16	8	10.86	5.34	0.17	182/256.0	38/32.0
109	8	8	10.86	5.34	0.17	92/256.0	13/32.0
153	4	8	10.67	5.43	0.17	182/256.0	9/32.0
153	8	8	10.64	5.45	0.17	182/256.0	19/32.0
109	16	8	10.53	5.51	0.17	92/256.0	27/32.0
96	4	4	10.34	5.61	0.16	72/256.0	6/32.0
109	4	8	10.29	5.64	0.16	92/256.0	6/32.0
96	8	4	10.28	5.64	0.16	72/256.0	12/32.0
102	4	8	10.23	5.67	0.16	81/256.0	6/32.0
96	16	4	9.85	5.89	0.16	72/256.0	24/32.0
128	16	4	9.66	6.00	0.15	128/256.0	32/32.0
128	8	4	9.42	6.16	0.15	128/256.0	16/32.0
102	16	96	9.31	6.23	0.15	81/256.0	25/32.0
102	8	8	8.55	6.78	0.13	81/256.0	12/32.0
102	4	4	8.00	7.24	0.13	81/256.0	6/32.0
102	8	4	7.87	7.37	0.12	81/256.0	12/32.0
153	4	4	6.80	8.52	0.11	182/256.0	9/32.0
109	16	4	6.78	8.55	0.11	92/256.0	27/32.0
153	16	4	6.77	8.57	0.11	182/256.0	38/32.0
153	8	4	6.74	8.61	0.11	182/256.0	19/32.0
128	4	4	6.53	8.88	0.10	128/256.0	8/32.0

109	8	4	6.28	9.24	0.10	92/256.0	13/32.0
102	16	4	5.83	9.95	0.09	81/256.0	25/32.0
109	4	4	4.49	12.91	0.07	92/256.0	6/32.0

Figures

Figure 1.

3D scatter plot of Gflops, nr, mr, and kc/mr sizes on Bell Cluster using O0 optimization flag

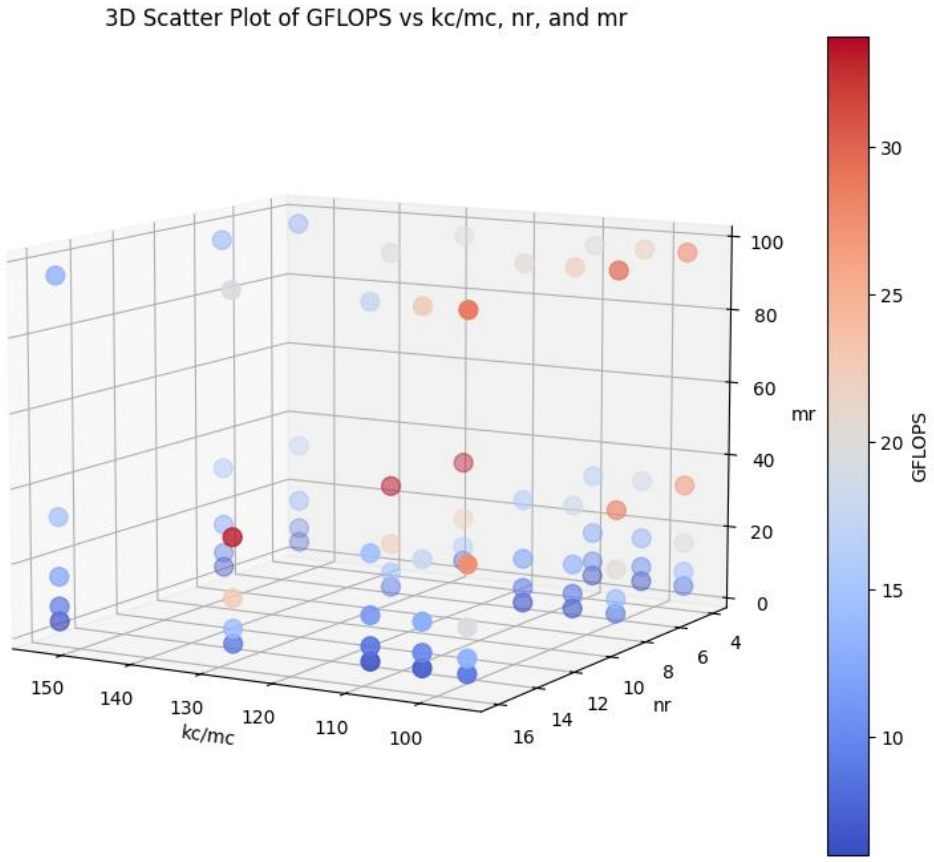

Figure 2.

3D scatter plot of Gflops, nr, mr, and kc/mr sizes on Bell Cluster using O0 optimization flag