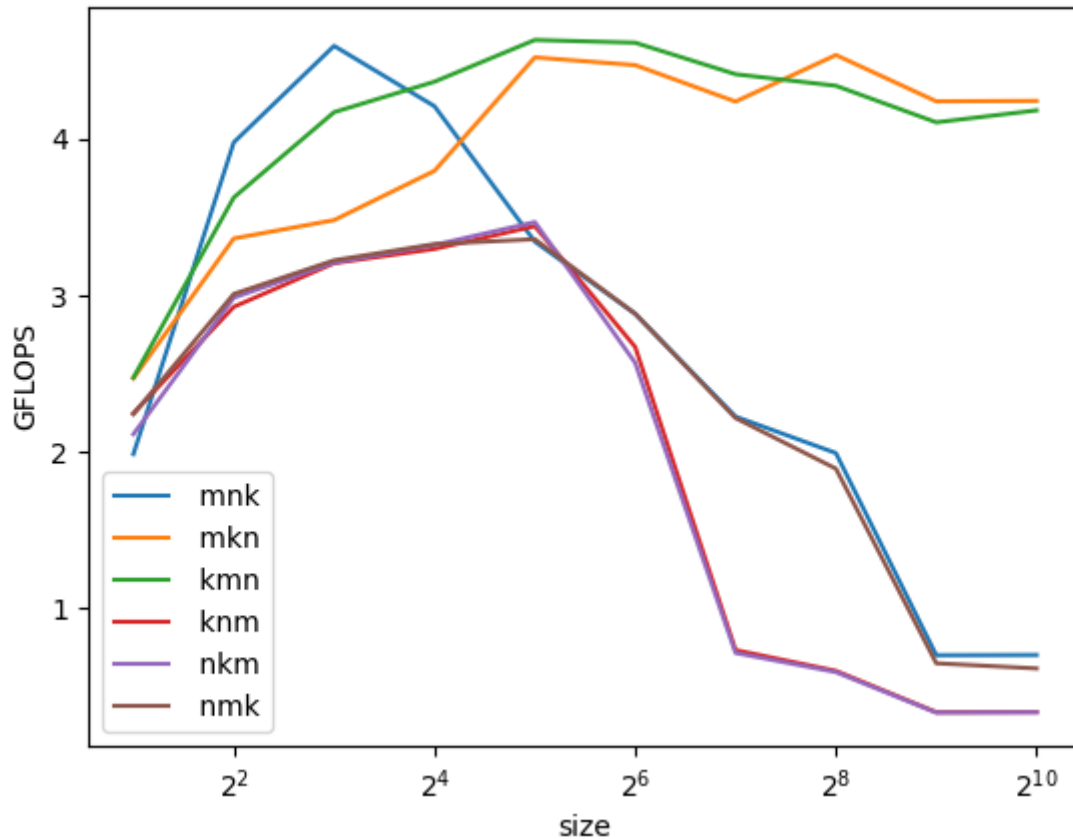


11 Data Locality

11.1 Matrix-Matrix Multiplication



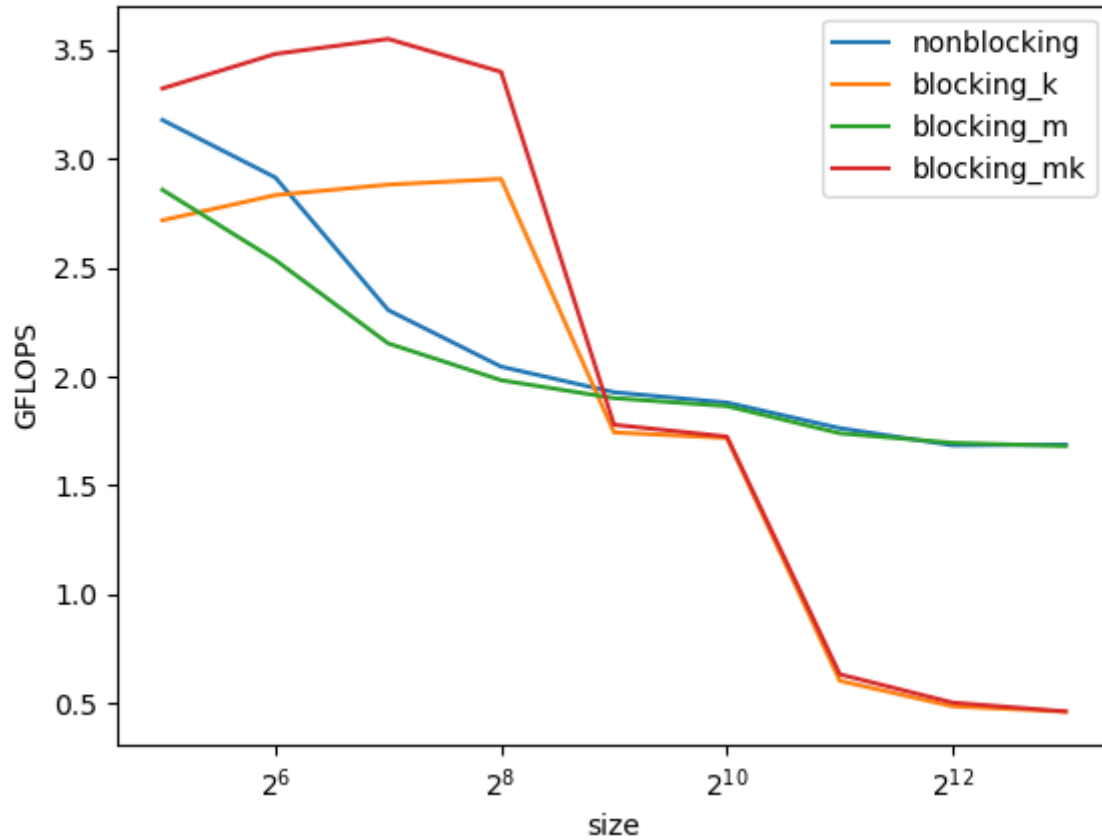
mkn and kmn perform significantly better than the other implementations for matrix-matrix multiplication. mkn is so good because we reuse the values in A in the inner most loop:

```
io_C[l_m * i_n + l_n] += i_A[l_m * i_k + l_k] * i_B[l_k * i_n + l_n];
```

$i_A[l_m * i_k + l_k]$ is reused for every l_n .

$i_B[l_k * i_n + l_n]$ is also decent as it is linear in memory so the memory lines get a lot of values at once.

11.2 Blocking



```
[gi24ken@node286 ~]$ getconf -a | grep CACHE
LEVEL1_ICACHE_SIZE      32768
LEVEL1_ICACHE_ASSOC     8
LEVEL1_ICACHE_LINESIZE  64
LEVEL1_DCACHE_SIZE      32768
LEVEL1_DCACHE_ASSOC     8
LEVEL1_DCACHE_LINESIZE  64
LEVEL2_CACHE_SIZE       1048576
LEVEL2_CACHE_ASSOC      16
LEVEL2_CACHE_LINESIZE   64
LEVEL3_CACHE_SIZE       25952256
LEVEL3_CACHE_ASSOC      11
LEVEL3_CACHE_LINESIZE   64
LEVEL4_CACHE_SIZE       0
LEVEL4_CACHE_ASSOC      0
LEVEL4_CACHE_LINESIZE   0
```

We observed the first drop at $i=2^9=512$, which means we load $2^9 * 8 * 4 + 4 * 8 = 16384 + 32$ bytes. This exceeds one half of the L1d cache size. This is problematic because we only load $8*4=32$ bytes of each array at a time but have a cachelinesize of 64 bytes, meaning we don't use half the cache.

The second drop is at $i=2^{11}=1024$, where we load $2^{11} * 8 * 4 + 4 * 8 = 65536 + 32$ bytes. This is exceeding the L1d size so it has to fallback to L2 cache.

We are not sure why 2 drops occur, because we are always within the L2 cache size $2^{13}=8192$
 $8192*8*4+4*8=524288+32 < 1048576$. And if 2 drops exist we are not sure why the second drop isn't
already at $i=2^{10}=2048$, where we load $2^{10} * 8 * 4 + 4 * 8 = 32768 + 32$ bytes. This should both exceed
the L1d size by 32 bytes and since the cache line size is not aligned we should also be unable to use
half the cache size. Nonetheless there isn't a drop here and we are not sure why.