1.

a)

Level 1: perform scalar and vector based operations.

Level 2: perform matrix-vector operations.

Level 3: perform matrix-matrix operations.

b)

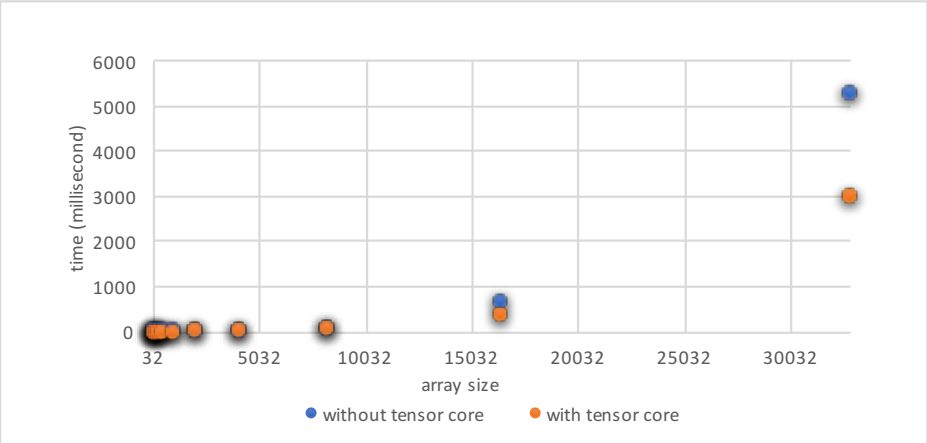
```
cublasStatus_t cublasSgbmv(cublasHandle_t handle, cublasOperation_t trans,
                           int m, int n, int kl, int ku,
                           const float           *alpha,
                           const float           *A, int lda,
                           const float           *x, int incx,
                           const float           *beta,
                           float           *y, int incy)
```

This function specifies matrix A to be column-major.

```
cublasStatus_t cublasSgemm(cublasHandle_t handle,
                           cublasOperation_t transa, cublasOperation_t transb,
                           int m, int n, int k,
                           const float           *alpha,
                           const float           *A, int lda,
                           const float           *B, int ldb,
                           const float           *beta,
                           float           *C, int ldc)
```

This function specifies matrix A, B, and C to be column-major.

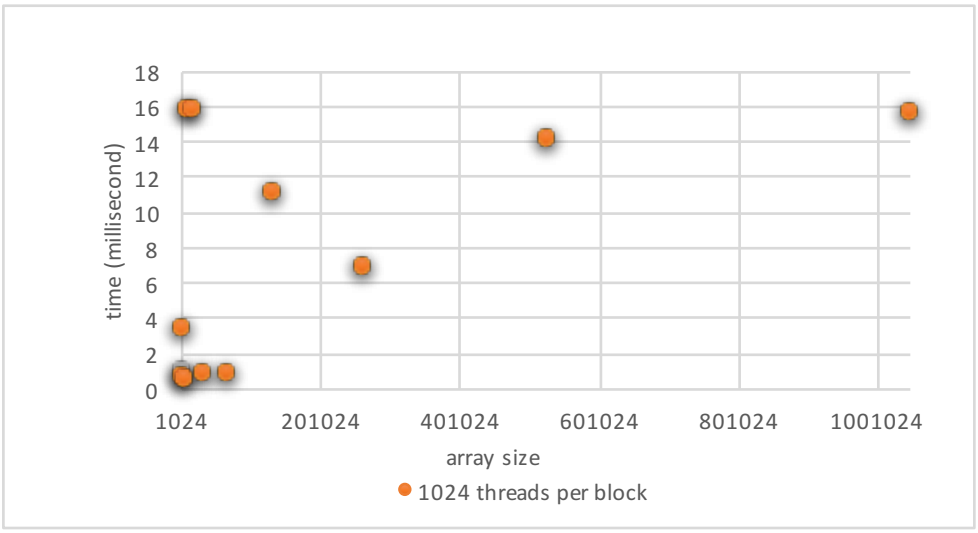
e), f)



g)

When utilizing tensor code, the performance is better since the computation can be done within matrix-multiply-and-accumulate units to enhance throughput.

2 c)


18
16
14
12
10
8
6
4
2
0
time (millisecond)
1024
201024
401024
601024
801024
1001024
array size
1024 threads per block