1a) BLAS libraries group functions into three levels. What do all Level 1 functions have in common? Level 2? Level 3? In other words, how did they decide how to group these functions?

Ans) In general, the grouping is done based on the type of operations performed at each level as mentioned below:

- Level 1: Scalar and Vector, Vector and Vector operations,
- Level 2: Vector and Matrix operations
- Level 3: Matrix and Matrix operations

1b) Some functions are specialized for performing their operations when the structure of the input matrix or vector is known. List and briefly explain two such functions which assume something about their input structure in order to optimize the computation.

Ans) Functions like cublasGetVector() and cublasGetMatrix are few of the function which assumes a certain structure for vector and matrices respectively. These functions copies elements from a vector/matrix in GPU memory into the host memory space. Elements in both vector/matrix are assumed to have size of elemSize bytes. The storage spacing between consecutive elements can be changed with increment of 1 for column and leading dimension for matrix respectively.

g) Comment on the differences that you preforms see in the scaling analyses and explain briefly what the cublasSetMathMode call changed about the computation.

Ans) The cublasSetMathMode enabled performs the operation faster compared to the disabled version in the scaling analyses especially for larger array lengths. The difference is that when the cublasSetMathMode is set, Tensor cores within the V100 architectures gets enabled. In this mode, Tensor cores are programmable matrix-multiply-and-accumulate units which primarily optimizes the GEMM (General matrix-matrix multiplication)operations since the cores are dedicated units which specialize to performs this operation faster.