Compute Platform Engineer Tech Task

1. Identify current GPU technologies for a Level 4/5 autonomous driving vehicle to be used as an accelerator engine in a compute platform. List pros and cons compared to other state-of-the-art accelerator engines.
    1. Which other hardware architectures beside GPUs could be suitable for a Autonomous Driving SW stack integration? Please explain your choice.
2. Design in a block diagram a central hardware compute platform for a Level 4/5 autonomous driving car for integrating a self-driving car SW stack. Please comment your selection and design.
    1. Which external interfaces and bandwidths are necessary?
    2. How could a deployment from a self-driving car SW stack look like based on this HW concept?
3. Assume we have an application running on GPU. We cannot meet the latency requirements and we decided to replace our GPU with a new one with more number of cores. Unfortunately, the new GPU didn't provide the performance improvement that we expected.
    1. What might be the reason of this?
    2. Is it possible to estimate that without HW replacement? How ?
4. Implement a general Matrix Multiplication [$A_{MxN}B_{NxK}$] function in C/C++/Assembly by using SIMD instructions and multi-core parallelism for CPU.
    1. What is the absolute performance of your implementation (flops) ?
    2. What is the relative performance compared to single-core and non-vectorized implementation ?
    3. What is the limiting factor for the performance and how can we improve it further?
    4. Note: Implementation can be performed on Arm/x86 architecture with any data type(single/double precision or fixed-point)
5. Implement a hardware model for matrix multiplication unit based on systolic arrays in C/C++/SystemC. The unit should consist of the following elements : 4x4 array of Processing elements (PE), Input and output buffers
    1. Test your solution with respect to a golden reference matrix multiplication function by multiplying two 16x16 matrices
    2. How can you use this HW for accelerating CNNs (Convolutional Neural Networks) ?
    3. How does simple ISA (instruction set architecture) instructions look like for this accelerator ?
    4. Note:
        1. You might ignore the concurrency, assume we would like to check if our algorithm is correct.
        2. You can use templates for data type, or just assume floating point data