1. AI accelerators like SambaNova, Cerebras, Graphcore, and Grog systems are suitable for AI workloads because they include key architectural features such as specialized hardware design, high memory bandwidth and larger amounts of on-chip memory, and scalability and parallelism. The specialized hardware design accelerates matrix multiplications and tensor operations. The high memory bandwidth and on-chip memory support the acceleration of memory intensive AI workloads. Scalability and parallelism include the parallel processing of data across many cores and processing units. This accelerates training and inference tasks speed.
2. SambaNova includes reconfigurable dataflow units that allows dataflow processing flexibility. It features a multi-tiered memory architecture with enough addressable memory for handling of large data. Cerebras includes a wafer-scale engine that consists of processing elements with its own memory. It operates independently. The system is highly parallel and scalable due to fine grained dataflow control mechanism within its processing elements. The intelligence processing unit of Graphcore includes many interconnected processing tiles. Each processing tile has its own core and local memory. The intelligence processing unit has two operational phases of computation and communication. It uses the phases through bulk synchronous parallelism. Grog's tensor streaming process focuses on

deterministic execution. It is advantageous for inference tasks of critical low latency.

3. The general workflow involved vendor specific implementation of ML frameworks like PyTorch to port model. Refactoring starts with training and testing. Next is accessing the testbed and downloaded the needed software. Then optimize the model by adjusting parameters such as batch size and memory usage. Lastly is to deploy the model on the testbed and refine it according to the results.
4. Autonomous driving systems development would benefit from AI accelerators. They rely on real-time processing of large amounts of data. This data comes from cameras, LiDAR, and other sensors. The accelerators provide rapid training and inference of advanced deep learning models that are commonly used for object detection, path planning, and decision making. This is done by proving the necessary scalability and parallelism. The results are improved response times and accuracy. The accelerators all provide latency reduction, necessary for timely reactions.