Technical Questions

Question 1]

A. 32 gpus in the most basic settings. There are some recent papers that explore how to fully utilize gpus. (e.g., ZeRO sharding, cpu-offloading, deepspeed, fairscale). I'll also use mixed precision if the gpu supports it.
B. Distributed data parallel, I will set up a port and address environment variables, then I'll use putorch multiprocessing and give it the train function.
C. I will probably first check the CUDA_VISIBLE_DEVICES environment variable and ensure that it is properly exported to the node performing the job.

Question 2]

- It all depends on the initial state of the model weights, at the first iteration, she could have started at a better place on the loss function manifold and started at a worse space the second time. If the loss function is convex, most likely they will both end up with similar results given the same hyperparameters/optimizer/lr scheduler are used. If the loss function is not convex, then the local/global minimum that she will arrive at is largely dependent on initialization since the gradient descent algorithm will not escape local minima.