Voltage Park - Tech Assessment

Objective

Build a text-to-video API (over HTTP). This API needs to take in a text prompt and convert it into a video. Underneath, you need to host an open-source text-to-video model: https://huggingface.co/genmo/mochi-1-preview

System environment:

You are provided with one worker node with ssh access. This worker node has:

1. 8XH100 GPUs
2. 18TB of cumulative NVMe storage (not mounted)
3. 2 CPUs with ~124 cores

The worker node is part of a K8s cluster. In addition to the worker node, it includes 1 control plane. The control plane is running in an EC2 instance to which you do not have access. You can only access the K8s cluster via `kubectl`.

Requirements:

Asynchronous Job Management

- **Asynchronous API Endpoints:** Implement a set of asynchronous API endpoints to manage the video generation workflow.
- We recommend that you start the following set of APIs (reasonable alternatives are acceptable):
    - **Submit Job API:** An endpoint to submit a new video generation request, returning a job ID immediately.
    - **Get Job Status API:** An endpoint to query the current status of a specific video generation job (e.g., pending, processing, completed, failed).
    - **List Jobs API:** An endpoint to retrieve a list of all submitted video generation jobs, potentially with filtering and pagination options.
    - **Get Generated Output File API:** An endpoint to retrieve the final generated video file once a job is successfully completed, using the job ID.

Video Generation

- Use the genmo mochi-1 model to generate videos. Please do not use any other model. https://huggingface.co/genmo/mochi-1-preview

Scalability and Concurrency

- **Concurrent Video Processing:** The system must be capable of processing multiple video generation requests concurrently. It should intelligently manage and distribute these workloads across all available GPUs to maximize resource utilization and throughput.

User Interface

- **Frontend Application:** Develop a basic frontend application that interacts with the backend API. This frontend should allow users to submit video generation prompts, monitor the status of their requests, and download completed videos.

Deployment

- **Kubernetes (K8s) Deployment:** The entire service, including the webserver and any associated worker processes, must be deployed using Kubernetes for container orchestration.
- **Resource Allocation:** Configure the deployment to ensure high availability and efficient resource utilization. This includes:
    - **Minimum Replicas:** Deploy at least two replicas of the video generation service to ensure redundancy and fault tolerance.
    - **GPU Allocation per Replica:** Each replica must be configured to utilize a minimum of two GPUs to handle computational demands effectively.

Constraints:

- You can reference any material and use any coding assistance tools of your choice
- Please DO NOT use your current employer's resources (e.g. AWS developers accounts, etc.) – personal accounts are fine, but should not be needed.

Assumption you are free to make

1. Length of the video
2. Resolution / quality of the video
3. API structure (so long as its some sensible JSON in and consumable format out) 4

Assessment Structure

1. You have until your scheduled review time to complete your assessment (please see calendar invite details).
2. The test is designed to be hard and the outcome is not binary
    a. Therefore, document all your work:
    b. Planning, debugging, options considered, tools used, etc.
3. No need to make a pretty doc or presentation – the call on Monday will be a discussion that will dive deep into the approach.

4. We recommend that you start with a basic working demo, then expand your solution

Useful Information:

1. We will need your public ssh key to authorize your access to worker nodes
2. In case you encounter any issues (e.g. broken GPUs or losing access to the cluster etc) please reach out to us over text (email over the weekend may not be fast)
    a. Jon Rafkind : 617-470-4281
    b. Melih Bercin : 206-406-2761
    c. Karan Kothari 480-277-9745
    d. Ragha Prasad 765-418-2068

Thank you and good luck!