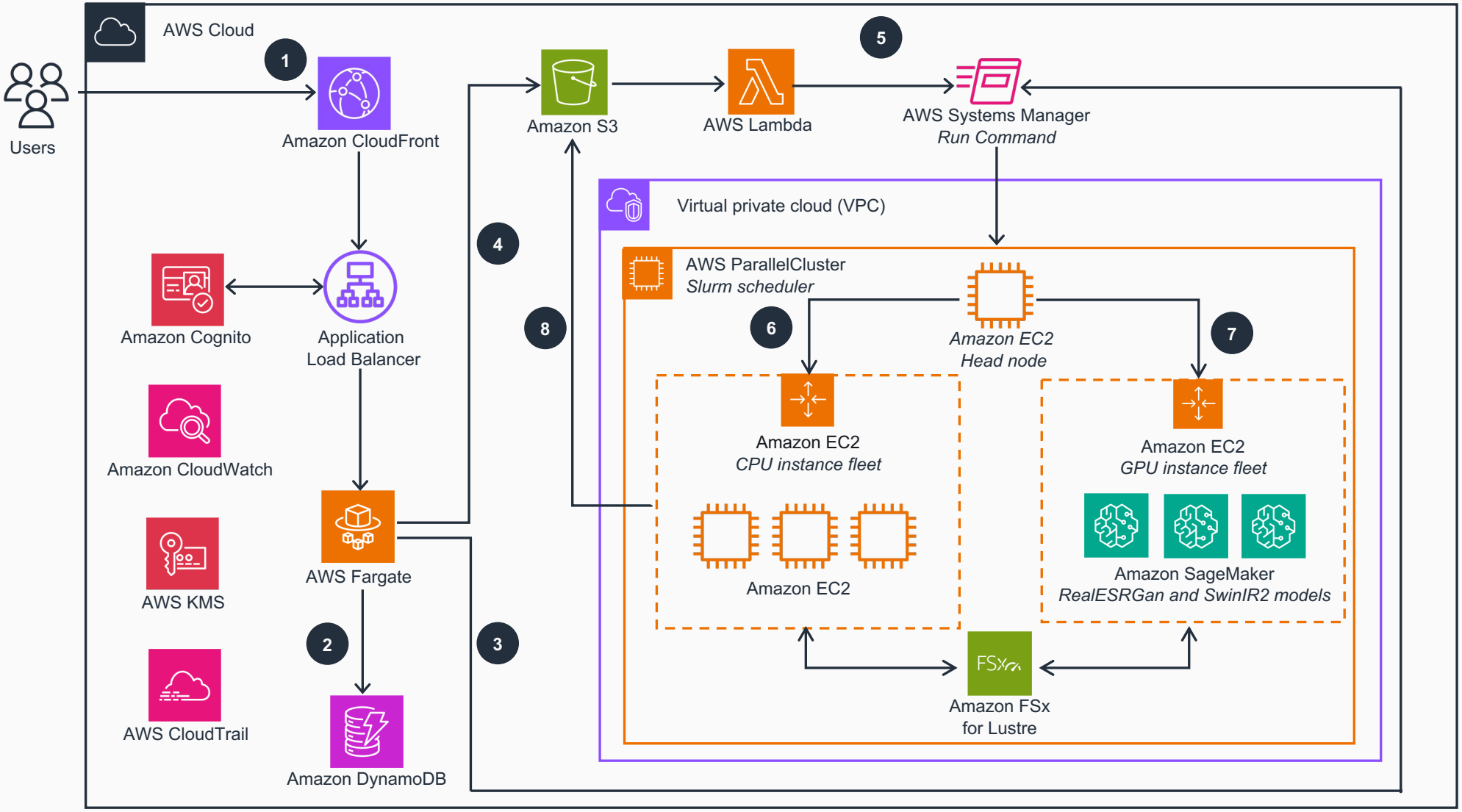
Guidance for Media Super Resolution on AWS

This architecture diagram shows you how you can enhance video content resolutions at scale in your AWS accounts using generative artificial intelligence (AI). This slide shows details on step 1-5. Fore more on steps 6-8, go to the next slide.



1. The user accesses the application, hosted on **AWS Fargate**, through an **Amazon CloudFront** distribution that is in front of an **Application Load Balancer (ALB)**. **ALB** redirects the user to **Amazon Cognito** for authentication for a new user session.

2. A task is registered in the task tracker table in **Amazon DynamoDB** for the user. This task tracking helps secure access to upscaled videos by associating pipeline tasks with the user who owns them.

3. App containers use the **AWS Systems Manager** Run Command to run scripts on the head node to allocate new tasks and get the status of tasks.

4. The user uploads or views upscaled video directly from **Amazon Simple Storage Service (Amazon S3).** This is done by requesting presigned URLs for uploading and downloading from the **Fargate** container that is hosting the app.

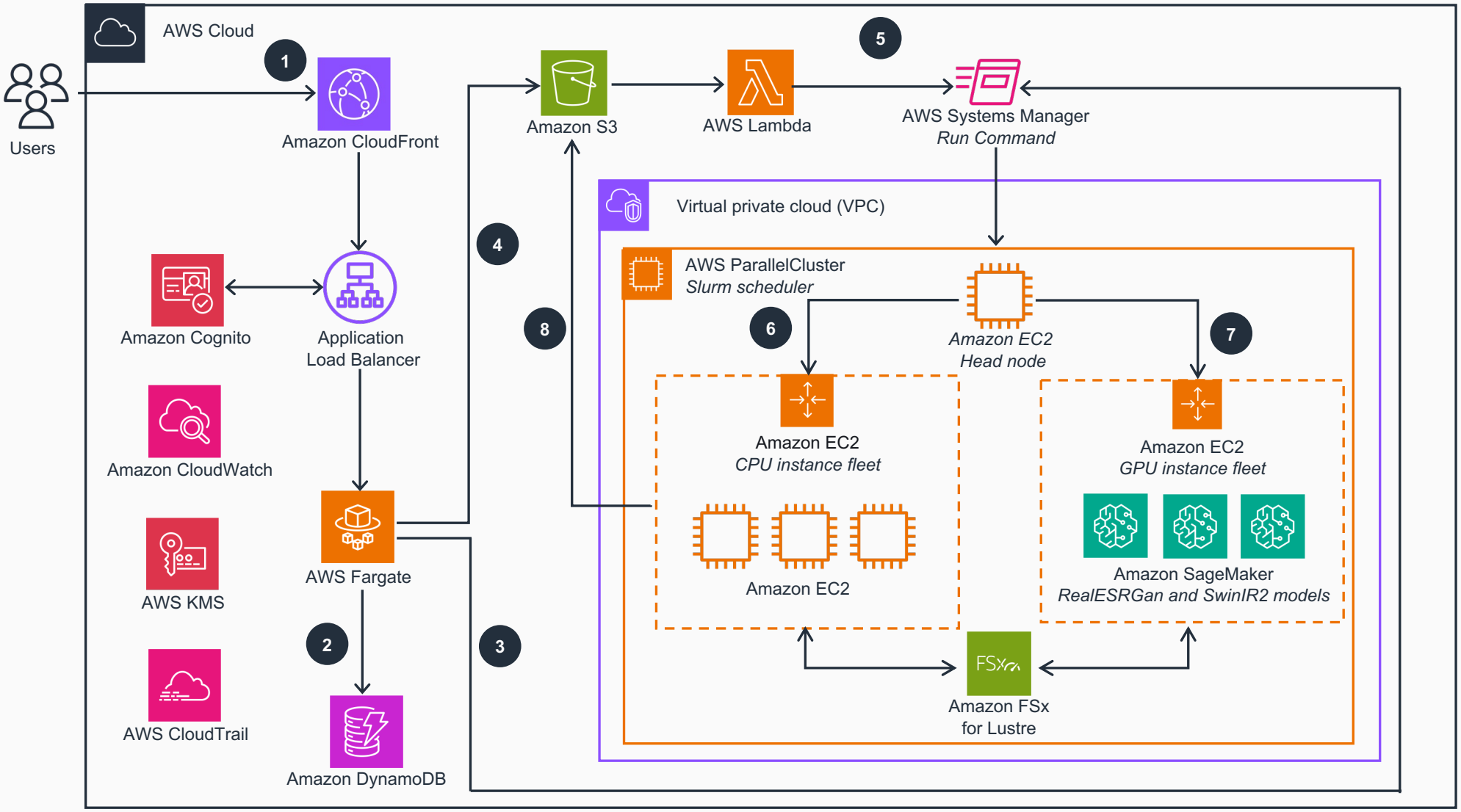
5. An **AWS Lambda** function is invoked upon a successful **Amazon S3** upload. This initiates a video upscaling workflow, which invokes a **Systems Manager** *Run Command.* A video upscaling pipeline is run by submitting the task into the Slurm job queue in **AWS ParallelCluster.**

Reviewed for technical accuracy March 4, 2024

AWS Reference Architecture

Guidance for Media Super Resolution on AWS

Steps 6-8



6 A scheduled task in **ParallelCluster** extracts video frames and writes images. It also extracts audio and media metadata, such as bitrate and frames per second (fps), into the shared **Amazon FSx for Lustre** file system for artificial intelligence (AI) super resolution tasks. The file system is encrypted with a key provided by **AWS Key Management System (AWS KMS).** The cluster automatically scales the central processing unit (CPU) compute fleet for video frame processing tasks.

7 **ParallelCluster** performs AI upscaling on each frame. It does this by invoking a generative AI model (*RealESRGan and SwinIR2)* through an **Amazon SageMaker** endpoint hosted in the Graphics Processing Unit (GPU) compute fleet. The output is written to the **FSx for Lustre** file system. The cluster automatically scales the GPU compute fleet for video upscaling tasks.

8 A **ParallelCluster** batch job encodes image frames to create new video content and uploads it to a given **Amazon S3** location. An **Amazon S3** presigned URL is created on-demand if an authorized user requests it.

Reviewed for technical accuracy March 4, 2024

AWS Reference Architecture