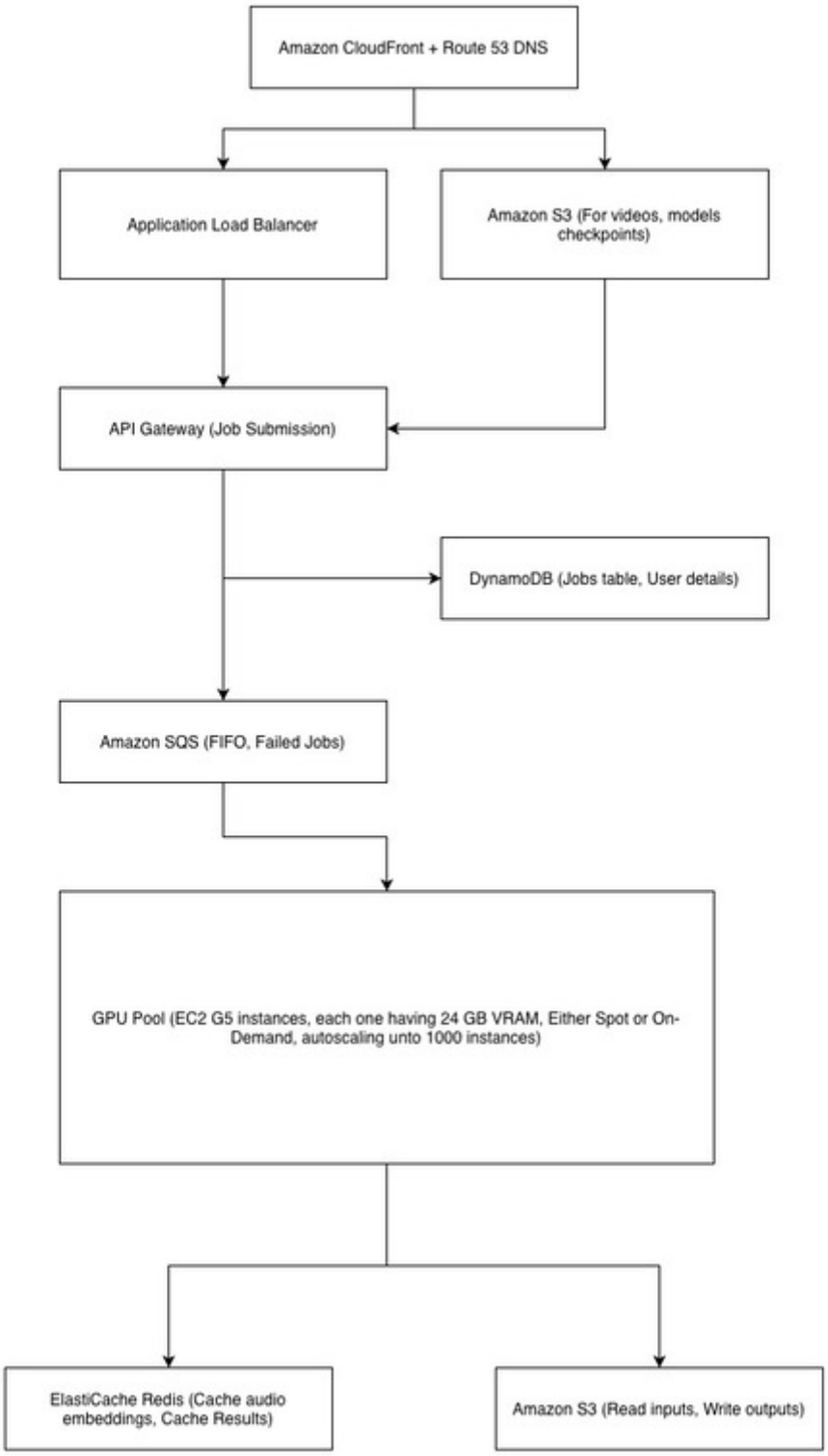
LatentSync AWS Deployment Plan

This documents outlines the technical architecture and operational considerations for deploying LatentSync at scale on AWS.

1. System Architecture Overview



- Core Design Decisions:

1. Async processing via SQS queue.
2. API layer (ECS Fargate) for handling job submissions, status checks.
3. GPU workers pull jobs from queue, process independently.
4. S3 for all video storage.
5. DynamoDB for job metadata and status tracking.
6. ElastiCache Redis for caching audio embeddings.
7. CloudWatch and X-Ray for monitoring.

2. Latency

- Targets

1. API should be less than 200ms.
2. Processing 90-120 seconds.
3. End-to-end should take 2-5 min.
4. Redis Cache for audio embedding with multi-region availability.
5. Keep 100 instances pre-loaded to avoid cold starts.

3. Memory

- Requirements

1. Atleast 24 GB of VRAM, as LatentSync v1.6 requires 18 GB VRAM minimum.
2. Opt for EC2 g5.xlarge or g6.xlarge with 70% Spot + 30% On-Demand.
3. If Queue depth > 100 for 2 minutes, scale up will be considered.
4. If Queue depth <= 50 for 10 minutes, scale down will be considered.

4. Failure Modes

- Resilience

1. With SQS, there should be a 15-min visibility timeout. After 3 tries, transfer to Dead Letter Queue.
2. Test inference every 5 min for health checks.
3. For spot interruption, a 2-min warning with graceful hand-off.
4. For GPU OOM, inference should be retried with lower `inference_steps`.

5. Monitoring

- Key Metrics & Alerting

1. Alert should be there if Queue depth is >500 or <10 for 10 minutes.
2. Alert should be there if processing time is taking > 180 seconds.