TikTok's ability to handle—and rapidly serve—billions of short videos rests on a highly scalable, cloud-native architecture built around distributed object storage, microservices, and a global content delivery network (CDN):

1. **Distributed Object Storage**
   – All uploaded videos are treated as immutable objects and stored in a massively scalable object store (e.g. AWS S3 or an in-house equivalent) that shards data across thousands of nodes and replicates it across multiple regions for durability and fault-tolerance. This lets TikTok grow to exabyte scale without hot spots or single points of failure

2. **Metadata in NoSQL & In-Memory Caches**
   – For each video object, TikTok writes metadata (owner, creation time, available renditions, segment indices, etc.) into a distributed NoSQL database such as Cassandra, with hot lookups served from a Redis cache. Microservices simply query by video ID to retrieve exactly which chunks to fetch, enabling sub-100 ms lookup times even at peak load

3. **On-the-Fly Transcoding & Segmenting**
   – At upload, the original video is transcoded into multiple bitrates and wrapped into HTTP-based streaming segments (e.g. HLS/DASH). These uniformly sized "chunks" allow adaptive bitrate streaming: the client can switch between different quality levels seamlessly according to network conditions. Pre-packaging into segments makes cache hits and parallel downloads trivial

4. **Global CDN & Edge Caching**
   – Once transcoded, each segment is pushed to a worldwide CDN. Edge servers in dozens of regions cache the most popular content (or the first few segments of newly uploaded videos), so that playback requests are answered from the nearest edge node rather than the origin store. This slashes latency and dramatically reduces backbone traffic—even when millions of users hit "play" simultaneously

By combining a cloud-scale object store with real-time metadata services, chunked adaptive streaming, and a far-reaching CDN, TikTok delivers snappy video uploads and near-instant playback at a global scale.