Azure Media Service Cloud Video Delivery

KILROY HUGHES

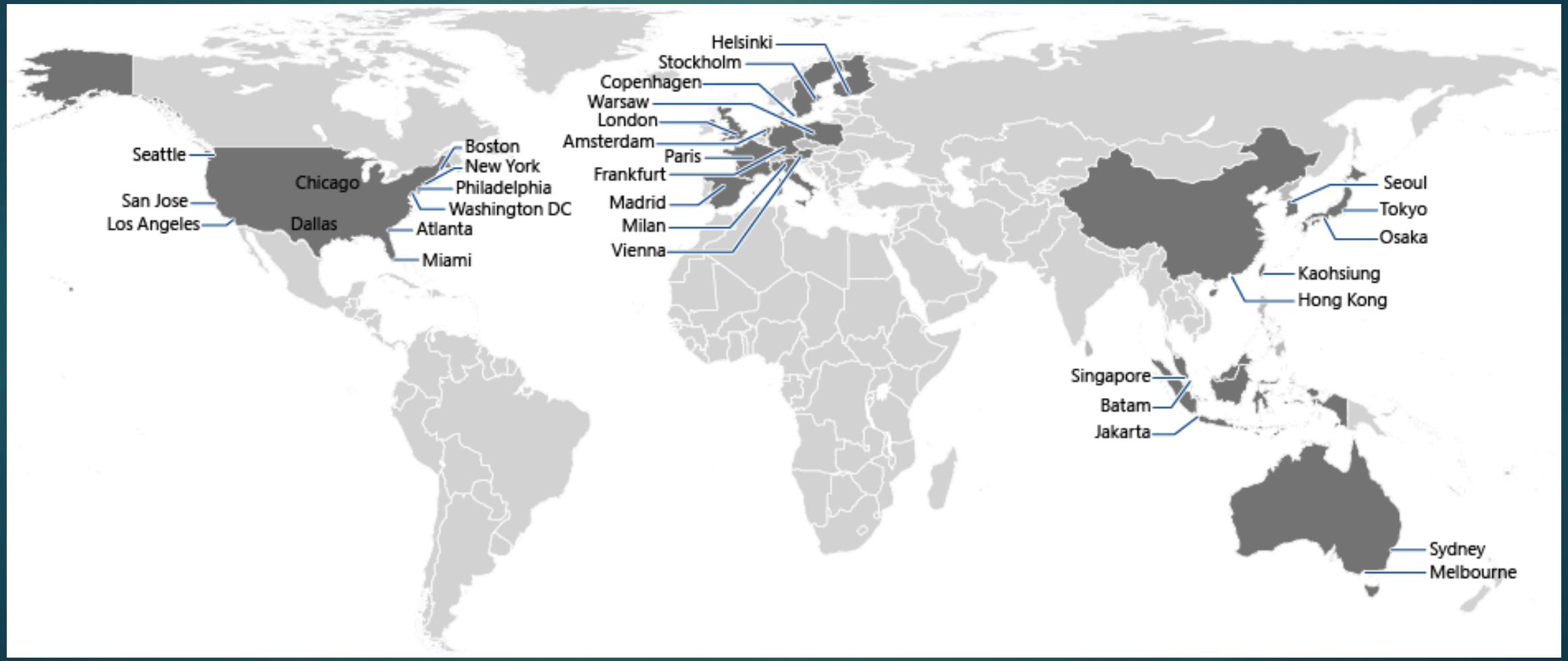
MICROSOFT AZURE MEDIA

2015.08.20

Azure Cloud Topology

- Public cloud providers such as Amazon Web Service, Google, IBM, Rackspace, etc. have similar topologies and offer infrastructure, platform, and/or software as a service
- Geographically distributed data centers – A computation and storage backbone
  - At least 2 physically separated Azure data centers in any Region, e.g. Australia, India, Brazil, China, East Asia, Europe (many in North America)
  - Dedicated backbone links between data centers (undersea fiber, etc.)
  - Live video ingest via "Express Route" uplinks using dedicated telco bandwidth; or ingest over Internet using Skype, RTMP, Smooth, RTP, etc.
- Data replication
  - Within physically isolated bays in a data center with different power, backbone carriers, etc.
  - Geographically separate data centers (protection against disasters, Internet outages, etc.)
  - Only elementary streams need be replicated because Segment packaging, delivery encryption, manifests, etc. are dynamically generated for each request origin request

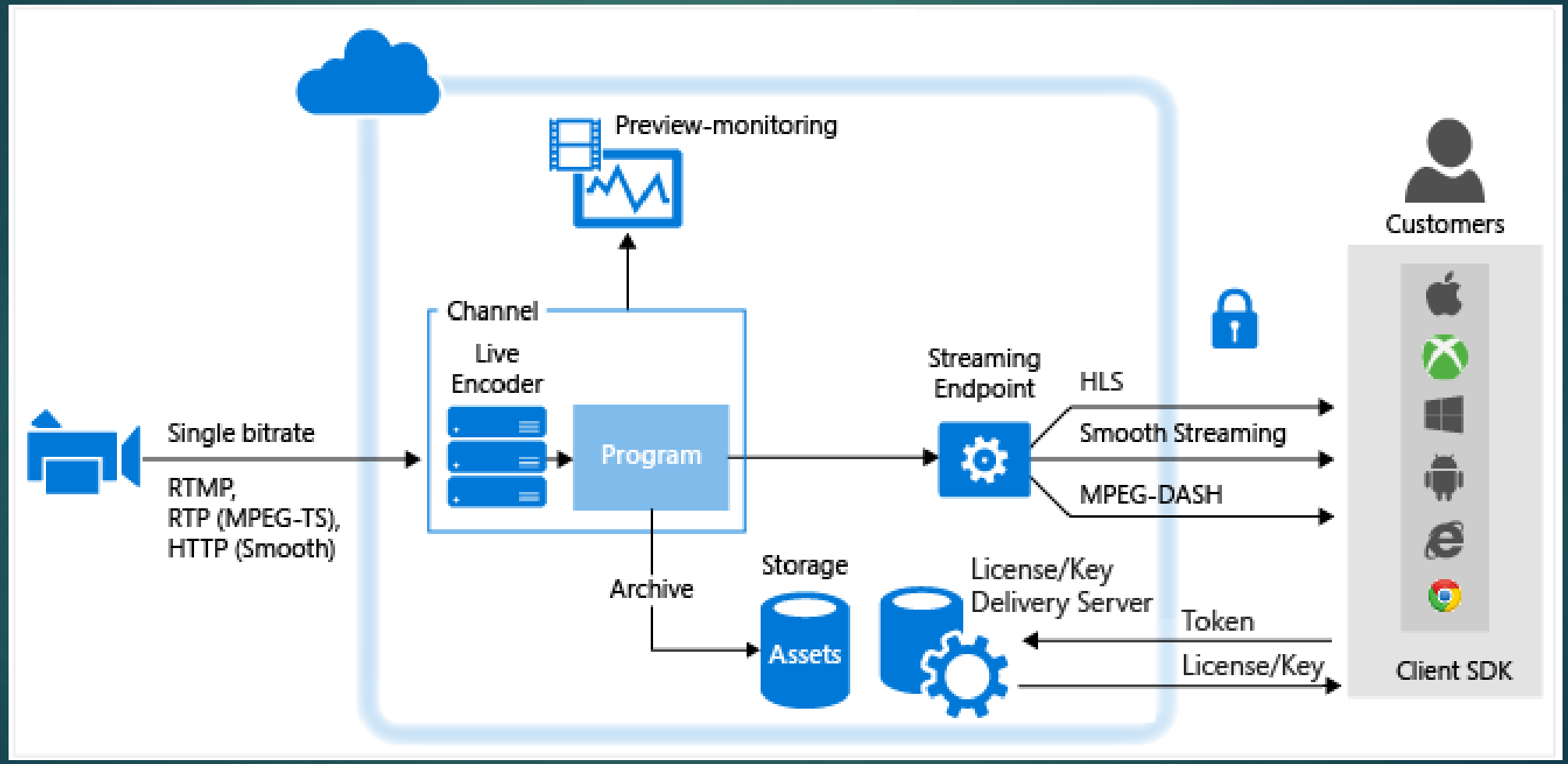
Microsoft Azure CDN point of presence (POP) locations (Brazil and India recently added)



Azure Cloud Media Processing

- Hardware abstraction layer with fault tolerant distributed processing
  - Resource management, event buss, table and database services, health monitoring, storage, streaming origins, etc. are Azure platform "microservices" that can be shared by all workloads
  - Azure Media Service hardware abstraction and fault tolerance: Always N+1 cores dynamically allocated to each task to allow rolling updates and recovery of media workflows, as well as addition of VMs to increase capacity
  - AMS hosted services include transcoding, video analysis, voice to subtitle generation, translation, transcoding, encryption, Segment packaging, clip editing, splice conditioning, ad insertion, presentation filtering (selected codecs, languages, bitrates, etc.), streaming, DRM licensing, database, web hosting, cross platform apps, CDN integration, etc.
  - Many third party services available, such as professional broadcast encoders, high speed upload over generic internet, watermarking, video and audio analytics, etc.; hosted and available for customers to use in their media workflows
  - General Azure services can be added to media services, such as Office, Exchange, SharePoint, Sequel, search, natural language processing, holographic analysis and rendering, software development, cross-platform mobile apps, Cosmos big data and analytics with machine learning, etc.; and the ability to run any software configured for Docker, etc.
- Live streaming replication and realtime reliability – "No single point of failure"
  - Geo-redundancy of uplink, compute, storage, origins, and CDNs for highest availability Service Level Agreement (SLA) for live streaming (Olympics, World Cup, Super Bowl, etc.)
  - Client transparent load balancing and failover with synchronized transcoding, Segmenting, and addressing

Simple Live Streaming Workflow



Live Operations

- Live scheduling, switching, and splicing of programs, ads, etc.
- Live monitoring of distributed work flows in several locations, with health alerts for processes exceed normal operating range
- Complete workflow event logs, player telemetry, user experience measurement, and errors in databases available to each customer for analysis
- Integration of delivery statistics with CDN load balancing and player behavior for network QOS management
- Live operations monitoring and management by humans to prevent or identify streaming problems anywhere in the workflow

CDNs and Adaptive Streaming

- CDNs solve a big part of the scaling and reliability challenges for large scale adaptive streaming, such as internet TV
  - Edge servers can store DASH manifests and segments near users for quick delivery of popular content (cached because it is frequently requested)
  - CDN can reduce origin server traffic by handling thousands of requests for the same resource, making dynamic packaging and intelligent processing on origin servers efficient
- VOD files can be stored near the edge by special arrangement, or addressed by byte range to pull efficient file page sizes from origin to edge that may pre-cache subsequent requests for that file and bitrate
- Live video "chunks" or Segments can be pushed to the edge for popular live video, e.g. live TV, to reduce latency. Supporting queued HTTP requests or HTTP/2 push-promise at the edge allows efficient low latency delivery of live chunks. Queued requests are also compatible with broadcast and multicast delivery, where many client requests are consolidated and resolved by delivering a single resource.
- Reliable live delivery depends on reliable realtime operation of every stage prior to the edge network, and tight integration of encoding, uplink, transcoding, packaging, etc.