Swinburne University of Technology

Project Report

Serverless/Event-driven Architectural Design Report

Unit code
COS20019

Unit name
Cloud Computing Architecture

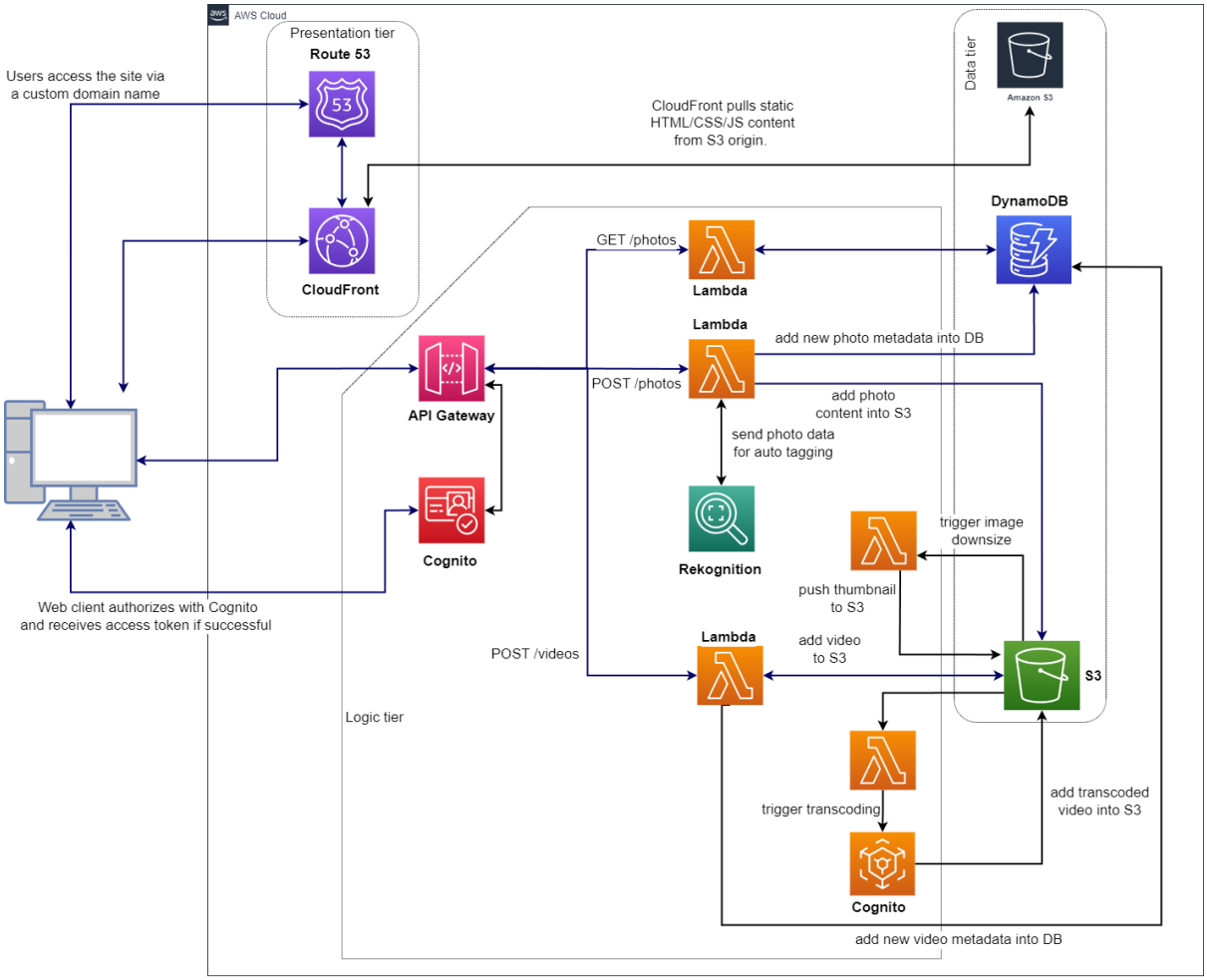
Students
Toan Nguyen (103797499)
Daniel Dang (103528453)
Chuong Ho (101921623)

Semeter
S2 - 2022

Table of contents

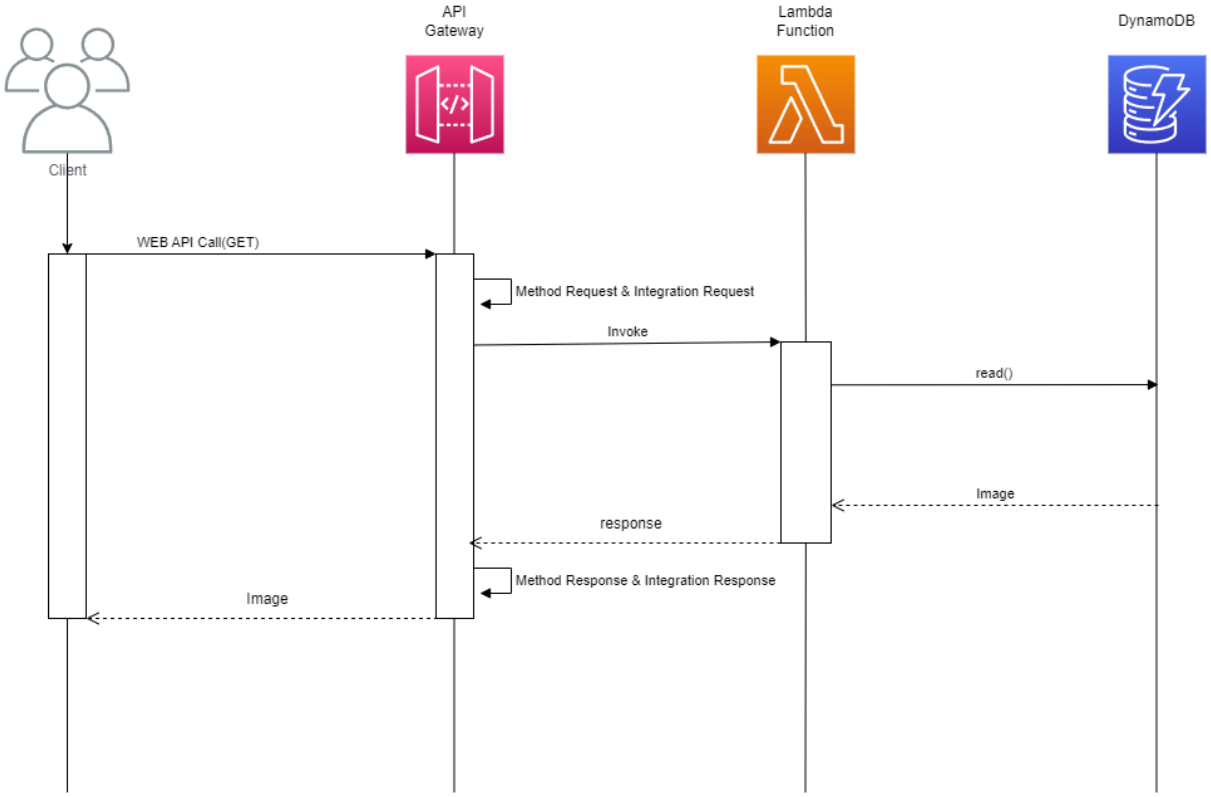
Architecture Design

Architectural Diagram

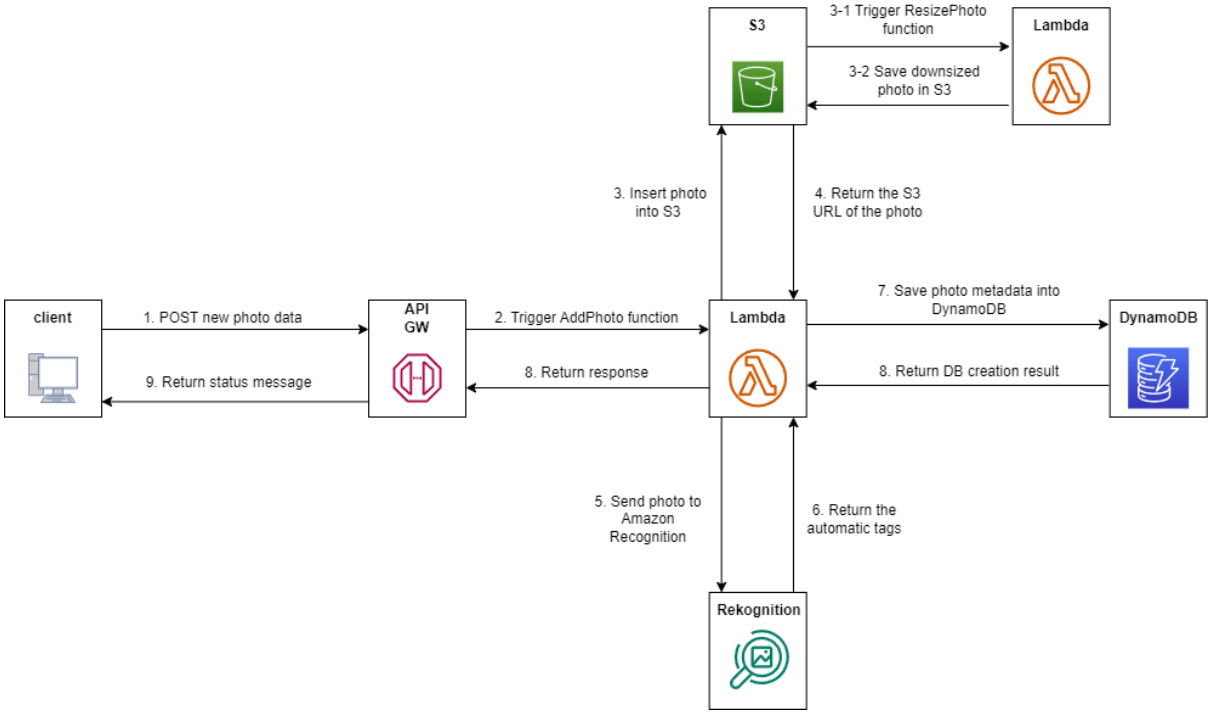


UML collaboration diagram

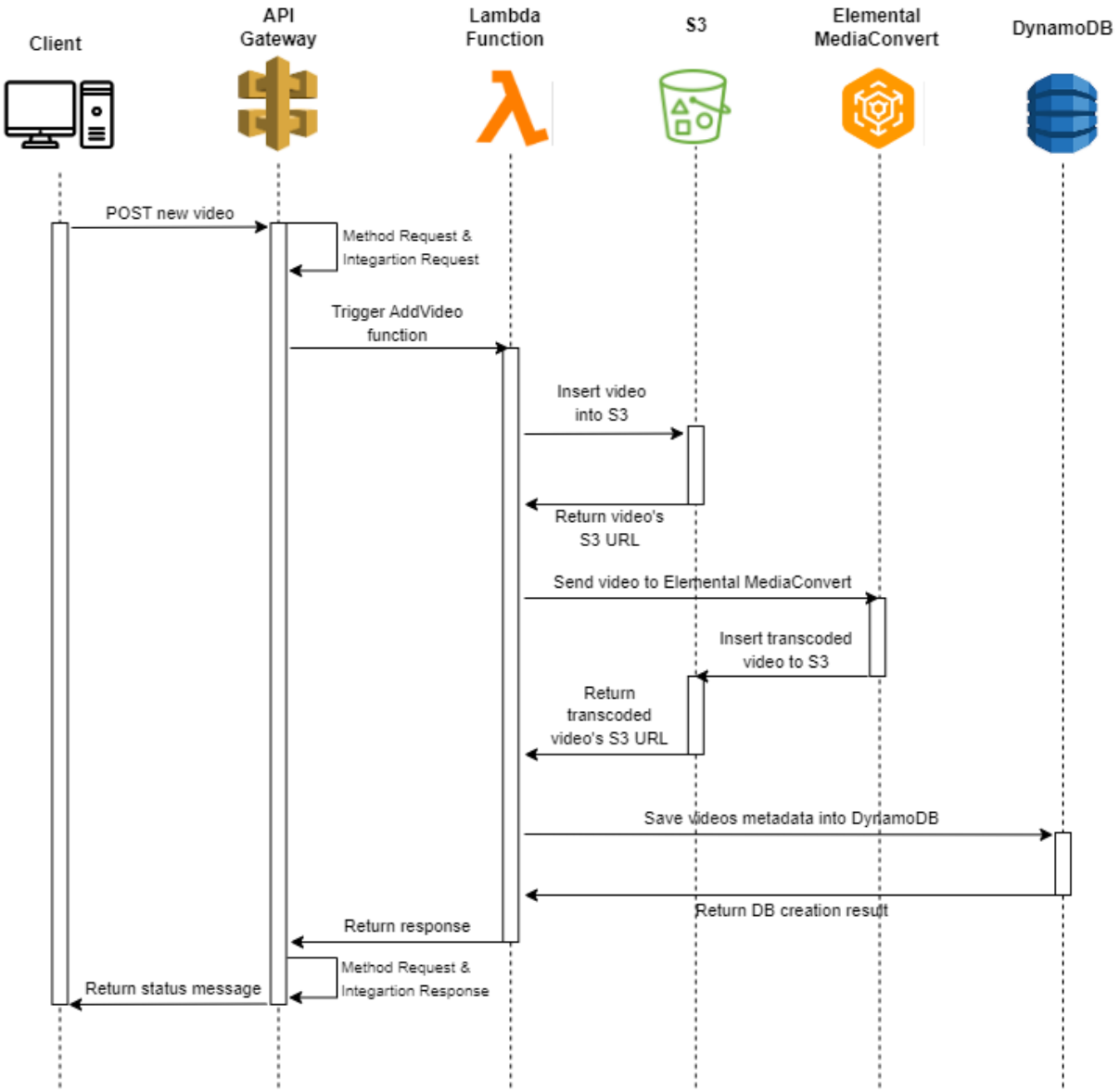
Use case 1. Retrieve the list of photos and their metadata



Use case 2. Post a photo



Use case 3. Post a video



Description of the services and how the business scenario is fulfilled using the proposed services

Route53

Route53 serves as a highly scalable and available Domain Name System that helps balance the traffic across many different regions.

Requirement fulfilment: With Route53's latency-based policy, the business can register a single domain name, and let Route53 direct the request to the location that entails the lowest latency. This contributes to reducing the global response time.

CloudFront

CloudFront helps the distribution of static and dynamic web content faster. CloudFront replicates the content in multiple geo locations across the world. Therefore, the response time

is boosted because the users receive the responses from the server cache that is closest to them instead of from the original region where the content is created.

Requirement fulfilment: along with Route53, CloudFront reduces the response time for users who reside outside Australia. CloudFront also helps decouple the presentation and the logic layer of the application.

API Gateway

API Gateway is the serverless alternative to EC2 instances. API Gateway supports the creation and maintenance of REST APIs without the need of a REST server.

Requirement fulfilment: API Gateway well aligns with the company's long term strategies to go serverless. API Gateway is automatically scaled in accordance with the traffic demand, therefore, the development team does not need to worry about provisioning the server when the demand is changing rapidly.

Lambda

Lambda is the serverless service that automatically invokes the code execution in response to events. Lambda can be integrated with API Gateway to implement the logic for each HTTP endpoint, or can be used to respond to photo and video insertion.

Requirement fulfilment: Along with API Gateway, Lambda is suitable for the company's scenario whose traffic demand is increasing rapidly. Lambda is versatile enough to support different functional use-cases, including retrieving media content, thumbnail creation and invocation of media transcoding.

S3

Amazon Simple Storage Service (Amazon S3) is a managed cloud storage solution that is scalable and offers 99.999999999% uptime. In addition, numerous management features are provided so that the business can easily organise and optimise the storage to meet its specific needs. Amazon S3 manages data using an object storage architecture that aims for scalability, high availability, low latency, and high durability. Objects in Amazon S3 are organised into buckets. Each object is identified by a unique, user-assigned key and can be as large as five terabytes in size.

Requirement fulfilment: Amazon S3 has a vast storage capacity and can store objects in any format. It can also be used in conjunction with other services, such as the Lambda function and Elemental MediaConvert, to automatically produce reformatted/transcoded versions of uploaded files.

DynamoDB

DynamoDB is a NoSQL database fully managed service offered by AWS. The company can build database tables using DynamoDB that can handle any volume of request traffic and store and retrieve any quantity of data. The throughput capacity of tables can be scaled up or down without a drop in performance.

Requirement fulfilment: DynamoDB helps the company to solve financial problems whose scenario is the cost of the relational database is high and has slow speed. The company can also reduce the cost of scaling a distributed database based on the horizontal scaling feature of DynamoDB.

Rekognition

Rekognition is the computer vision service to extract insights from media content such as photos and videos. Rekognition offers pre-trained and customised AI models that are ready to be integrated into the cloud architecture.

Requirement fulfilment: Rekognition helps the company cope with the future use-case to extract tags from photos, leveraging AI technology. Rekognition reduces the cost of building, training and deploying the models.

Cognito

Amazon Cognito is the user identity and access management service. Cognito allows the developers to add user login and signup features into the application. Cognito also allows fine grain access to different AWS services in accordance with the user roles.

Requirement fulfilment: With the AWS Free Tier, Amazon supports 50000 active users per month without a fee. The company can leverage the security provided by Amazon Cognito instead of building and maintaining their own access management service.

Elemental MediaConvert

Elemental MediaConvert is the transcoding service that supports multiple input and output formats. MediaConvert processes videos with the on-demand price models based on the duration of the processed videos.

Requirement fulfilment: MediaConvert is a suitable candidate to implement the video transcoding features. It removes the need to provision and manage the video processing infrastructure. There is no upfront cost, and the company only needs to pay for the video output duration.

Design Rationale

Alternative solutions

Number of tiers

In terms of tier-based architecture, a web service can have either two tiers (presentation-data) or (presentation-logic-data). In accordance with the business requirement on low coupling, the three-tiered architecture is favourable as it enables clear separation between logic and data layer.

Computing Services

Table 2.1 summarises the functional requirements of the album application.

Table 2.1. Photo album functional requirements

Task number	Description	Level of frequency	Acceptable latency

1	Respond GET requests from clients to retrieve the photo metadata (title, data creation, photo url and tags) from the database.	High frequency	Must be fast
2	Handling POST requests from clients containing the metadata of the new photos and the photos' content. If all the metadata is valid, the computing server performs a CREATE request to append the metadata into the database, and triggers S3 to upload the photo content into the bucket.	Low frequency	Reasonably fast
3	After each photo is uploaded into the storage, the thumbnail is created by reducing its size.	Low frequency	Reasonably fast
4	Customers sometimes upload media instead of photos. In such a case, a transcoding task must be initiated to convert the original video format into the targeted format.	Very low frequency	A video uploading task can be time-consuming

TASK 1. WEB CONTENT DELIVERY

The first initiative is to decouple the website delivery away from the other application tasks, admitting the current approach has several drawbacks. Currently, EC2 instances are used to both serve website content delivery as well as handle photo upload and other future tasks, making it highly coupling. Secondly, EC2 instances reside in a single region (us-east-1), thus making a number of users suffer from slow access to the application. The developer team also needs to manage access control and security on their own which increases the overall operational cost.

CloudFront can be used to promote low latency for global access and application low coupling. With CloudFront, the web content is available across the AWS content delivery networks (CDN), and the client only needs to request from the nearest edge location, hence reducing latency and improving reliability. CloudFront improves low coupling by taking the website delivery task away from the computing resources, therefore the instances can be used to focus on other tasks. In terms of security, CloudFront provides an automatic shield against cyber attacks, reducing the burden from the operations team.

TASK 2. SERVERLESS LOGIC LAYER

Currently, the application logic to retrieve photo metadata and upload photo metadata is handled by EC2 instances. The approach might not be well-aligned with the future strategy to go serverless. Some other alternatives include

- **API Gateway:** API Gateway delegates the task of creating and maintaining servers to AWS, and the development team only needs to worry about the application logic. API Gateway allows stateless client-server communication and fully supports all REST methods, including GET, POST, PUT, PATCH and DELETE. To create the API gateway, developers need to specify the endpoint, the RESTFUL methods and which service that the API gateway is calling. The most suitable implementation for the

company's scenario is to let API gateway call different Lambda functions, which further utilise other AWS resources such as databases, S3 or transcoding services.

- **Fargate:** Fargate is another AWS serverless service that allows developers to run container-based applications without managing servers or clusters of EC2 instances. With Fargate, instead of maintaining the instances on its own, the development team focuses on the development of the application containers and lets AWS manage the deployment and scaling task. It is worth noting that Fargate still requires load balancers.

Given the functional requirement being relatively simple, API Gateway is more favourable as it does not require the infrastructure management, such as the creation and development of the containers. API Gateway is inherently highly available and scalable, and the application use cases such as retrieving photo metadata or uploading photos hardly exceeds the time limit of the API Gateway (15 seconds).

TASK 3. THUMBNAIL CREATION

The current task to create a thumbnail is implemented by Lambda. This implementation already satisfies the serverless/event-driven requirements and is reasonably scalable in the future.

TASK 4. VIDEO TRANSCODING TASK

There are three alternatives to handle the video transcoding task, including using EC2, Elastic Transcoder and MediaConverter.

EC2 instances can be used to transcode videos. As transcoding is a computationally heavy task, more powerful instance types must be used. Developer teams have to manage the transcoding logic on their own. These overall increase the cost for provisioning more expensive instances and also on development.

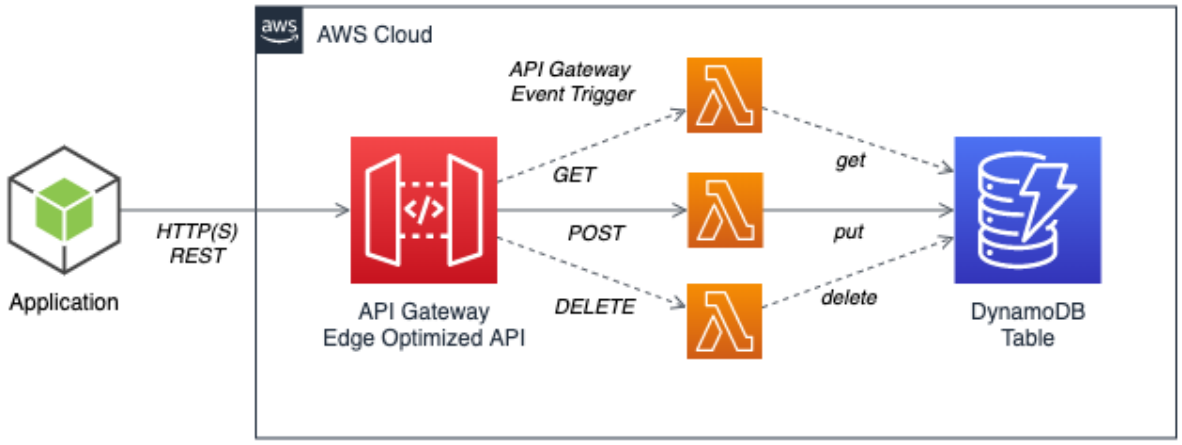
Elastic Transcoder is an AWS dedicated service for transcoding tasks. This decouples away the video transcoding features from the EC2 instances. The customers pay for the durations of the video output.

MediaConvert is another new AWS transcoding service that supports more input and output formats than Elastic Transcoder. MediaConverter shares the same pricing model as the Elastic Transcoder, and has a cheaper base price.

As per the application requirement, MediaConvert is the most suitable as it supports more input and output formats while maintaining a lower cost.

Database

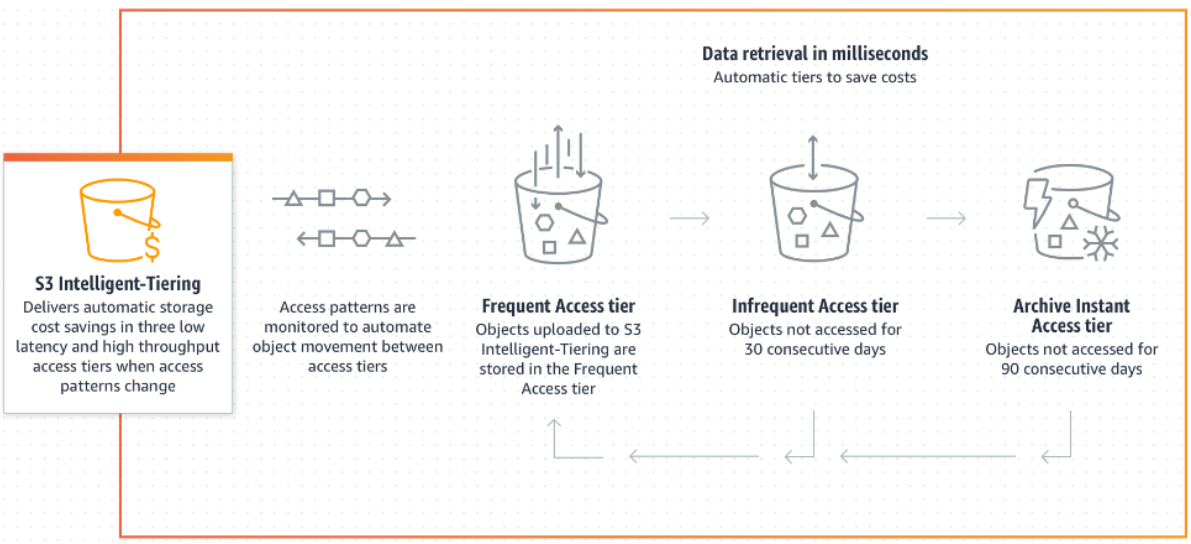
DynamoDB should be used to store information of the image as well as the video which are uploaded by users. Since the system scales up to double every 6 months and is expected to keep continuing for the next 3 years, which means it will double 6 times in total. Horizontal scaling would be a perfect option, making it possible to scale with less downtime. While vertical scaling is limited to the capacity of just one machine, which means in order to scale it up, 2^6 = 64 original hardwares is required, which makes it not sustainable. Furthermore, cost-effectiveness is also one of the target criteria. Compared to RDS, DynamoDB is a low-cost database with scalability, which is an affordable solution for the company. Which is why NoSQL database **(Amazon DynamoDB**) is the optimal option.



File Storage

Amazon S3 should be used to store media uploaded to the site. Choosing the most cost-effective storage class for S3 is difficult, as it depends on the access pattern. If we use S3 Standard, we will have millisecond-fast access and a high throughput. As more and more files are uploaded to the storage, the infrequently accessed ones will cost the company a significant amount of money. Alternatively, if we use classes with infrequent access (such as S3 One Zone-IA and S3 Glacier), we can optimise the cost, but the access time will be very slow. **Amazon S3 Intelligent-Tiering** provides the optimal solution. This is the world's only cloud storage class that automates cost savings for data with variable access patterns. It is designed to store objects in three access tiers and optimised for both frequent and infrequent access. S3 Intelligent-Tiering monitors access patterns and optimises cost on granular object level. Objects that have not been accessed frequently will be moved to lower-cost access tier. Furthermore, it has no impact on performance, no operational overhead and no retrieval fee.

When objects are uploaded to the Intelligent-Tiering class, they will be stored in Frequent Access Tier. Intelligent-Monitoring will then monitor every stored object's access for a small monthly monitoring and automation fee. Objects that have not been accessed in over 30 days will be moved to the Infrequent Access tier. They will be moved to the Archive Instant Access tier if they have not been accessed for over 60 days. If any of them are subsequently accessed, they will be moved back to the Frequent Access tier. The Infrequent Access tier and the Archive Instant Access tier can reduce storage costs by up to 40% and 68%, respectively.



Justification

Operational excellence

The architecture is designed with loose coupling principles in mind. The presentation, logic and data layer are effectively decoupled. For example, if the website content is updated, only CloudFront content is modified without changing the logic layer.

The architecture is highly scalable, and can cope with the changing demand automatically without manual operational efforts.

Performance

While being serverless, the overall architecture does not sacrifice performance. Regarding the presentation logic, there has been lots of successful evidence that uses the combination of Route53 and CloudFront to minimise the global latency. A successful example of using these two stacks is Slack. With CloudFront, Slack has accelerated the average global response time from 90 ms to 15 ms.

Regarding the logic layer, API Gateway and Lambda are capable of handling requests simultaneously. While the burst concurrency limits for the us-east and the asia-pacific are 3000 and 1000, respectively, these do not pose any issues, because current Lambda functions are designed to execute in just a few milliseconds. Therefore, it will be unlikely for the application to reach the burst limit.

Regarding the data layer, DynamoDB is automatically replicated across the globe, therefore also contributing to reducing the latency time across the globe.

Reliability

Amazon S3 provides a storage system built for mission-critical and primary data storage that is highly durable. It stores objects across a minimum of three Availability Zones on numerous devices. Amazon S3 is designed to provide 99.999999999% durability and 99.99% availability of objects over a given year.

Security

To maintain the security of the files, the objects in an S3 bucket should not be accessible to the public and their access should be restricted using S3 bucket policy. In addition, rather than storing the credentials for the S3 API, an IAM role will be created and assigned to the web server so that only the website can call the API. An IAM permission must be configured so that only the Lambda function has read/write access to the DynamoDB database.

For data protection purposes, each individual user account is configured with AWS Identity and Access Management, so that each user is granted only the permissions necessary to fulfil their job duties.

Cost Optimisation

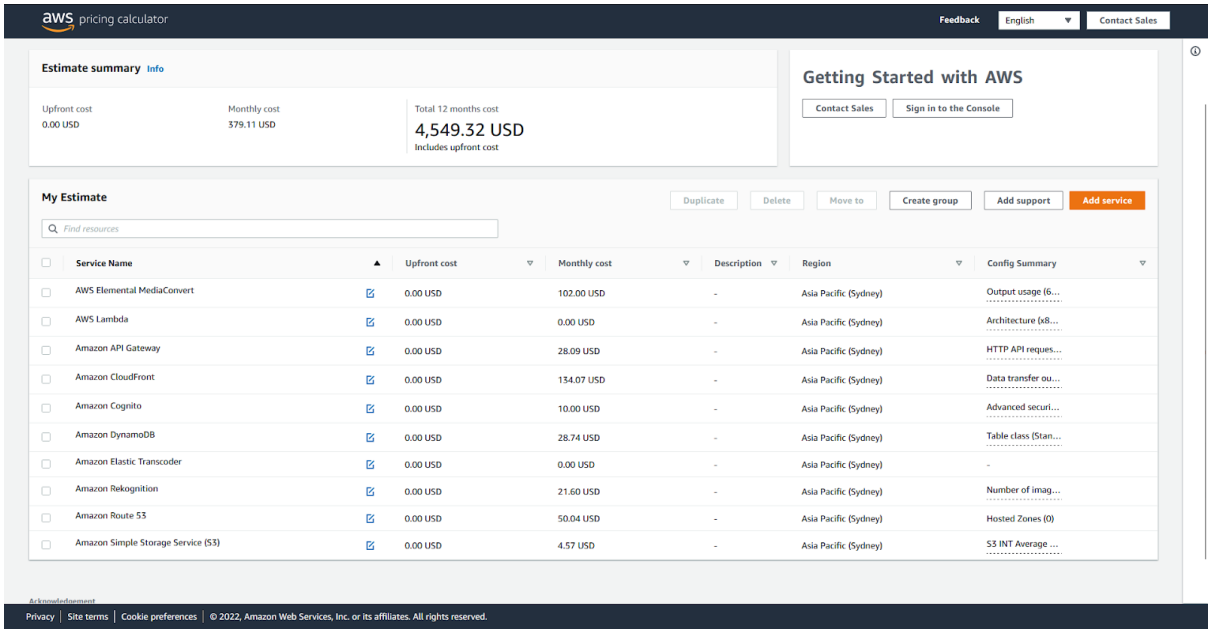
Amazon S3 Intelligent-Tiering automates storage cost savings when data access patterns by transferring data to the most economical access tier. For a small monthly object monitoring and automation charge, S3 Intelligent-Tiering can help businesses save up to 68% of storage cost.

Estimated Cost

The estimate is performed on the following assumptions

Parameter	Value
Number of current users	200
Number of read operations per day per user	10
Number of write (photo) operations per day per user	2
Number of write (video) operations per day per user	0.5
Average photo size	3MB
Average	

Regarding to the demand of users, the estimated cost for the architecture using all services which are provided in the diagrams was calculated as below:



Estimate summary

Upfront cost	Monthly cost	Total 12 months cost	Currency
0	379.11	4549.32	USD
		* Includes upfront cost	

Detailed Estimate

Group hierarchy	Region	Description	Service	Upfront	Monthly	First 12 months total	Currency
My Estimate	Asia Pacific (Sydney)		Amazon Elastic Transcoder	0	0	0	USD
My Estimate	Asia Pacific (Sydney)		AWS Elemental MediaConvert	0	102	1224	USD
My Estimate	Asia Pacific (Sydney)		DynamoDB on-demand capacity	0	28.74	344.88	USD
My Estimate	Asia Pacific (Sydney)		Rekognition Image	0	21.6	259.2	USD
My Estimate	Asia Pacific (Sydney)		S3 Intelligent - Tiering	0	4.57	54.84	USD
My Estimate	Asia Pacific (Sydney)		Amazon CloudFront	0	134.07	1608.84	USD
My Estimate	Asia Pacific (Sydney)		Amazon Route 53	0	50.04	600.48	USD
My Estimate	Asia Pacific (Sydney)		Amazon Cognito	0	10	120	USD
My Estimate	Asia Pacific (Sydney)		Amazon API Gateway	0	28.09	337.08	USD
My Estimate	Asia Pacific (Sydney)		AWS Lambda	0	0	0	USD

Acknowledgement

* AWS Pricing Calculator provides only an estimate of your AWS fees and doesn't include any taxes that might apply. Your actual fees depend on a variety of factors, including your actual usage of AWS services.

References

Amazon 2021, Amazon API Gateway, <https://docs.aws.amazon.com/apigateway/latest/developerguide/welcome.html>

Hava 2021, *What is AWS CloudFront and does it make a difference?*, <https://www.hava.io/blog/what-is-aws-cloudfront-and-does-it-make-a-difference>, accessed on 23 Oct 2022.

(n.d.)., *Data protection in Amazon S3 - Amazon Simple Storage Service*. , <https://docs.aws.amazon.com/AmazonS3/latest/userguide/DataDurability.html>, accessed on 23 OCt 2022.

Amazon 2021, *What is Amazon DynamoDB?,*

<https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Introduction.html>, accessed on 23 Oct 2022.

Amazon 2021, *Relational (SQL) or NoSQL? <https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/SQLtoNoSQL.WhyDynamoDB.html>*, accessed on 23 Oct 2022.