AWS Reference Architectures

Amazon EC2

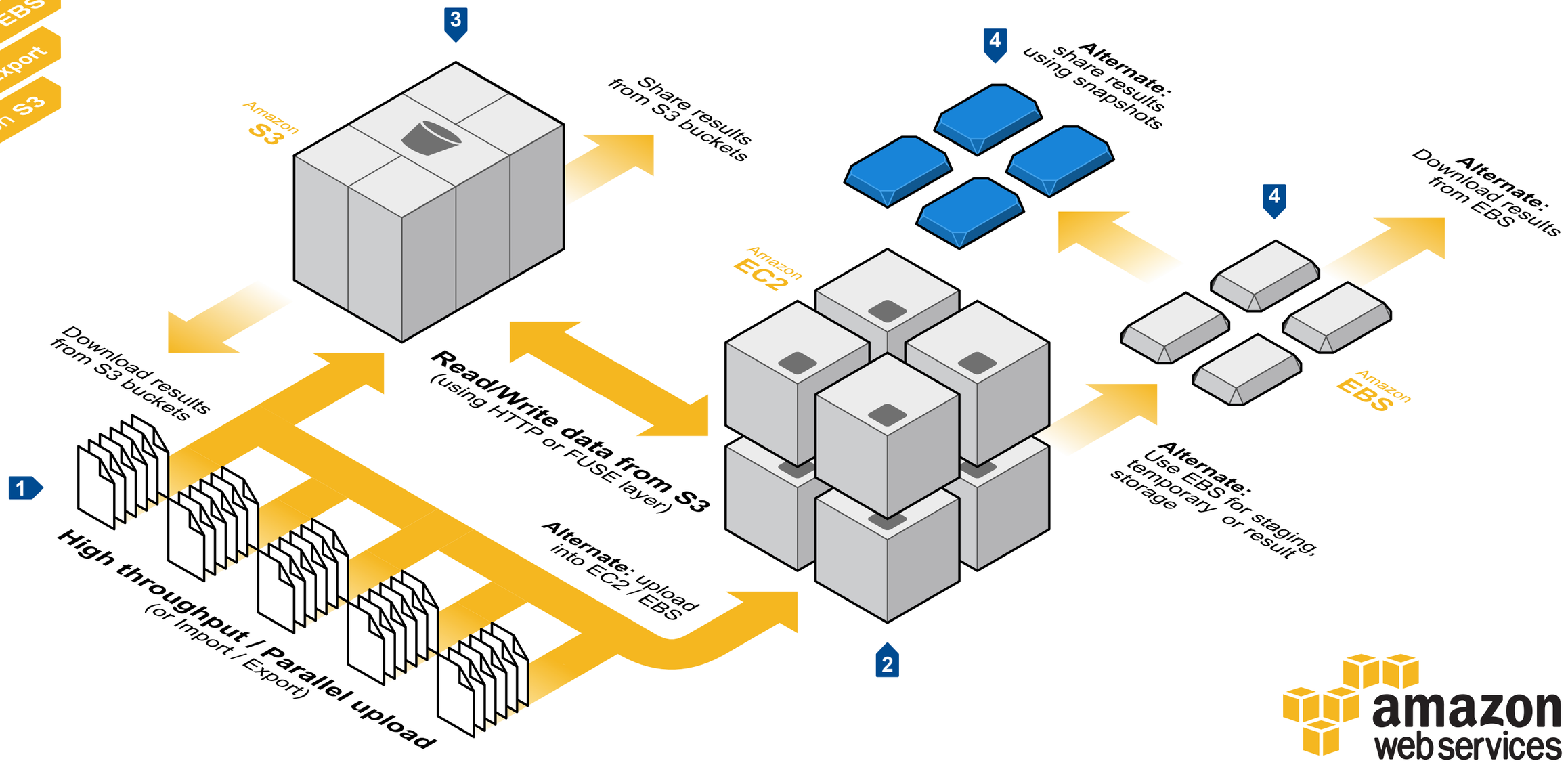
Amazon EBS

AWS Import / Export

Amazon S3

LARGE SCALE COMPUTING & HUGE DATA SETS

Amazon Web Services is very popular for large-scale computing scenarios such as scientific computing, simulation, and research projects. These scenarios involve huge data sets collected from scientific equipment, measurement devices, or other compute jobs. After collection, these data sets need to be analyzed by large-scale compute jobs to generate result data sets. Ideally, results will be available as soon as the data is collected. Often, these results are then made available to a larger audience.



amazon web services

System Overview

1 To upload large data sets into AWS, it is critical to make the most of the available bandwidth. You can do so by uploading data into **Amazon Simple Storage Service (S3)** in parallel from multiple clients, each using multithreading to enable concurrent uploads or multipart uploads for further parallelization. TCP settings like window scaling and selective acknowledgement can be adjusted to further enhance throughput. With the proper optimizations, uploads of several terabytes a day are possible. Another alternative for huge data sets might be **Amazon Import/Export**, which supports sending storage devices to AWS and inserting their contents directly into **Amazon S3** or **Amazon EBS** volumes.

2 Parallel processing of large-scale jobs is critical, and existing parallel applications can typically be run on multiple **Amazon Elastic Compute Cloud (EC2)** instances. A parallel application may sometimes assume large scratch areas that all nodes can efficiently read and write from. S3 can be used as such a scratch area, either directly using HTTP or using a FUSE layer (for example, s3fs or SubCloud) if the application expects a POSIX-style file system.

3 Once the job has completed and the result data is stored in **Amazon S3**, **Amazon EC2** instances can be shut down, and the result data set can be downloaded The output data can be shared with others, either by granting read permissions to select users or to everyone or by using time limited URLs.

4 Instead of using **Amazon S3**, you can use **Amazon EBS** to stage the input set, act as a temporary storage area, and/or capture the output set. During the upload, the concepts of parallel upload streams and TCP tweaking also apply. In addition, uploads that use UDP may increase speed further. The result data set can be written into EBS volumes, at which time snapshots of the volumes can be taken for sharing.