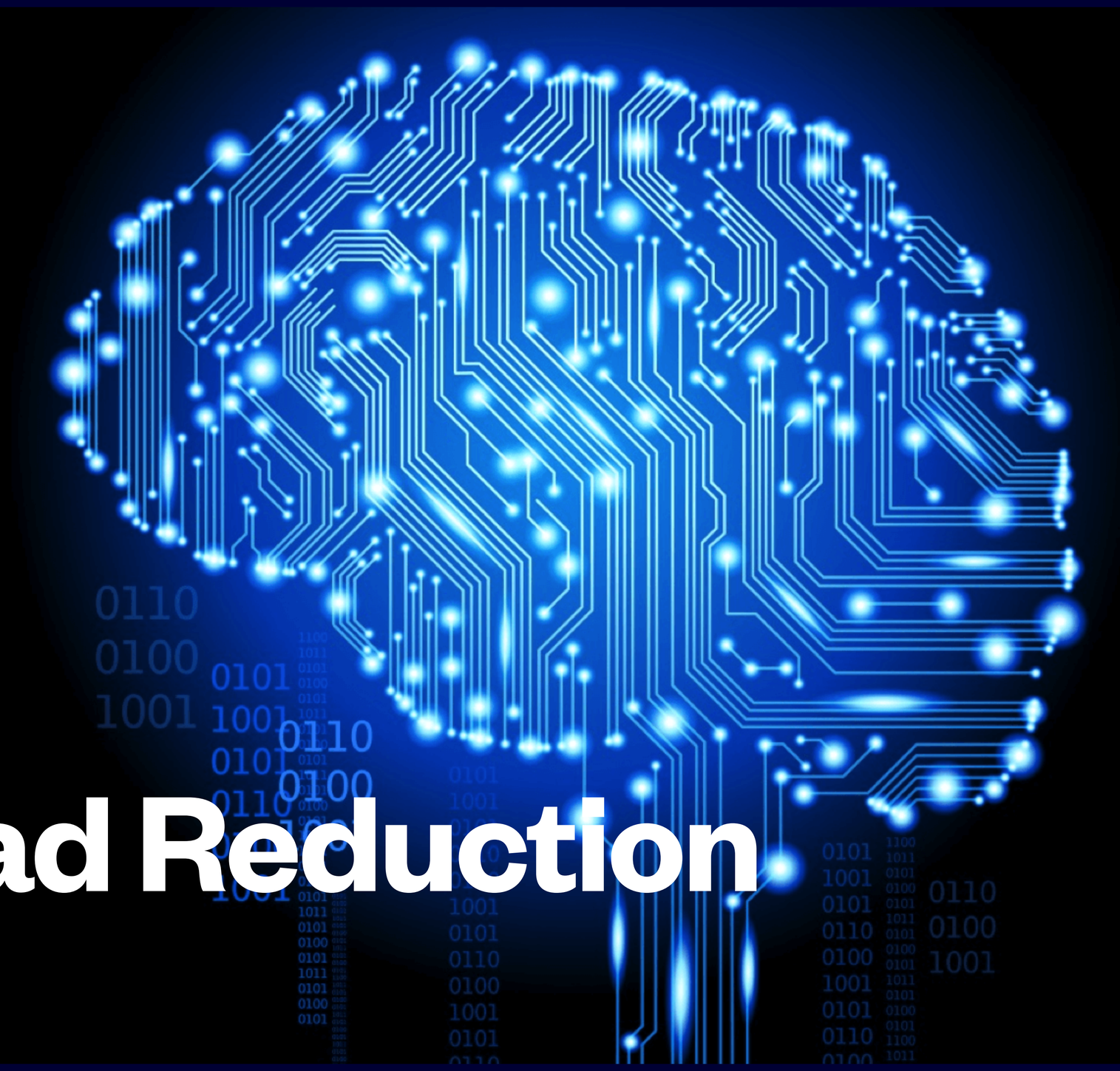
Four methods



AI Cloud Load Reduction

MAO YI

1: PREPROCESS

Client and preprocess data which will be sent to cloud.

This basically transfer some computation from cloud to client.

2: COMPRESSION

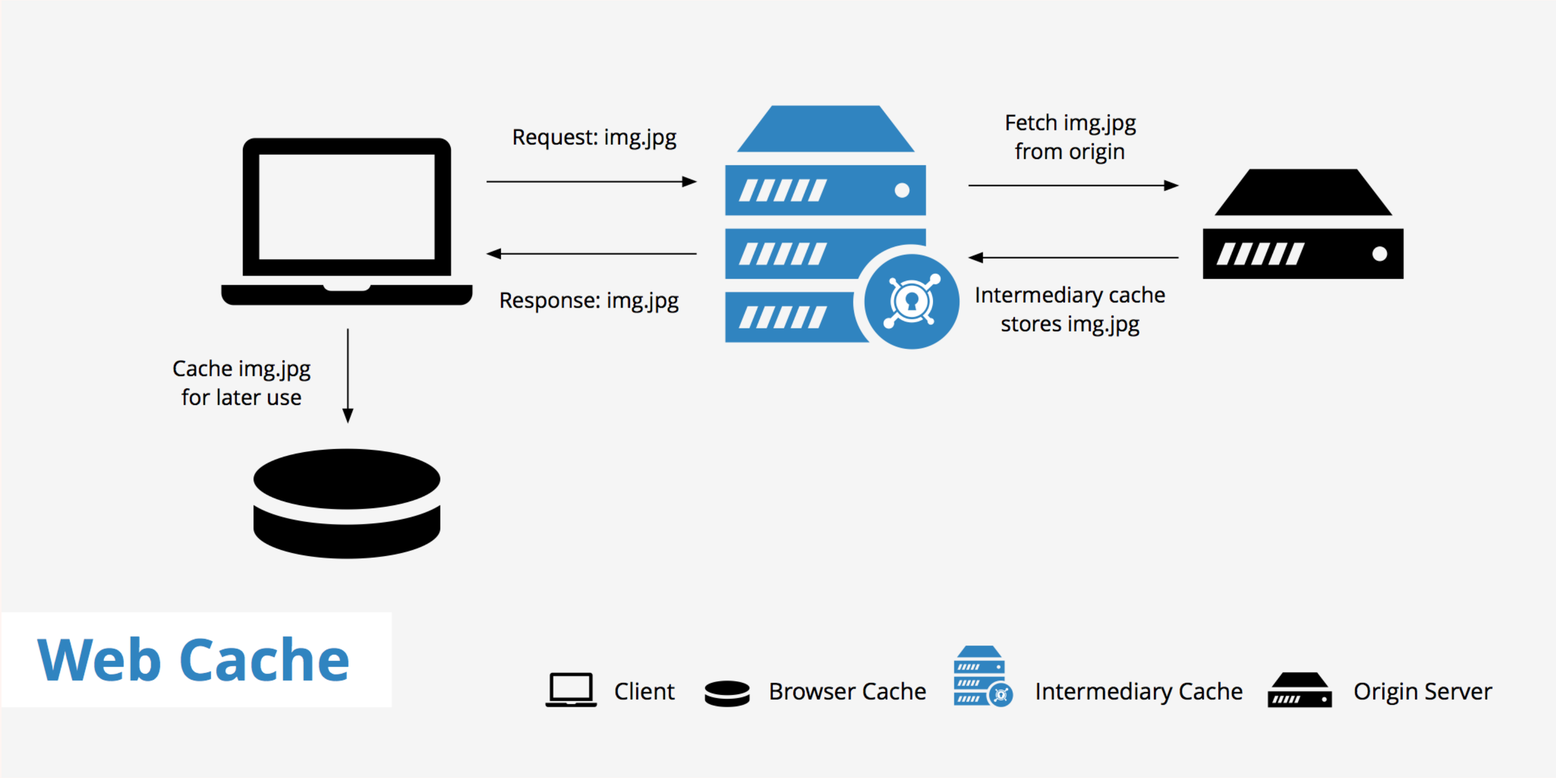
Cloud could compresses some big files, like models, in advance.

This would reduce the load of transmitting files.

3: CACHING

Cache some files like models would also be helpful to reduce the load.

This would reduce the load of transmitting files.



4: CDN

A lot of cloud server company provide CDN service(Content Delivery Network).

This would free AI servers, make them focus on computation.