SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Video Processing with Serverless Computing: A Measurement Study

Miao Zhang, Yifei Zhu, Cong Zhang, Jiangchuan Liu

OUTLINE

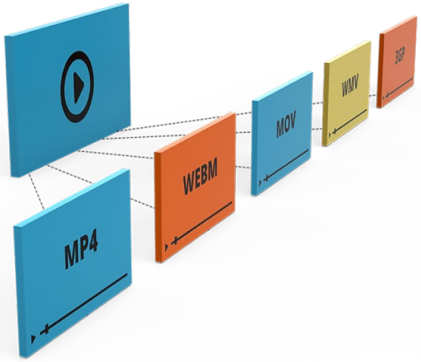
- A Brief Introduction of Video Processing
- Why Serverless Computing?
- Existing Efforts
- A Measurement Study
- Summary & Future Directions

BCKGROUND

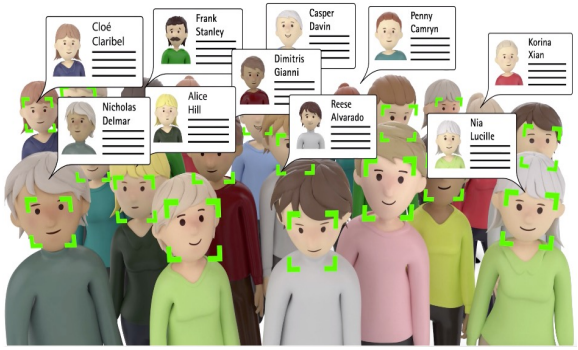
Video Processing:

Video conversion tasks (e.g., compression, transcoding, editing)

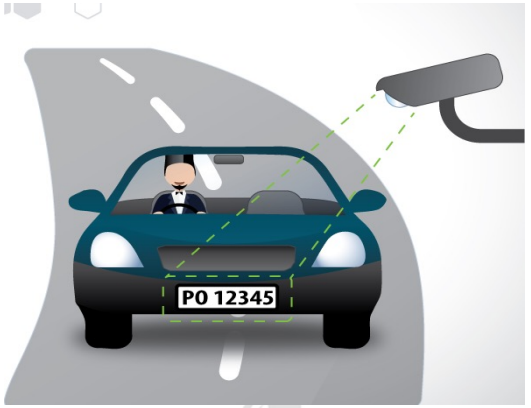
Video analytics tasks (e.g., scene recognition, face detection)



Video format conversion



Identification



Intelligent traffic system

BCKGROUND

Current Trends

- **More cameras.**

  More videos. More opportunities.

  Extracting information and properly responding are increasingly difficult.

- **Higher quality (Ultra HD videos, e.g., 4K, 8K) .**

  Better viewing experiences.

  Higher burden on video conversion tasks.

- **Advances in computer vision algorithms.**

  Higher accuracy [1].

  Higher cost. A object detector [2] processes only 1.2 frames/s on a GPU.

How to achieve low-latency and cost-effective video processing?

[1] J. Jiang et al. Chameleon: scalable adaptation of video analytics. In ACM SIGCOMM 2018.

[2] X. Zhu et al. Flow-guided feature aggregation for video object detection. In IEEE ICCV 2017.

OUTLINE

- A Brief Introduction of Video Processing
- Why Serverless Computing?
- Existing Efforts
- A Measurement Study
- Summary & Future Directions

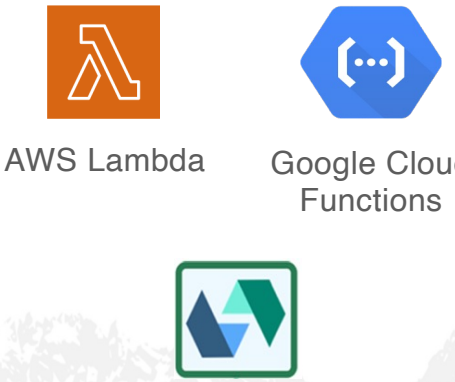
MOTIVATION

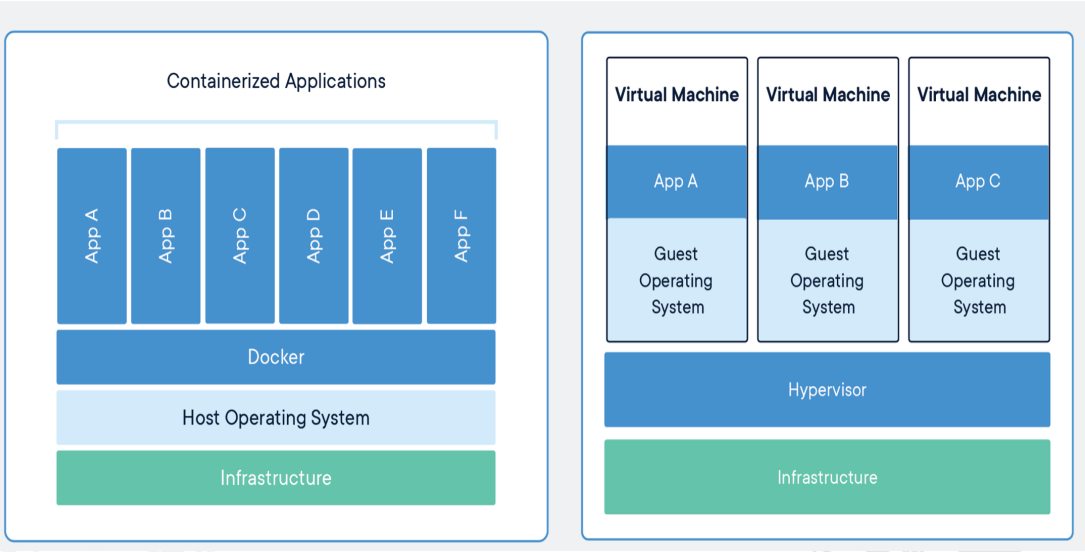
Why serverless computing?

Light-weight Implementation

Map each function instance into its own container [1].

Launch thousands of parallel function instances in milliseconds.





Comparing Containers and Virtual Machines [2]

[1] I. E. Akkus et al. SAND: Towards High-Performance Serverless Computing. In USENIX ATC 2018.

[2] https://www.docker.com/resources/what-container.

MOTIVATION

Why serverless computing?

Reduced Cost

Pay-as-you-go pricing strategy.

Fine-grained billing (e.g., 100ms).

Table 1: Pricing Schemes of Serverless Computing Platforms (beyond free tiers)[α]

	Symbol	AWS Lambda	GCF
Price per invocation	I	\$0.0000002	\$0.0000004
Memory (CPU)	M (P)	{128(p), ..., 3008(23.5p)}	{128(200), 256(400), 512(800),1024(1400), 2048(2400)}
Price per 100ms	C	$10^{-10} * 16.28M$	$10^{-10} * (2.44M + 10P)$

[α] : The unit of memory size is MB, the unit of CPU is MHz and the unit of price is US dollar; p is unknown to users.

Reduced Operational Management

Automatic scaling and monitoring are provided by cloud providers.

OUTLINE

- A Brief Introduction of Video Processing
- Why Serverless Computing?
- **Existing Efforts**
- A Measurement Study
- Summary & Future Directions

EXISTING EFFORTS

Academia

Excamera [1]:

- provides a framework **mu** to run 5,000-way parallel jobs.
- designs a video codec for massive fine-grained parallelism.

Sprocket [2]:

- orchestrates video pipelines with a domain-specific language.
- exploits intra-video parallelism to achieve low-latency.

[1] S. Fouladi et al. Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads. In USENIX NSDI 2017.

[2] L. Ao et al. Sprocket: A Serverless Video Processing Framework. In ACM SoCC 2018.

EXISTING EFFORTS

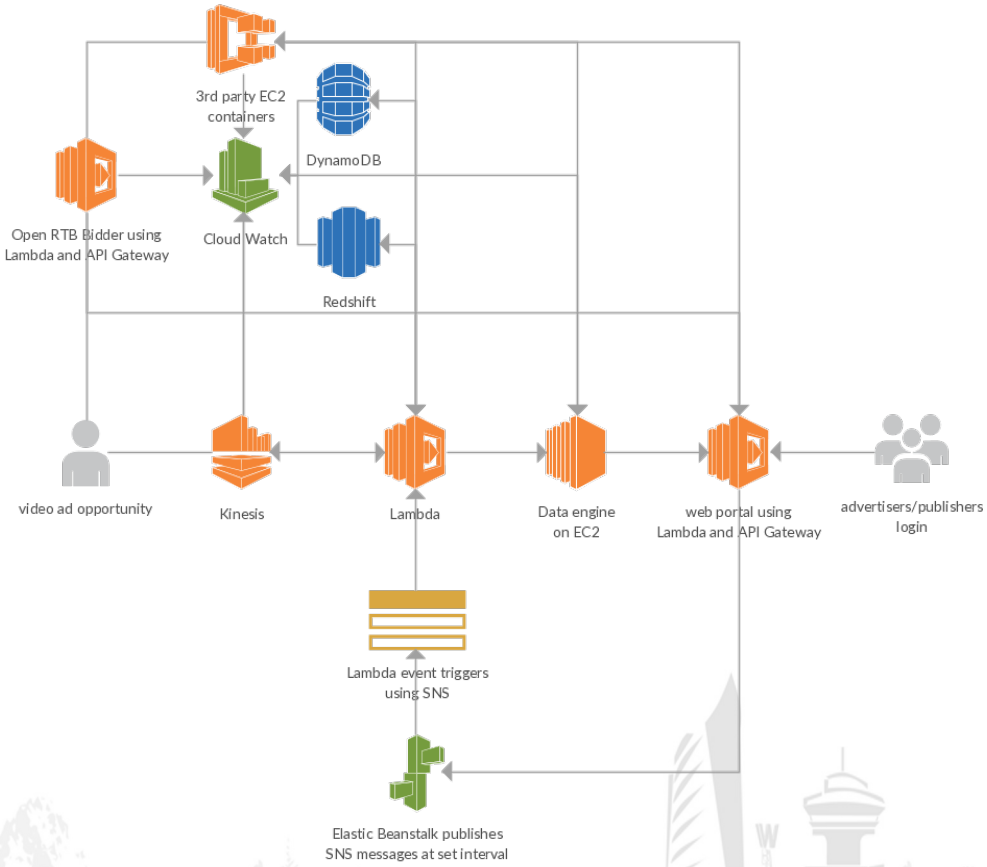
Industry

Vidroll [1] :

- real-time ads bidding.
- real-time ads transcoding.

Netflix [2] :

- self-managing infrastructure.
- replace inefficient processes.



VidRoll Architecture on AWS

[1] https://aws.amazon.com/solutions/case-studies/vidroll/.

[2] https://aws.amazon.com/solutions/case-studies/netflix-and-aws-lambda/.

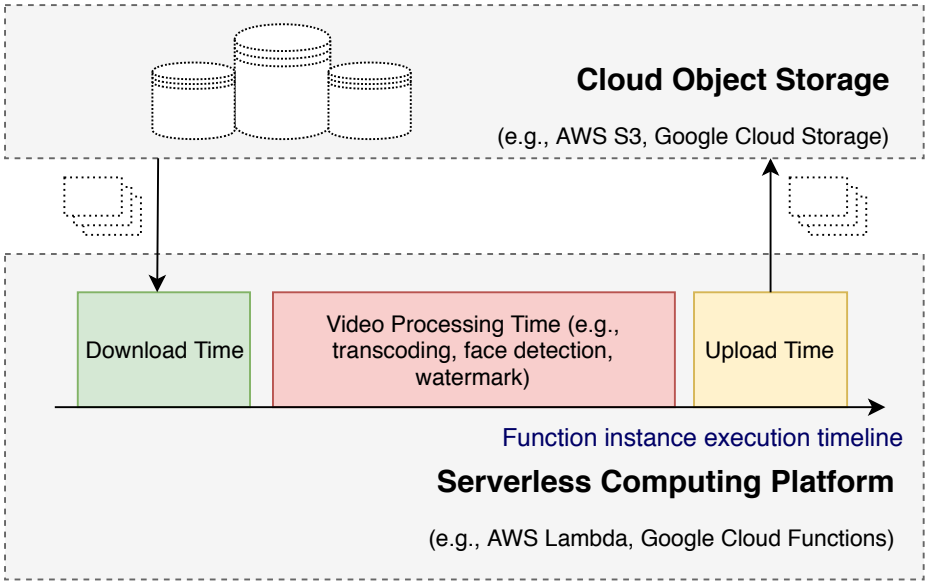
OUTLINE

- A Brief Introduction of Video Processing
- Why Serverless Computing?
- Existing Efforts
- **A Measurement Study**
- Summary & Future Directions

MEASUREMENT

Setup

- **Platform:**
  - AWS Lambda
  - Google Cloud Functions
- **Runtime:**
  - Python3.7
- **Applications:**
  - Transcoding (FFmpeg [1])
  - Face detection (MTCNN [2])
- **Metrics:**
  - Function execution duration
  - Monetary cost



Measurement function framework

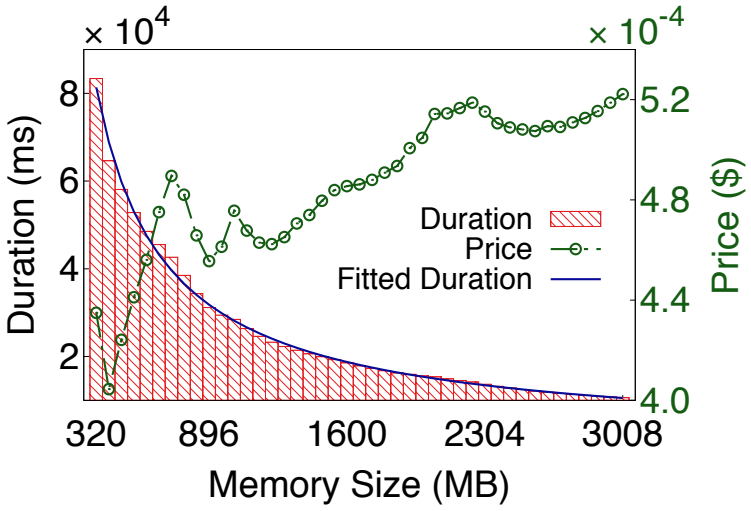
[1] https://ffmpeg.org.

[2] K. Zhang et al. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23, 10 (2016), 1499-1503.

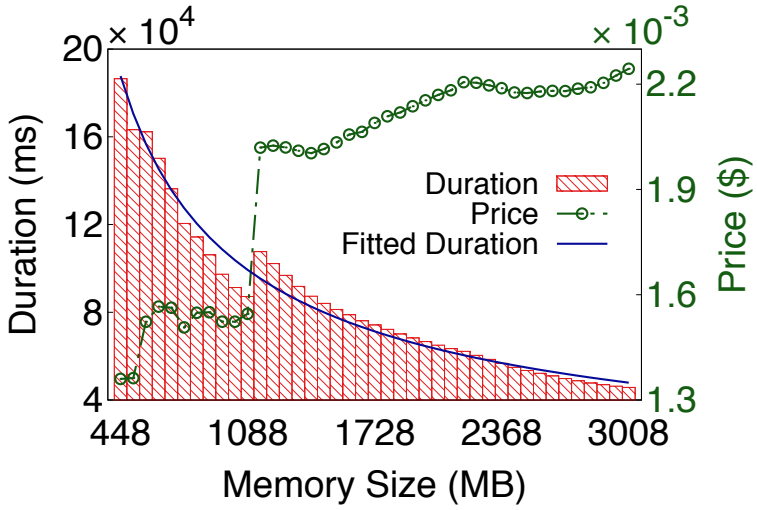
MEASUREMENT

Results

Function Configuration



(a) *Transcoding* function deployed with AWS Lambda



(b) *Face detection* function deployed with AWS Lambda

MEASUREMENT

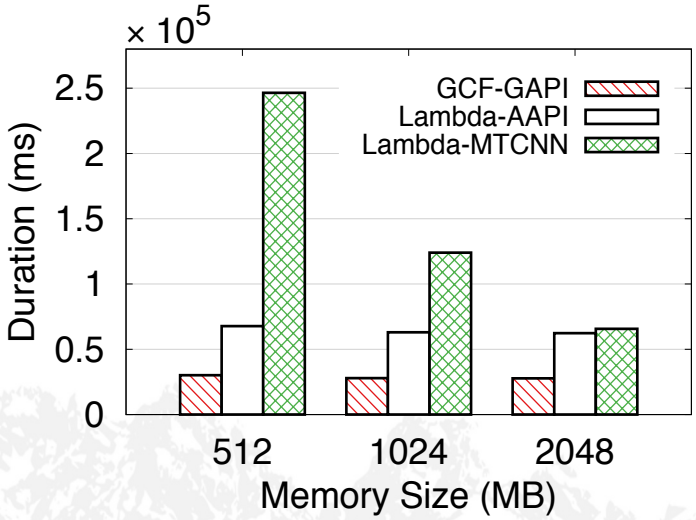
Results

Function Implementation Scheme

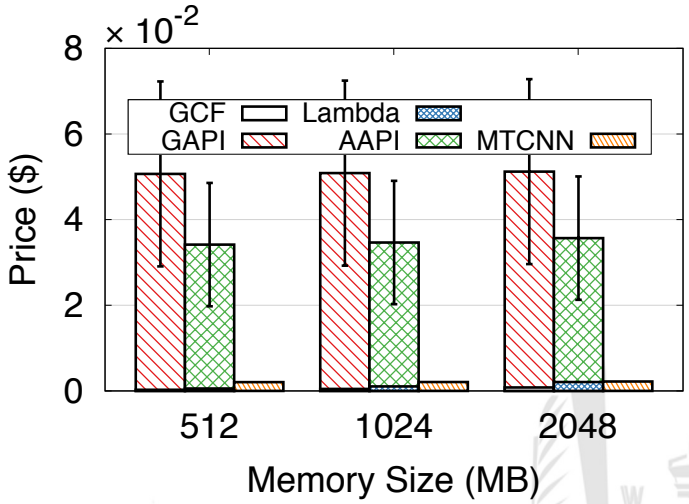
GCF-GAPI: combing Google Cloud Functions with Google Cloud Vision API [1].

Lambda-AAPI: combing AWS Lambda with Amazon Rekongnition Image API [2].

Lambda-MTCNN: a MTCNN model deployed with AWS Lambda function.



(a) Face detection execution duration



(b) Monetary cost

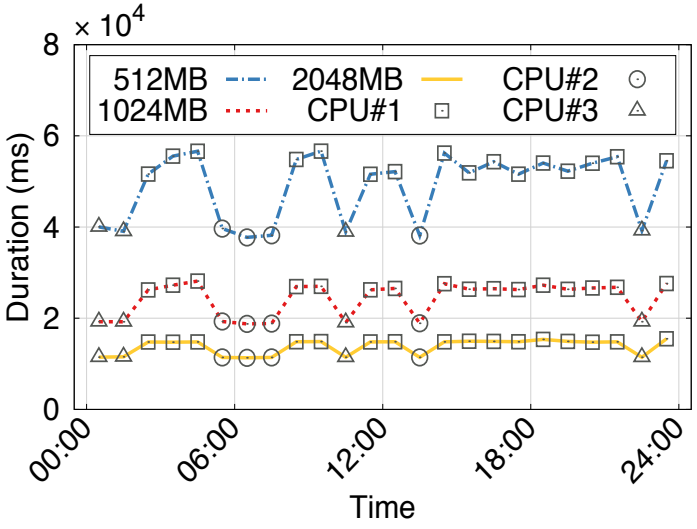
[1] https://aws.amazon.com/rekognition/image-features/.

[2] https://cloud.google.com/vision/.

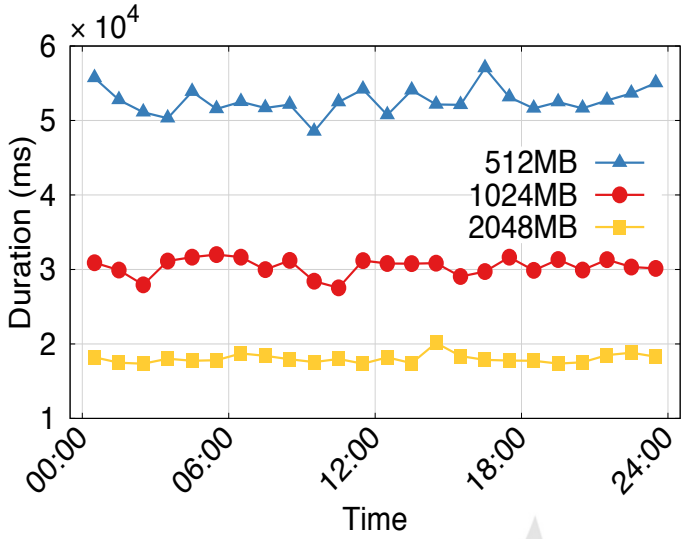
MEASUREMENT

Results

Insights into System Factors



(a) *Transcoding* function execution duration changes in one day (AWS Lambda).

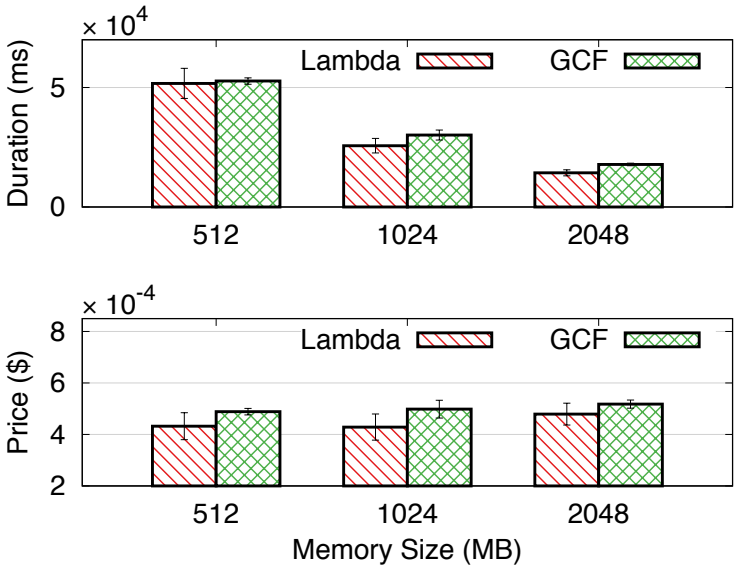


(b) *Transcoding* function execution duration changes in one day (GCF).

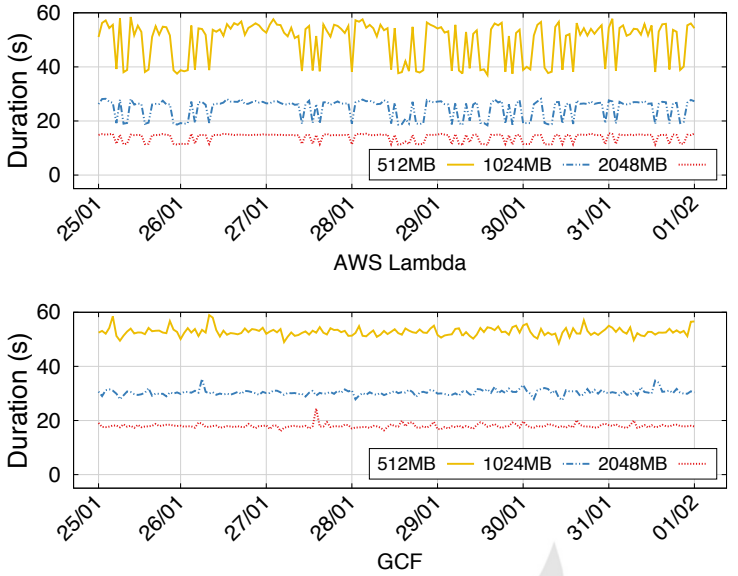
MEASUREMENT

Results

Platform Comparison



(a) Execution duration and monetary cost of *transcoding* function deployed with AWS Lambda and GCF.



(b) *Transcoding* function execution duration changes within one week (25/01/2019-31/01/2019).

OUTLINE

- A Brief Introduction of Video Processing
- Why Serverless Computing?
- Existing Efforts
- A Measurement Study
- Summary & Future Directions

SUMMARY

- Serverless computing is a good fit for building low-latency and cost-effective video processing applications.
- Dynamic profiling of workloads is necessary for finding the best resource configuration of video processing functions.
- Running pre-trained models in serverless functions locally has latency and cost advantages over calling external APIs.
- The performance of serverless video processing applications is platform dependent.

FUTURE DIRECTIONS

Serverless Function Configuration Optimization:

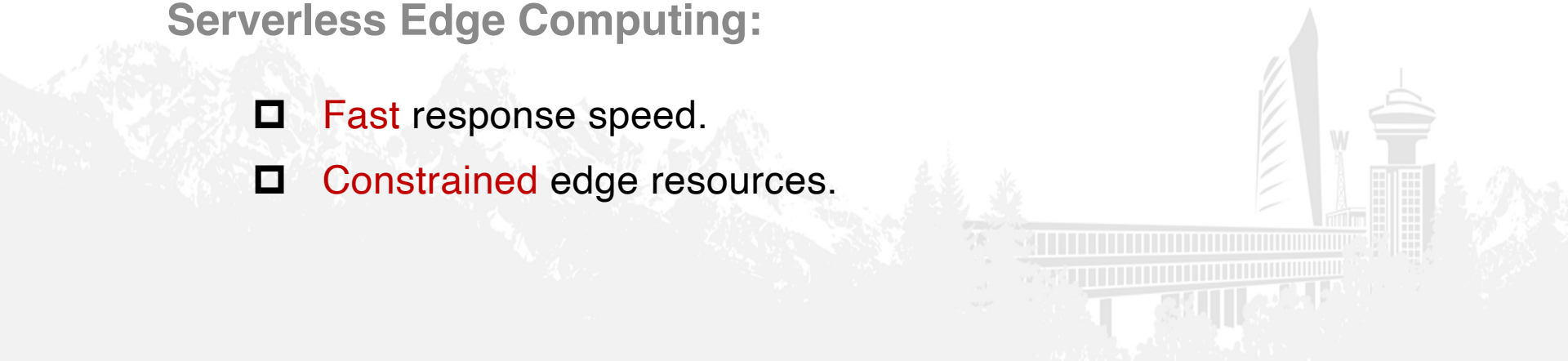
- Cost-efficient and scalable applications.
- Large configuration space.

Serverless Deep Learning:

- High video processing performance.
- Constrained resources (no GPU support).

Serverless Edge Computing:

- Fast response speed.
- Constrained edge resources.



THANK YOU