URU

Video Understanding at Scale

Deep Learning in a Serverless Infrastructure

@uruvideo

@uruvideo

uruvideo.com

Who am I?

Brunno Attorre

Co-Founder & CTO at Uru

- ML and Distributed Systems at J.P. Morgan and Brazilian startup Buscapé.
- ML, AI and Big Data writer for the Brazilian Java Magazine.
- Masters in Computer Science from Cornell University.
- Bachelors in Computer Science from Mackenzie University in Brazil.

URU java Magazine buscapé J.P.Morgan CORNELL UNIVERSITY FOUNDED A.D. 1865 M

What we do at Uru.

What we do at Uru.

BRAND SAFETY API

STORYBREAK API

CONTENT RECOGNITION API

PRODUCT LISTINGS API

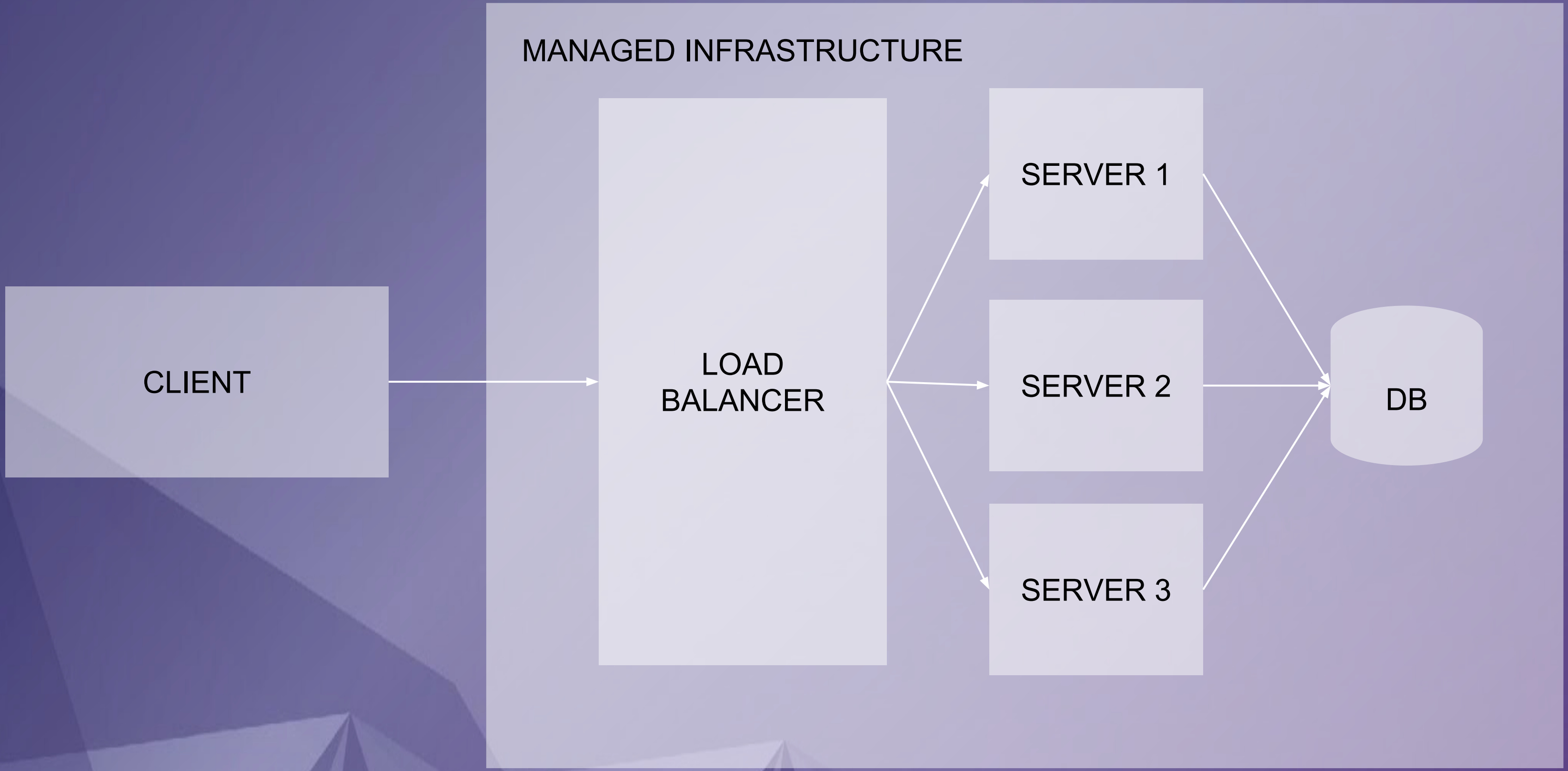
What is this talk about

How we used a serverless infrastructure to scale our Deep Learning models (and what we learned from that).

Agenda

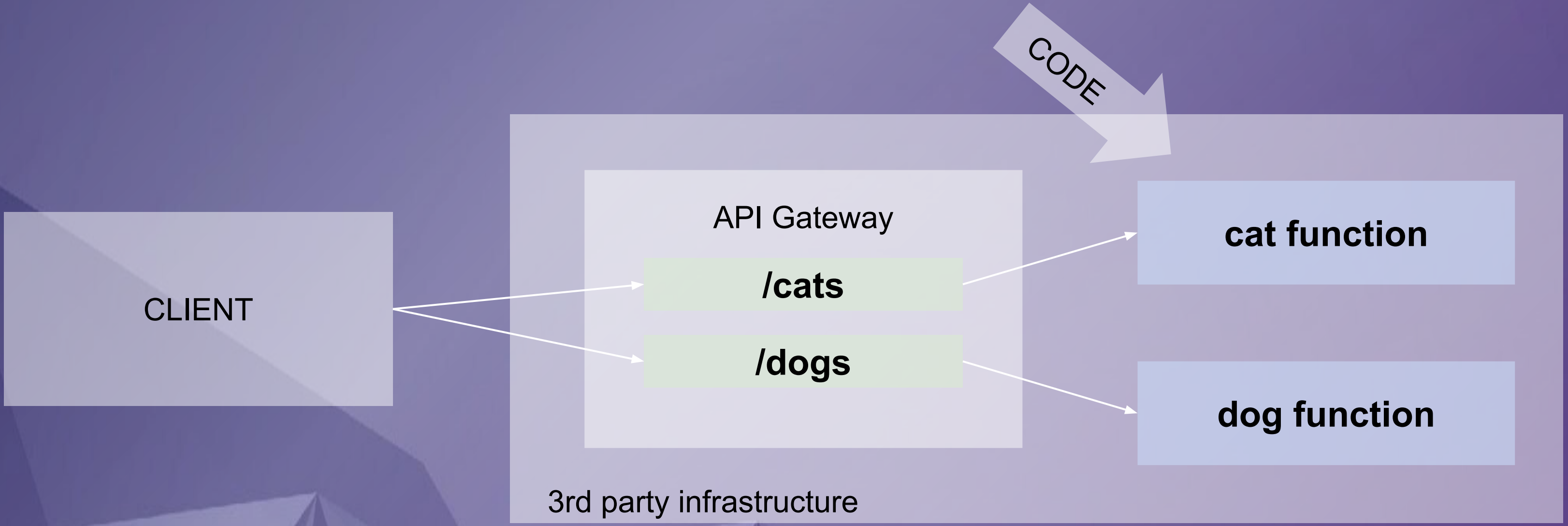
1. What is a serverless architecture?
2. Understanding video with AI.
3. Our experience at Uru:
    a. Conception of a sequential pipeline.
    b. Scaling it up with a parallel serverless infrastructure.
4. Results and what we learned from going serverless.

What is a serverless architecture?

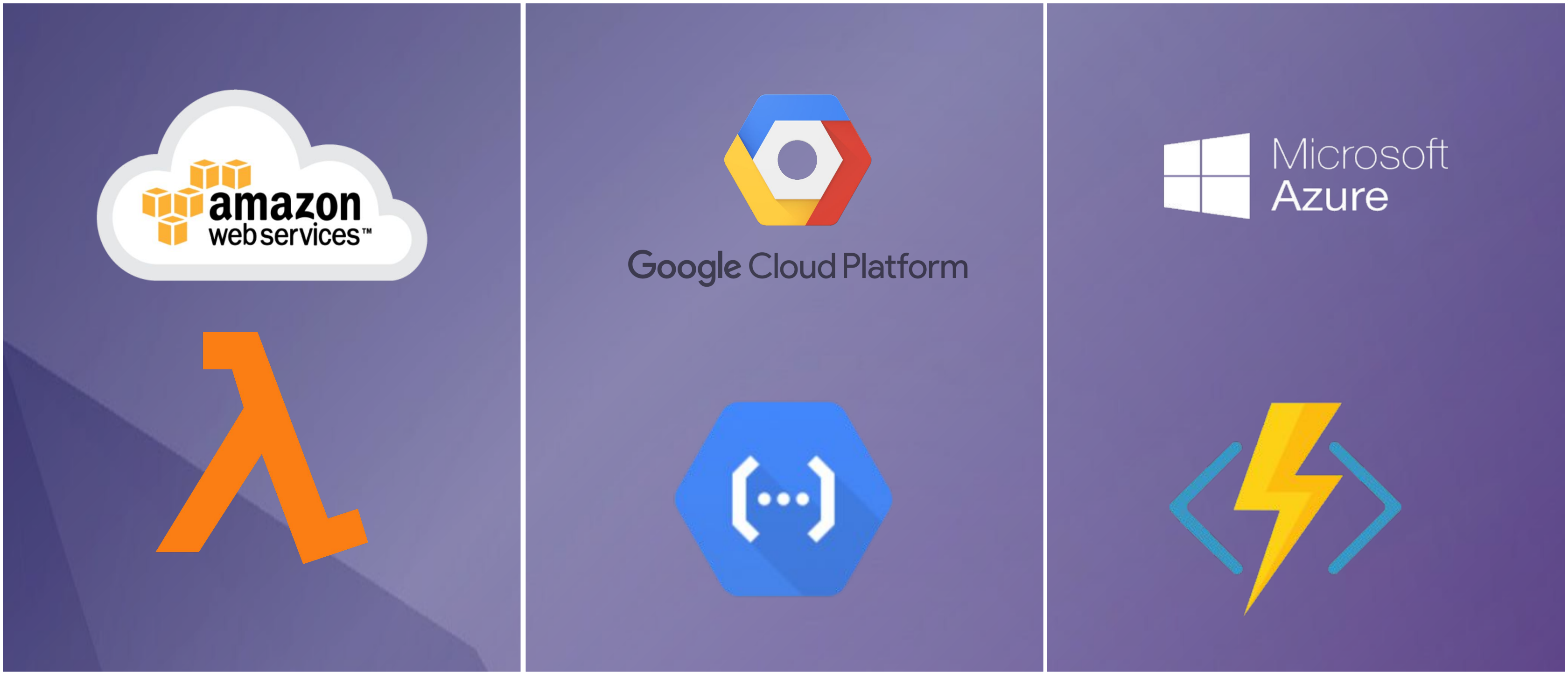


What is a serverless architecture?

Stateless compute containers that are event-triggered, ephemeral (may only last for one invocation), and fully managed by a 3rd party.



What is a serverless architecture?



Understanding video with AI.

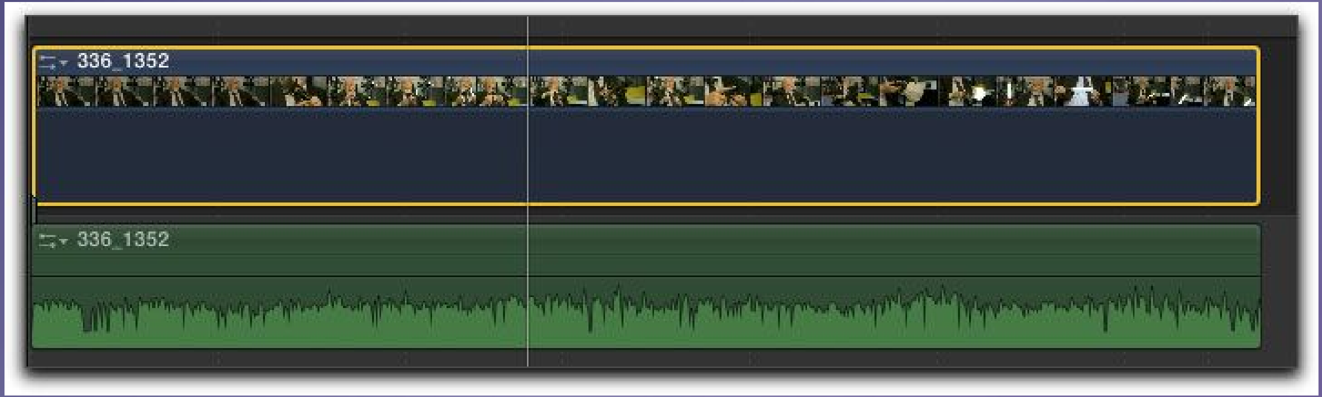
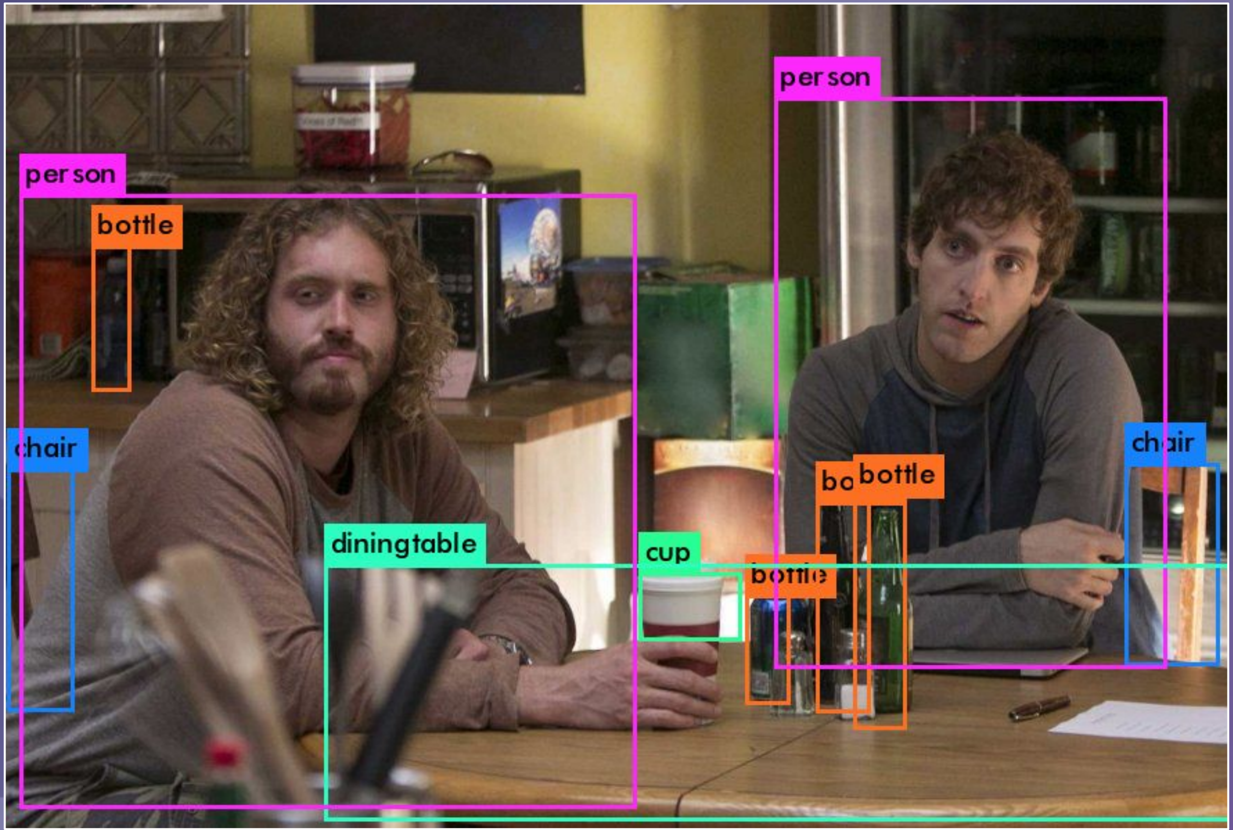


IMAGE DATA

TEXTUAL DATA

AUDIO DATA

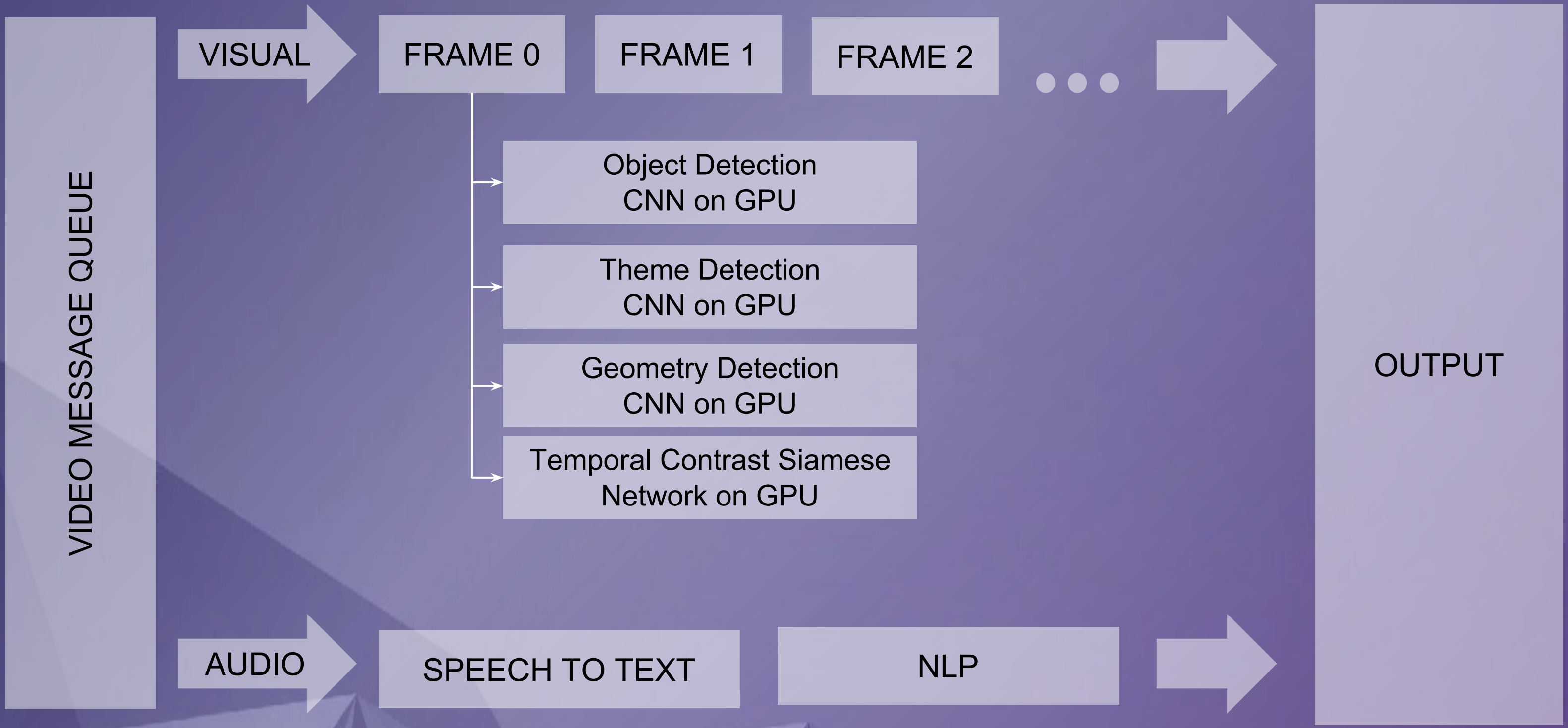
Understanding video with AI



Our experience at Uru

SEQUENTIAL PIPELINE

Our experience at Uru



VIDEO MESSAGE QUEUE

VISUAL

FRAME 0

FRAME 1

FRAME 2

…

Object Detection
CNN on GPU

Theme Detection
CNN on GPU

Geometry Detection
CNN on GPU

Temporal Contrast Siamese
Network on GPU

AUDIO

SPEECH TO TEXT

NLP

OUTPUT

Our experience at Uru

Sequential pipeline advantages:

- **Simple flow**
- **Temporal structure**
- **Simple error handling**

Our experience at Uru

Sequential pipeline disadvantages:

- Speed
- Expensive
- Hard to scale
- Monolithic architecture

Our experience at Uru

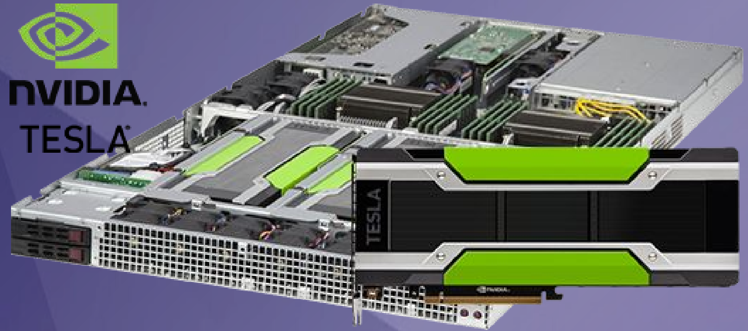
Our experience at Uru

PARALLELIZING OUR WORKFLOW

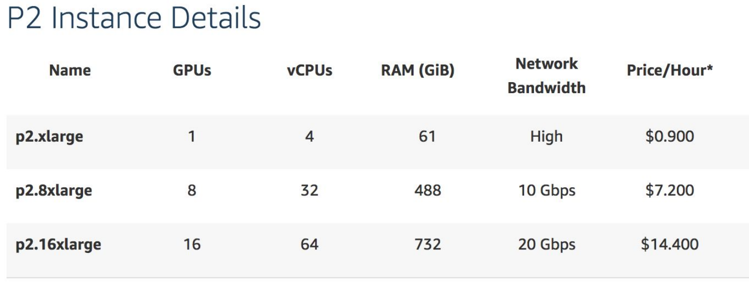
Our experience at Uru

Deep Learning on Lambda vs GPU.

BASELINE - CNN Inception V3 trained on Imagenet:



Latency for 1 image on GPU is < 0.8 seconds.



P2 Instance Details

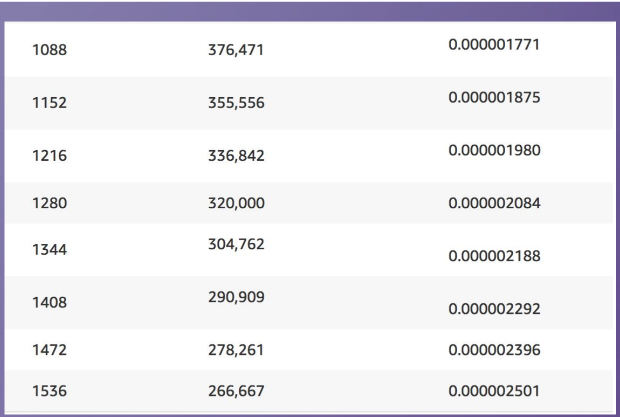
Name	GPUs	vCPUs	RAM (GiB)	Network Bandwidth	Price/Hour*
p2.xlarge	1	4	61	High	$0.900
p2.8xlarge	8	32	488	10 Gbps	$7.200
p2.16xlarge	16	64	732	20 Gbps	$14.400

Our experience at Uru

Deep Learning on Lambda vs GPU.

BASELINE - CNN Inception V3 trained on Imagenet:

Latency for 1 image on lambda is around 5-10 seconds.



1088	376,471	0.000001771
1152	355,556	0.000001875
1216	336,842	0.000001980
1280	320,000	0.000002084
1344	304,762	0.000002188
1408	290,909	0.000002292
1472	278,261	0.000002396
1536	266,667	0.000002501

Our experience at Uru

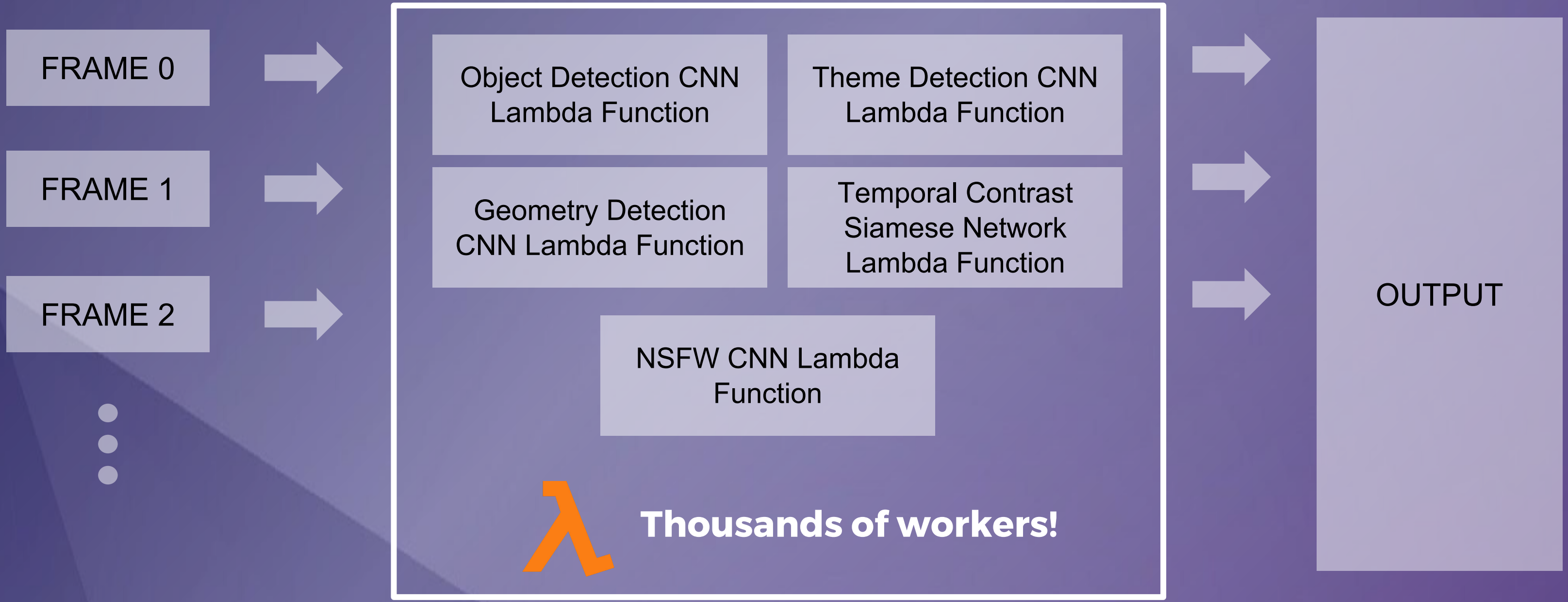
Deep Learning on Lambda vs GPU.

Serverless wins!

- **Distribute the workload**
- **It's cheaper**
- **Speed is capped by the number of serverless workers**
- **Maintenance is basically 0**

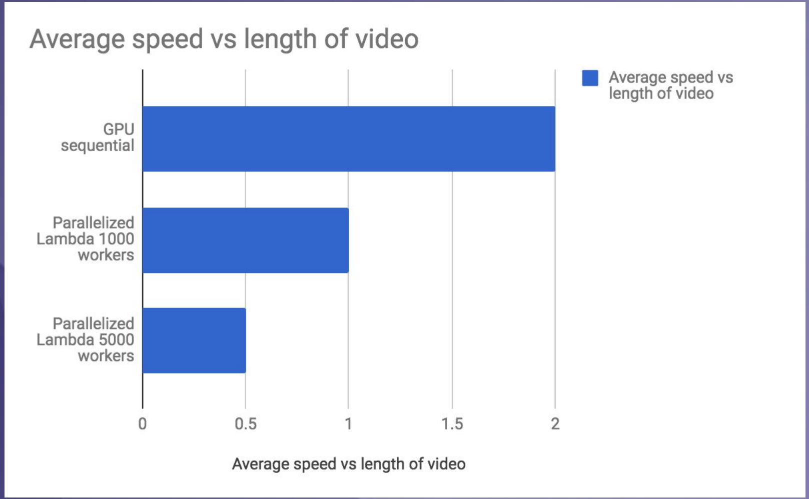
Our experience at Uru

Serverless Workflow.

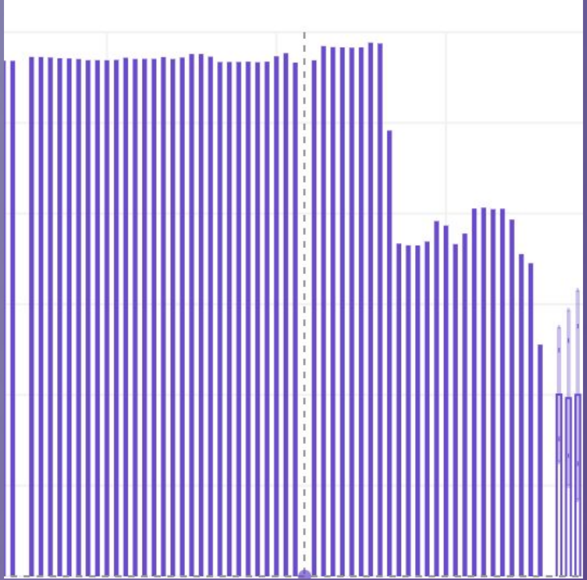

FRAME 0
FRAME 1
FRAME 2
Object Detection CNN Lambda Function
Theme Detection CNN Lambda Function
Geometry Detection CNN Lambda Function
Temporal Contrast Siamese Network Lambda Function
NSFW CNN Lambda Function
λ
Thousands of workers!
OUTPUT

RESULTS AND CONSIDERATIONS

Results



Speed



Cost

Results

What we learned

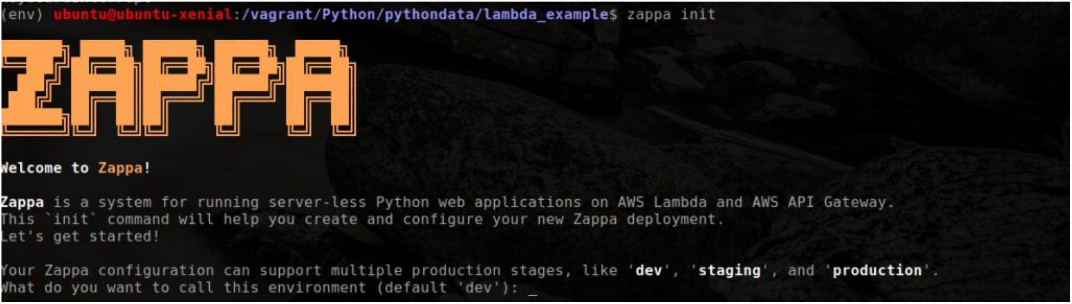
Disk space is a luxury:

- **500 to 512 MB of disk space.**
- **Code to be deployed needs to be small: 50MB compressed!**

Results

What we learned

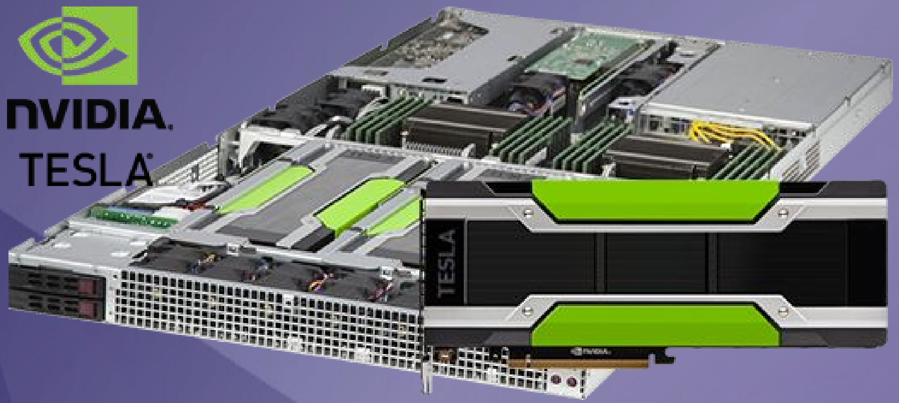
Useful frameworks:



```
(env) ubuntu@ubuntu-xenial:/vagrant/Python/pythondata/lambda_example$ zappa init

Welcome to Zappa!

Zappa is a system for running server-less Python web applications on AWS Lambda and AWS API Gateway.
This `init` command will help you create and configure your new Zappa deployment.
Let's get started!

Your Zappa configuration can support multiple production stages, like 'dev', 'staging', and 'production'.
What do you want to call this environment (default 'dev'): _
```

Results

What we learned

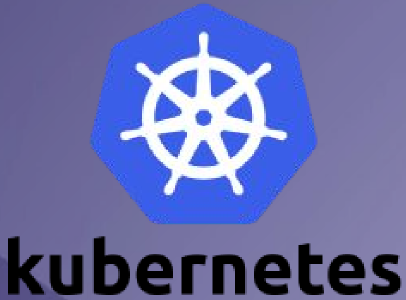


TRAIN

EXPERIMENTATION

Results

What we plan to do on the future:



URU

Video Understanding at Scale

Deep Learning in a Serverless Infrastructure

@uruvideo

@uruvideo

uruvideo.com