From CPU to GPU to ASIC: Our Transcoding Journey

Ilya Mikhaelis
Streaming Backend Tech Lead
Mayflower

Plan

1. Introduction: Who we are
2. Transcoding pipeline
3. Software transcoder
4. GPU transcoder
5. ASIC transcoder
6. Results

NETINT

Who we are

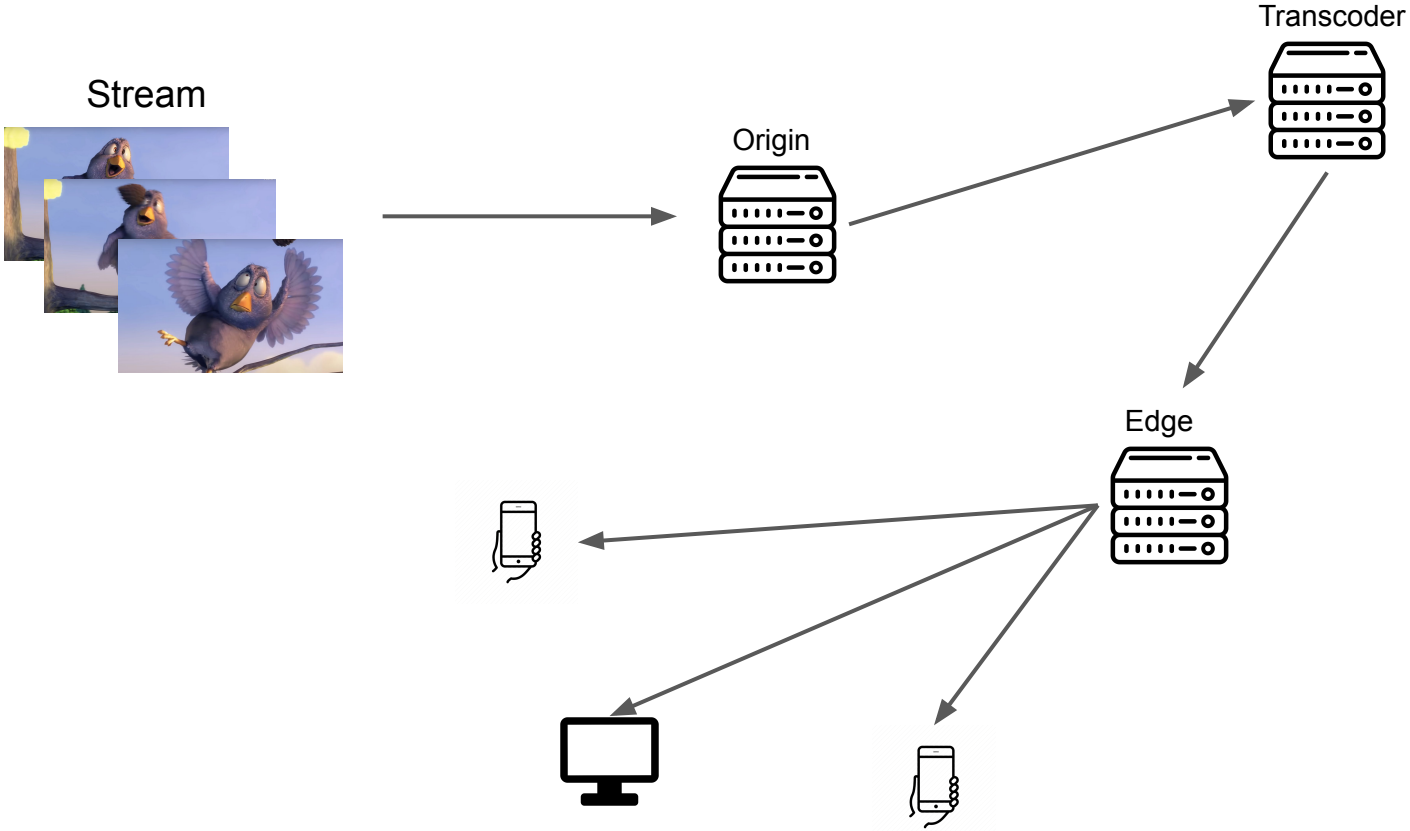
Internal Mayflower project

- 1B+ total visits per month
- 100M+ total registered users
- Top 50 of the most visited sites worldwide

- Own low latency CDN
- 10K+ incoming simultaneous streams
- 1M+ outgoing simultaneous streams
- Try to keep minimal possible latency

NETINT

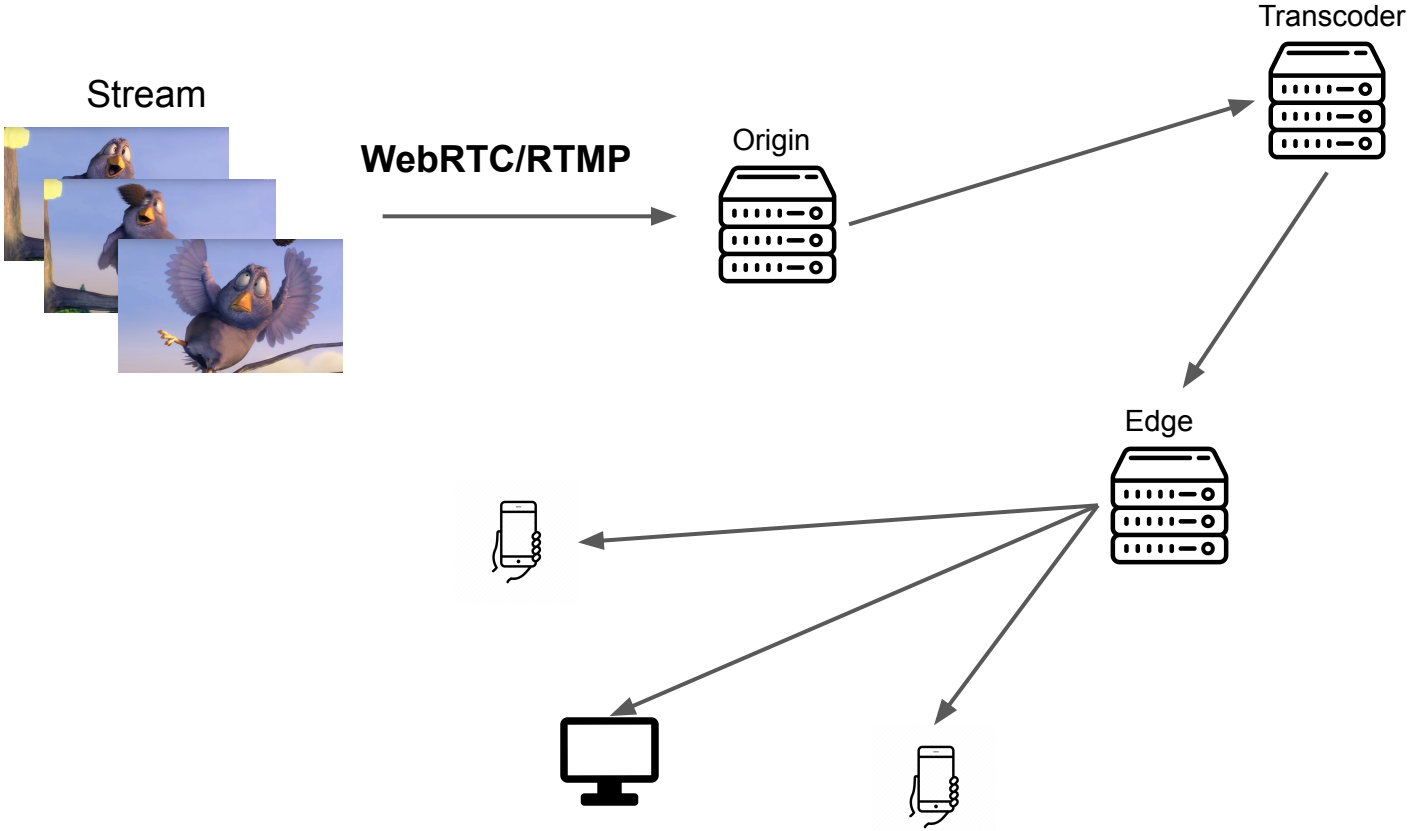
Low latency CDN



MOUFLON
MEDIA ENGINE

NETINT

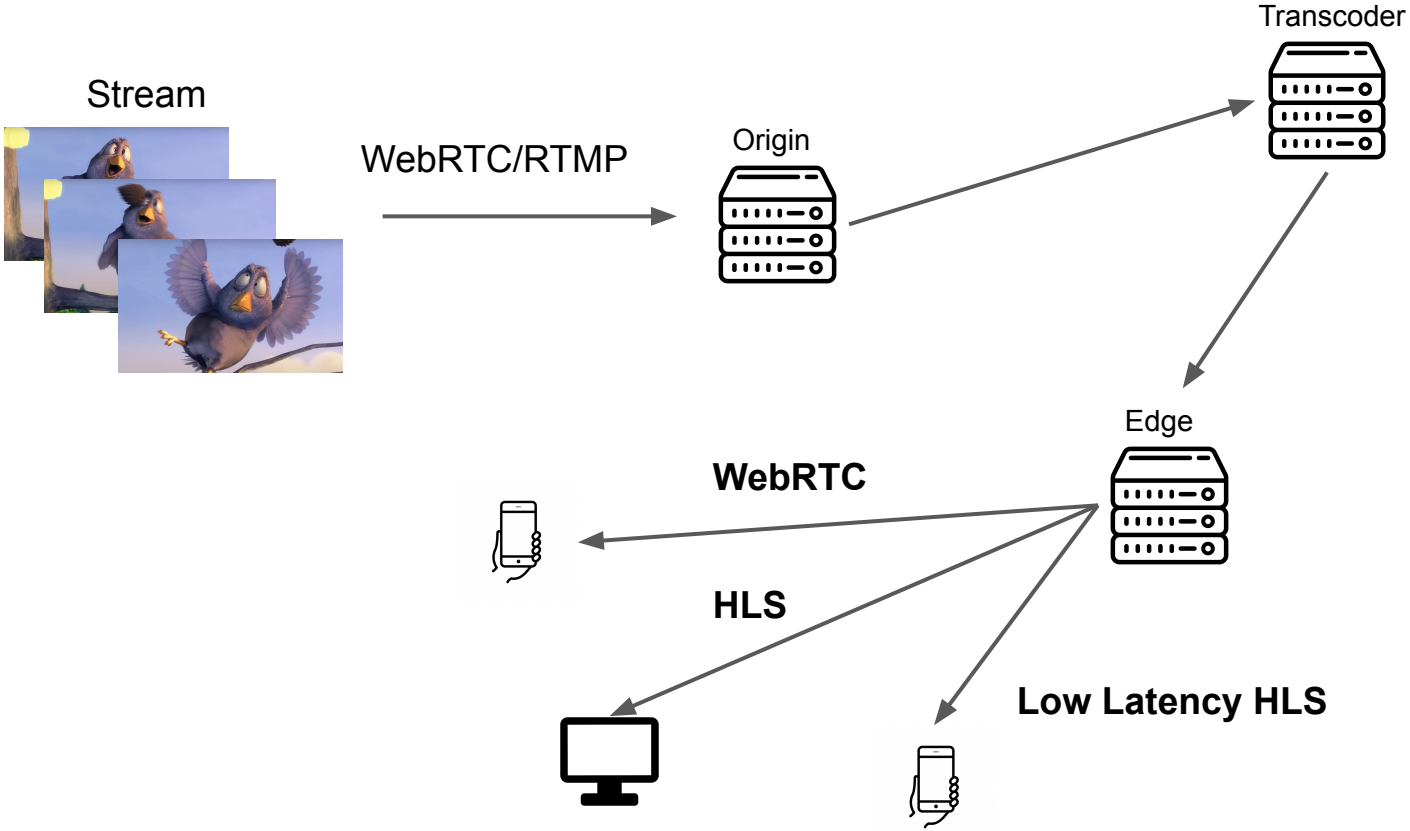
Low latency CDN



MOUFLON
MEDIA ENGINE

NETINT

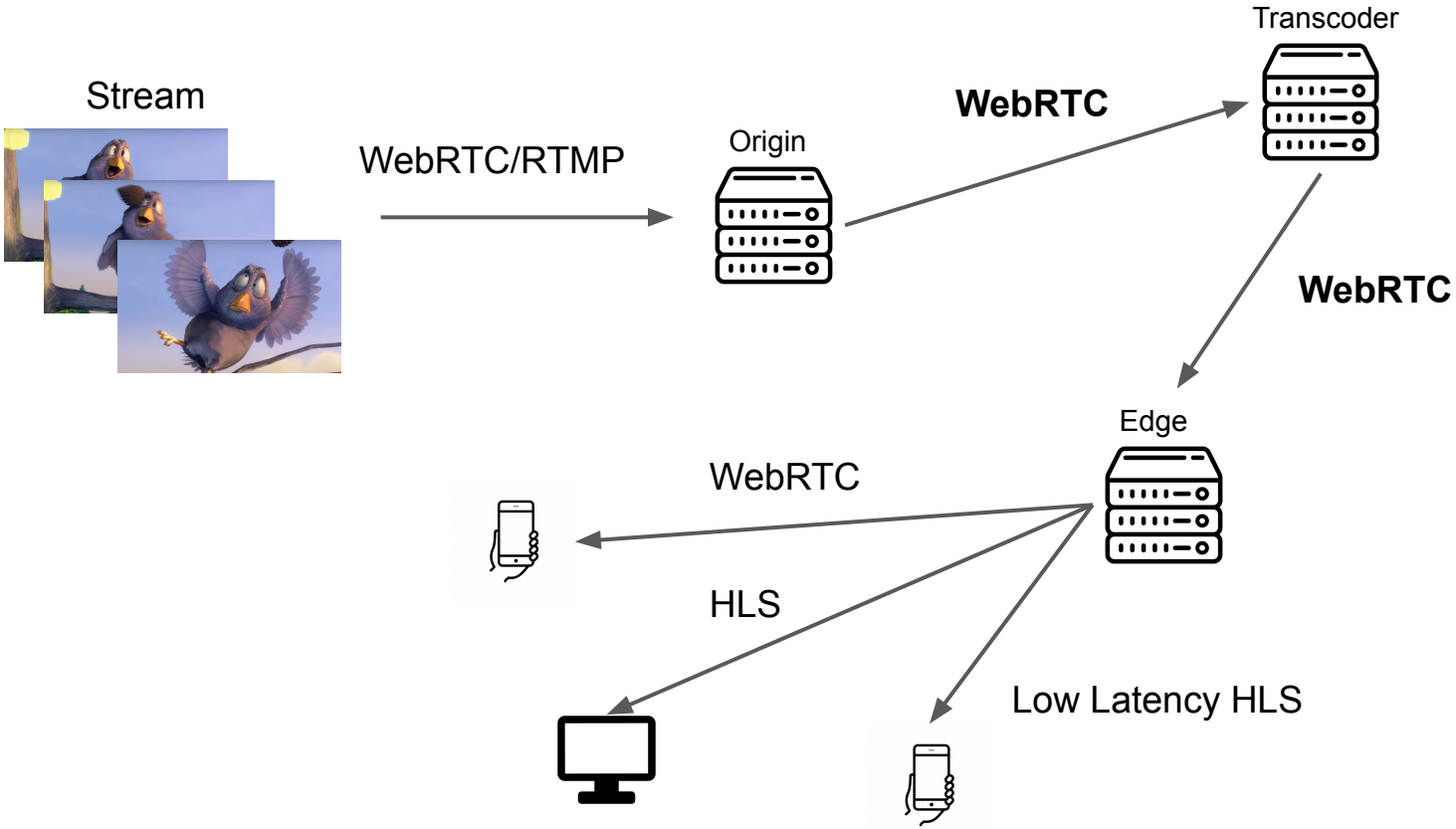
Low latency CDN



MOUFLON
MEDIA ENGINE

NETINT

Low latency CDN



MOUFLON
MEDIA ENGINE

NETINT

Transcoding pipeline

Our approach:

- using only bare metal servers to achieve maximum network and resource utilization

FFMPEG

MOUFLON
MEDIA ENGINE

NETINT

Transcoding pipeline

Our approach:

- using only bare metal servers to achieve maximum network and resource utilization
- using high performance system profile for our servers

FFMPEG

MOUFLON
MEDIA ENGINE

NETINT

Transcoding pipeline

Our approach:

- using only bare metal servers to achieve maximum network and resource utilization
- using high performance system profile for our servers
- reducing resource wastage by implementing of all network parts (e.g. protocols for receiving and publishing streams) by ourselves

FFMPEG

MOUFLON
MEDIA ENGINE

NETINT

Transcoding pipeline

Our approach:

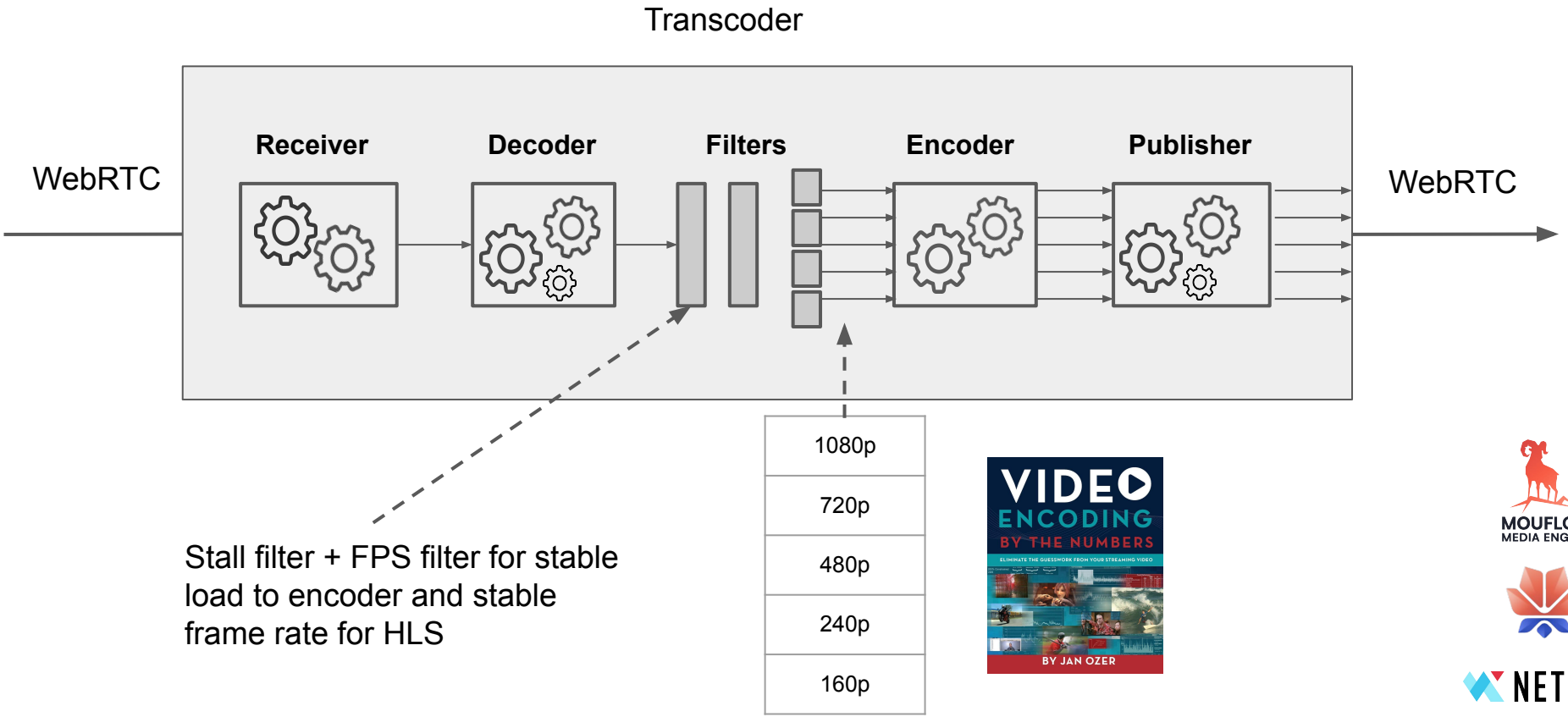
- using only bare metal servers to achieve maximum network and resource utilization
- using high performance system profile for our servers
- reducing resource wastage by implementing of all network parts (e.g. protocols for receiving and publishing streams) by ourselves
- using libav (which is part of ffmpeg) as a framework for transcoding inside our transcoder servers

FFMPEG

MOUFLON
MEDIA ENGINE

NETINT

Transcoding pipeline



Software transcoder

Legacy servers park

Libx264 software transcoder platform

MOUFLON MEDIA ENGINE

NETINT

Software transcoder

Base	Dell PowerEdge R940xa
Processor	(Intel Xeon Gold 6248 2.50Ghz 20 cores) x 4
Memory	24x16GB DDR4 3200MHz ECC Reg
Storage	2x480GB SSD
Network	2x10G SFP+ (X710DA2)

MOUFLON MEDIA ENGINE

NETINT

Software transcoder

Transcoder	x264
Stream density 720p30 without source transcode	60-70

MOUFLON
MEDIA ENGINE

NETINT

Software transcoder

Problems:

- Number of full hd streams increased dramatically

MOUFLON
MEDIA ENGINE

NETINT

Software transcoder

Problems:

- Number of full hd streams increased dramatically
- Streamers started to use B-frames. One of our view protocols - WebRTC doesn’t support B-frames

MOUFLON
MEDIA ENGINE

NETINT

Software transcoder

Problems:

- Number of full hd streams increased dramatically
- Streamers started to use B-frames. One of our view protocols - WebRTC doesn’t support B-frames
- Streamers started to increase GOP interval up to 10 second

MOUFLON
MEDIA ENGINE

NETINT

Software transcoder

Problems:

- Number of full hd streams increased dramatically
- Streamers started to use B-frames. One of our view protocols - WebRTC doesn't support B-frames
- Streamers started to increase GOP interval up to 10 second

Need to transcode source

MOUFLON MEDIA ENGINE

NETINT

Software transcoder

Transcoder	x264
Stream density per server 1080p30	20
Power consumption W	650
Server cost $	20`000
Stream cost $	1`000
Production transcoding cost $ (10K incoming streams)	10`000`000

MOUFLON
MEDIA ENGINE

NETINT

Software transcoder

Transcoder	x264
Stream density per server 1080p30	20
Power consumption W	650
Server cost $	20`000
Stream cost $	1`000
Production transcoding cost $ (10K incoming streams)	10`000`000

MOUFLON MEDIA ENGINE

NETINT

Production transcoding cost

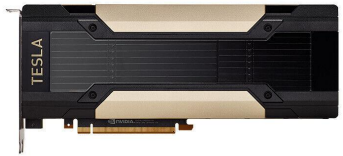
C_S - server cost (with transcoding units, e.g. GPU cards)
N_S - total number of streams
N_{TS} - capacity of the server
Round - round up to the integer
C_{TP} - production transcoding cost

C_{TP} = *Round*(N_S / N_{TS}) * C_S

MOUFLON
MEDIA ENGINE

NETINT

GPU transcoder



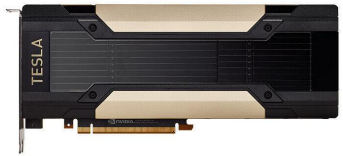
Card	Stream density per card (1080p30)	Power consumption W	Card cost $
Nvidia Tesla V100	20	250	8`000

- decoder/encoder queues must be around zero
- the cost of cards was valid at the time of tests (2020 year)

MOUFLON
MEDIA ENGINE

NETINT

GPU transcoder



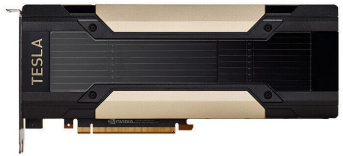
Card	Stream density per card (1080p30)	Power consumption W	Card cost $
Nvidia Tesla V100	20	250	8`000
Nvidia Tesla P100	19	250	7`000

- decoder/encoder queues must be around zero
- the cost of cards was valid at the time of tests (2020 year)

MOUFLON
MEDIA ENGINE

NETINT

GPU transcoder



Card	Stream density per card (1080p30)	Power consumption W	Card cost $
Nvidia Tesla V100	20	250	8`000
Nvidia Tesla P100	19	250	7`000
Nvidia Tesla T4	**16**	**70**	**3`000**

- decoder/encoder queues must be around zero
- the cost of cards was valid at the time of tests (2020 year)

MOUFLON
MEDIA ENGINE

NETINT

GPU transcoder

Transcoder	NVIDIA T4
Stream density per server 1080p30	96
Power consumption W	1`070
Server cost $	38`000
Stream cost $	395
Production transcoding cost $ (10K incoming streams)	3`990`000

MOUFLON
MEDIA ENGINE

NETINT

GPU transcoder

Transcoder	NVIDIA T4 (corrected to current market cost)
Stream density per server 1080p30	96
Power consumption W	1`070
Server cost $	**32`000** (~2`000 per card)
Stream cost $	**333**
Production transcoding cost $ (10K incoming streams)	**3`360`000**

MOUFLON
MEDIA ENGINE

NETINT

ASIC transcoder

NETINT

MOUFLON
MEDIA ENGINE

NETINT

T432 transcoder

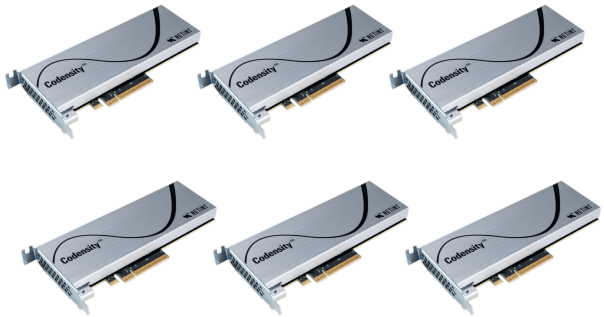
Card	T432
Stream density per card (1080p30)	25
Power consumption W	27
Cost $	2`000

- decoder/encoder queues must be around zero
- the cost of cards was valid at the time of tests (2021 year)

MOUFLON
MEDIA ENGINE

NETINT

T432 transcoder

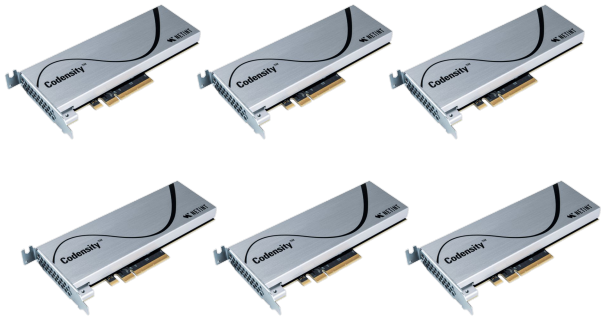


Transcoder	T432
Stream density per server 1080p30	150
Power consumption W	812
Server cost $	32`000
Stream cost $	213
Production transcoding cost $ (10K incoming streams)	2`144`000

MOUFLON
MEDIA ENGINE

NETINT

T432 transcoder



Transcoder	T432 (corrected to current market cost)
Stream density per server 1080p30	150
Power consumption W	812
Server cost $	**29`600** (1`600 per card)
Stream cost $	**197**
Production transcoding cost $ (10K incoming streams)	**1`983`200**

MOUFLON
MEDIA ENGINE

NETINT

T432 transcoder

Game changer?

Disadvantages:

- No hardware scaler.
  Need to scale in software.

MOUFLON
MEDIA ENGINE

NETINT

T432 transcoder

Game changer?

Disadvantages:

- No hardware scaler. Need to scale in software.
- Can’t reach real time (60 fps) for h264 4K60 fps.

MOUFLON
MEDIA ENGINE

NETINT

T432 transcoder

Game changer?

Disadvantages:

- No hardware scaler. Need to scale in software.
- Can’t reach real time (60 fps) for h264 4K60 fps

Even with software scaler on Dell R940 CPUs load is 40%. Cards limit is a bottleneck, NOT a CPU. Good solution if you have already had a park of servers with empty PCIe slots.

MOUFLON
MEDIA ENGINE

NETINT

Quadra T2 transcoder

With hardware scaler inside!

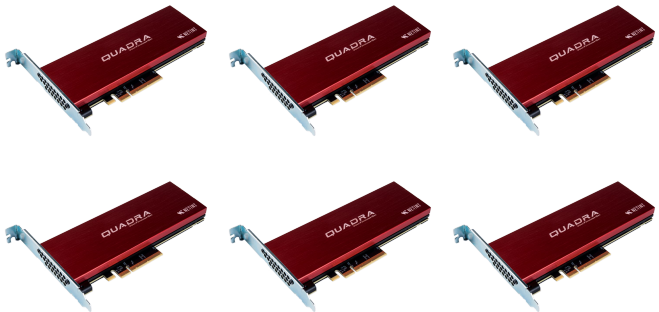
Card	Quadra T2
Stream density per card (1080p30)	45
Power consumption W	40
Cost $	3`000

- decoder/encoder queues must be around zero

MOUFLON
MEDIA ENGINE

NETINT

Quadra T2 transcoder



Transcoder	Quadra T2
Stream density per server 1080p30	270
Power consumption W	890
Server cost $	38`000
Stream cost $	141
Production transcoding cost $ (10K incoming streams)	1`444`000

MOUFLON
MEDIA ENGINE

NETINT

Results: streams density

Type	Card stream density	Server stream density (1080p30)	Number of servers
Software	0	20	500
NVIDIA T4	16	96	105
T432	25	150	67
Quadra T2	**45**	**270**	**38**

MOUFLON
MEDIA ENGINE

NETINT

Results: power consumption

Type	Card power consumption W	Server power consumption W	Production power consumption W
Software	0	**650**	325`000
NVIDIA T4	70	1070	112`350
T432	27	812	54`404
Quadra T2	**40**	**890**	**33`820**

MOUFLON
MEDIA ENGINE

NETINT

Results: cost

Type	Card cost	Server cost $	Stream cost $	Production transcoding cost $
Software	0	20`000	1`000	10`000`000
NVIDIA T4	2`000	32`000	333	3`360`000
T432	1`600	29`600	197	1`983`200
Quadra T2	**3`000**	**38`000**	**141**	**1`444`000**

MOUFLON
MEDIA ENGINE

NETINT