ateme

Transforming Video Delivery

ATEME TRANSCODING ON AZURE:

IMPACT OF HIGH-PERFORMANCE CLOUD COMPUTE ON FILE AND LINEAR PROCESSING

White Paper - October 2021

Table of contents

I.INTRODUCTION

ATEME works with content providers and Multichannel Video Programming Distributors (MVPDs) to provide video transcoding solutions for a wide range of projects. A common thread with our customers is the intention of relying on public cloud technology and infrastructure for video transcoding.

In this white paper, we will focus on the benefits provided by Azure cloud for the following use cases:

- File video transcoding
- Linear video transcoding.

Both use cases are deployed today and in production on all three major cloud vendors.

ATEME compression solutions are designed to take advantage of the latest CPU technology. In March of 2021, Microsoft announced the availability of HBv3 series Virtual Machines (VMs). These high-performance compute instances feature up to 120 third-generation AMD EPYC 7003-series (Milan) CPU cores. This means faster job completion for file transcoding and the ability to input more channels per instance for linear transcoding.

This paper describes the testing and performance of ATEME transcoding solutions on Azure high-performance compute instances.

II. CHALLENGES OF CLOUD MEDIA PROCESSING

Time and cost are significant challenges for file-based video transcoding workloads. Video file assets can be long and the time it takes to complete a file transcoding job is proportional to the duration of the file, compute resources available, and video output profile specifications. Infrastructure requirements and costs continue to go up with the exponential growth of content production and the need to support higher resolutions such as UHD (Ultra High Definition).

ATEME addresses these challenges by providing a file-transcoding solution designed for the cloud environment, based on a micro-services architecture, with native support of object-storage technologies such as Azure blobs. The solution relies on a proprietary encoding core and offers advanced workflow features such as distributed transcoding. The results of our tests show that in UHD use cases, we can **reduce processing costs by up to 75%** using Azure high-performance compute instances (compared to more traditional cloud compute instances), and by 35% to 45% in HD (High Definition) use cases.

Linear video workloads have their own unique challenges. They require a permanent, highly available infrastructure and the capacity to ingest and transcode high-bitrate, high-resolution streams in a variety of input formats and compression codecs. Advanced codecs such as HEVC require more compute resources to transcode than AVC, while higher resolutions such as UHD will require more compute than traditional 1080p resolution.

The results of our tests demonstrate that Azure high-performance compute instances enable more linear channels processed on a single instance, reducing infrastructure costs, but also enable support for far more complex channels (more profiles, higher resolution, HDR support, etc.) on a single instance, simplifying solution design. We observed up to a 60% improvement in density with HBv3 instances over the previous-generation compute-optimized instances (96 cores based on second-generation AMD EPYC 7002 series).

The tests described in this paper were designed to measure the following metrics using the newly introduced Azure HBv3 virtual machines for processing:

1. For file video transcoding, single-job transcoding speed and single-server throughput
2. For linear video transcoding, channel density for a single server and the highest supported channel configuration.

III. FILE-BASED VIDEO TRANSCODING TESTS

The purpose of these tests is to determine the optimal number of jobs to process simultaneously on a single computing instance to achieve the best performance, for both HD and UHD workflows.

The metrics defined to evaluate the performance of the solution are the following:

- Turnaround speed, represented by the xRT value, which is a comparison of file transcoding speed to real-time, defined by the following formula:

xRT = AvgEncodeTime / SourceDuration

- Server throughput, represented by the xRT server value, which indicates the speed at which a server can process content compared to real-time, defined by the following formula:

xRT Server = xRT / #ofParallelJobs

These two metrics directly impact each other, and solution load-balancing is a trade-off between the two. The more jobs that are processed simultaneously on a server, the fewer resources are allocated per job, the slower a single job can be transcoded – but the more content a server will be able to process in a given period. Conversely, the fewer jobs that are processed simultaneously, the more resources are allocated per job, and the faster a single job can be transcoded – but the amount of content a server will process in that same given period will be less.

For UHD processing:

Figure 1: Input asset used

Input	
Source Format	MOV
Video Codec	ProRes 422 LT
Resolution	3840x2160
Framerate	25 fps
Source Bitrate	311 Mbps
Audio Codec	6 x PCM channels
Source Duration	30 minutes

Figure 2: Output Adaptive Bit Rate (ABR) profiles

Outputs	
6 Profiles HEVC video, 1 AAC audio, 1 DD Audio	
Resolution	Bitrate
3840x2160	18 Mbps
2560x1440	12.2 Mbps
1920x1080	5.8 Mbps
1280x720	4.25 Mbps
960x540	1.85 Mbps
960x540	1.2 Mbps

Figure 3: Transcoding results

	HB120rs_v3						
	#ofParallelJobs	**Cores/Job**	**Total Cores**	**AvgEncodeTime (s)**	**xRT**	**xRT Server**	**CPU %**
1	1	118	118	5977	3.32	3.32	45
2	2	59	118	7316	4.06	2.03	75
3	3	39	117	9014	5.01	1.67	85
4	4	29	116	11431	6.35	1.59	91
5	5	23	115	13145	7.30	1.46	92
6	6	19	114	15392	8.55	1.43	92

As illustrated in the results above, the performance of the application is a trade-off between job turnaround time (xRT) and server throughput (xRT server). For the UHD workflow, using the HB120rs_v3 instances, it seems the best compromise between the two is with three jobs encoding in parallel (row #3) on the machine. Fewer jobs will significantly negatively impact the server throughput, while more jobs would only bring marginal server throughput improvement while greatly increasing job transcoding speed.

For distributed UHD processing:

Figure 4: Input asset used

Input	
Source Format	MOV
Video Codec	ProRes 422 LT
Resolution	3840x2160
Framerate	25 fps
Source Bitrate	311 Mbps
Audio Codec	6 x PCM channels
Source Duration	44 minutes 20 seconds

Figure 5: Output Adaptive Bit Rate (ABR) profiles

Outputs	
6 Profiles HEVC video, 1 AAC audio, 1 DD Audio	
Resolution	**Bitrate**
3840x2160	18 Mbps
2560x1440	12.2 Mbps
1920x1080	5.8 Mbps
1280x720	4.25 Mbps

Figure 6: Transcoding results

2 x HB120rs_v3							
	#ofParallelJobs	**Cores/Job**	**Total Cores**	**AvgEncodeTime (s)**	**xRT**	**xRT Server**	**CPU %**
1	1	6 parts x 39	234	2501	0.94	1.88	85

ATEME's file video-transcoding solution offers the possibility to transcode a single source file faster by splitting it into several segments and having those segments transcoded in parallel before stitching them in the end, thanks to its micro-services architecture, with a feature called distributed transcoding. This is a powerful tool, especially in a public cloud environment.

This test is performed using the optimal configuration defined from the previous test. A single UHD source job was distributed across two HB120rs_v3 instances, with each instance processing three segments of the source file in parallel, for a total of six segments. In this configuration, the ATEME file-transcoding solution can transcode the content faster than real-time (xRT < 1).

This achievement opens up many benefits for file video-transcoding use cases. For example, it could help to meet the requirements of a Service-Level Agreement on the processing time for specific content, such as sports highlights or show replays.

For HD processing:

Figure 7: Input asset used

Input	
Source Format	MXF OP1a
Video Codec	AVC
Resolution	1920x1080
Framerate	25 fps
Source Bitrate	114 Mbps
Audio Codec	16 x PCM channels
Source Duration	30 minutes

Figure 8: Output Adaptive Bit Rate (ABR) profiles

Outputs	
5 Profiles HEVC video, 1 AAC audio, 1 DD Audio	
Resolution	**Bitrate**
1920x1080	5.8 Mbps
1280x720	4.25 Mbps
960x540	1.85 Mbps
960x540	1.2 Mbps
640x360	0.5 Mbps

Figure 9: Transcoding results

HB120rs_v3							
	#ofParallelJobs	**Cores/Job**	**Total Cores**	**AvgEncodeTime (s)**	**xRT**	**xRT Server**	**CPU %**
1	1	118	118	4465	2.48	2.48	19
2	2	59	118	4599	2.56	1.28	40
3	3	39	117	4878	2.71	0.90	58
4	4	29	116	5204	2.89	0.72	73
5	5	23	115	5676	3.15	0.63	83
6	6	19	114	6304	3.50	0.58	89
7	7	16	112	7132	3.96	0.57	90
8	8	14	112	7819	4.34	0.54	91
9	9	13	117	8498	4.72	0.52	94
10	13	9	117	11842	6.58	0.51	96

As in the case of the results for UHD, for HD processing the performance of the application is again a trade-off between turnaround time for a single job and server throughput. In this case, it seems that the best compromise is to have the HB120rs_v3 transcode six jobs (row #6) in parallel to guarantee high server throughput while keeping a fast single-job transcoding speed.

As demonstrated by the results above, the ATEME file-transcoding solution was able to take full advantage of the new Azure high-performance compute instances.

Processing optimization is a trade-off between server throughput and job turnaround time. The sweet spot for the ATEME application seems to be when CPU usage is around 90% on HB120rs_v3 instances.

Comparing this with the throughput achieved using the previous generation of CPUs and other instance types, we see a significant reduction in the infrastructure cost:

- 75% reduction for UHD
- 35%-45% reduction for HD.

The ATEME file-transcoding solution also unlocks the ability to process UHD content faster than real-time by distributing the transcoding workload over two instances.

IV. LINEAR VIDEO TRANSCODING TESTS

The purpose of these tests is to determine the channel density of a single computing instance for different typical linear workflows.

The channel density metric is simply defined as the number of channels an instance can process without any stability issues over a long period of time. Each channel configuration is based on an input stream and a certain number of output profiles.

For this linear video-transcoding test, we established a set of output profiles depending on the input stream specifications for the channel, and we measured the maximum number of channels that could be supported on a particular instance without compromising on video quality, and with the instance being stable over a long period of time (channels must stay live for 24 consecutive hours without any issues for the density to be fully validated).

We used sources and profiles commonly used by tier-one service and content providers:

Figure 10: UHD channel configuration

3840x2160p 59.94 Source			
Codec	**Resolution**	**Framerate**	**Bitrate**
HEVC	3840x2160	59.94	18 Mbps
HEVC	2560x1440	59.94	12 Mbps
HEVC	1920x1080	59.94	6 Mbps
HEVC	1280x720	59.94	4 Mbps
HEVC	960x540	59.94	2 Mbps
HEVC	960x540	59.94	1 Mbps

Figure 11: 1080p channel configuration

1920x1080p 59.94 Source			
Codec	**Resolution**	**Framerate**	**Bitrate**
AVC	1920x1080	59.94	8 Mbps
AVC	1280x720	59.94	6 Mbps
AVC	1280x720	29.97	2 Mbps
AVC	640x360	29.97	0.75 Mbps
AVC	480x288	29.97	0.5 Mbps

Figure 12: 720p channel configuration

1280x720p 59.94 Source			
Codec	**Resolution**	**Framerate**	**Bitrate**
AVC	1280x720	59.94	5 Mbps
AVC	1280x720	29.97	2 Mbps
AVC	640x360	29.97	0.75 Mbps
AVC	480x288	29.97	0.5 Mbps

All channels were ingesting an IP input and generating a DASH output.

The reference instance size for this test to compare with the performance of the HB120rs_v3 instance is a 96-core machine based on second-generation AMD EPYC 7002-series processors. This instance size/type is available on all three public cloud providers and is commonly used to run the ATEME linear-transcoding solution.

Figure 13: Channel density results

Configuration	Instance Type	Channel Density
UHD	96	0
UHD	120	1*
1080p	96	9
1080p	120	11
720p	96	13
720p	120	22

*The UHD stack without the 1440p profile can be run using ATEME Constant Quality and High VQ setting on a single HB120rs_v3 instance, for premium video quality.

As could be expected, the HBv3 high-performance compute instances enable ATEME's linear solution to achieve better channel density for all three tested configurations.

For the UHD channels, the 120-core machines unlocked support of the full UHD output profiles ladder on a single instance, which was not the case on the previous generation 96-core machines. This achievement greatly simplifies the support of this kind of channel in the cloud environment by eliminating the need to split a channel over multiple machines, with all the synchronization complexity that creates. As noted, a single machine can run the full UHD stack but for the 1440p profile, using ATEME Constant Quality rate control and High UHD VQ setting to deliver top-rate quality on this workflow.

For the 1080p and 720p channels, the increase in channel density enabled by the new high-performance compute instance results in reduced infrastructure costs per channel. For the 720p workflow the density increases by 60%, significantly exceeding the simple ratio of the number of cores between the two instances (120/96).

V. CONCLUSION

These tests demonstrate the benefits of the new Azure high-performance compute instances for video transcoding in the cloud.

For file video transcoding use cases, using HBv3 instance, the ATEME solution can achieve better throughput for both UHD and HD test cases, reducing infrastructure cost up to 75% for UHD and 35%-45% for HD. Taking advantage of ATEME's distributed transcoding feature, the new high-performance compute instances also enable faster-than-real-time transcoding, at high quality, for UHD.

For linear video-transcoding use cases, the HBv3 instances enable ATEME's solution to achieve even higher channel density than on previous-generation instances – up to 60% for the 720p workflow. This reduces cloud infrastructure costs and unlocks new workflows in the cloud supporting large UHD channels on a single processing instance, thereby making it simpler to implement the solution in the cloud and providing greater flexibility.

Authored by

Amilcar Padilla, Head of Solutions Engineering

Ryan Nolan, Deployment Engineering

Julien Carrat-Perrin, Solution Engineering