Using hardware acceleration for video decoding in XProtect

Introduction

Selecting the right graphics cards for your XProtect installation is a prerequisite for successful deployment, however not all graphics cards are equal. This document is intended to provide some insight for the new or existing customer looking into graphics card procurement.

Why graphics card acceleration matter in XProtect

Despite Central Processing Units (CPUs) becoming faster and more efficient every year, Graphics Processing Units (GPUs) still have a significant performance benefit over CPUs because GPU's decode video frames in hardware, while CPUs rely on generic operations that are not optimized for the sole purpose of video decoding. While XProtect does support CPU-based software video decoding, using GPUs for decoding videos instead of CPUs is strongly recommended as GPUs can help increase the efficiency of XProtect.

XProtect supports GPUs from Intel and Nvidia. Please see section Nvidia Micro Architectures and Intel Quick Sync video for more details.

Device drivers

Just like all other software, GPU device drivers are updated regularly from manufacturers like Intel and Nvidia. Updated device drivers usually provide improved performance and functionality, but the opposite may also inadvertently occur. Milestone therefore recommends appropriate change control is executed to pinpoint when a possible device driver regression is introduced in your XProtect environment.

Be aware that Windows Update also provides device driver updates. If Windows Update keeps installing a device driver with regression, you can change "Device installation Settings" from the System Properties in Microsoft Windows®.

The Milestone XProtect Video Load Balancer.

XProtect utilizes a Milestone-developed load balancer, enabling XProtect to utilize multiple graphics cards for a higher aggregate amount of video streams.

The XProtect video load balancer supports a combination of technologies listed below in parallel:

- Nvidia graphics cards
- Intel® Quick Sync Video technology
- CPU-based software decoding.

The settings are:
OnlyNvidia – XProtect will only use Nvidia GPUs for decoding video.
OnlyIntel – XProtect will only use Intel GPUs for decoding video.
AutoNvidia – XProtect will use Intel and Nvidia GPUs and software decoding to handle the load.

See : How to modify the default CPU/GPU settings in the XProtect Smart Client

! For improved performance, Milestone Systems recommends connecting monitors to only one graphics card, preferably the graphics card with the best performance. If multiple monitors are connected to different graphics cards on the same workstation, performance may be affected.

Verifying GPU acceleration is working

By default, hardware acceleration is set to automatic (Auto) and will automatically detect and enable GPUs from both Intel and Nvidia.

To verify if GPU acceleration is enabled, please refer to: Enabling hardware acceleration

! Enabling the Video Diagnostic Overlay may result in a performance reduction, and it is recommended to disable the Video Diagnostic Overlay when returning to normal operation.

Nvidia Micro Architectures

Each generation of GPUs from Nvidia is built on an underlying micro architecture. This micro architecture is deployed across product lines and models, with variations in memory, clock speed and durability. Because of this, Milestone does not state support for individual Nvidia graphics cards, but instead state supports for the underlying micro architecture.

To map an Nvidia graphics model number (or marketing name) to its micro architecture, use this article from Wikipedia : List of Nvidia graphics processing units

Milestone supports GPUs from Nvidia using the Pascal micro architecture, as well as newer micro architectures e.g., Turning, Amper

General recommendation for procuring Nvidia cards:

- A card's ability to create 3D graphics (polygon count) does not matter for video processing.
- Nvidia SLI is not supported, nor does it provide any performance benefit for video processing.
- Prefer cards with multiple NVDEC (Nvidia Decoder) chips. Multiple NVDEC chips provide a greater aggregated video decoding capability than cards with only one NVDEC chip.
- Graphics card memory matter, especially when working with high-definition video
- A good resource to view the number of NVDEC and their generation (NDEC generation) is Nvidia's video encode and decode GPU support matrix. See https://developer.nvidia.com/video-encode-and-decode-gpu-support-matrix-new#Encoder

Example from Nvidia video encode and decode GPU support matrix.

From image below, we can see all GPUs are using 5th generation video encoder/decoder chips, but the Nvidia RTXA5000, A5500 and A6000 are the best professional choices for video decoding because of two NVDEC chips per card. A500 and A1000 are laptop mobile solutions.

BOARD	FAMILY	CHIP	NDEC Generation	Desktop/ Mobile	# OF CHIPS	# OF NVDEC /CHIP	Total # of NVDEC
NVIDIA RTX A500 Laptop GPU	Ampere	GA107	5th Gen	M	1	2	1
NVIDIA RTX A1000 Laptop GPU	Ampere	GA107	5th Gen	M	1	2	2
NVIDIA RTX A4000	Ampere	GA104	5th Gen	D	1	1	1
NVIDIA RTX A4500	Ampere	GA102	5th Gen	D	1	1	1
NVIDIA RTX A5000	Ampere	GA102	5th Gen	D	1	2	2
NVIDIA RTX A5500	Ampere	GA102	5th Gen	D	1	2	2
NVIDIA RTX A6000	Ampere	GA102	5th Gen	D	1	2	2

Datacentre grade cards: The A series of cards uses both 4th and 5th generation of the NVDEC chips. As each card of the A series has a specific purpose, the number of decoding chips also vary. The clear winner – and the most expensive – is the A100 with 5 decoder chips.

BOARD	FAMILY	CHIP	NDEC Generation	# OF CHIPS	# OF NVDEC /CHIP	Total # of NVDEC
NVIDIA A100	Ampere	GA100	4th Gen	1	5	5
NVIDIA A40	Ampere	GA102	5th Gen	1	2	2
NVIDIA A30	Ampere	GA100	4th Gen	1	4	4
NVIDIA A16	Ampere	GA107	5th Gen	4	2	8
NVIDIA A10	Ampere	GA102	5th Gen	1	2	2
NVIDIA A2	Ampere	GA107	5th Gen	1	2	2

Nvidia Micro Architectures supported by XProtect

Prof. Product	XProtect Support
Hopper	Supported
Ampere (2020)	
Volta (2017)	

Consumer Product	XProtect Support
Blackwell	Supported
Lovelace (2022)	
Ampere (2020)	
Turing (2018)	
Pascal (2016)[1]	
Maxwell (2014)	Not Supported
Kepler (2012)	
Fermi (2010)	
Tesla (2006)	
Curie (2004)	
Rankine (2003)	
Kelvin (2001)	
Celsius (1999)	
Fahrenheit (1998)	

Please also see the section Use case recommendation for graphics card in XProtect.

[1] Quadro P500/P520) do not have encoding/decoding capabilities and are therefore not supported.

Intel Quick Sync Video

Milestone supports Intel Quick Sync Video, from Intel® CPU generation 6 or newer.

Motherboards utilizing Intel chipset often have a built-in GPU. The built-in GPU is normally disabled when discrete graphics cards are added to the system but can be re-enabled in the BIOS.
By enabling both graphics cards, XProtect can balance the video decoding load on both GPUs.

Intel CPU Generation	Quick Sync Video Versions	Supported in XProtect
Generation 12 – Alder Lake	Version 8	Supported – see note*
Generation 11 – Tiger Lake/Rocket Lake	Version 8	Supported
Generation 10 – Ice Lake/Comet Lake	Version 7	Supported
Generation 9 Coffee (refresh)	Version 6	Supported
Generation 8 Coffee Lake/Whisky Lake	Version 6	Supported
Generation 7 Kaby Lake/Cannon Lake	Version 6	Supported
Generation 6 Sky Lake	Version 5	Supported
Generation 5 Broadwell	Version 4	Not Supported
Generation 4 Haswell	Version 3	Not Supported
Generation 3 Ivy Bridge	Version 2	Not Supported
Generation 2 Sandy Bridge	Version 1	Not Supported
Generation 1 Nehalem	N/A	Not Supported

* XProtect 2022 R3 and newer in combination with Intel® Graphics – Windows DCH Drivers version 31.0.101.3413/31.0.101.2111 (release date 01/09-2022) or newer, is required for stable operation on Intel CPUs generation 12 and newer.

Intel's UHD GPU is produced in different sizes, depending on CPU model. For example, the Alder Lake 12th generation CPU is produced with both the UHD 770, UHD730, and UHD 710 variants.

UHD710 and UHD730 has 16 and 23 Execution Units and 1 Multi-Format Codec Engines, while the UHD770 with its 32 execution and 2 Multi-Format Codec Engines units takes the performance lead.

A detailed comparison can be found on Intel Ark

Milestone does not currently support the Intel Arc lines of graphics cards, internal verification process is currently being undertaken.

Use case recommendation for graphics card in XProtect.

	Smart Client	Smart Wall	Recording Server	Management Client
Nvidia Quadro Nvidia RTX AXXXX	Recommended	Recommended	Recommended	Recommended
Nvidia GeForce	Limited usage	Not recommended	Not recommended	Limited usage
Nvidia Tesla Nvidia Data Center Nvidia AXXX	Not Recommended	Not Recommended	Recommended	Not Recommended
Intel Quick Sync Video	Recommended	Recommended	Recommended	Recommended
Intel Arc	Not supported – Pending verification	Not supported – Pending verification	Not supported – Pending verification	Not supported – Pending verification

Nvidia Quadro / Nvidia RTX AXXXX series of cards is the professional line of GPUs, intended for enterprise usage.

Nvidia GeForce GPU is suitable for the occasional use of the smart client and management client, but is not recommended for sustained usage, as these graphics' cards are intended for consumer use. Also, Nvidia state in their manufacturer's warranty that enterprise usage of these cards will void the guarantee. See https://www.nvidia.com/en-us/support/warranty/

Nvidia Tesla, Nvidia Data Centre or Nvidia AXX cards are target computational workloads, are passively cooled, have specific hardware requirements, and do not have any display outputs, which renders them less than ideal for use in XProtect Smart Client and Management Client, but they are good candidates for motion detection in the recording server, where no display output is required. Perquisite for the use of computational cards, is the built-in video decoder (NVDEC). If a card does not have decoder block, it is not supported by XProtect. See manufactures product specifications for more details.

Intel Quick Sync Video is motherboard integrated and a good all-round choice.

Intel Arc discrete graphics cards are pending verification by Milestone and are therefore not currently supported.

Related ARTICLE: FAQ - Using hardware acceleration (Nvidia, Intel) in XProtect Smart Client 2018 R1 and all newer versions

FAQ section:

Question: Where can I find an overview of which generations of NVIDIA cards support H.264 and H.265?
Answer: The official NVIDIA CODEC SDK PAGE lists the NVIDIA generations beginning with Kepler and their respective H.264 and H.265 decoding support capabilities. An even more detailed overview of NVdec and NVenc (decoding and encoding capabilities) on the Nvidia chips is this: https://developer.nvidia.com/video-encode-decode-gpu-support-matrix

Question: Do I need to specify that NVIDIA hardware acceleration should be used?
Answer: No, the Smart Client can automatically detect NVIDIA card models and their capabilities, and it will automatically use them if the option "Hardware acceleration" is set to Auto.

Question: [updated] Does the NVIDIA hardware acceleration take advantage of multiple video cards installed?
Answer: Yes. We use all GPUs for decoding, so more GPUs mean better performance.
(*) But please also check the next question/answer as we currently only use one GPU for display/rendering optimally.

Question: [updated] How many physical displays(*) can I attach to my graphics cards?
Answer: Currently, we recommend attaching multiple monitors to only one GPU (one graphics card) on one computer system; typically, this would mean up to 4 displays attached to one graphics card.
(If you attach multiple monitors to more than one graphics card on the same system, the performance starts to decrease drastically as you start doing that. The issue is still being actively investigated.)
Note: For more details, please check KB 19947 / 33840 "Possible workarounds for the low FPS issue in the Smart Client when using multiple monitors (white paper) ".

Question: What will happen if I do not have an NVIDIA card installed or if it doesn't support the type of video stream received?
Answer: The Smart Client will automatically detect such an issue and will fall back to Intel Quick Sync or to software decoding.

Question: Are there any limitations on the operating system architecture?
Answer: NVIDIA only supports 64-bit versions of Windows. A 32-bit Windows version will fall back to Intel Quick Sync or to software decoding.

Question: Do JPEG streams take advantage of hardware acceleration?
Answer: JPEG streams are decoded using NVIDIA hardware acceleration if possible; otherwise decoding will fall back to software (as Intel Quick Sync does not improve performance in this case).

Question: I have issues with the video not working/displaying, what are the first things that I should check?

Answer: Ensure that your machine meets the minimum specifications for hardware acceleration.
See the "XProtect Smart Client - Hardware acceleration quick guide" available in the Milestone downloads section — https://www.milestonesys.com/support/help-yourself/manuals-and-guides/?prod=3&type=13&lang=27.
If it does meet the minimum requirements, next ensure that you are using the latest Intel GPU and/or NVIDIA display drivers and see if the problem persists even after you update to the latest drivers.

Question: Some of my streams have "Hardware acceleration: Off"?
Answer: H.264, H.265, and JPEG can be hardware accelerated via Nvidia but only H.264 and H.265 can be hardware accelerated via Intel Quick Sync.
Therefore, if the stream is MPEG-4, MxPEG, or JPEG (Intel Quick Sync only), the display will show "Hardware acceleration: Off."
Additionally, both NVIDIA and Intel Quick Sync can perform hardware acceleration of H.264 streams up to 4K resolution (3840 x 2160) and H.265 (HEVC) up to 8K resolution (7680 x 4320). This means that if any resolution (vertical or horizontal) is greater than these limitations, we will fall back to software decoding, and you will again see "Hardware acceleration: Off."

Question: Smart client turns Hardware acceleration off for all the cameras when using a camera with a resolution higher than 4K resolution (3840 x 2160) for H.264 streams and up to 8K resolution (7680 x 4320) for H.265 (HEVC), why is that?
Answer: Smart client by design will revert to software decoding and turns the hardware acceleration off because the Nvidia cards cannot support a higher resolution than the ones mentioned above.
Note: There is an issue in all the Smart Client versions prior to 2020 R1 that the Smart Client will turn the hardware acceleration off for all the cameras regardless of their resolution even if only one of the cameras is using an unsupported resolution.
This issue has been resolved in smart client 2020 R1 and later, which means that the Smart Client will only turn the hardware acceleration off for a camera with the unsupported resolution and other cameras that are using a supported resolution will continue using Nvidia Hardware acceleration.

Question: Nvidia offers different driver editions for my graphics card: "DCH", "Studio" and "Standard". Which one should I choose?
Answer: The DCH type (Declarative Componentized Hardware Supported apps) uses the new Windows universal driver type, which is recommended for XProtect hardware acceleration on Windows 10, Server 2016, and Server 2019 and the same for the "Studio" type. You may read more in this article from Nvidia: "NVIDIA DCH/Standard Display Drivers for Windows 10 FAQ".
The DCH-driver installers are smaller and faster to install because they do not offer the Nvidia Control panel, and in most cases, it will also work with the standard-type drivers.

Question: Mostly the Smart Client is mentioned above, but hardware accelerated decoding is also used in the Management Client, Recording Server, and Mobile Server, right?
Answer: Yes, the same rules apply for hardware acceleration in the Smart Client, as well as the Management Client, Recording Server, Mobile Server, and MIP SDK.

Related articles:

- KB 8771 "Smart Client NVIDIA hardware acceleration not working when the "NVSMI" folder is missing"
- KB 6212 "How to run the XProtect Smart Client on Nvidia Optimus hybrid graphic cards"
- KB 40572 "Hardware acceleration stops working after upgrade to 2021 R2 (troubleshooting)"
- KB 34607 "Minimum requirements for Hardware Acceleration (GPU decoding)"