CATEGORY: VISUALIZATION - LARGE SCALE & MULTI-DISPLAY - VD01
POSTER P5174
CONTACT NAME Rudolfs Bundulis: rudolfs.bundulis@lu.lv

NVENC Based H.264 Encoding for Virtual Machine Based Monitor Wall Architecture

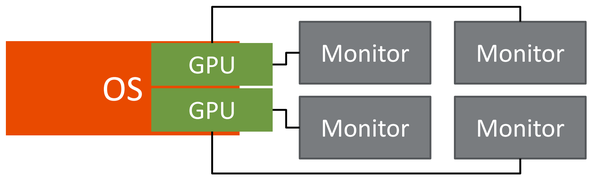
RIGA HIGH TECH UNIVERSITY

R. Bundulis (rudolfs.bundulis@lu.lv), G. Arnicans (guntis.arnicans@lu.lv), and R. Gailums (rihards.gailums@rhtu.edu.lv)
University of Latvia / Riga High Tech University, Latvia

Introduction

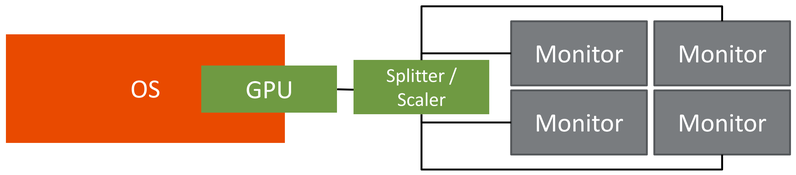
- IT industry experiences an increasing growth for display surfaces with high resolution
- Use cases for such surfaces include satellite and map data, x-ray and microscope images, multimedia, CCTV, etc.
- Existing solutions are not scalable, do not offer hardware abstraction, suffer from wiring limitations

Currently Popular Monitor Wall Architectures



Display wall architecture where each output of the GPU maps to a tile on the display wall

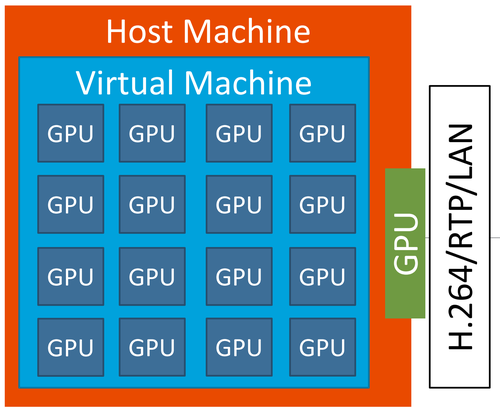
Pro: OS can natively manage the displays
Con: Power consumption, supported monitor count limited by the output count of the GPUs and expansion slots for the GPUs on the motherboard, deployment is limited by wiring



Display wall architecture where each output of the GPU is split/upscaled among the tiles on the display wall

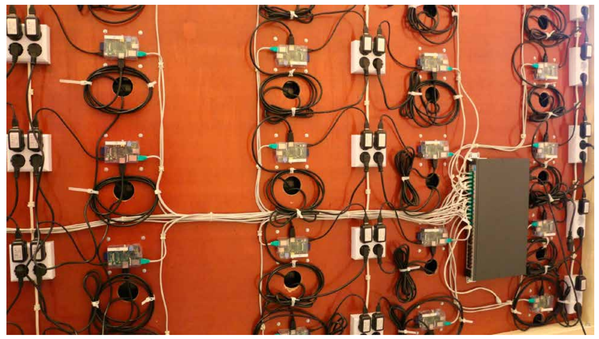
Pro: Software complexity is reduced since it does not have to be multiple monitor aware
Con: Small resolution and DPI, visualization is not displayed in it's native resolution
Con: Expensive

Proposed Virtual Machine Based Monitor Wall Architecture



The proposed display wall architecture where each output of a virtual GPU maps to a tile on the display wall and is transmitted as a H.264 stream over LAN

Virtual machine based monitor wall running Google maps inside Chrome web browser on 16 tiles at 1920x1080 pixels each giving a total resolution of 32 megapixels



Each tile has a dedicated LAN connection and H.264 decoder

- The host machine collects the frame buffer data from the virtual machine GPUs and performs H.264 encoding of the video stream on the physical host GPU thus the architecture heavily relies on a fast hardware H.264 encoder allowing the hosted virtual machines to fully use the CPU
- Non FPS intensive use cases allow a great number of virtual monitors to be hosted on a single physical GPU thus reducing the power consumption

Pro: Scalable, host machine can run multiple virtual machines, multiple virtualized GPU's map to physical GPU's to maximize efficiency
Pro: LAN connection to the display wall removes wire length limitations forced by DVI/HDMI cables
Pro: Total resolution of the wall goes beyond the ones that can be achieved using physical hardware
Con: Lossy compression
Con: No Direct3D, OpenGL support

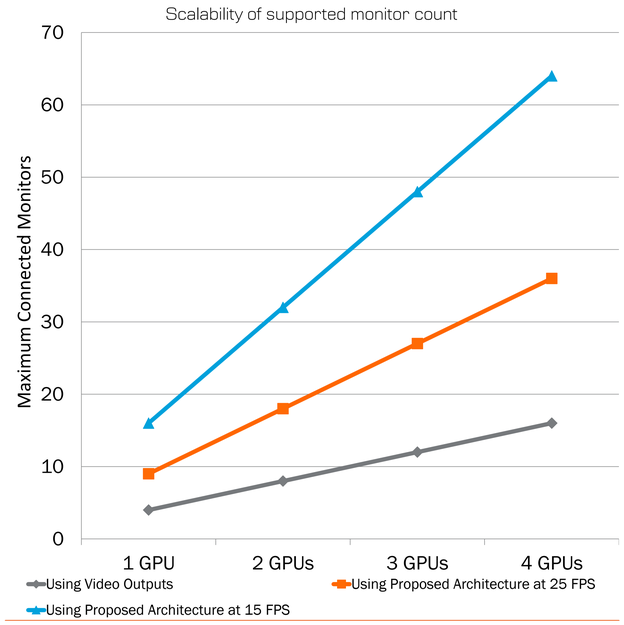
Why NVENC?

Currently there are two main alternatives to be used as the H.264 encoder in this architecture – Intel Quick Sync and NVENC. NVCENC is more feasible because:

- The total encoding power can be increased by stacking up multiple GPUs that support NVENC without penalties while not all Intel Quick Sync GPUs have built in video memory so scaling these cards introduce a performance penalty of using system memory
- NVENC does not put any limitations on other components of the system, while Intel Quick Sync supports a limited amount of CPUs
- Current benchmarks seem to show that the overall FPS performance for a single GPU (which is the main criteria for this architecture) is better for NVENC than Intel Quick Sync

Scalability

The graph below demonstrates the scalability possibilities in terms of possible maximal amount of connected monitors for the traditional architecture versus the proposed one on a Quadro K4000 card that has 4 outputs.



Conclusions

- The current experiments show that this architecture is very feasible for non FPS intensive use cases where the display wall can be driven by a single physical GPU
- The total resolution provided by this architecture even using the currently available compression technology greatly exceeds the resolutions of existing solutions, it would be expected for the resolution to grow in the future
- The architecture itself scales very good, it is limited mainly by OS support for multiple monitors (this can be overcome by simulating a single high resolution display in the virtual machine that spans the whole resolution of the physical wall) and the possibility to stack multiple GPU's in the host system
- Future work should focus on the ability to virtualize OpenGL and Direct3D to remove the advantages of non-virtualized architectures