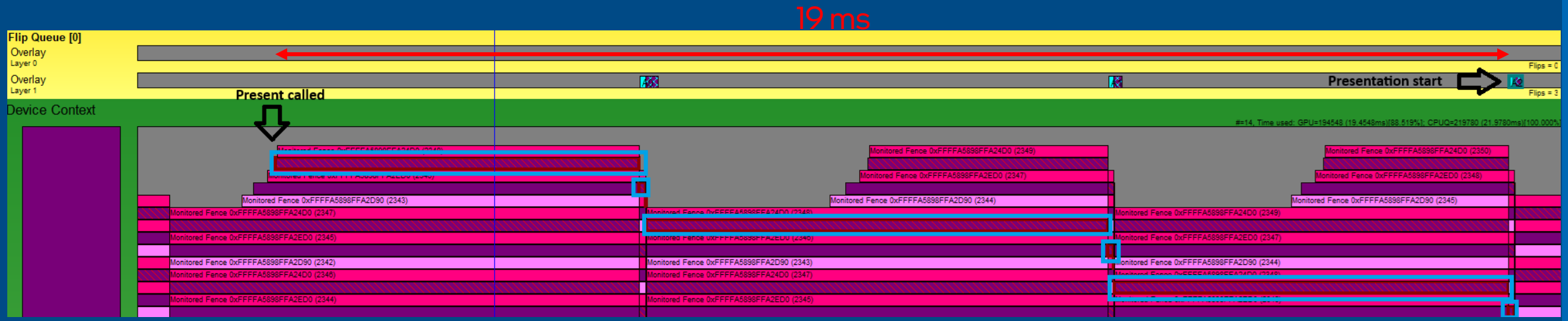
Latency Reduction Examples

Latency

- Depends on multiple factors:
  - application's design and performance,
  - monitor's specification (variable or fixed refresh rate),
  - software settings (e.g. VSync).
- Metrics:
  - Input-to-render
    - Delay between application's input sampling and rendering start on a GPU.
  - Render-to-screen
    - Delay between a finish of GPU rendering and actual presentation on a screen.
  - **Input-to-screen**
    - **Time between input sampling and frame presentation on a screen.**
- Tools
  - GpuView can be used to analyze application's latency and find bottlenecks.

Render queue – no latency reduction

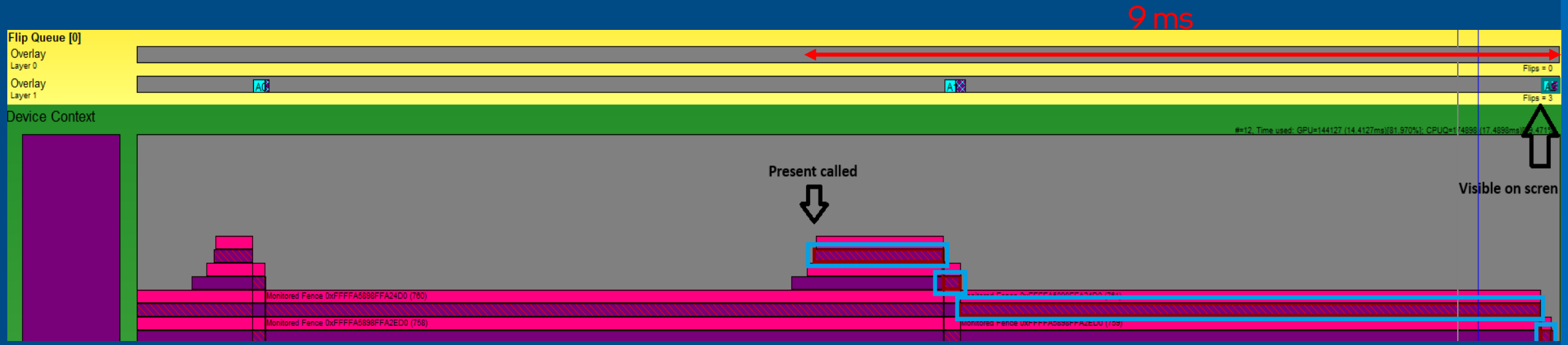
- Multiple frames in the queue
  - Two previous frames shown in between Present call and actual presentation start
- Not related to Display configuration and mode
- Causes input lag due to prematurely sampled input from the user
  - GPU can't proces the frame yet due to previous frames being in the queue
- Present packet spent 19 ms in a queue



60Hz FRR monitor, VSync disabled

Render queue – latency reduction

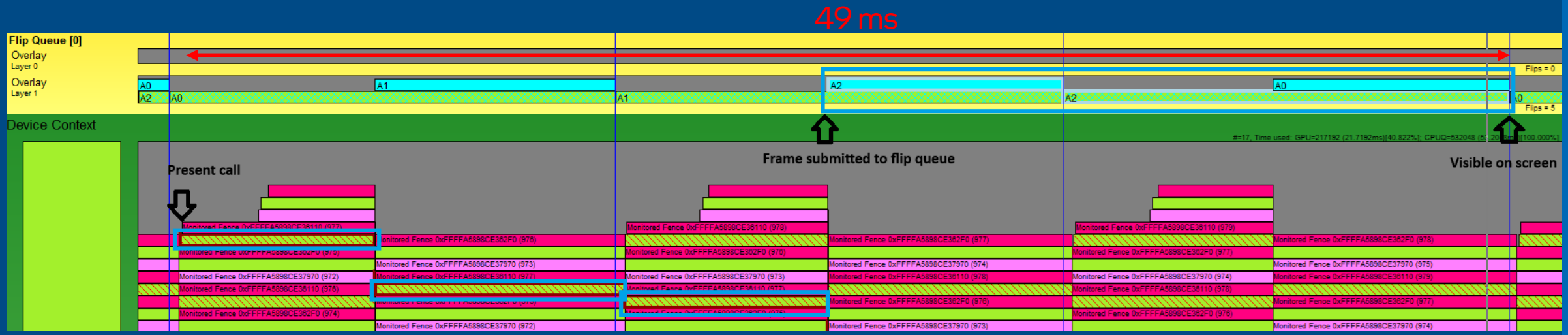
- Frames are submitted around a time when previous frame finishes execution on a GPU
- Present packet spent 9 ms in a queue (over 50% reduction)
- GPU utilization not affected



60Hz FRR monitor, VSync disabled

Display optimizations

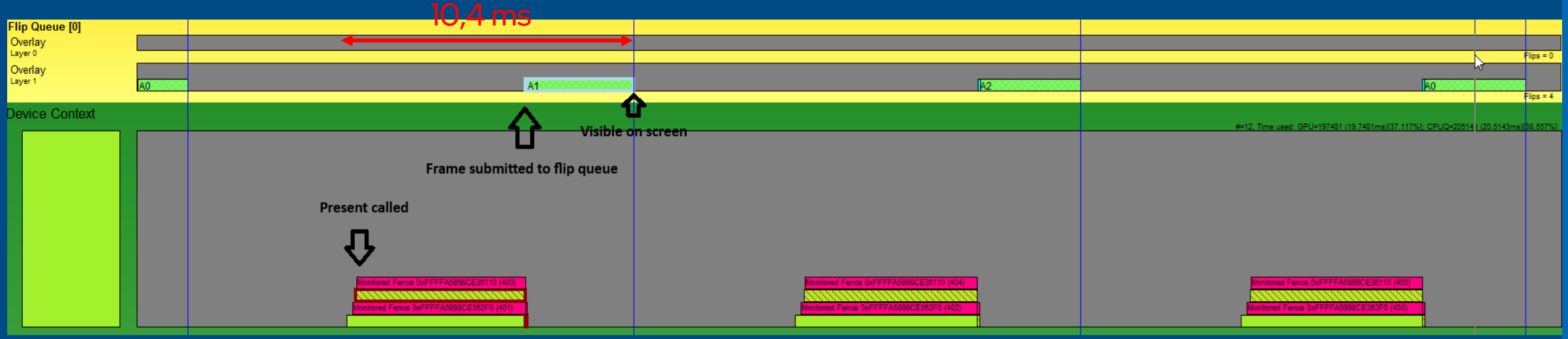
- Available in "No tearing" (aka "VSync ON") display modes
- 49 ms from Present submission until actual presentation
- Delay caused by Render Queue and Flip Queue



60Hz FRR monitor, VSync enabled

Display optimizations

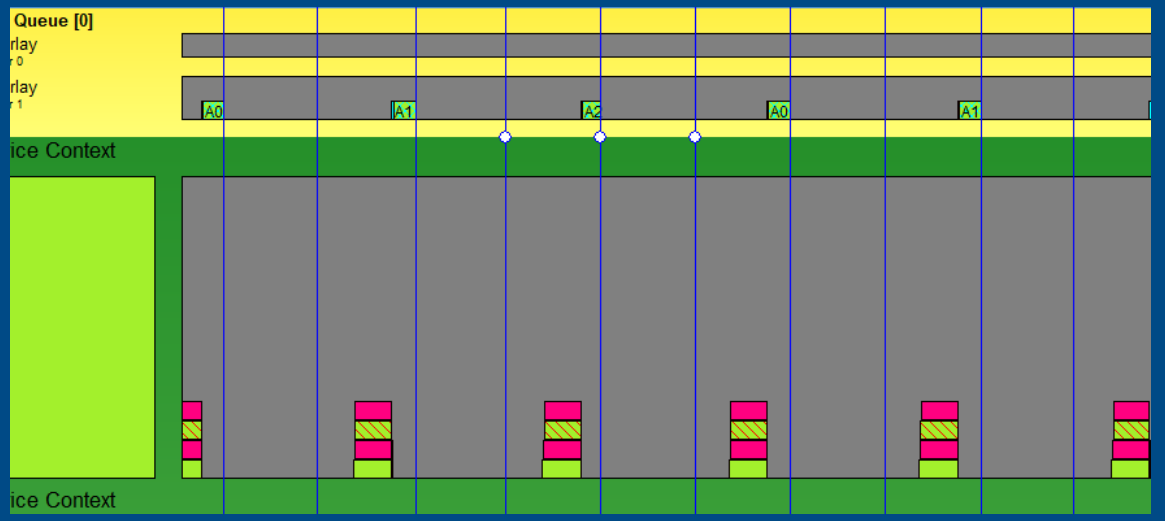
- Possible if performance is sufficient (FPS >= display refresh rate)
- 10.4 ms from Present submission until actual presentation
- Render queue is also reduced
  - Gaps created due to V-Sync FPS limit do not affect performance
- No Input-to-render latency
- Render-to-screen minimized



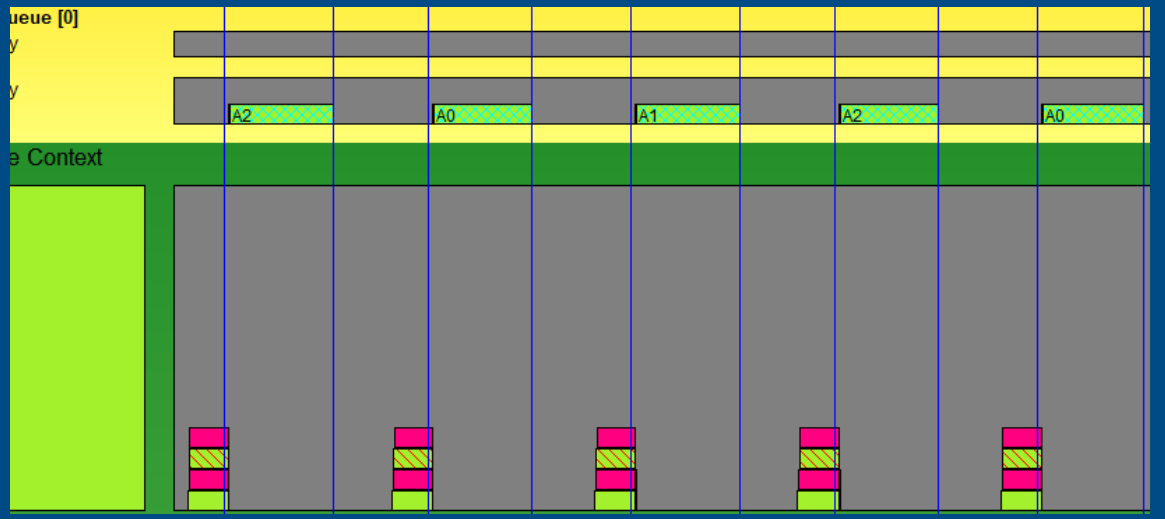
60Hz FRR monitor, VSync enabled

Frame limiter and latency reduction

- 60 Hz display, VSync on
- Frame limiter set to 30 FPS
  - Means 1 frame every other Vsync period
- Frame cap interval is aligned with Vsync intervals



Latency reduction on, mnimal input-to-screen time



Latency reduction off, frames are ready to early, present packets wait in a flip queue

intel