ODR: OnDemand Rendering

Improving Resource and Energy Efficiency for Cloud 3D through Excessive Rendering Reduction

Tianyi Liu*

*Jerry Lucas** *Sen He*‡ *Tongping Liu* *Xiaoyin Wang**

*Wei Wang**

University of Texas at San Antonio*

University of Arizona ‡

University of Massachusetts Amherst

UTSA Computer Science

THE UNIVERSITY OF ARIZONA

UMASS AMHERST

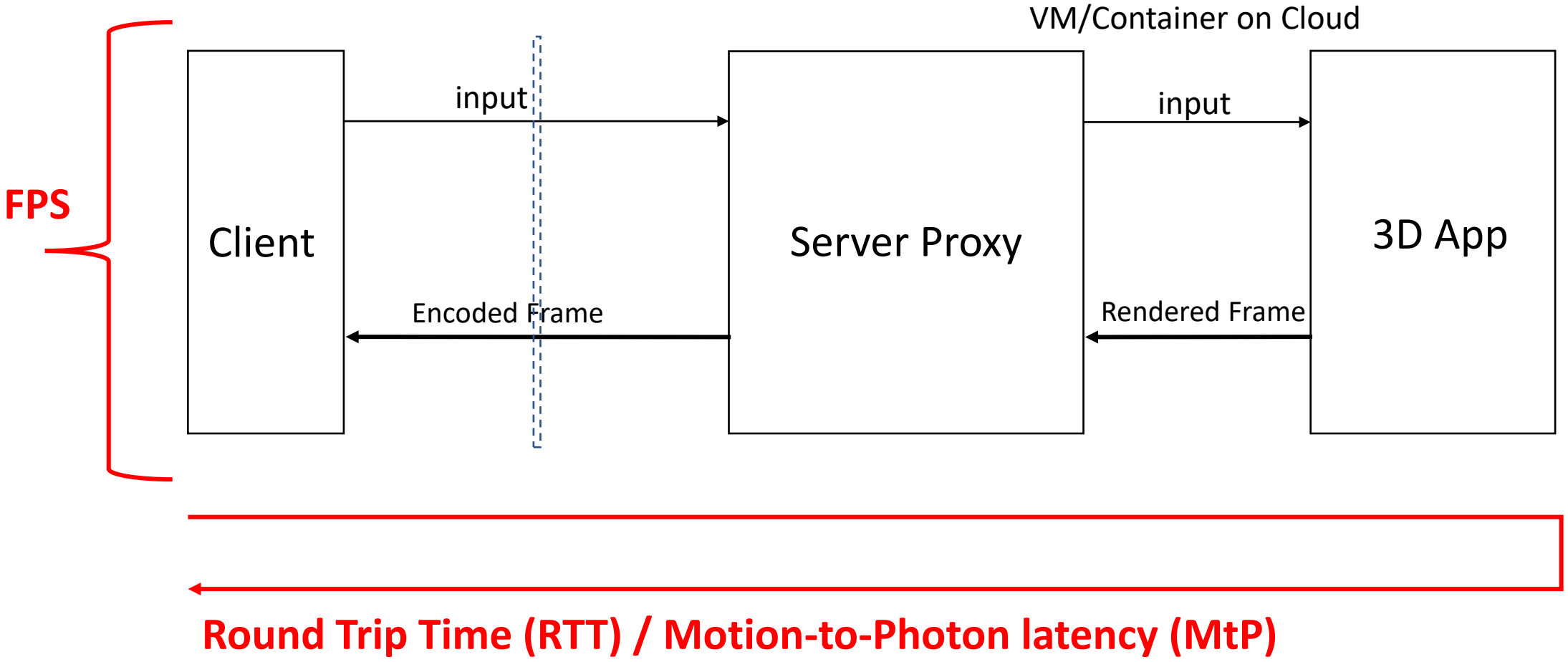
Outline

1. Background: Cloud 3D
2. Research Problem
3. Related Work
4. ODR Design
    - OnDemand Rendering
    - Priority Frame
5. Evaluation
6. Summary

Cloud 3D System (An idea)

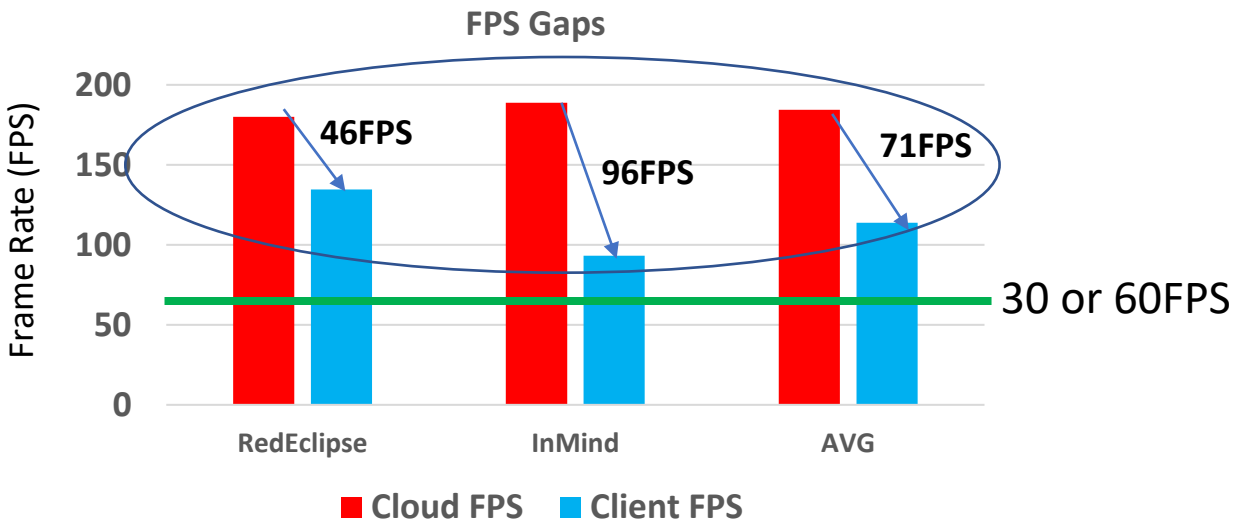
- As User (Benefits)
  - Via a thin-client
  - Anywhere & Anytime
  - No hardware update
  - No download
  - ……
- Further Abstraction →

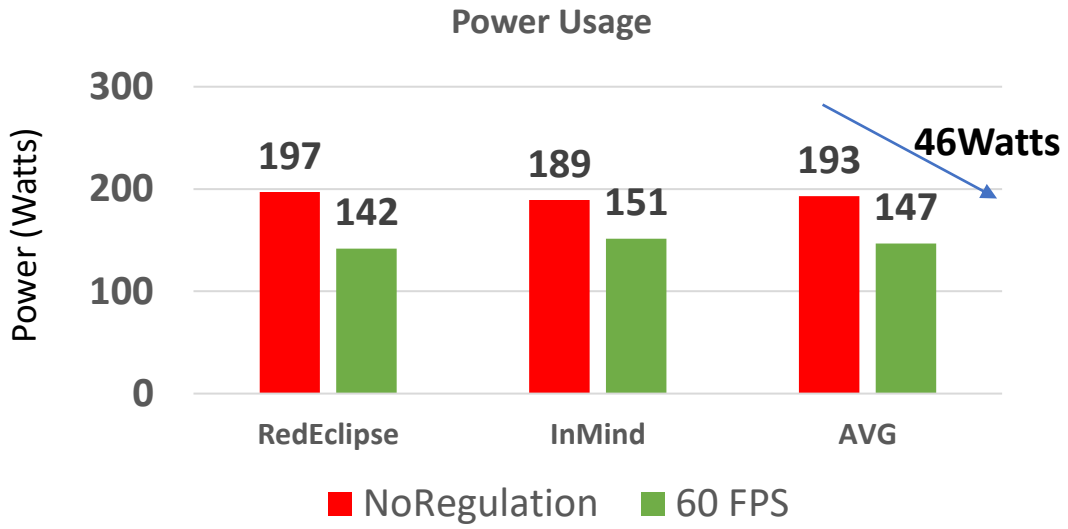
Cloud 3D System Overview



Research Problem: Low System Efficiency (1/2)

System efficiency is an important design metric for cloud3d.





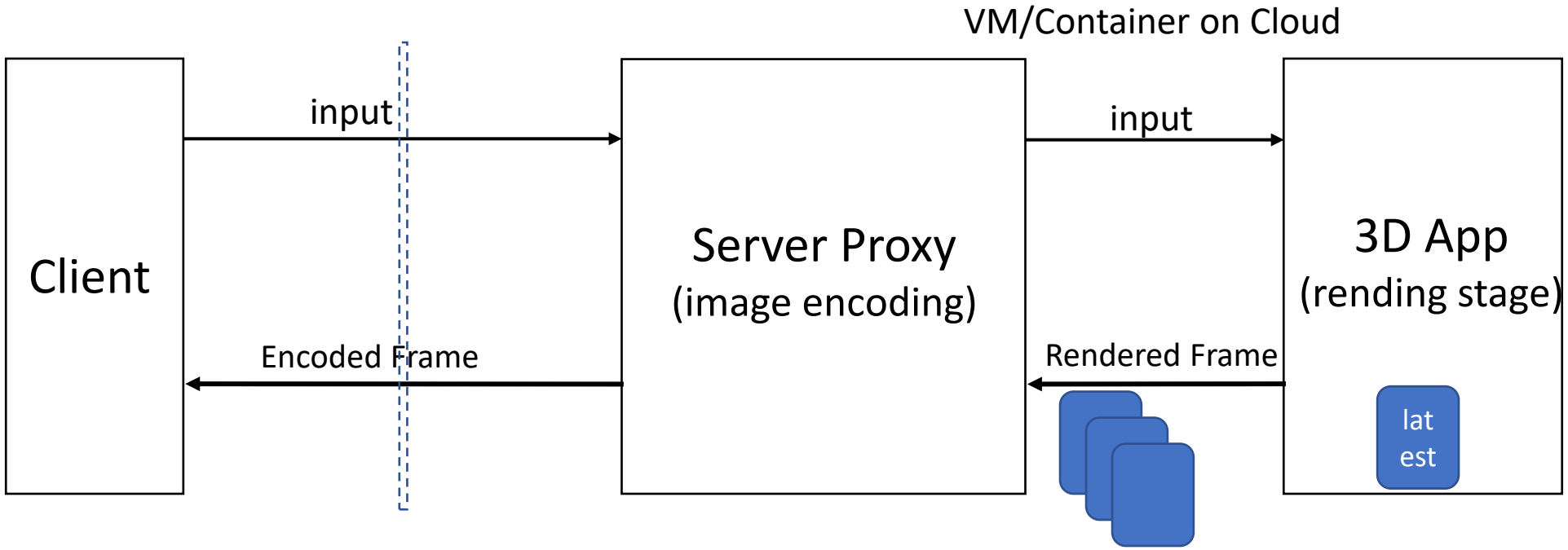
- Frame rate gap leads to low system efficiency.
- Now, let's explore how this frame rate gap happens?

Research Problem: Low System Efficiency (2/2)

Real Time Goal: Each component in cloud3D system usually work in parallel, and they are designed to provide latest images for next pipeline stage to ensure low MtP latency.

Root cause:

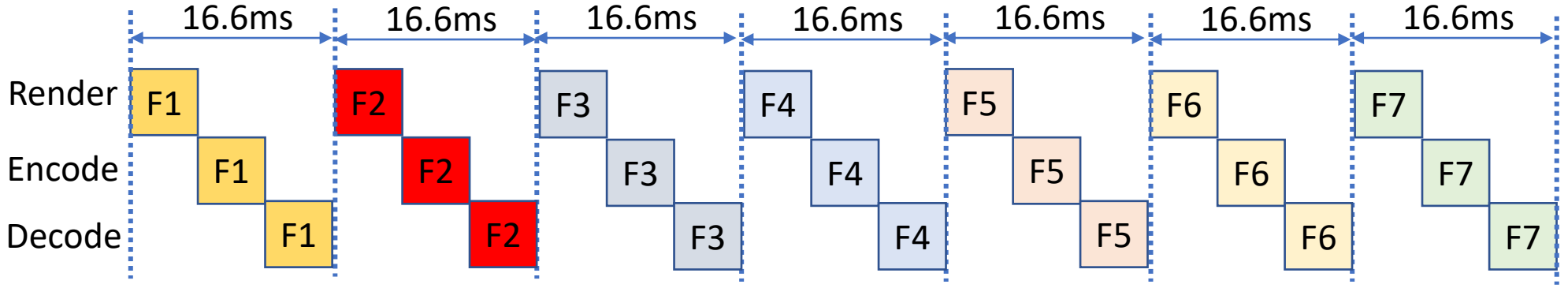
- Speed mismatch between cloud 3D stages causes frame dropping.
- Pipeline synchronization would violate real-time requirement of cloud3D.



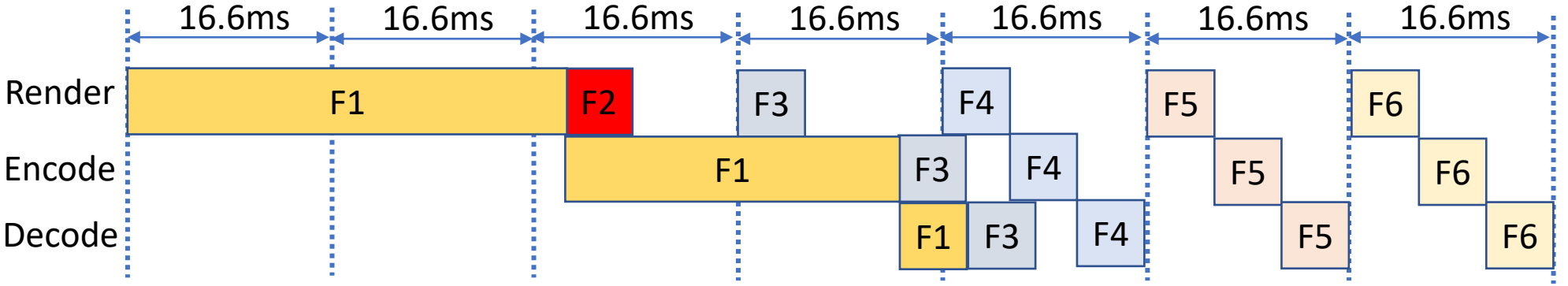
Related Work: FPS regulation (1/3)

Solution1: Interval-Based FPS Regulation (Int~) [1,2]

Ideal Pipeline of Interval-Based (Int~) FPS regulation



Actual Pipeline w. Interval-Based (Int~) FPS regulation



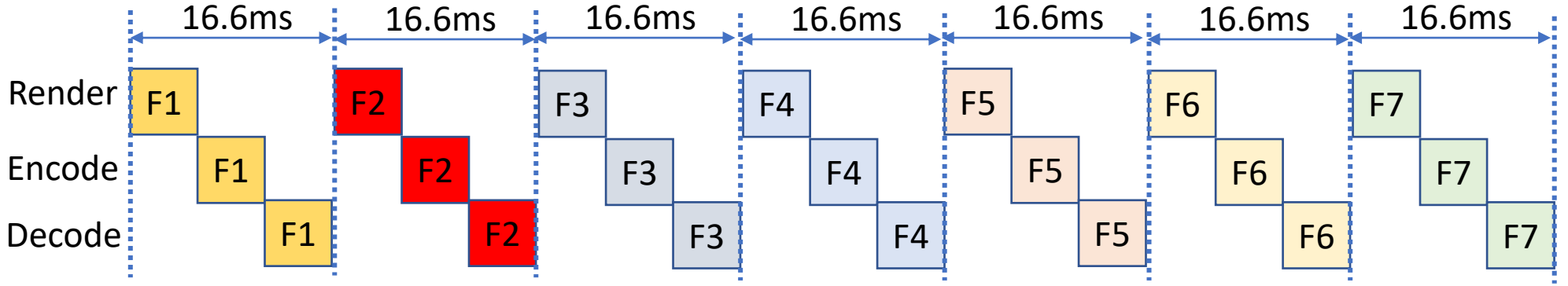
Cons: Still have FPS Gap & FPS is low.

[1] Dan Ginsburg, Budirijanto Purnomo, Dave Shreiner, and Aaftab Munshi. *OpenGL ES 3.0 Programming Guide*. Addison-Wesley Professional, 2014.
[2] Andrew Mulholland and Glenn Murphy. *Java 1.4 Game Programming*. Wordware Publishing, Inc., 2003.

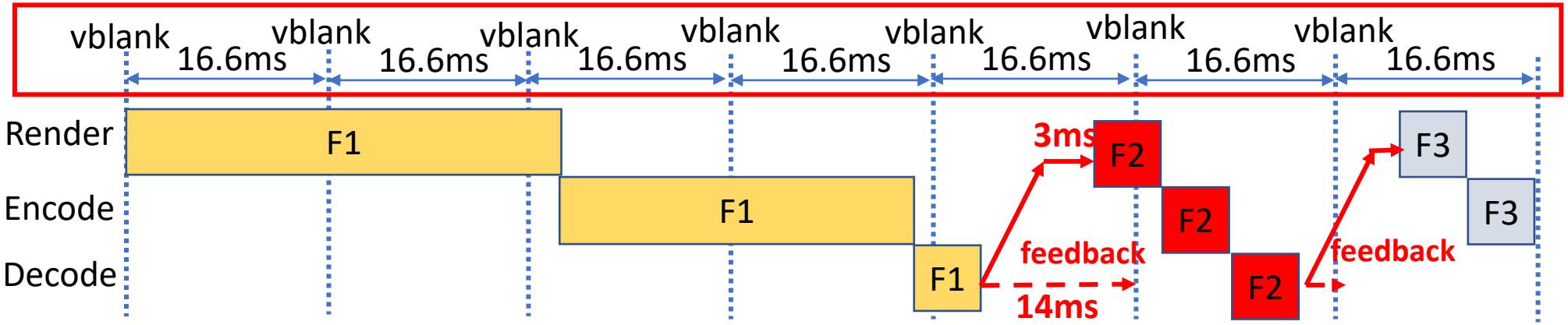
Related Work: FPS regulation (2/3)

Solution2: Remote-Vsync (RVS)[3]

Ideal Pipeline



Actual Pipeline w. Remote-Vsync (RVS) FPS regulation



Pros: No FPS Gap; Cons: But FPS is low.

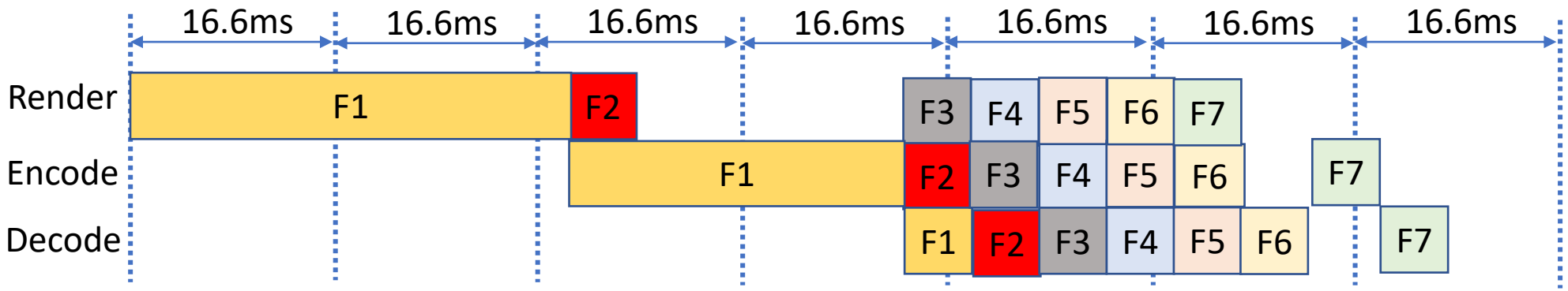
[3] Luyang Liu, Ruiguang Zhong, Wuyang Zhang, Yunxin Liu, Jiansong Zhang, Lintao Zhang, and Marco Gruteser. Cutting the Cord: Designing a High-Quality Untethered VR System with Low Latency Remote Rendering. In *Proc. of Int'l Conf. on Mobile Systems, Applications, and Services*, 2018.

Related Work & Challenges Summary(3/3)

Expect a Better Pipeline

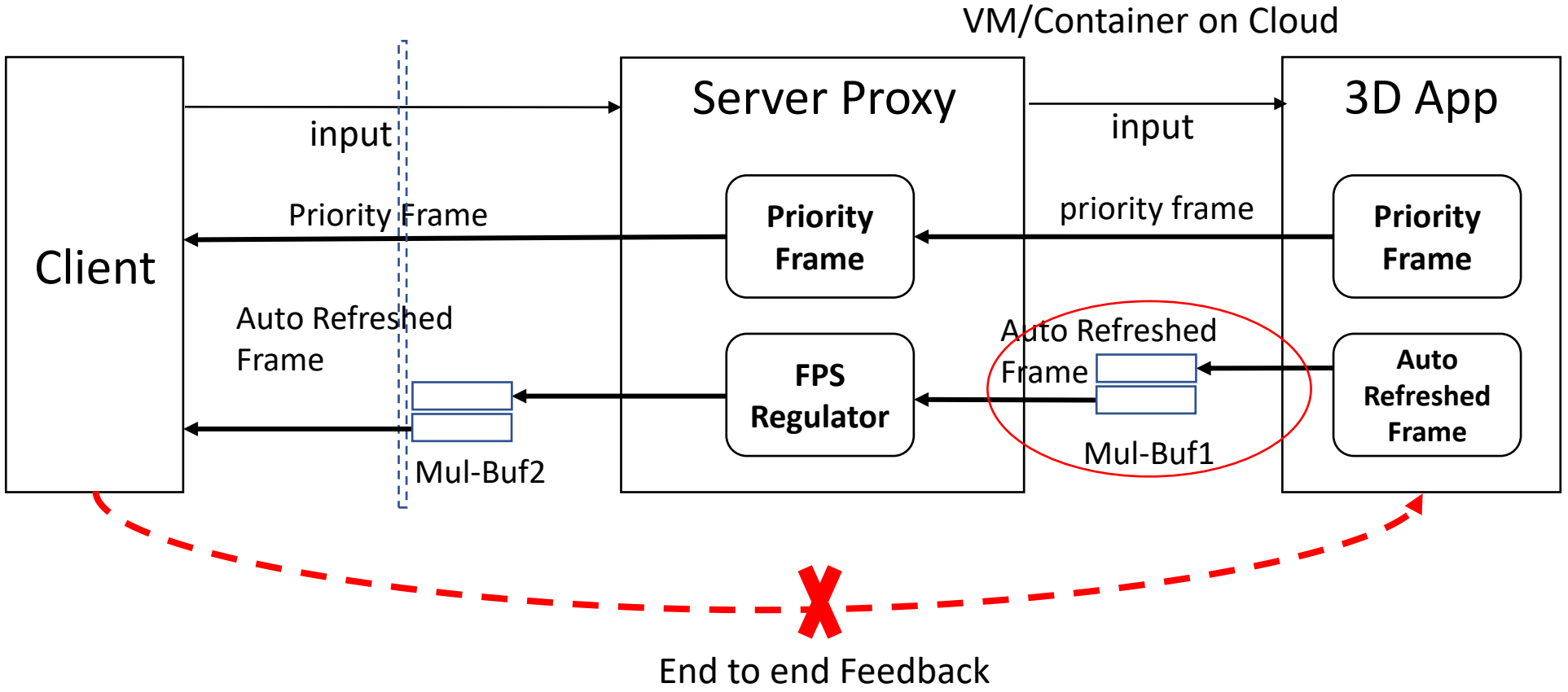
> Interval-based Regulation (Int ~)

> Remote V-Synch (RVS ~)



OnDemand Rendering: Two Multi-Buffers

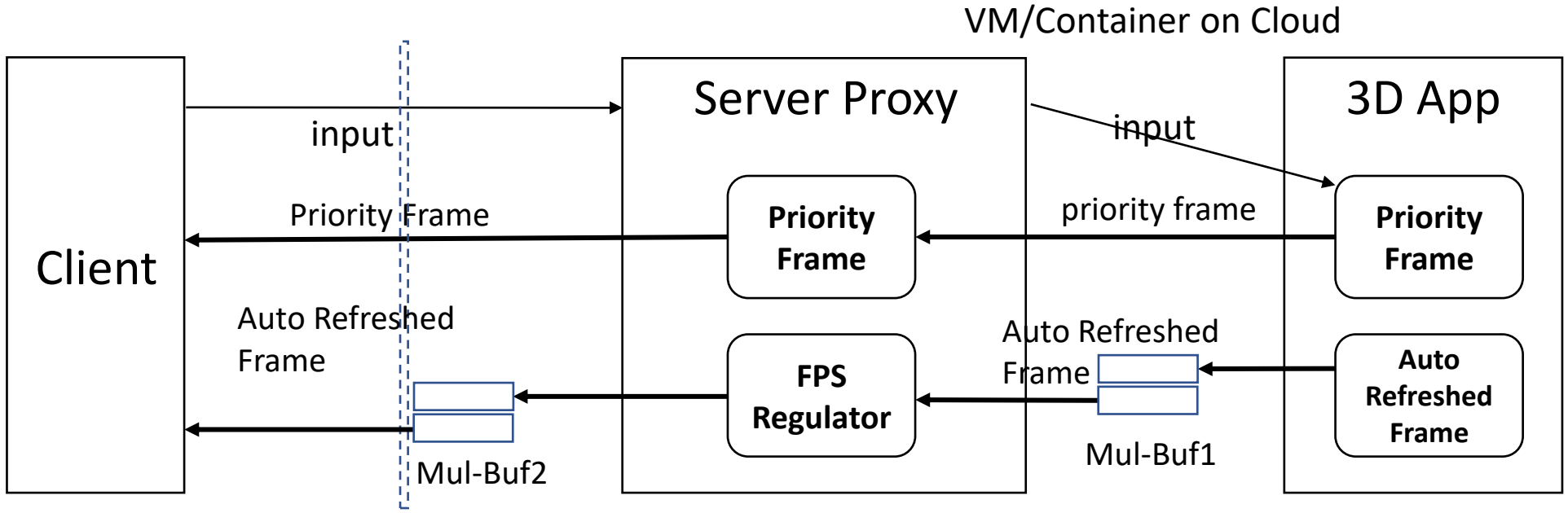
- Multi-Buffer: synchronization & parallelization



Pros & Cons: Synchronization between producer and consumer can eliminate framerate gap while maintaining high fps. **However**, it would violate the real-time feature of cloud3d, because the faster stage needs to wait for the slower stage.

OnDemand Rendering: Priority Frame & FPS Regulator

Key Observation: Two kinds of frames: 1) input-triggered frames; 2) frames generated by the application's internal updates. & **Input-triggered frames determines the user experience**. So, input-triggered frames can be prioritized.

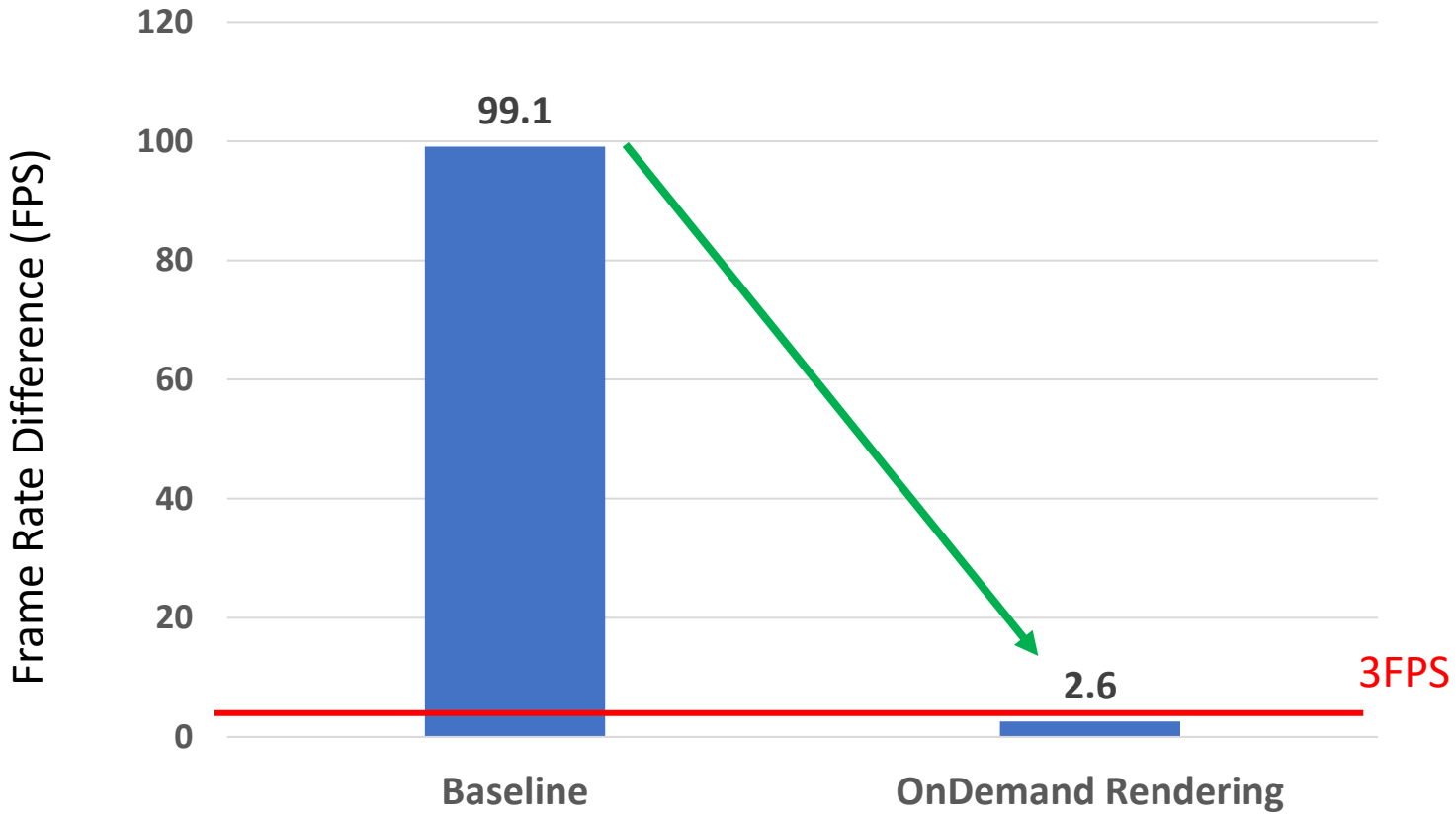


- Priority frame: guarantee real-time requirement of cloud 3D.
- FPS Regulator: accelerate or slowdown frame processing.

Evaluation Setups:

- **Platform**
  - Pictor Benchmarking Framework
- **Experiments**
  - Cloud3D evaluation on Private cloud with 720p & 1080p
  - Cloud3D evaluation on Google Cloud with 720p & 1080p
- **Metrics**
  - FPS gap (FPS).
  - Average FPS & MtP latency.
  - 99%Tail performance.
  - Micro-architectural level behaviors & Energy consumption.

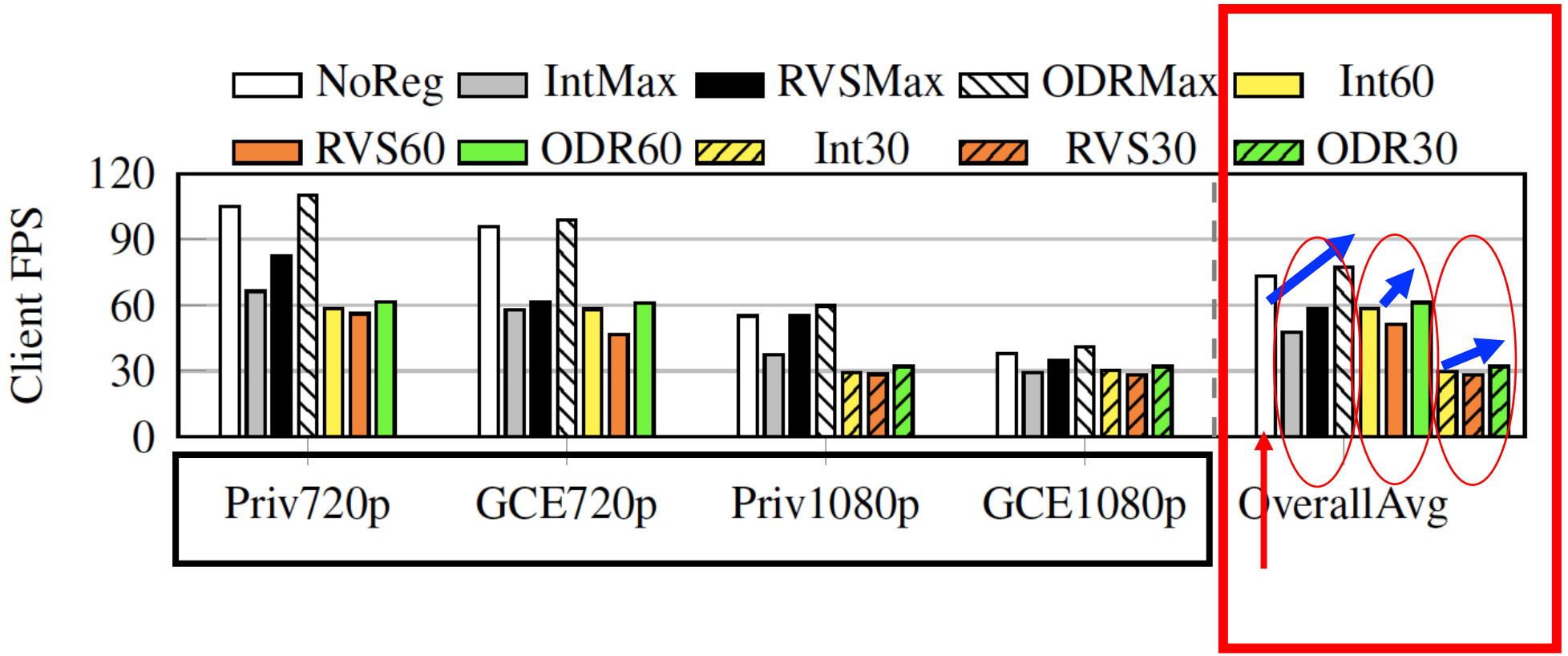
Evaluation: FPS Gaps



OnDemand Rendering can effectively bridge the frame rate gap.

Evaluation: Average Frame Rate (FPS)

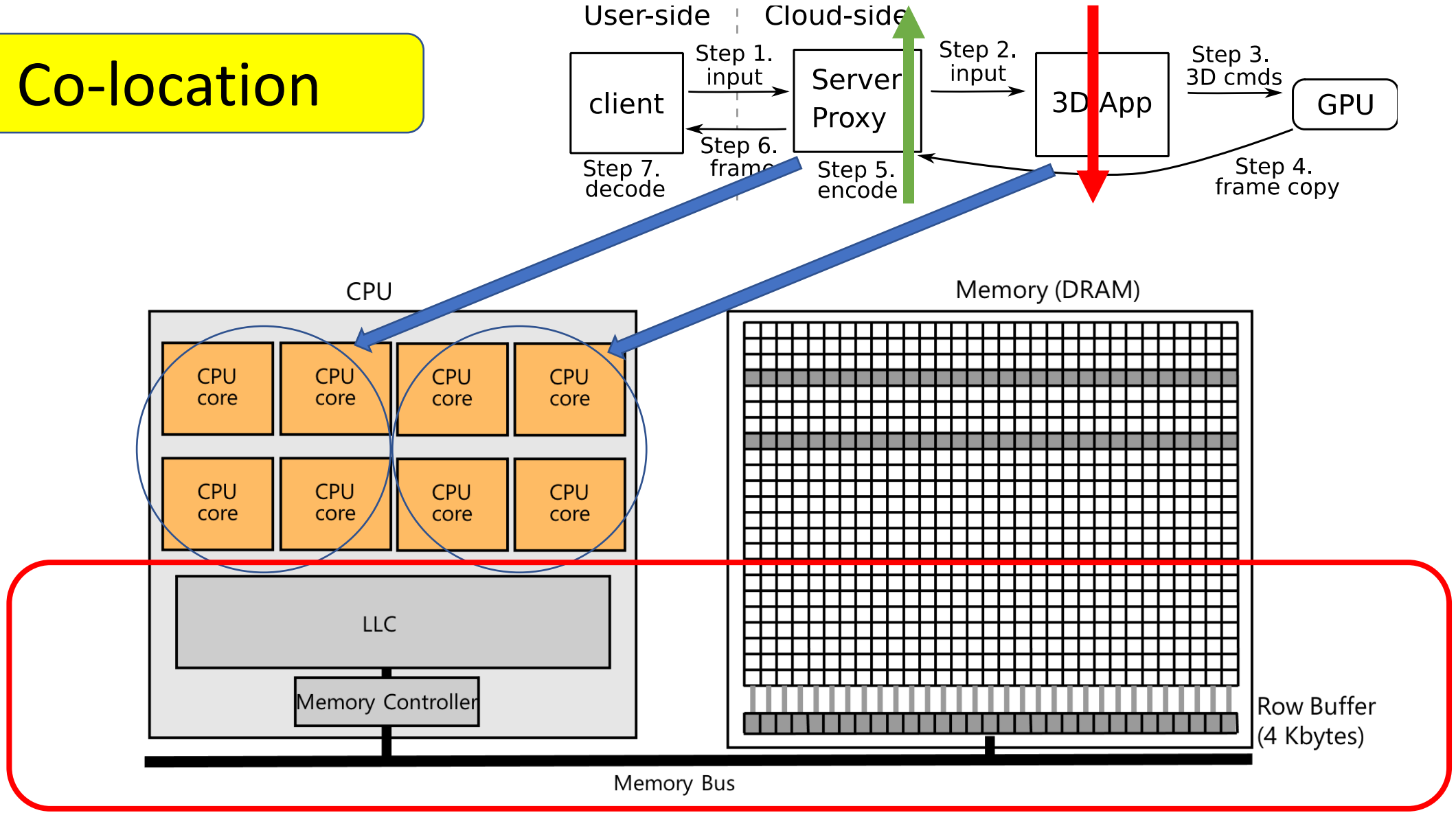
1) Private Cloud: 720p or 1080p 2)Google Cloud: 720p or 1080p



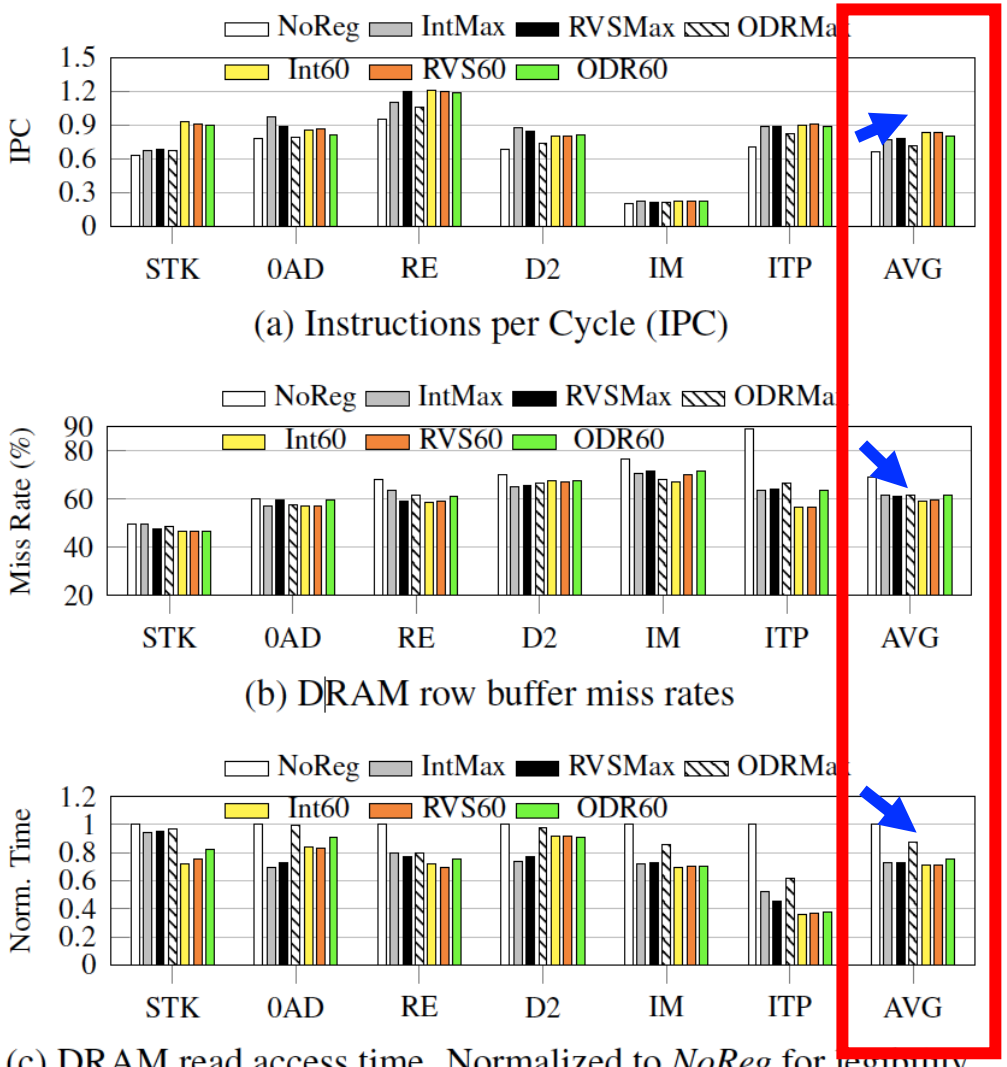
ODR has HIGHER average FPS than SOTA solutions.

Root cause: Hardware Contention

Co-location



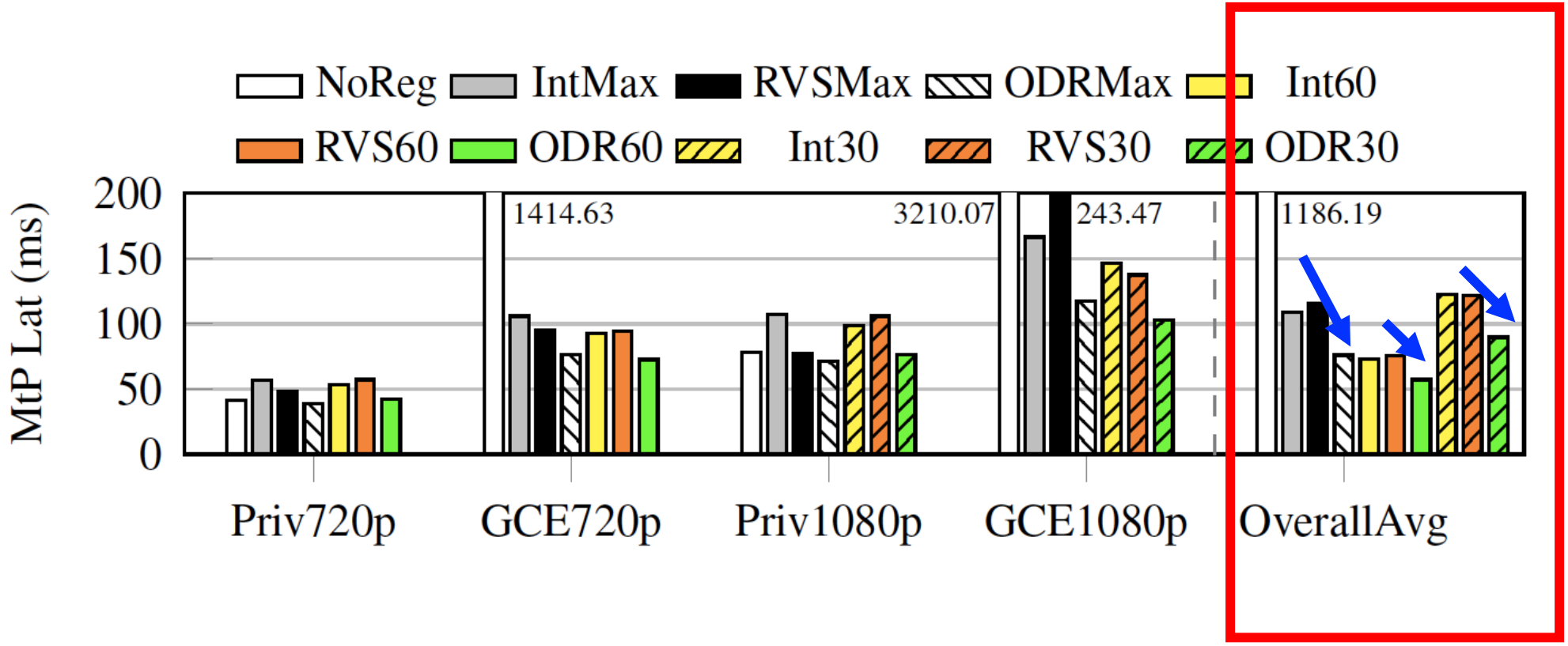
Micro-Architectural Results



ODR has LESS hardware contentions.

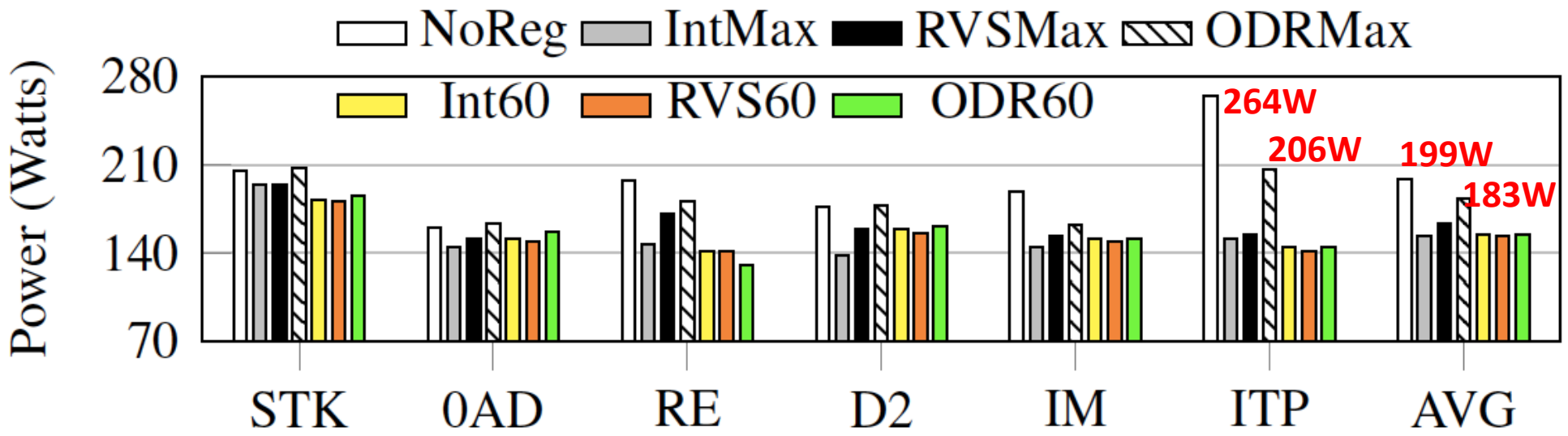
Evaluation: Average Motion-to-Photon Latency (MtP)

1) Private Cloud: 720p or 1080p 2)Google Cloud: 720p or 1080p



ODR has LOWER average MtP latency than SOTA solutions, because of Priority Frame.

Evaluation: Power Consumption



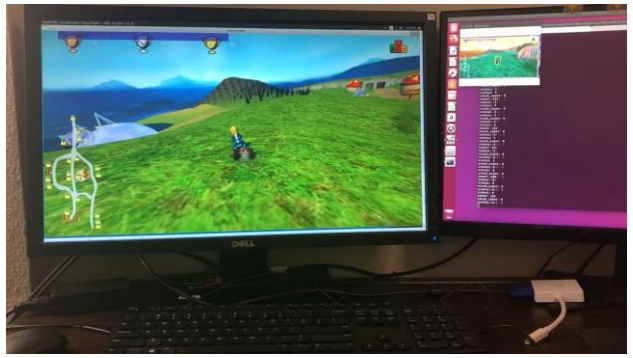
ODR has BETTER energy and resource efficiency.

Demo Cloud3D in our LAB:

https://www.youtube.com/watch?v=4VG0KgFgc_c
https://www.youtube.com/watch?v=-BnYlKonxJI
https://www.youtube.com/watch?v=mgz5tWt2_rc
https://youtu.be/gfoEGBjE6XA
https://youtu.be/ADh-vgHiO7M

User Experience Study

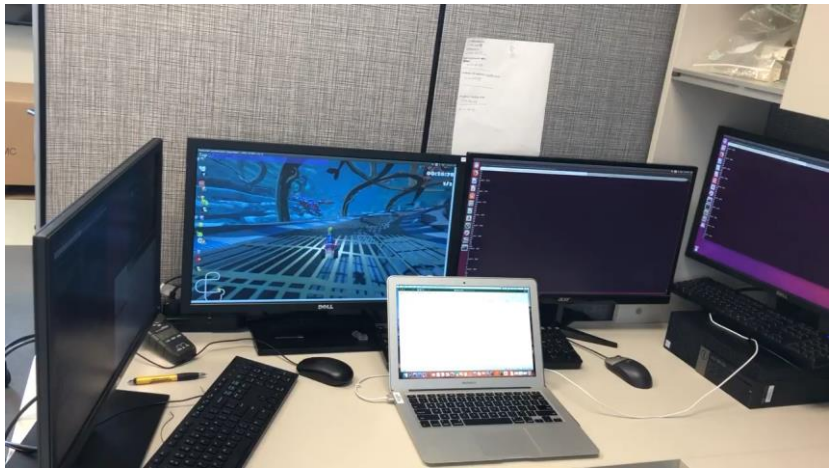
1. AI Bot Example:



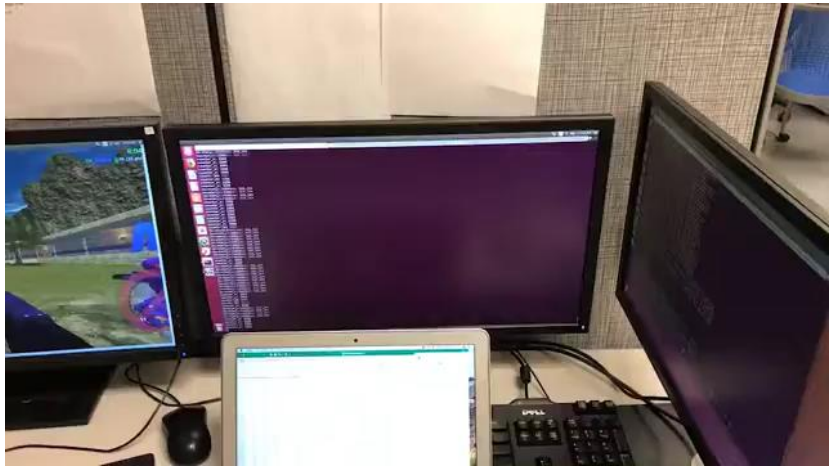
2. Local & Edge

3. Google Cloud (Public)

4. Four 3D Game Run on Edge.



5. Another Four 3D Game Run on Edge.



Conclusion

- A novel FPS regulation solution, OnDemand Rendering (ODR),
  - Multi-buffering
  - Priority frame
  - Dynamic delay/acceleration

  to reduce excessive rendering and ensure QoS satisfaction.
- Compared to no FPS regulation
  - ODR improved DRAM performance by 19%
  - Reduced power usage by 16.0%
  - Increased client FPS by 5.5%
  - Reduced MtP latency by 92.0%
  - ODR also outperformed existing SOTA solutions (Interval-Based/Remote-Vsynch).

ODR: OnDemand Rendering

Improving Resource and Energy Efficiency for Cloud 3D through Excessive Rendering Reduction

Tianyi Liu*

*Jerry Lucas** *Sen He*‡* *Tongping Liu* *Xiaoyin Wang**

*Wei Wang**

University of Texas at San Antonio*

University of Arizona‡

University of Massachusetts Amherst

UTSA Computer Science

THE UNIVERSITY OF ARIZONA

UMASS AMHERST