nsdi '26

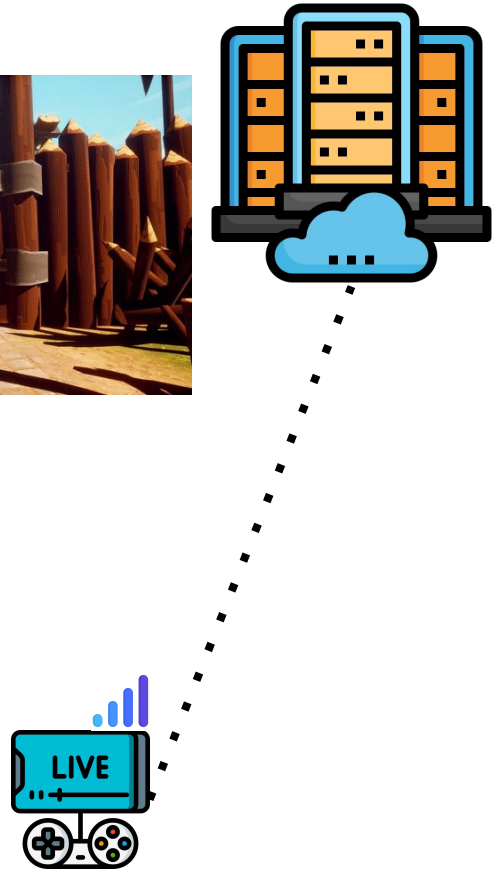
Stimpack: An Adaptive Rendering Optimization System for Scalable Cloud Gaming

Jin Heo[1], Vic Wang[2], Ketan Bhardwaj[2], Ada Gavrilovska[2]

[1] Dolby [2] Georgia Tech

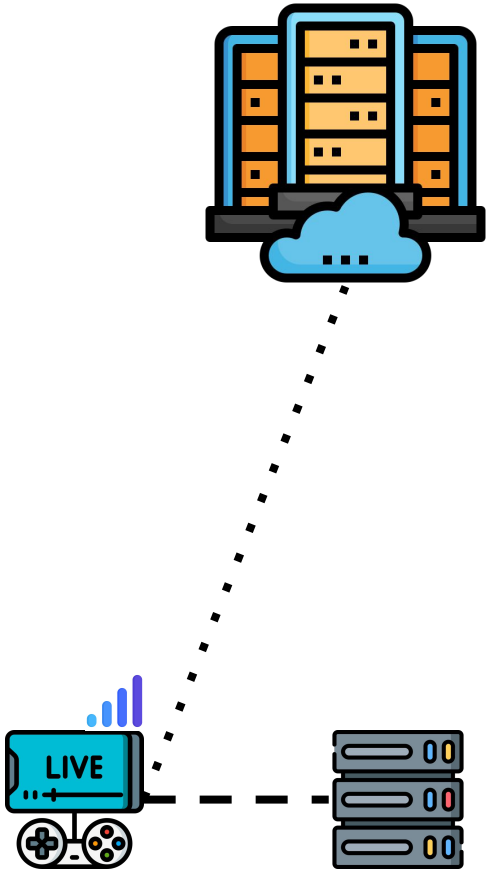
Cloud Gaming

- Runs a game on the remote server and streams the rendered frames.
- FPS + Visual Quality + Low Latency → Gaming Experience



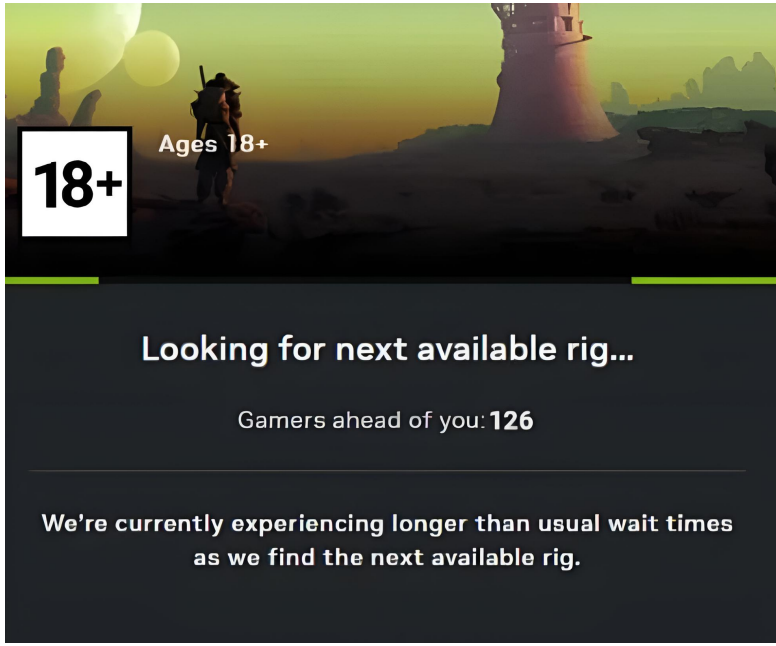
Fundamental Dilemma

- Location Trade-off
  - Cloud servers: high latency, but high scalability → Poor Service Accessibility
  - Edge servers: low latency, but limited resources (Scalability Bottleneck)



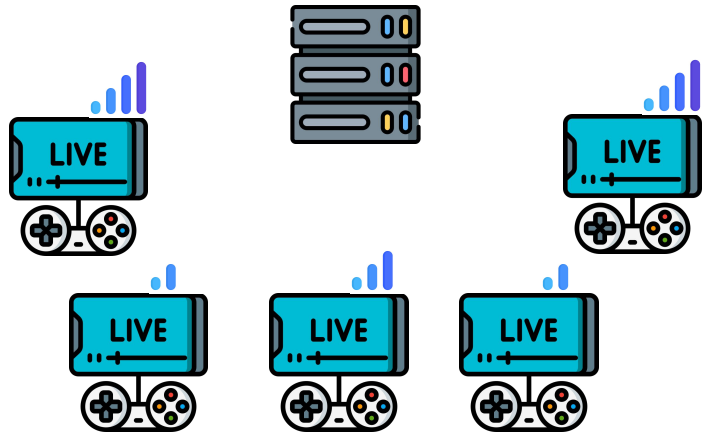
Fundamental Dilemma

- Location Trade-off
  - **Cloud servers**: high latency, but high scalability → **Poor Service Accessibility**
  - **Edge servers**: low latency, but limited resources (Scalability Bottleneck)
- Existing Approach
  - Strategies: playtime limits & user queuing for serving users' QoS
  - Low Service Availability



Fundamental Dilemma

- Location Trade-off
  - Cloud servers: high latency, but high scalability → Poor Service Accessibility
  - Edge servers: low latency, but limited resources (Scalability Bottleneck)
- Existing Approach
  - Strategies: playtime limits & user queuing for serving users' QoS
  - Low Service Availability
- Goal: Increasing per-GPU scalability while maintaining QoS



Insight: Inefficient Rendering Effort

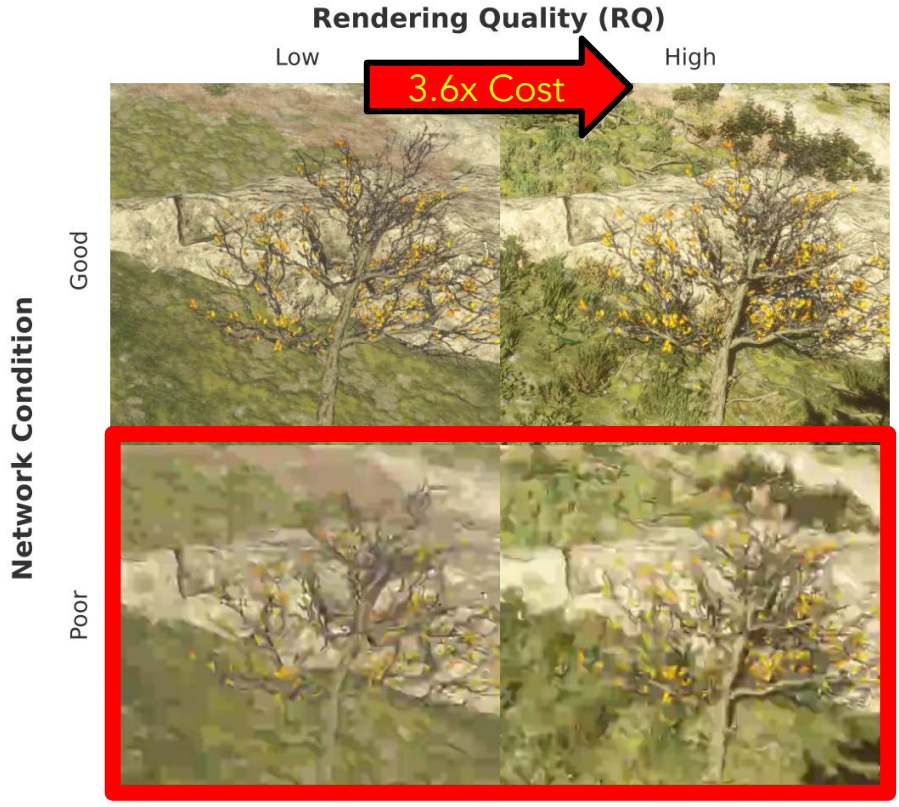
- User-perceived Frame Quality
  - Determined by both **Rendering Quality (RQ)** and **Compression Quality (QP)**

Insight: Inefficient Rendering Effort

- User-perceived Frame Quality
  - Determined by both **Rendering Quality (RQ)** and **Compression Quality (QP)**
- Wasted Effort in Poor Network Condition
  - Diminishing return of rendering efforts due to compression loss
  - GPU resource usage ≠ User-perceived quality gain

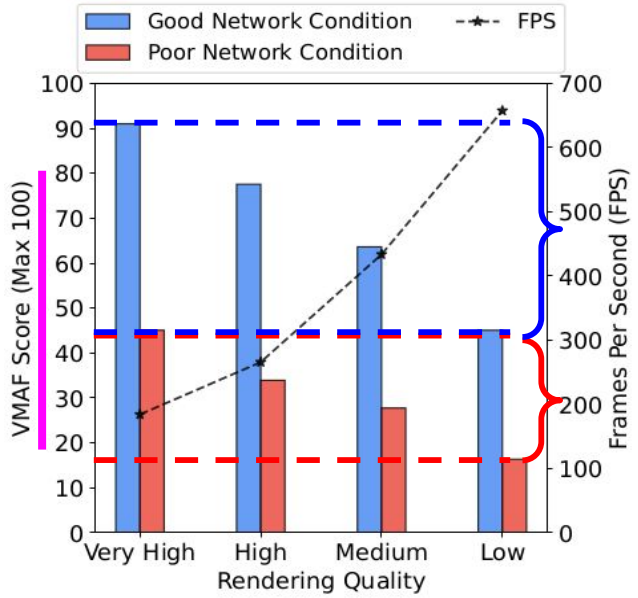
Insight: Inefficient Rendering Effort

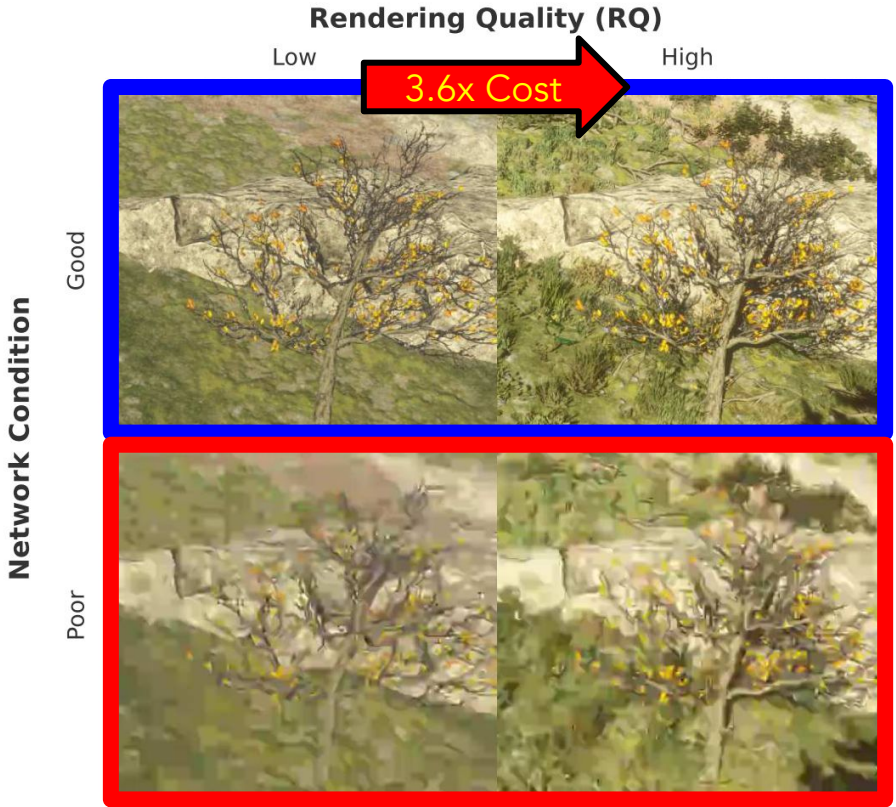
- User-perceived Frame Quality
  - Determined by both **Rendering Quality (RQ)** and **Compression Quality (QP)**
- Wasted Effort in Poor Network Condition
  - Diminishing return of rendering efforts due to compression loss
  - GPU resource usage ≠ User-perceived quality gain



Insight: Inefficient Rendering Effort

- User-perceived Frame Quality
  - Determined by both **Rendering Quality (RQ)** and **Compression Quality (QP)**
- Wasted Effort in Poor Network Condition
  - Diminishing return of rendering efforts due to compression loss
  - GPU resource usage ≠ User-perceived quality gain





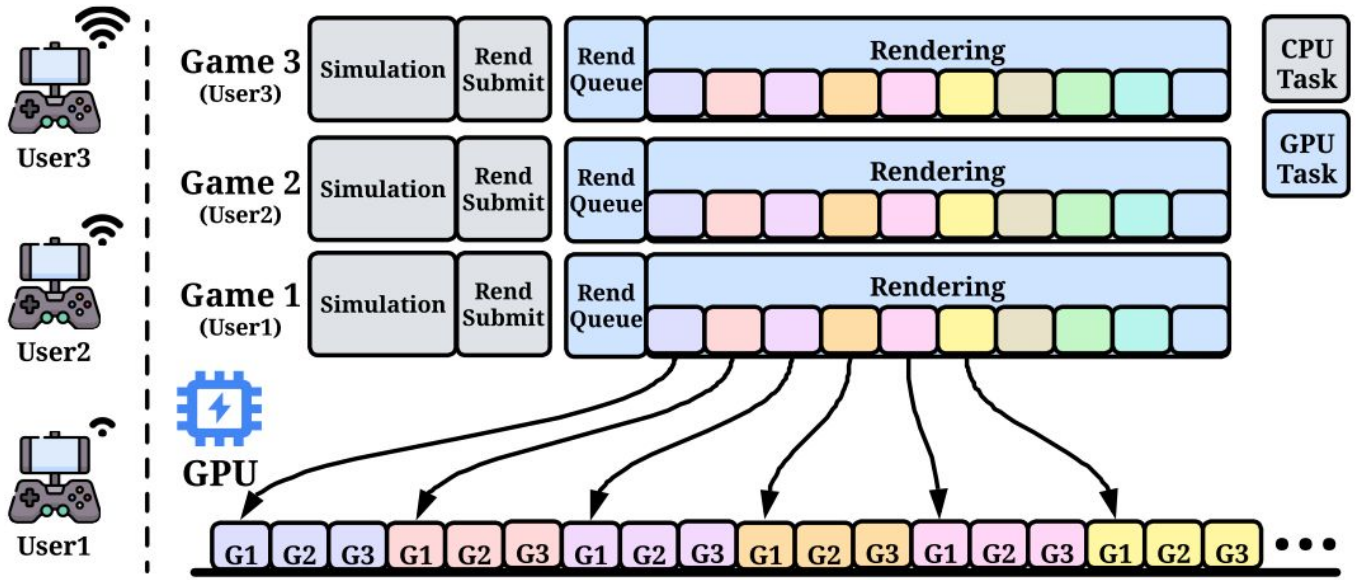
Insight: Inefficient Rendering Effort

- User-perceived Frame Quality
  - Determined by both **Rendering Quality (RQ)** and **Compression Quality (QP)**
- Wasted Effort in Poor Network Condition
  - Diminishing return of rendering efforts due to compression loss
  - GPU resource usage ≠ User-perceived quality gain

High quality rendering not considering user's network condition can be resource-inefficient;
Optimize resource allocation based on network-aware, user-perceived quality.

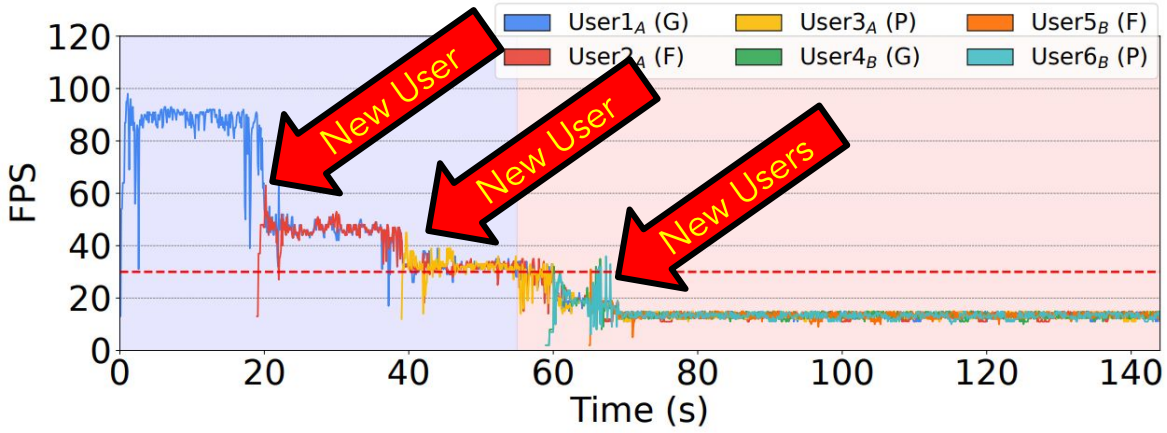
Stimpack: High-level Approach Overview

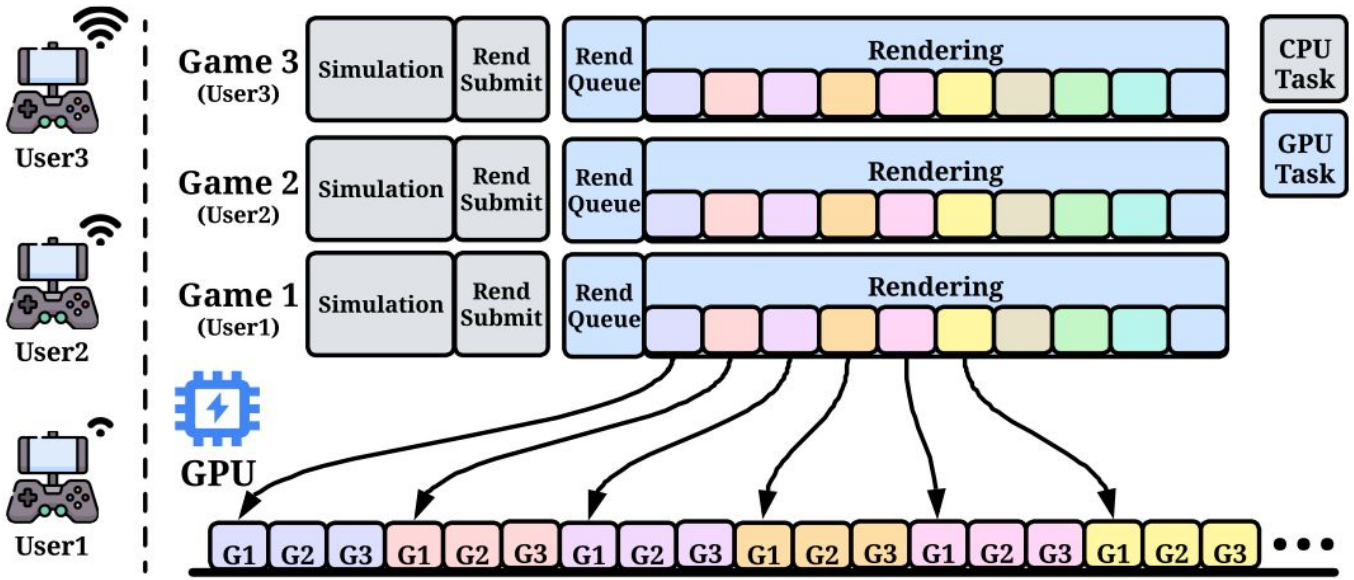
- GPU Resource Sharing in Cloud Gaming
  - Multiple concurrent users → Increased frame rendering latency → FPS drop



Stimpack: High-level Approach Overview

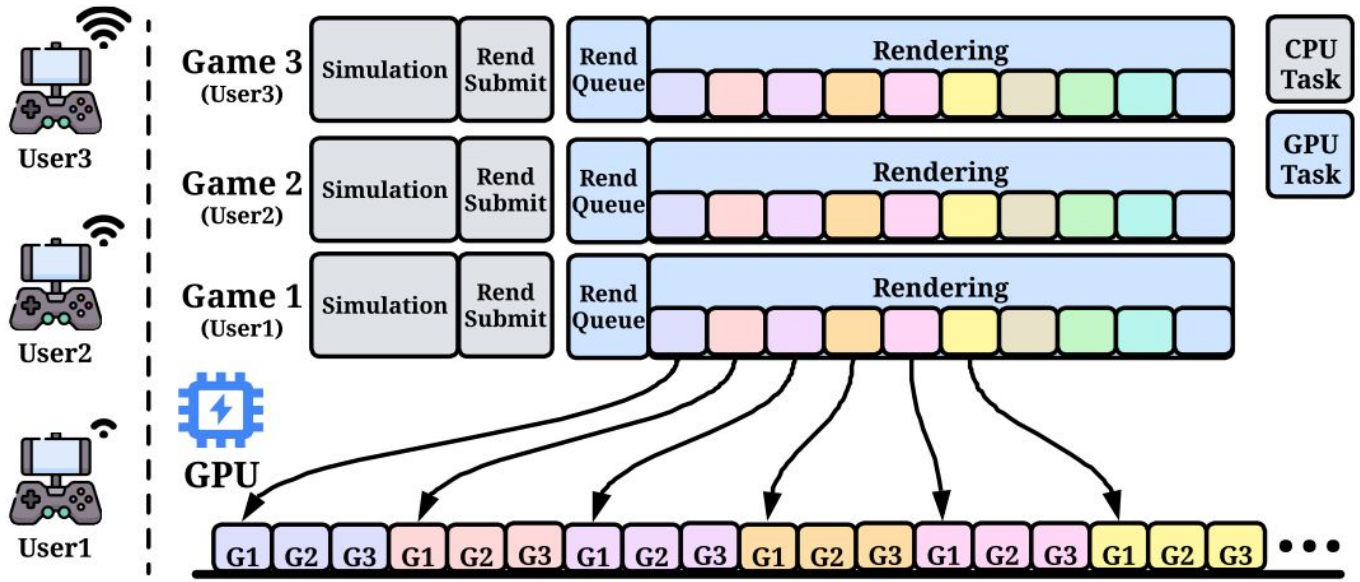
- GPU Resource Sharing in Cloud Gaming
  - Multiple concurrent users → Increased frame rendering latency → FPS drop





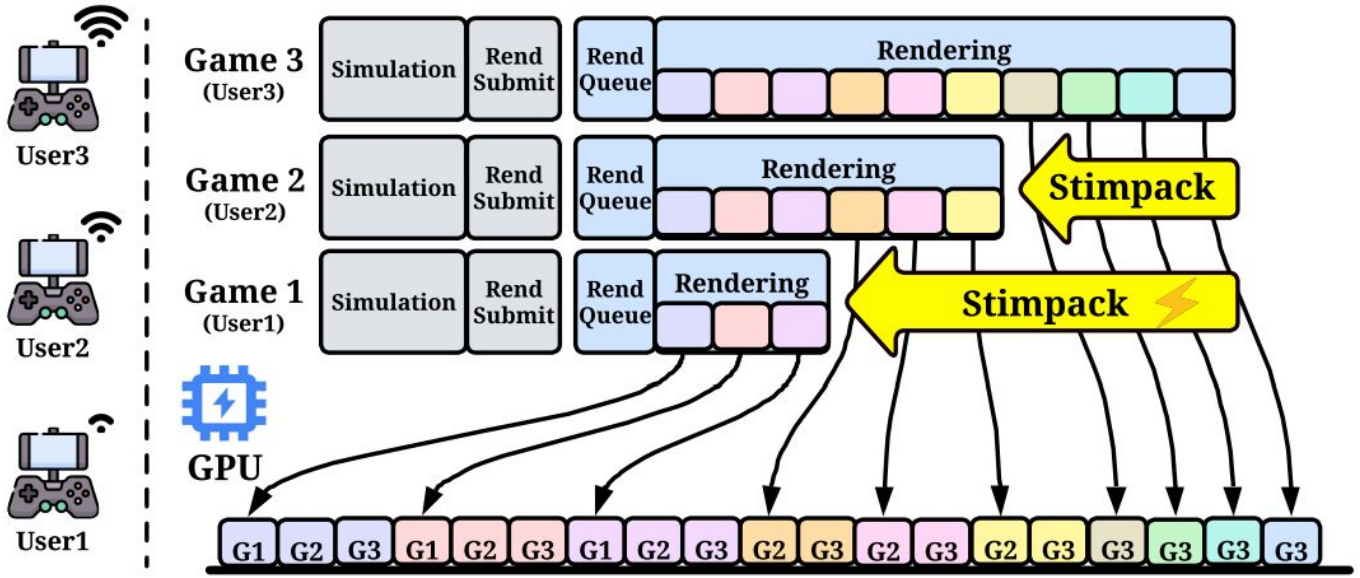
Stimpack: High-level Approach Overview

- Adaptive Workload Optimization
  - Optimizes RQ balancing user-side frame quality and server-side cost
  - Quantifies resource efficiency with compression-aware, use-side frame quality prediction



Stimpack: High-level Approach Overview

- Adaptive Workload Optimization
  - Optimizes RQ balancing user-side frame quality and server-side cost
  - Quantifies resource efficiency with compression-aware, use-side frame quality prediction



Stimpack: High-level Approach Overview

- Adaptive Workload Optimization
  - Optimizes RQ balancing user-side frame quality and server-side cost
  - Quantifies resource efficiency with compression-aware, use-side frame quality prediction
- Optimization Objectives
  - Playability: maintain playable FPS
  - Visual Quality: preserve visual quality
  - System Utility: maximize system-wide resource efficiency via prioritized RQ adjustments

Stimpack: High-level Approach Overview

- Adaptive Workload Optimization
  - Optimizes RQ balancing user-side frame quality and server-side cost
  - Quantifies resource efficiency with compression-aware, use-side frame quality prediction
- Optimization Objectives
  - Playability: maintain playable FPS
  - Visual Quality: preserve visual quality
  - System Utility: maximize system-wide resource efficiency via prioritized RQ adjustments

$$\max_{RQ} \sum_{u} Score_u,$$

$$\text{s.t. } FPS_{thresh} \leq FPS_u \leq FPS_{upper} \quad \forall u$$

$$RQ_{min} \leq RQ_u \leq RQ_{max} \quad \forall u$$

$$Score_u = \alpha\,(VQ\_Score_u) + (1-\alpha)\,(FPS\_Score_u),$$

$$\text{where} \quad 0 \leq \alpha \leq 1$$

Stimpack: High-level Approach Overview

- Adaptive Workload Optimization
  - Optimizes RQ balancing user-side frame quality and server-side cost
  - Quantifies resource efficiency with compression-aware, use-side frame quality prediction
- Optimization Objectives
  - Playability: maintain playable FPS
  - Visual Quality: preserve visual quality
  - System Utility: maximize system-wide resource efficiency via prioritized RQ adjustments

Efficiency score balancing server-side rendering cost and user-side quality gain

$$\max_{RQ} \sum_{u} Score_u,$$

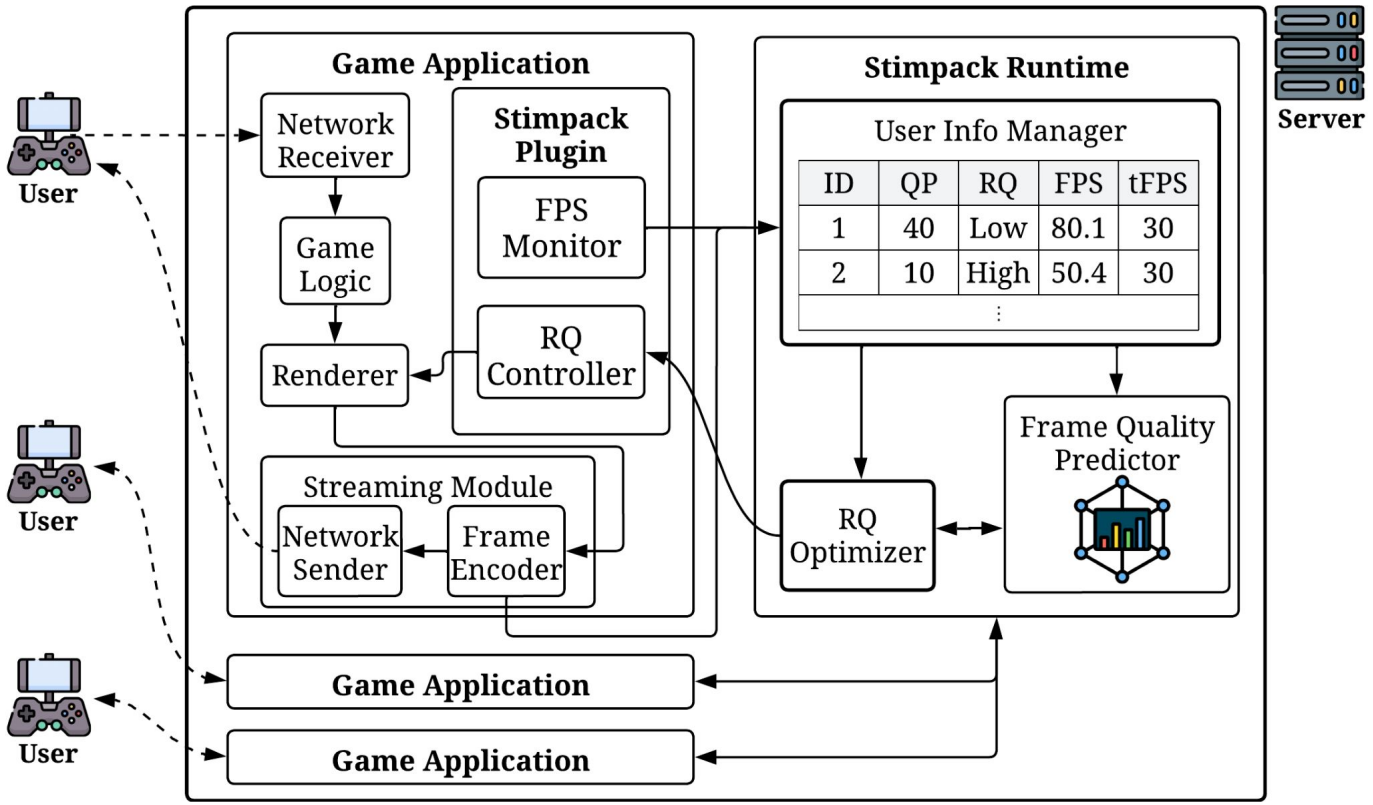
$$\text{s.t. } FPS_{thresh} \leq FPS_u \leq FPS_{upper} \quad \forall u$$

$$RQ_{min} \leq RQ_u \leq RQ_{max} \quad \forall u$$

$$Score_u = \alpha\,(VQ\_Score_u) + (1-\alpha)\,(FPS\_Score_u),$$

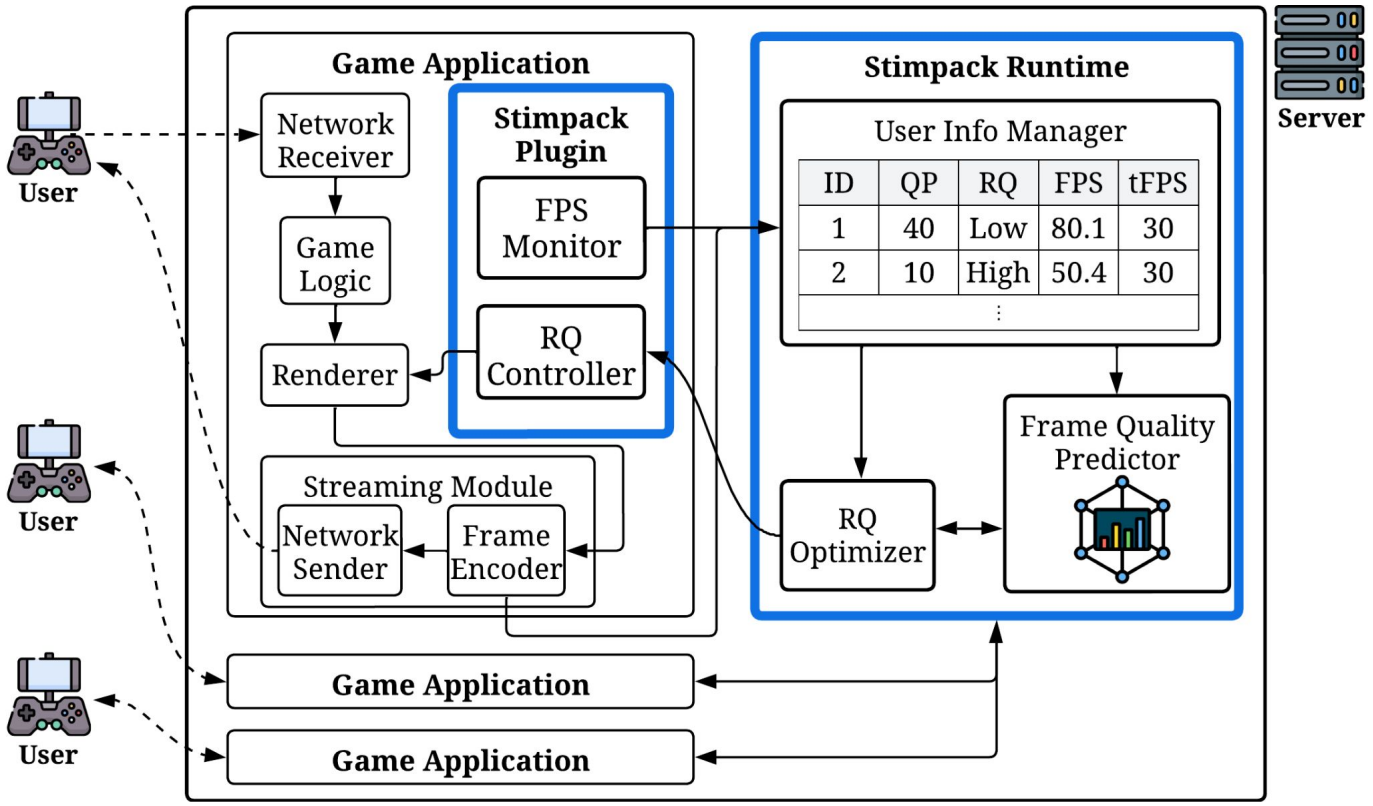
$$\text{where} \quad 0 \leq \alpha \leq 1$$

Stimpack: System Architecture



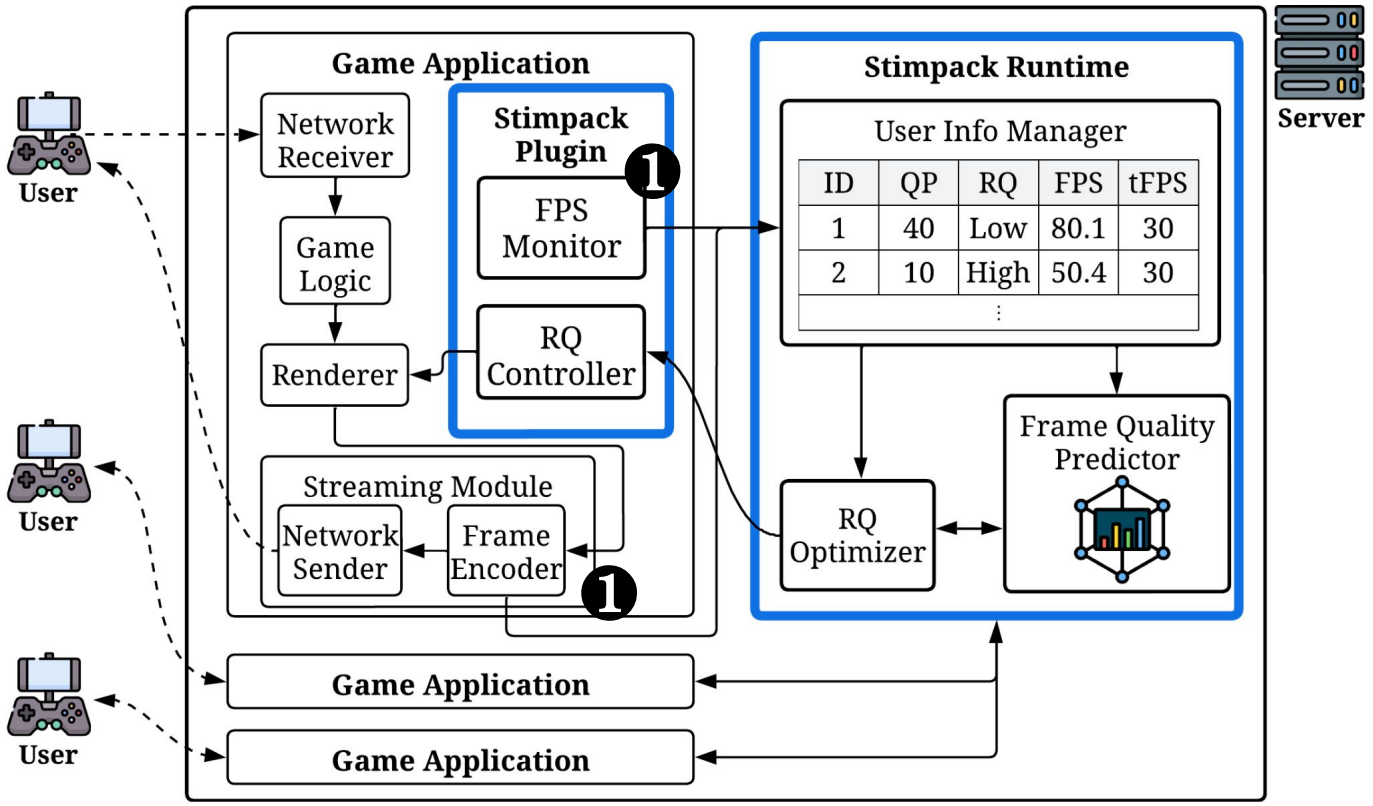
Stimpack: System Architecture

Plugin: Integrated into each game monitoring FPS and controlling RQ



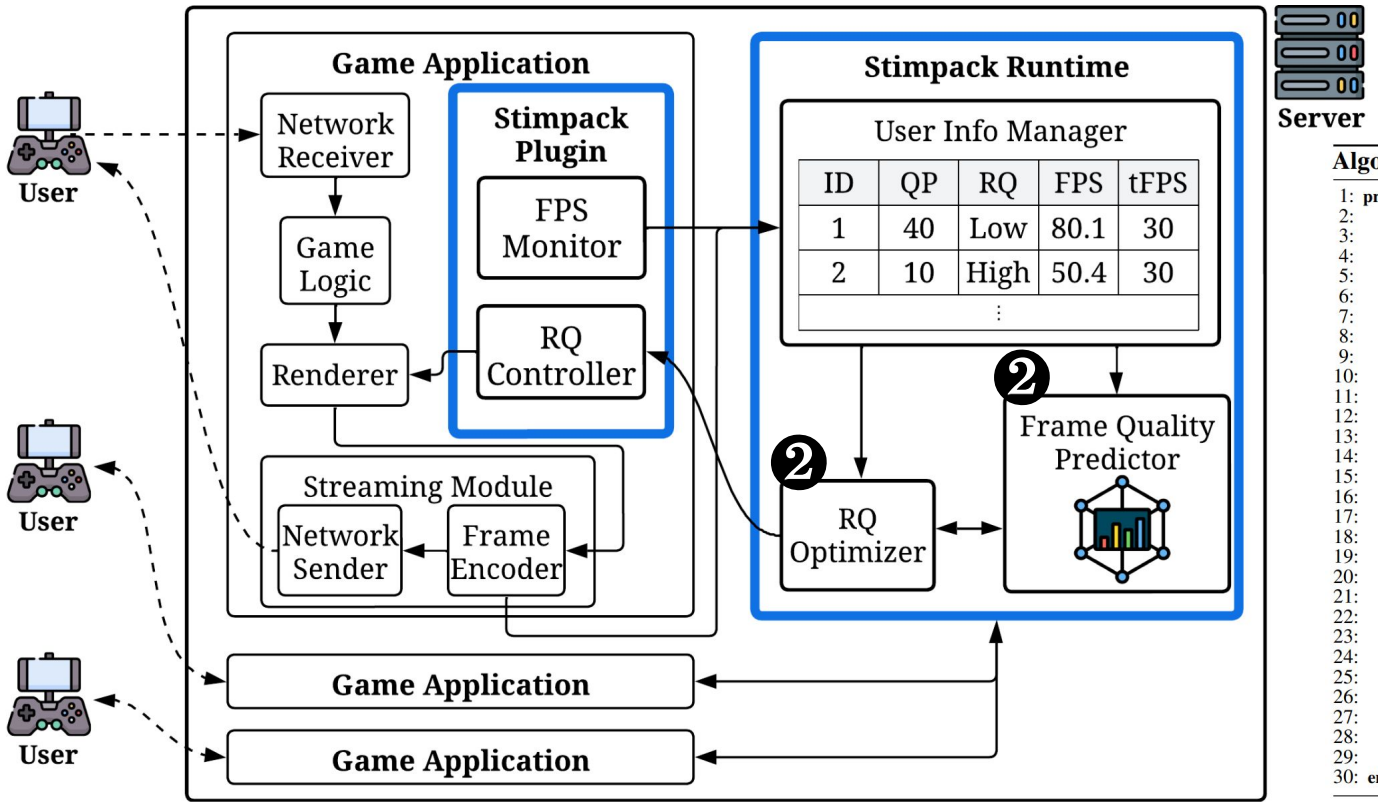
Runtime: An engine running optimization process and coordinating RQ adjustment for multiple users

Stimpack: Operational Workflow



❶ Monitoring: Captured FPS and compression parameter (QP) are reported to the Runtime.

Stimpack: Operational Workflow

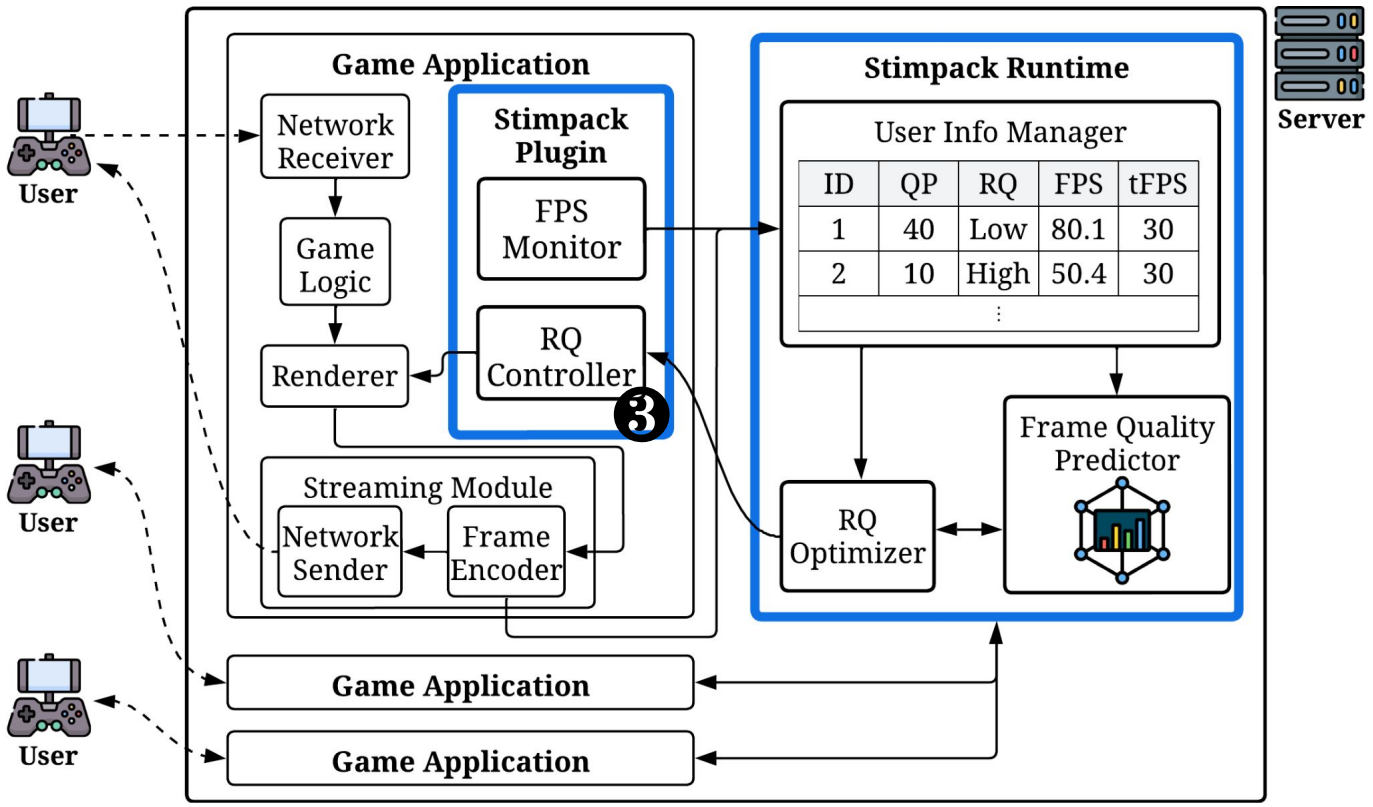


Algorithm 1 The RQ optimization process of Stimpack

```
procedure OPTIMIZE_RQ(user_table, N = current round)
    if N < Backoff Round then
        return
    end if
    demote_candidates ← users (RQ_min < RQ_u ≤ RQ_max)
    promote_candidates ← users (RQ_min ≤ RQ_u < RQ_max)
    if users (FPS_u < FPS_thresh) exist in user_table then
        /* RQ Demotion */
        for candidate in demote_candidates do
            Calculate Score_u (Eq. 2)
        end for
        demote_user ← candidate of minimum Score_u
        Demote_RQ(demote_user)
        adjusted_user ← demote_user
    else
        /* RQ Promotion */
        if candidates (FPS_u < FPS_thresh + FPS_buffer) exist then
            return
        end if
        for candidate in promote_candidates do
            Calculate Score_u (Eq. 2)
        end for
        promote_user ← candidate of maximum Score_u
        Promote_RQ(promote_user)
        adjusted_user ← promote_user
    end if
    if Is_RQ_oscillating(adjusted_user) then
        Backoff Round ← Get_backoff_round(N)
    end if
end procedure
```

❷ Decision Making
- Calculates efficiency scores using FPS and predicted user-side frame quality.
- Runs the optimization process to maximize the system-wide efficiency.

Stimpack: Operational Workflow



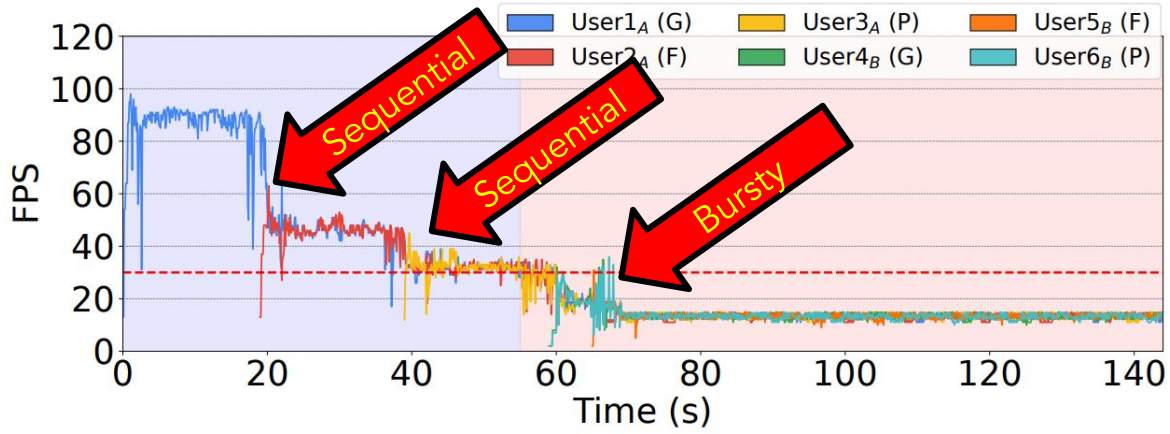
❸ Execution: commands are sent back to the plugin controller, adjusting RQ.

Evaluation Setup

- Testbed Environment
  - Ryzen 9 7950X, 64 GB RAM, RTX 4090 (24GB VRAM)
- Two Sample Games powered by Unreal Engine 5
  - Village Shooter: moderate-load, cartoon-style graphics
  - Mountain Hiker: intense-load, realistic and complex graphics

Evaluation Setup

- Testbed Environment
  - Ryzen 9 7950X, 64 GB RAM, RTX 4090 (24GB VRAM)
- Two Sample Games powered by Unreal Engine 5
  - Village Shooter: moderate-load, cartoon-style graphics
  - Mountain Hiker: intense-load, realistic and complex graphics
- Evaluation Scenarios
  - Arrival Patterns of 6 users: Gradual arrival (sequential) & Bursty arrival (simultaneous)
  - Network Conditions: Static/Dynamic QP traces

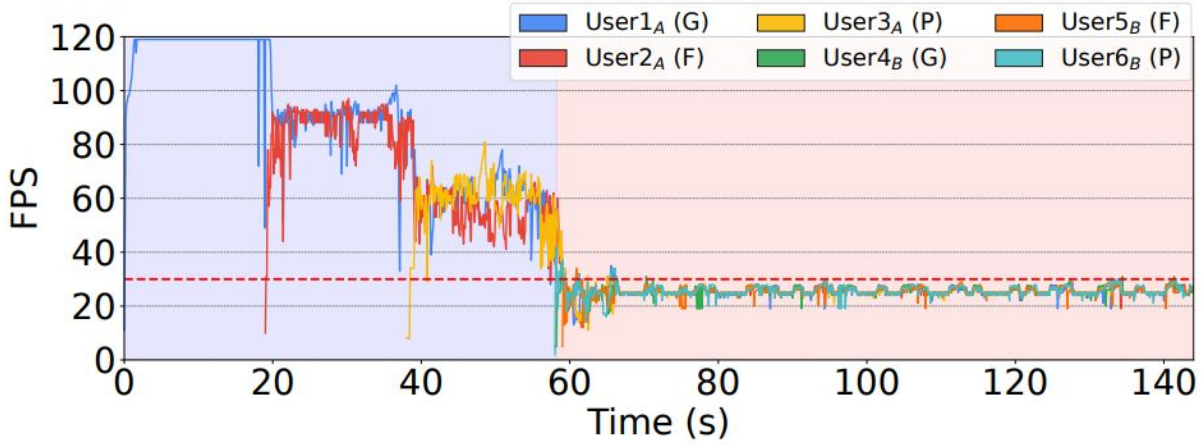


Performance Evaluation: Scalability & FPS Stability

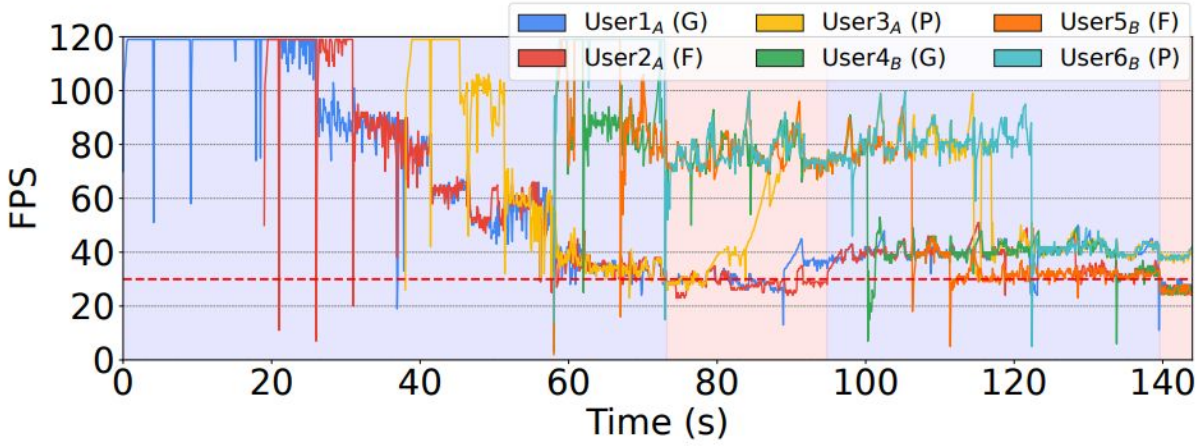
- Experiment Goal
  - Measure the maximum capacity with same resource footprint.
  - Compare Stimpack against a naive baseline (Fixed Very High RQ).
- The Baseline Approach
  - Existing service policies: Assigns maximum RQ to each user by default, not assuming GPU sharing.
  - Serves as a proxy to evaluate how highest-only approach scales under GPU sharing.

Performance Evaluation: Scalability & FPS Stability

- Key Findings: Preventing FPS Collapse
  - Baseline: GPU saturation leads to FPS drops below the playable FPS (30).
  - Stimpack: Maintains playable FPS by adaptively shedding workloads.



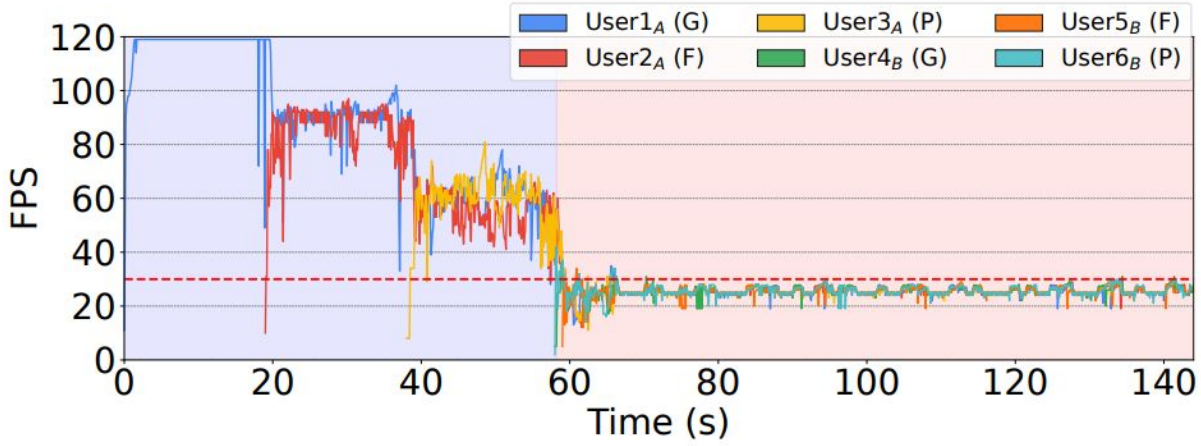
(a) The highest-only case: all users with `Very High` RQ



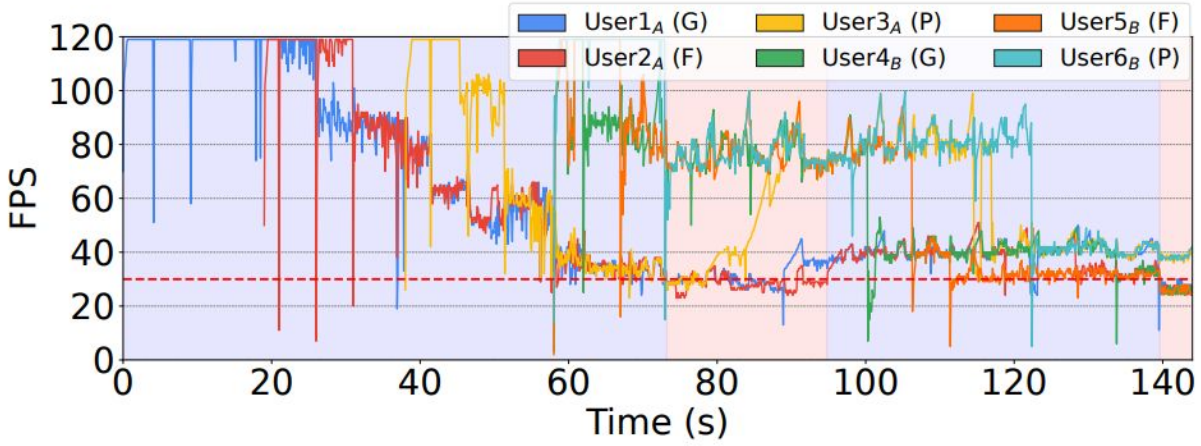
(b) Stimpack: users with adaptive RQ with its optimization

Performance Evaluation: Scalability & FPS Stability

- Key Findings: Preventing FPS Collapse
  - Baseline: GPU saturation leads to FPS drops below the playable FPS (30).
  - Stimpack: Maintains playable FPS by adaptively shedding workloads.
- Capacity Gains
  - Village Shooter (moderate): users with playable FPS 5→6 (+20%)



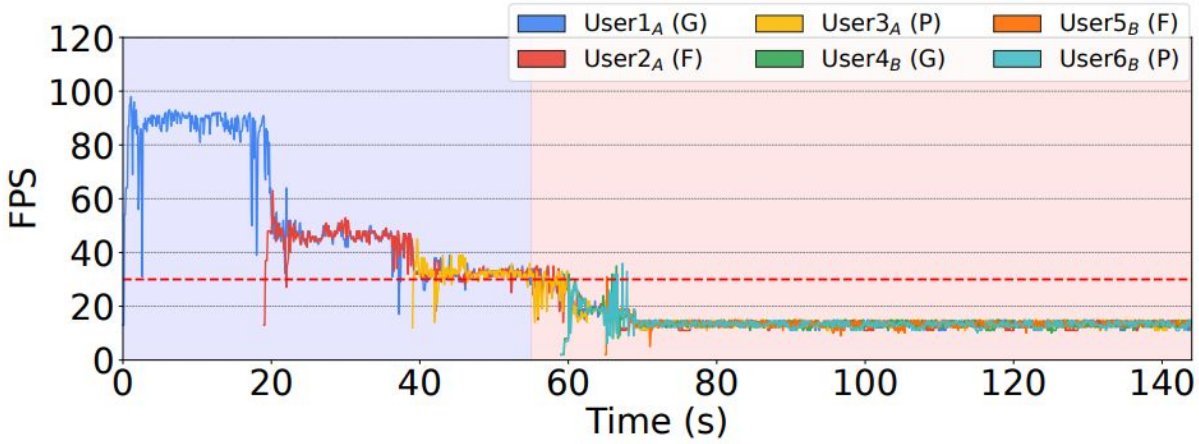
(a) The highest-only case: all users with `Very High` RQ



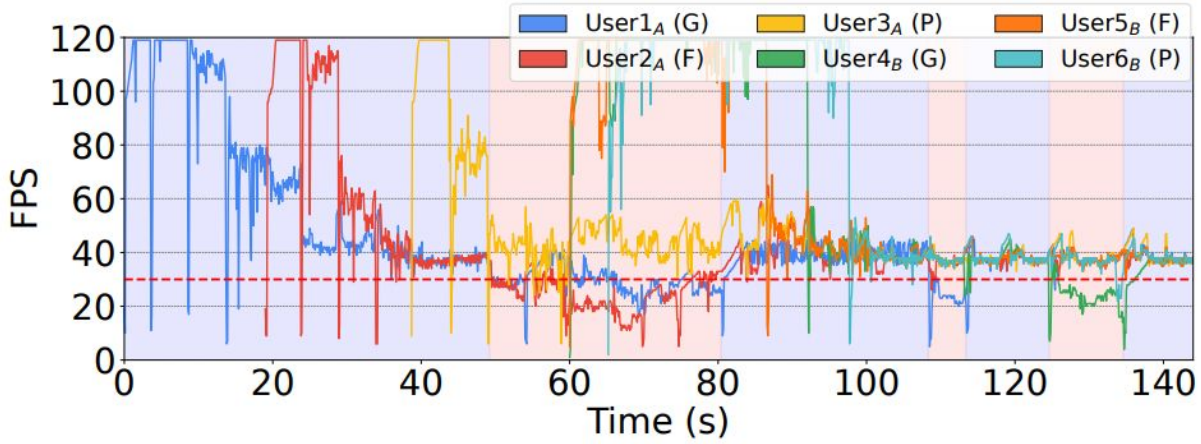
(b) Stimpack: users with adaptive RQ with its optimization

Performance Evaluation: Scalability & FPS Stability

- Key Findings: Preventing FPS Collapse
  - Baseline: GPU saturation leads to FPS drops below the playable FPS (30).
  - Stimpack: Maintains playable FPS by adaptively shedding workloads.
- Capacity Gains
  - Village Shooter (moderate): users with playable FPS 5→6 (+20%)
  - Mountain Hiker (intense): users with playable FPS 3→6 (+100%)



(a) The highest-only case: all users with `Very High` RQ



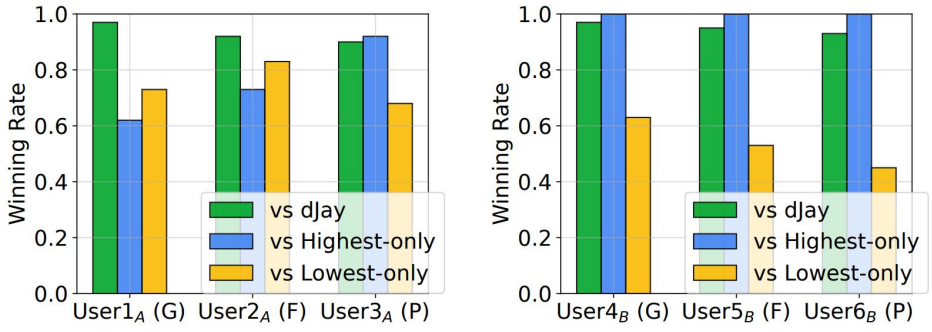
(b) Stimpack: users with adaptive RQ with its optimization

Gaming Experience Evaluation: User Study

- Not only for FPS, visual quality affects play's experience.
- Methodology
  - 30 participants (Blind A/B testing) comparing Stimpack vs. Baselines.
  - Metric: Winning Rate (Includes Wins + 0.5*Draws)

Gaming Experience Evaluation: User Study

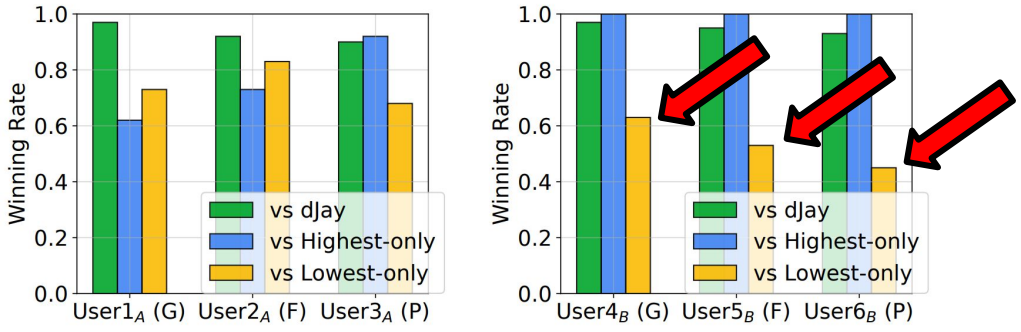
- Baselines
  - Highest-only: highest RQ for all users (compromising FPS for visual quality)
  - Lowest-only: lowest RQ for all users (compromising visual quality for FPS)
  - dJay: adjusts RQ for adapting server loads, not considering user's network condition.



(a) Village Shooter (b) Mountain Hiker

Gaming Experience Evaluation: User Study

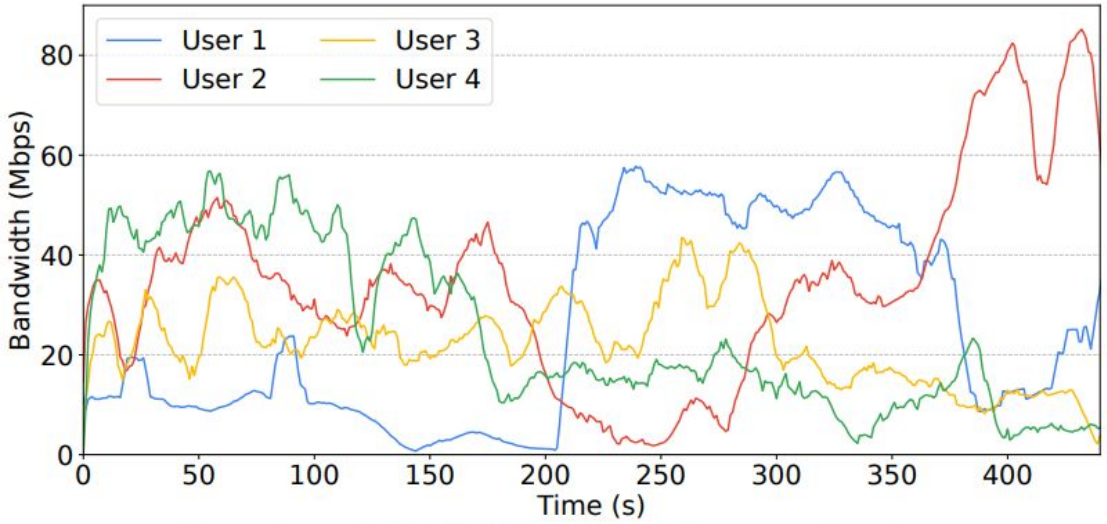
- Results Analysis
  - Stimpack outperformed in most cases, confirming its effectiveness in balancing gaming experience (FPS and VQ).
  - Subjective Preference in Gaming Experience
    - Some users preferred smoothness even with lower visual quality to moderate visual quality and smoothness.



(a) Village Shooter (b) Mountain Hiker

Real-world Adaptability: Dynamic Network Conditions

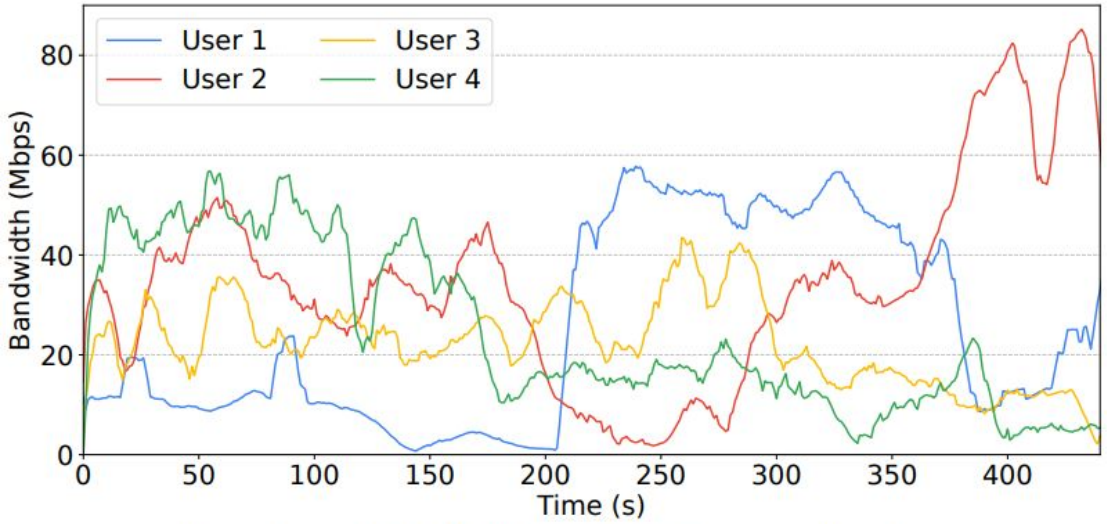
- Experimental Setup
  - Applied real-world 4G/LTE bandwidth traces to 4 concurrent users



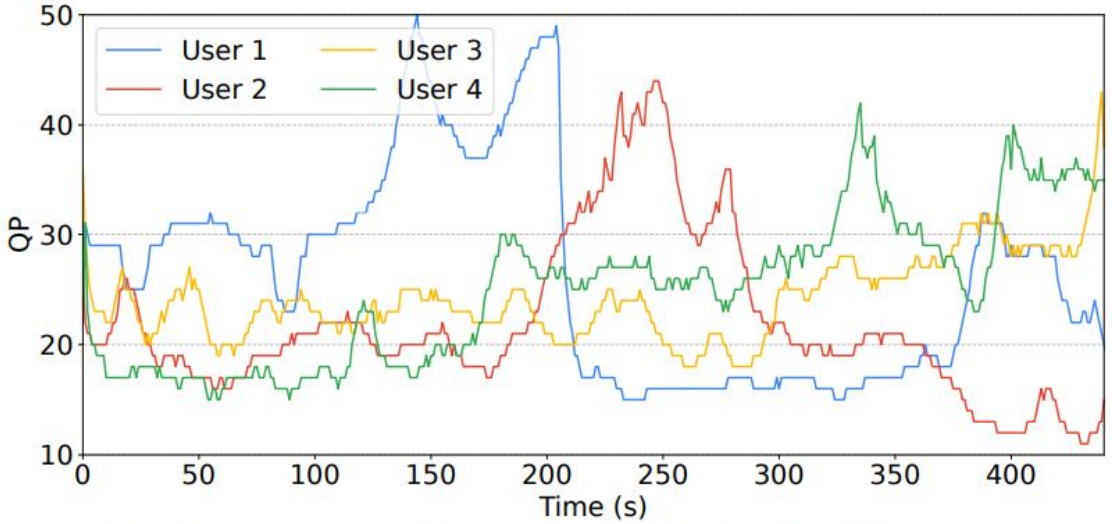
(a) The selected 4G/LTE bandwidth traces for 4 users

Real-world Adaptability: Dynamic Network Conditions

- Experimental Setup
  - Applied real-world 4G/LTE bandwidth traces to 4 concurrent users
  - Simulated continuous network fluctuations to validate Stimpack's effectiveness



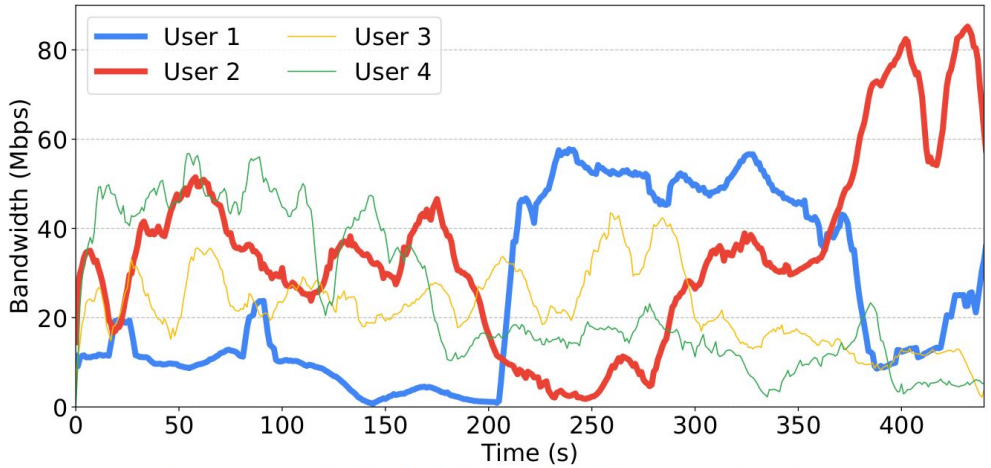
(a) The selected 4G/LTE bandwidth traces for 4 users



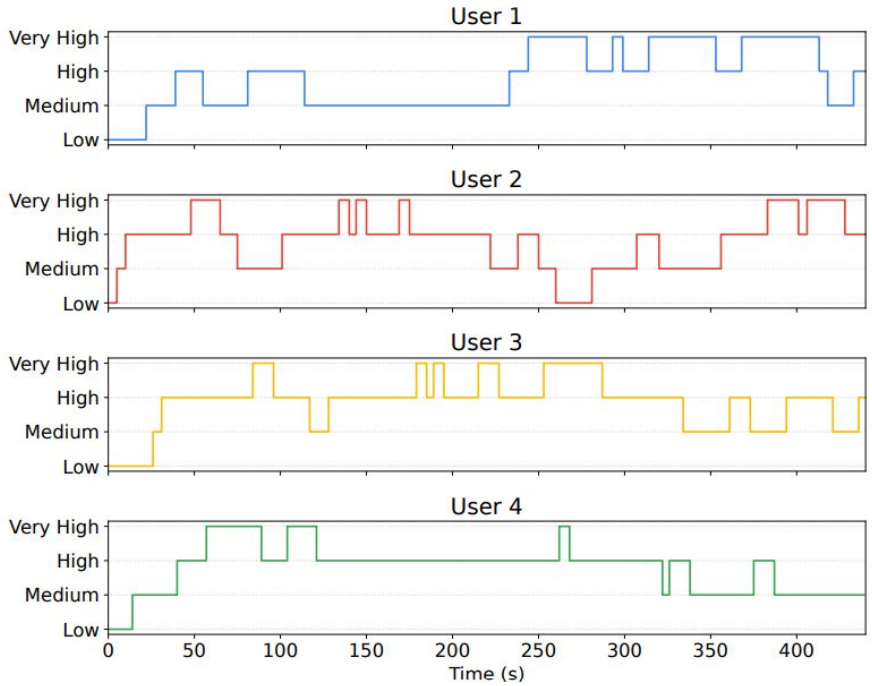
(b) The generated QP traces from the bandwidth traces

Real-world Adaptability: Dynamic Network Conditions

- Key Observation: Dynamic Prioritization
  - Network-Aware RQ: Stimpack's RQ adjustments closely mirror bandwidth fluctuations.

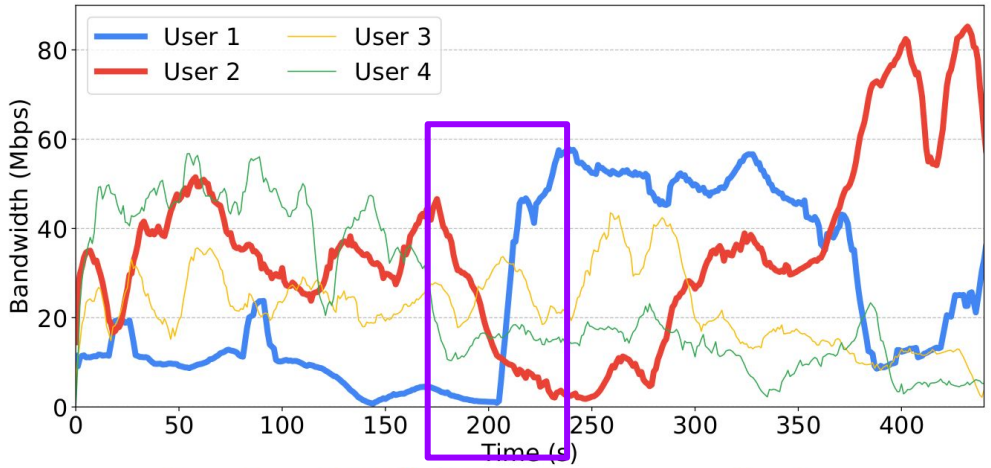


(a) The selected 4G/LTE bandwidth traces for 4 users

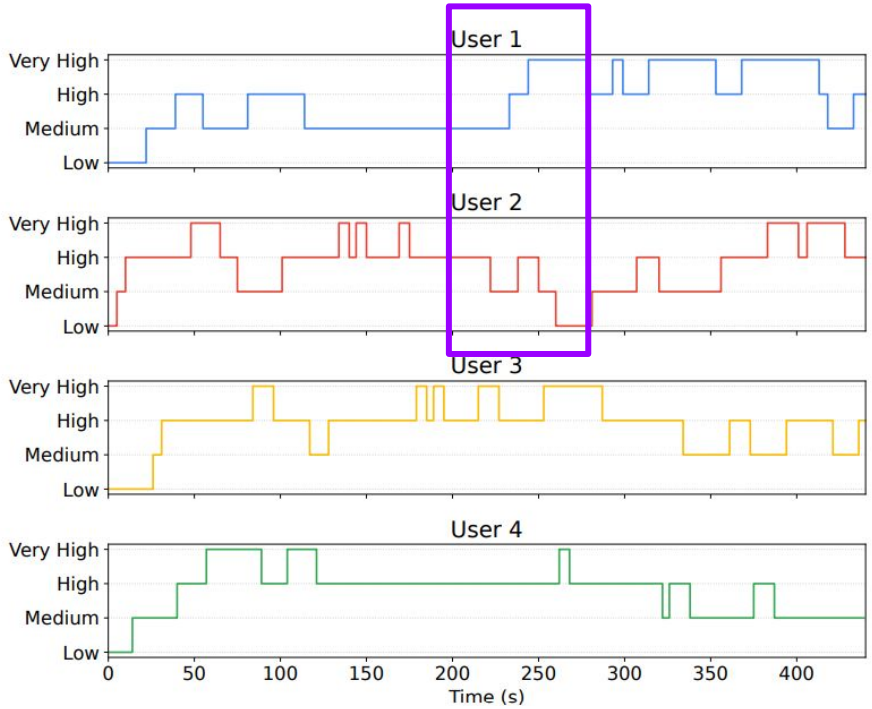


Real-world Adaptability: Dynamic Network Conditions

- Key Observation: Dynamic Prioritization
  - Network-Aware RQ: Stimpack's RQ adjustments closely mirror bandwidth fluctuations.
  - Resource Reallocation: At points of network crossover (e.g., 200s, 400s), Stimpack changes resource priority between users.

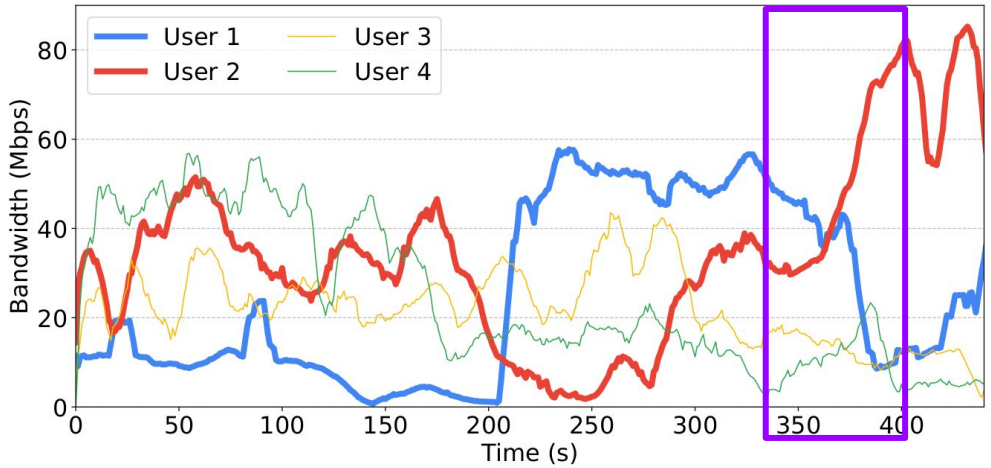


(a) The selected 4G/LTE bandwidth traces for 4 users

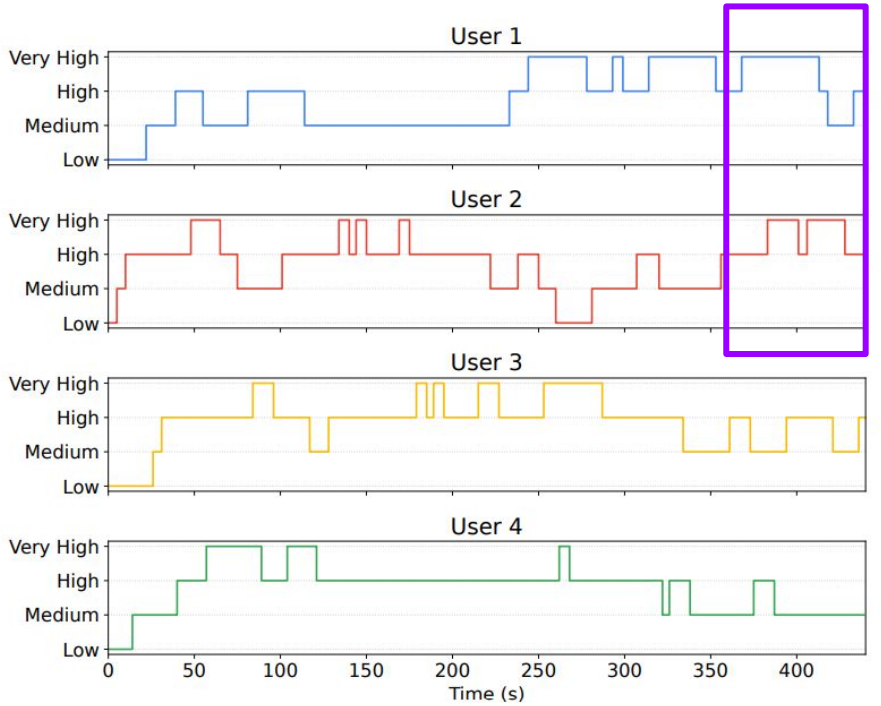


Real-world Adaptability: Dynamic Network Conditions

- Key Observation: Dynamic Prioritization
  - Network-Aware RQ: Stimpack's RQ adjustments closely mirror bandwidth fluctuations.
  - Resource Reallocation: At points of network crossover (e.g., 200s, 400s), Stimpack changes resource priority between users.



(a) The selected 4G/LTE bandwidth traces for 4 users



Summary

- Holistic System-level Optimization for Cloud Gaming
  - Successfully bridged the gap between server-side rendering costs and user-side quality.
  - Improved per-GPU scalability - 2x user capacity while maintaining high QoE.

Summary

- Holistic System-level Optimization for Cloud Gaming
  - Successfully bridged the gap between server-side rendering costs and user-side quality.
  - Improved per-GPU scalability - 2x user capacity while maintaining high QoE.
- Stimpack’s Core Insight
  - Perception-Aware Efficiency: Recognizing that rendering "more" does not always mean the user "see" more.
  - In networked systems, server-side effort can be ineffective due to downstream bottlenecks (e.g., content delivery with compression).

—

THANK YOU

Project Repo