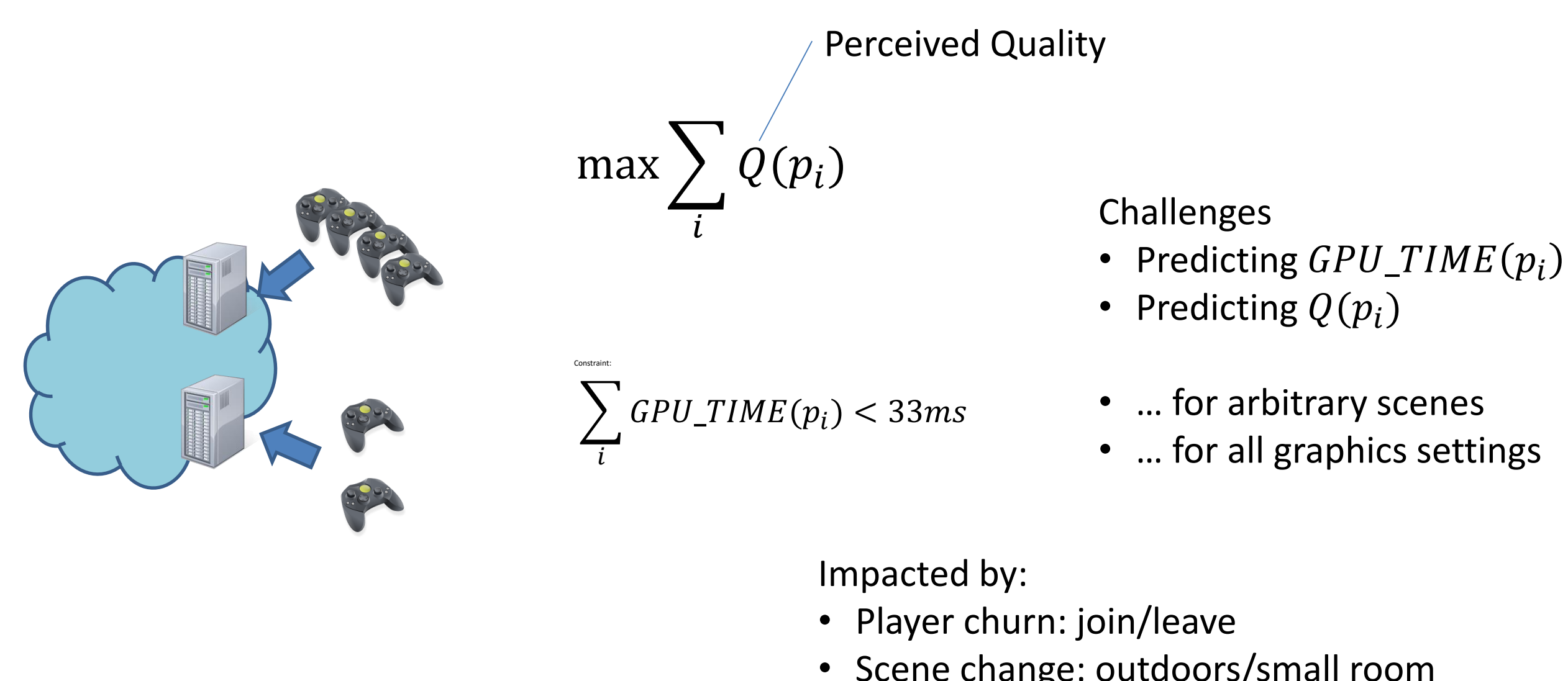# dJay: Enabling High-density Multi-tenancy for Cloud Gaming Servers with Dynamic Cost-Benefit GPU Load Balancing

Sergey Grizan, David Chu, Alec Wolman, Roger Wattenhofer
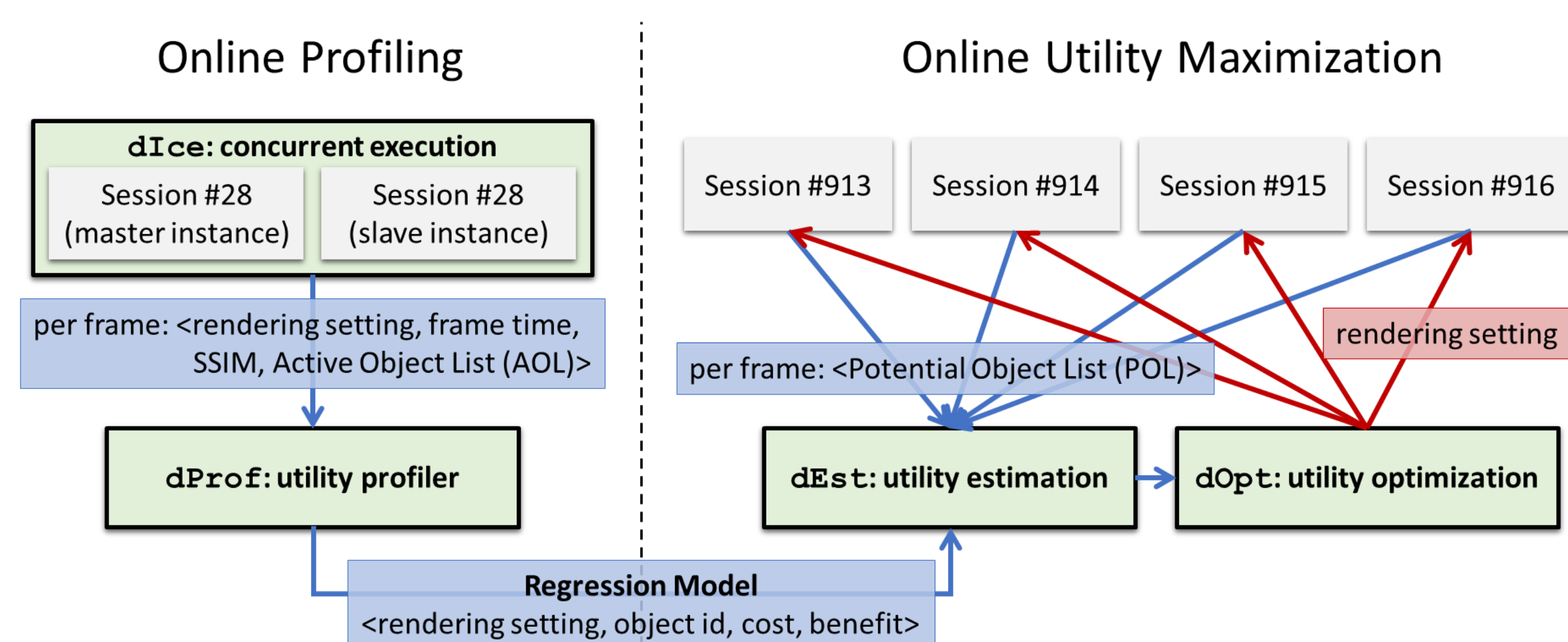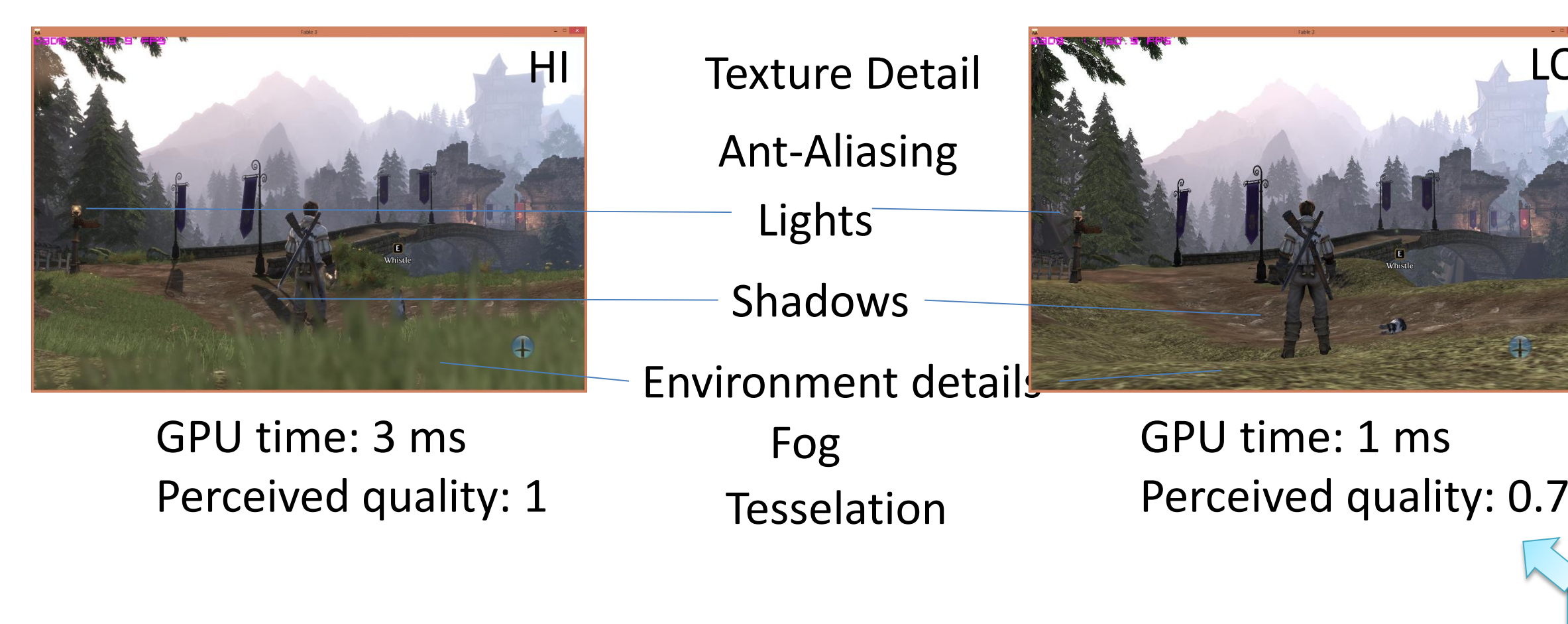
Density Optimization Problem

Perceived Quality

$$\max \sum_i Q(p_i)$$
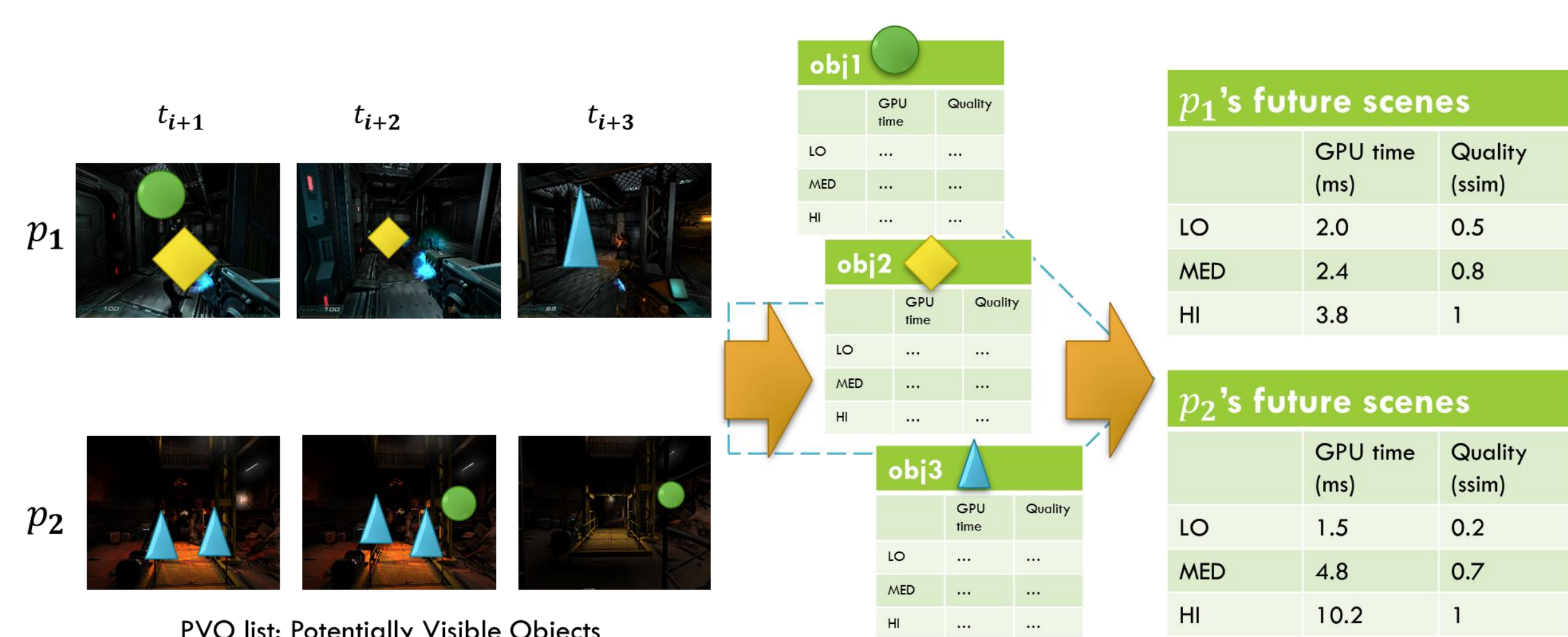
$$\sum_i GPU\_TIME(p_i) < 33ms$$

Challenges
- Predicting $GPU\_TIME(p_i)$
- Predicting $Q(p_i)$

- ... for arbitrary scenes
- ... for all graphics settings

Impacted by:
- Player churn: join/leave
- Scene change: outdoors/small room

Challenge

Much lower CapEx and OpEx

Christmas Morning: GPU oversubscription!!

System Architecture

Online Profiling

dIce: concurrent execution

| Session #28 (master instance) | Session #28 (slave instance) |

per frame: <rendering setting, frame time, SSIM, Active Object List (AOL)>

dProf: utility profiler

Regression Model <rendering setting, object id, cost, benefit>

Online Utility Maximization

| Session #913 | Session #914 | Session #915 | Session #916 |

rendering setting

per frame: <Potential Object List (POL)>

dEst: utility estimation → **dOpt: utility optimization**

Key Insight: Tunable Visual QoS

HI — Texture Detail, Ant-Aliasing, Lights, Shadows, Environment details, Fog, Tesselation — LO

GPU time: 3 ms
Perceived quality: 1

GPU time: 1 ms
Perceived quality: 0.7

PSNR, SSIM

Utility Estimating: Cost and Benefit

SOLVING FOR OPTIMAL SETTINGS

t_{i+1} t_{i+2} t_{i+3}

p_1 p_2

obj1

	GPU time	Quality
LO		
MED		
HI		

obj2

	GPU time	Quality
LO		
MED		
HI		

obj3

	GPU time	Quality
LO		
MED		
HI		

p_1's future scenes

	GPU time (ms)	Quality (ssim)
LO	2.0	0.2
MED	2.4	0.8
HI	3.8	1

p_2's future scenes

	GPU time (ms)	Quality (ssim)
LO	1.5	0.2
MED	4.8	0.7
HI	10.2	1

PVO list: Potentially Visible Objects

BUILDING OBJECT TABLES

t_{i+1} t_{i+2} t_{i+3}

p_1 LO

p_1 HI

quality clone

p_1's trace

	GPU time (ms)	Quality (ssim)
LO	2.0	0.5
HI	3.8	1

obj1

	GPU time	Quality
LO		
MED		
HI		

obj2

	GPU time	Quality
LO		
MED		
HI		

obj3

	GPU time	Quality
LO		
MED		
HI		
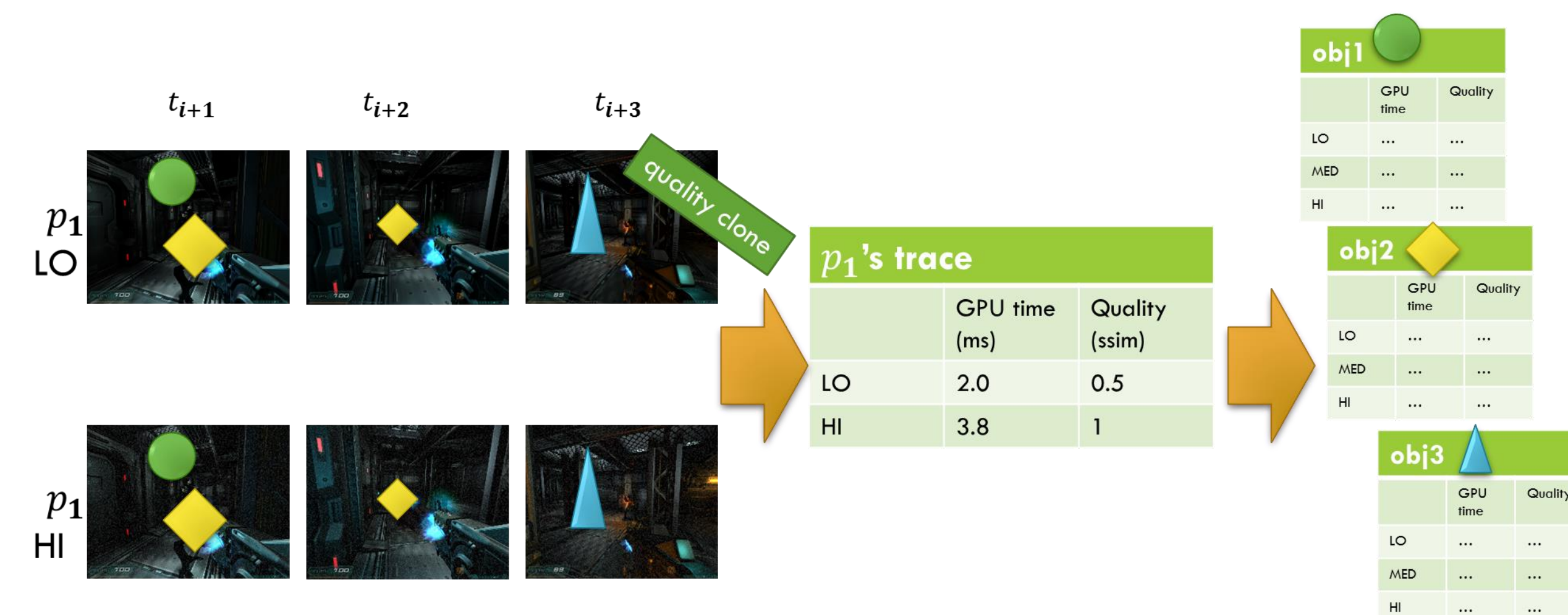
OBJECT TABLES BY CLUSTERING

Indistinct cluster, Insignificant object

Distinct cluster, Significant object

- HQ time   · LQ time
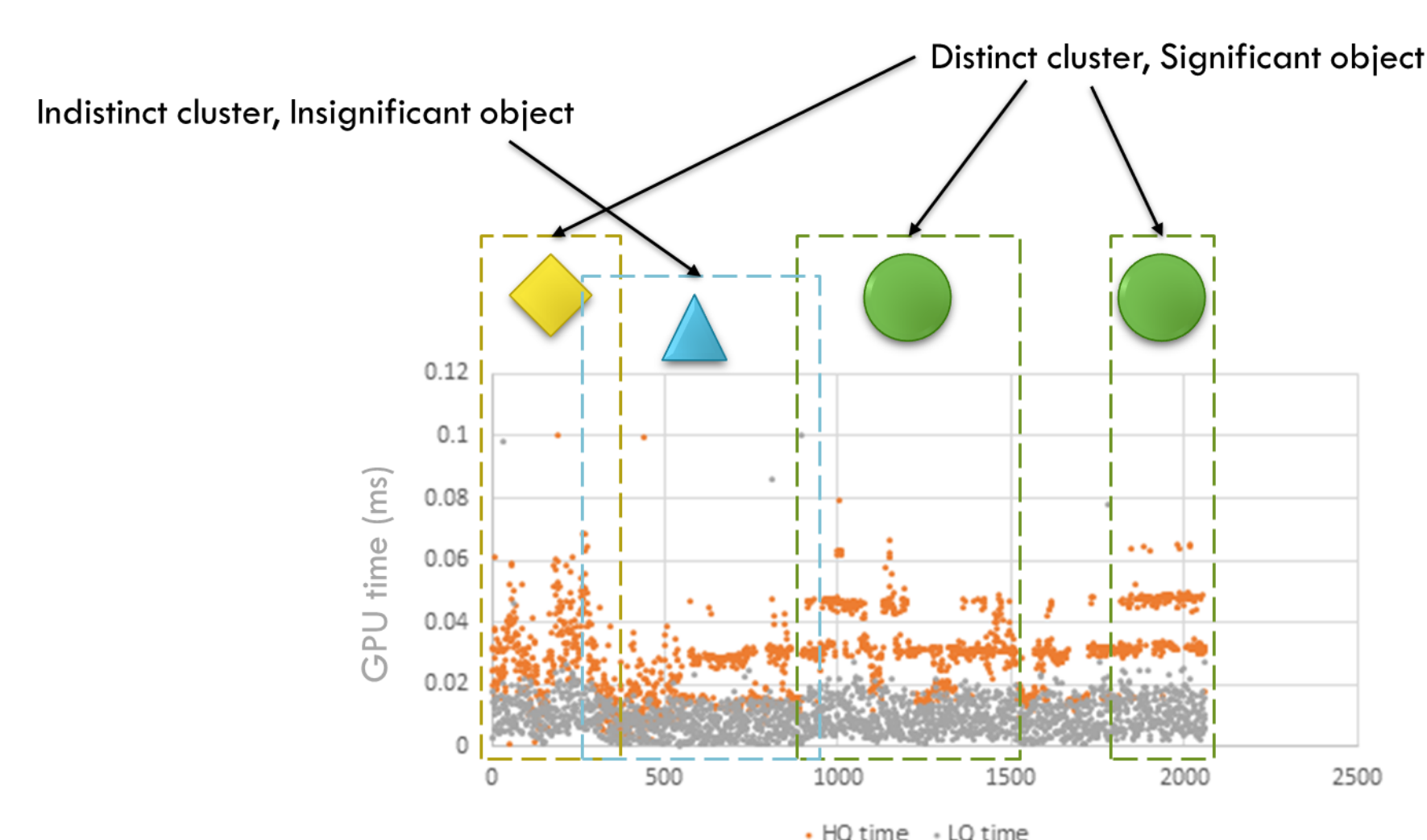
Results on Fable 3

Increasing multi-tenancy

SSIM score

All HI   All ME   All LO   dJay

1 player: 75 fps, 82 fps, 83 fps, 75 fps
2 players: 40 fps, 78 fps, 80 fps, 40 fps
3 players: 29 fps, 76 fps, 78 fps, 29 fps
4 players: 16 fps, 18 fps, 51 fps, 51 fps
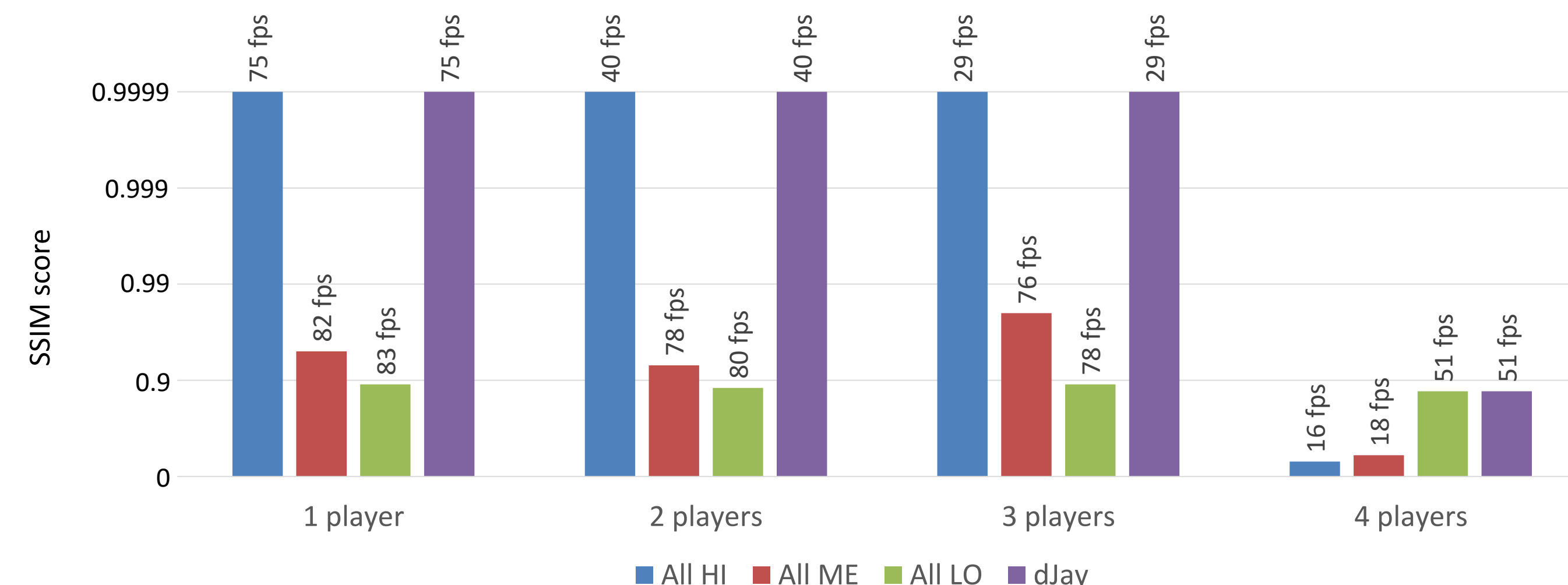
Increasing scene complexity

SSIM score

All HI   All ME   All LO   Auto

Simple scene: 30 fps, 49 fps, 83 fps, 30 fps
Complex scene: 17 fps, 20 fps, 51 fps, 40 fps

Responsiveness

critical level

Frame time, ms

Frame number

Player 1   Player 2