Hairpin: Rethinking Packet Loss Recovery in Edge-based Interactive Video Streaming

Zili Meng, *Xiao Kong, Jing Chen, Bo Wang✉, Mingwei Xu✉,*
Rui Han, Honghao Liu, Venkat Arun, Hongxin Hu, Xue Wei

清華大學
Tsinghua University

腾讯
Tencent

TEXAS
The University of Texas at Austin

University at Buffalo

Background
Interactive Video Streaming

Immersive interaction over the Internet is the future.

Ultra-low and consistent latency is the key factor of user's experience.

*Video from RED

Background
Interactive Video Streaming

Next-generation applications involve life-or-death decisions!

AR-assisted driving

Remote surgery

They all need continuous operations of up to **10+ hours**, where a single stall can be **fatal**!

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Motivation: Latency Variation

What does latency variation mean for us
when we say we want a latency of lower than xx msec?

A 0.3 second stall

*Video source: https://www.youtube.com/watch?v=hfySDsMW8BU

0.1% Stall rate

⇕

Such a 0.3 sec stall happens every 300 secs (5 min)

Motivation: Latency Variation

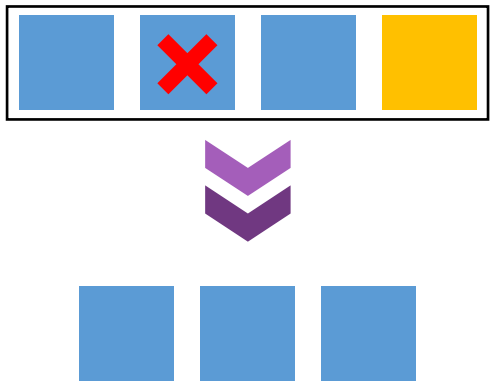
Taming the network latency variation for interactive video streaming.

$$latency_{tail} = (1 + RTX_{tail}) \cdot RTT_{tail}$$

- RTT_{tail}: Congestion Control Mechanism
  - Queueing delay, Propagation delay, etc.
  - Not the scope of this paper.
- RTX_{tail}: Loss Recovery Mechanism

Packet Loss Recovery

- Forward error correction (FEC)



FBRA [MMSys'14]

FracTal [MM'17]

RLAFEC [MMSys'22]

OptFEC [TIT'19]

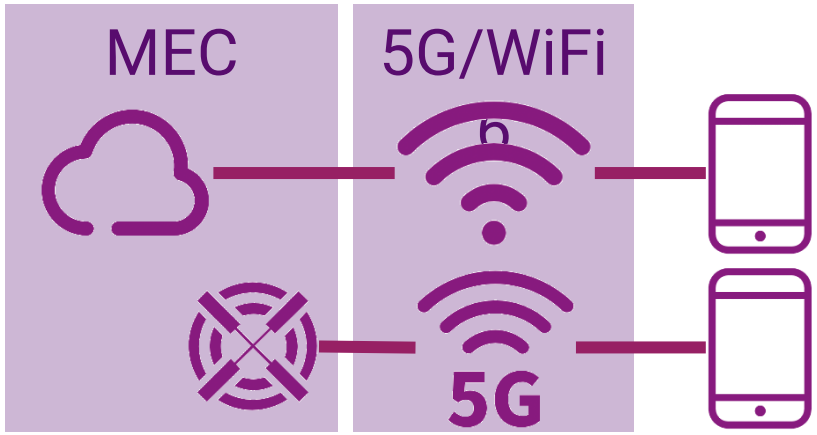
Tambur [NSDI'23]

FlexFEC [WebRTC]

Packet Loss Recovery

Now the latency goes down…

- Low network RTT (10-20ms) due to Multi-access Edge Computing (MEC), 5G/WiFi6.



This is not the ping latency…
Low latency even with load!

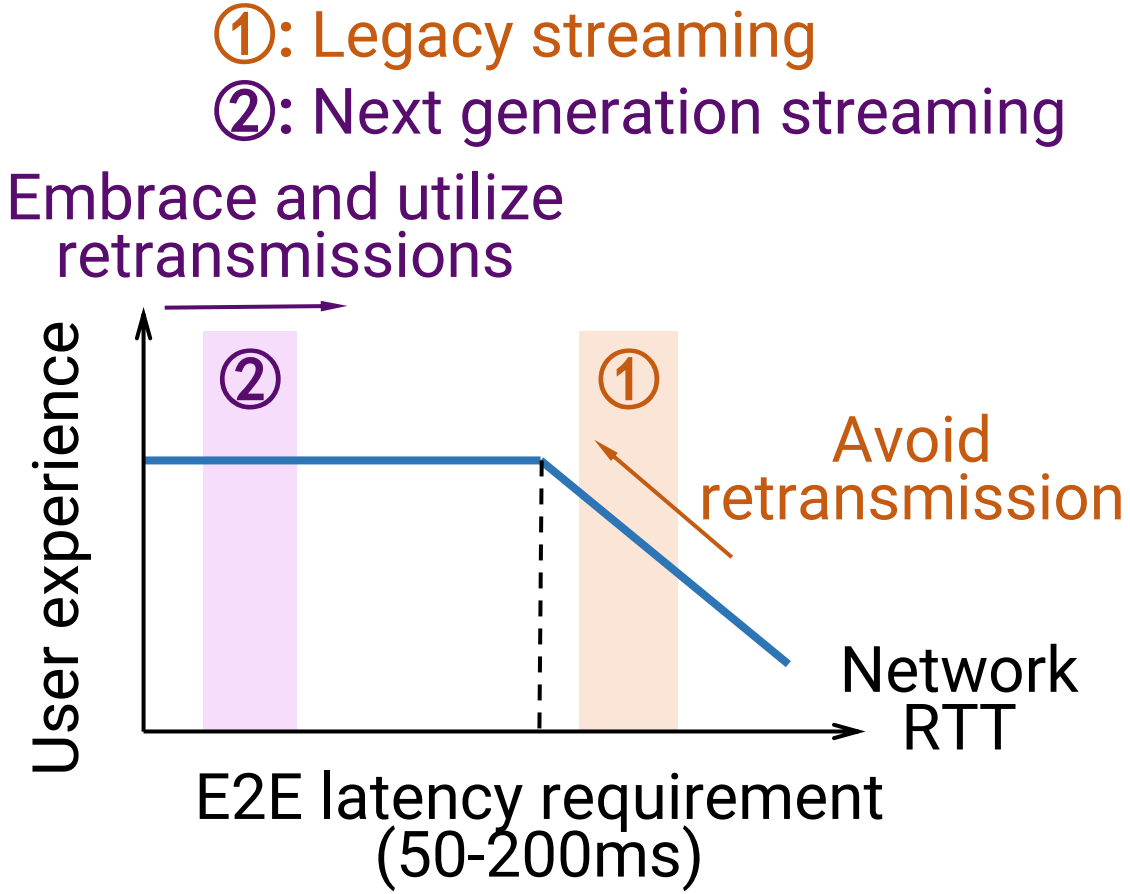
Capacity	Stadia	GeForce	Luna
15 Mb/s	16.0 (1.7)	16.8 (1.5)	17.2 (2.1)
25 Mb/s	16.6 (2.2)	16.8 (1.6)	17.0 (1.5)
35 Mb/s	17.1 (1.4)	18.2 (1.8)	16.4 (1.6)

RTT (ms) measured by Xu et al [IMC'22]

Packet Loss Recovery

Now the latency goes down…

- Low network RTT (10-20ms) due to Multi-access Edge Computing (MEC), 5G/WiFi6.
- Human's perception ability of interactive latency is bounded by 50-200ms.

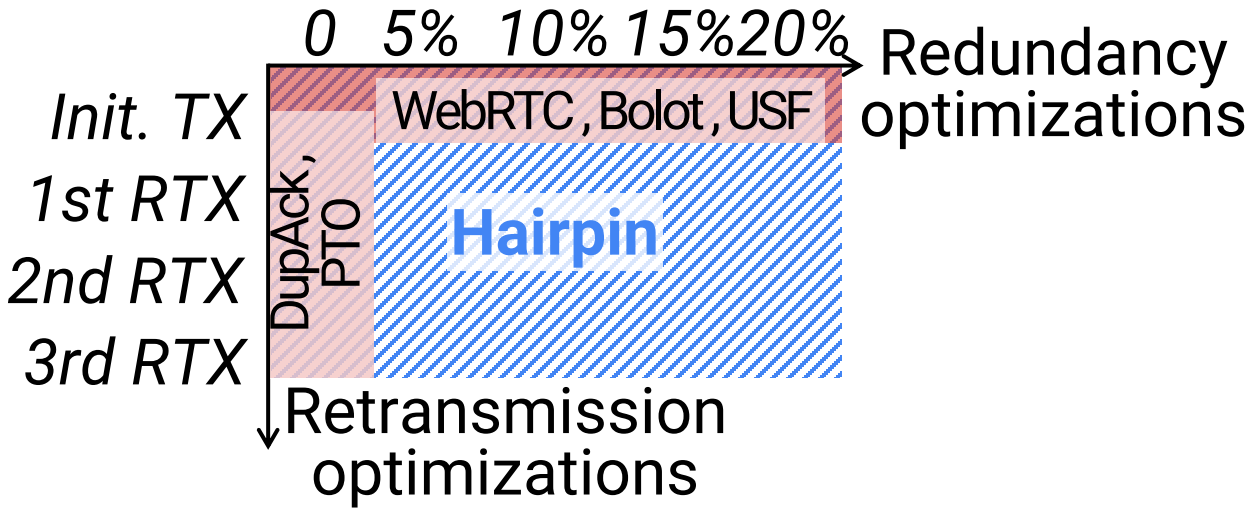


香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Packet Loss Recovery

Insight: co-optimize redundancy and retransmission.

- Network RTT < Human's perception ability



When the network is lossy…

Should I add 10% redundancy?
Should I rely on retransmissions?

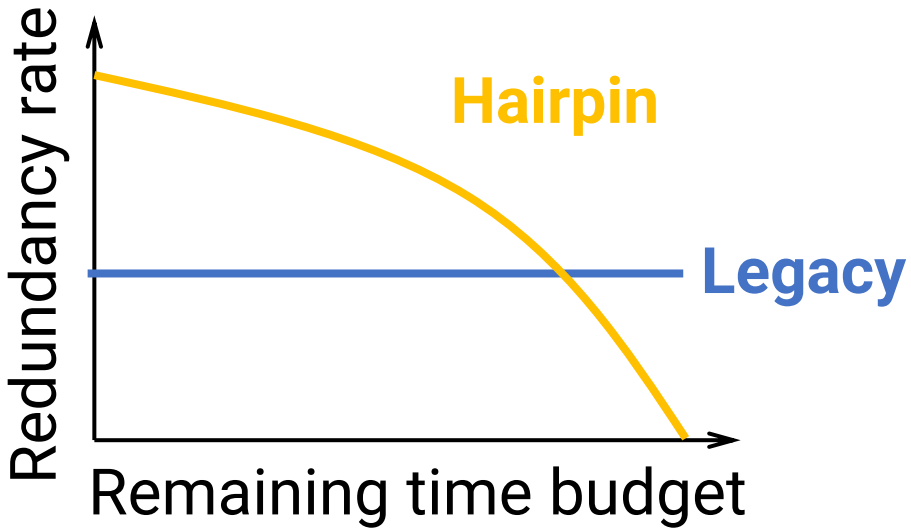
You can think further!
Differentiate retransmissions!

Packet Loss Recovery

Insight: co-optimize redundancy and retransmission.

Solution: Differentiating retransmission

- When there are many chances to transmit, do not add redundancy.
- When there are few chances to transmit, aggressively add redundancy.



Packet Loss Recovery

Insight: co-optimize redundancy and retransmission.

Solution: Differentiating retransmission

Suppose loss rate = 20%
(RTT: 20 ms; deadline: 50 ms)

Legacy

1st (20%)

2nd (20%) *All delivered*

Hairpin

1st (0%)

2nd (50%)

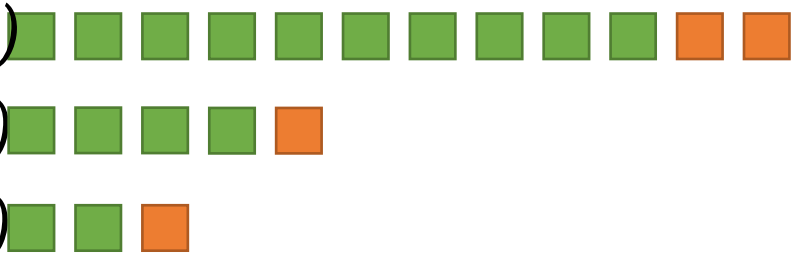
Loss rate rises to 40%

Legacy

1st (20%)

2nd (20%)

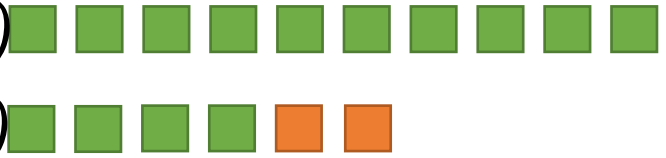
3rd (20%)



Hairpin

1st (0%)

2nd (50%)

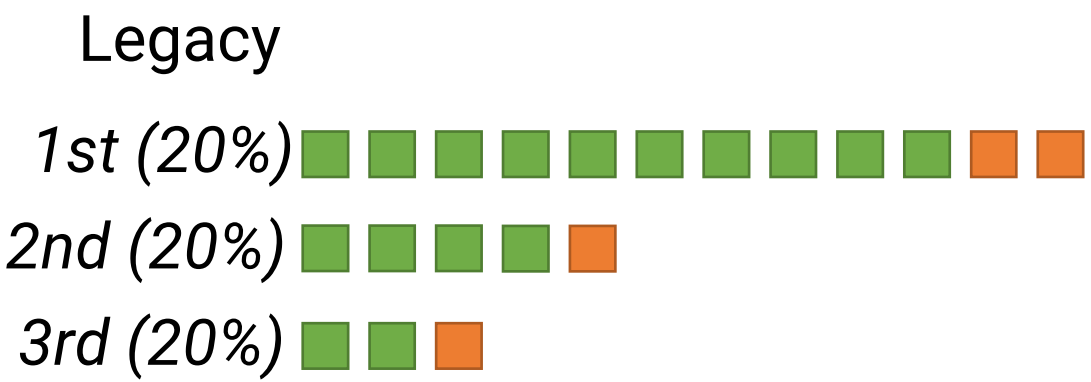


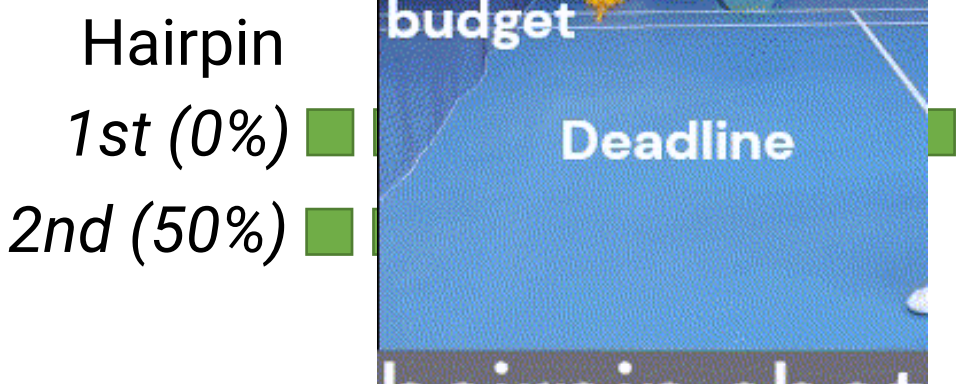
Packet Loss Recovery

Insight: co-optimize redundancy and retransmission.

Solution: Differentiating retransmission

- Benefit 1: Fewer deadline misses
  - Higher protection rate for retransmissions helps to ha losses.
- Benefit 2: Save bandwidth
  - *Protecting all packets by 10%* costs the same as *protec y 100%*





*Video from Bilibili.

Packet Loss Recovery

Insight: co-optimize redundancy and retransmission.

Challenges

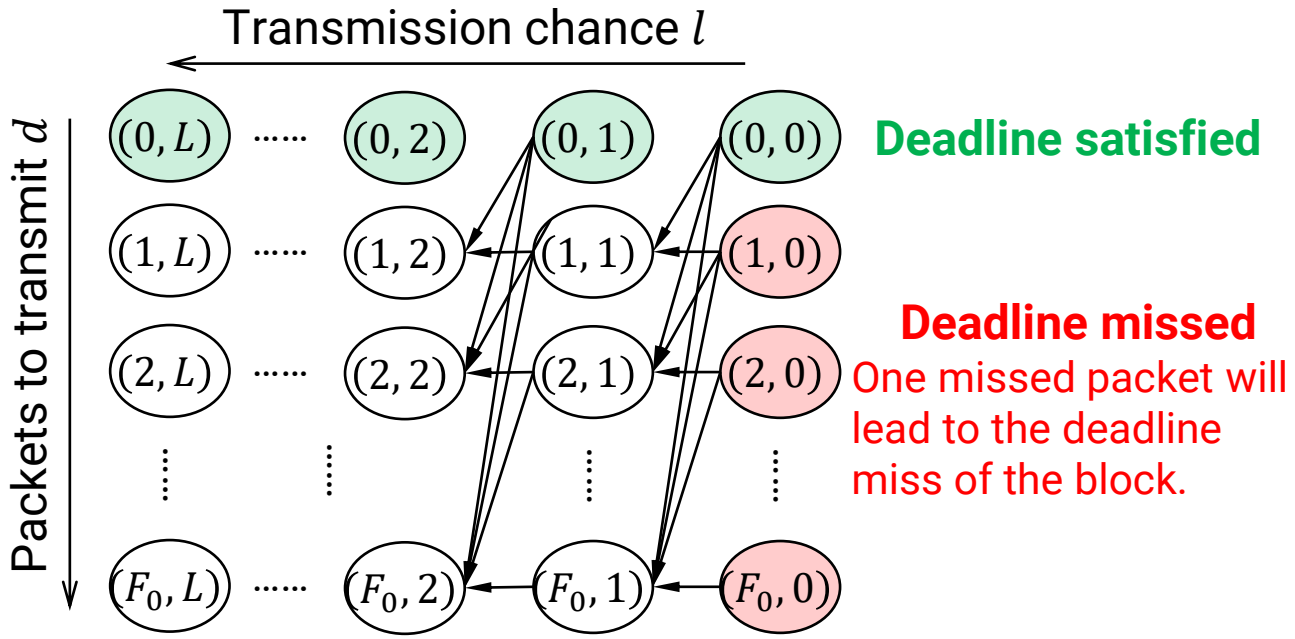
- Temporal dependency
  - One decision will affect the outcome of the next round.
- Spatial dependency
  - Redundancy rate and block size in each transmission are coupled.
- Convoluted goals
  - Deadline miss rate and bandwidth cost are non-trivial to estimate at tail.

Mathematical Formulation

Insight: co-optimize redundancy and retransmission.

Solution

- Formulate the redundancy-retransmission joint optimization with Markov Chains.

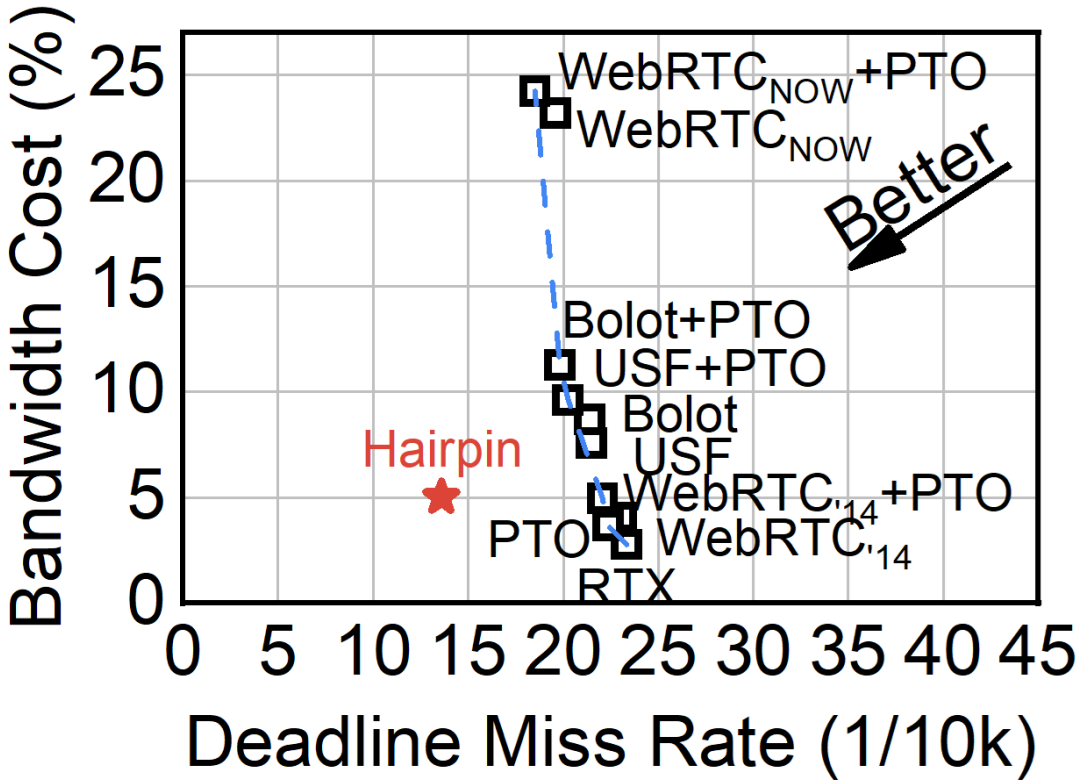


Evaluation

Insight: co-optimize redundancy and retransmission.

Evaluation

- NS-3 simulation
- Application in Zhuge [SIGCOMM'22]:
  - WebRTC (UDP) with GCC
- 3 sets of bandwidth traces:
  - WiFi, Ethernet, cellular
- 10 baselines
- Metrics
  - **Deadline miss rate**
  - **Bandwidth cost**

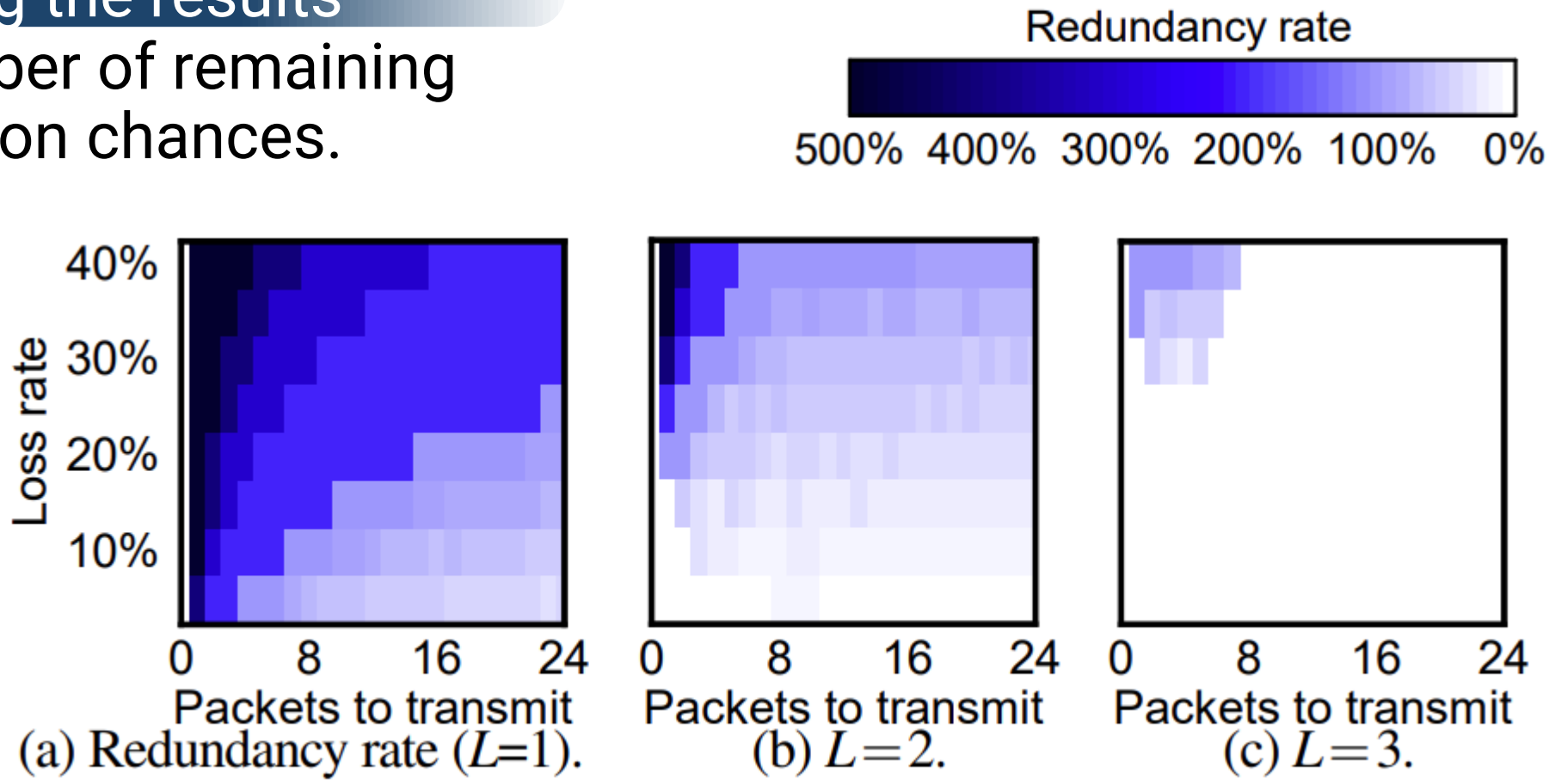


Evaluation

Insight: co-optimize redundancy and retransmission.

Understanding the results

- L is the number of remaining retransmission chances.



(a) Redundancy rate (L=1). (b) L=2. (c) L=3.

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

More evaluation

- Working with different congestion control algorithms…
- Application-level metrics (stalls, frame delays, …)
- Network-level metrics (delays, loss rates, …)
- Parameter sensitivity and more!
- Source codes:
  - NS-3 simulation (compatible with ns-3.33 version): https://github.com/hkust-spark/hairpin
  - WebRTC patch (compatible with M119 release): https://github.com/hkust-spark/hairpin-webrtc

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Takeaway

- Packet loss recovery is no longer “the more redundancy, the better performance”.
- When sufficient time budget, rely on retransmission; when deadline approaching, rely on redundancy.
- This improves both bandwidth cost and deadline miss rate simultaneously.
- Source codes:
  - NS-3 simulation (compatible with ns-3.33 version): https://github.com/hkust-spark/hairpin
  - WebRTC patch (compatible with M119 release): https://github.com/hkust-spark/hairpin-webrtc
- Thank you!!