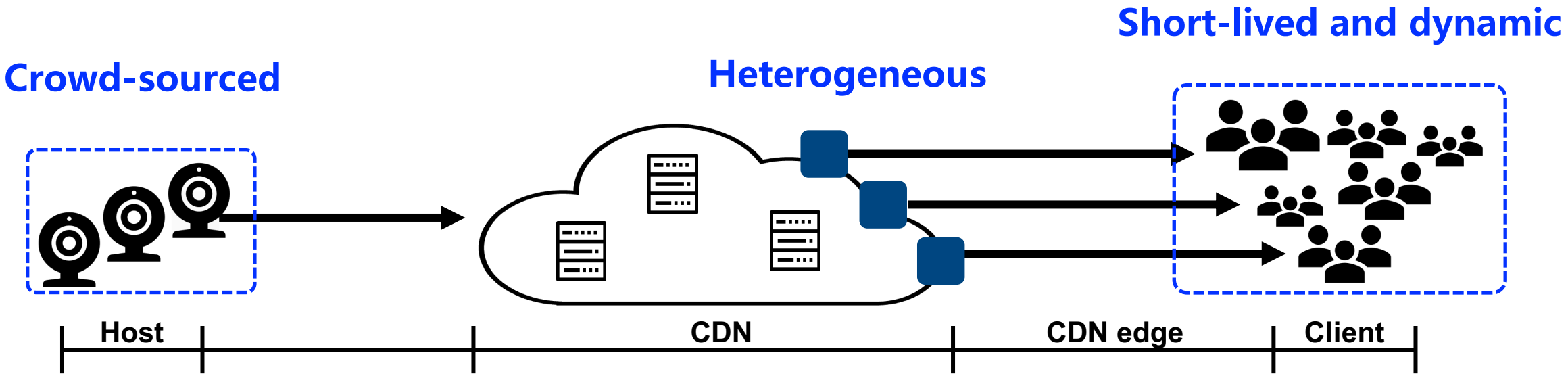
Cost-efficient Request Mapping for Large-scale Live Streaming Services

Yu Tian, Zhenyu Li, Matthew Yang Liu, Qinghua Wu, Zhaoxue Zhong, Ao Li, Jiaxing Zhang, Gerui Lv, Chuanqing, Xi Wang, Jian Mao, Gareth, Jie Xiong, Zhenhua Li, Gaogang Xie

Background

❖ Live streaming evolution

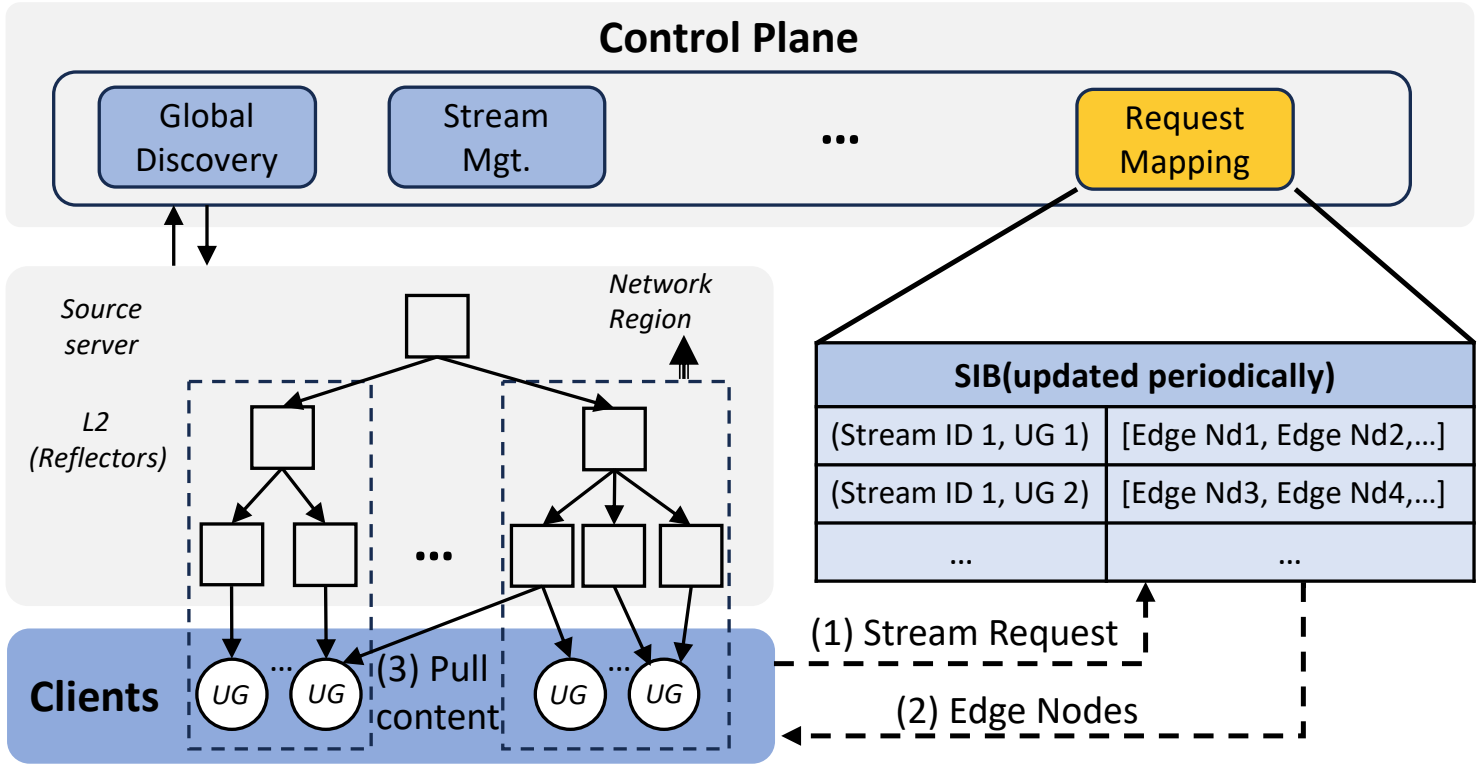


Background

- Live CDN requirement
  - SLA guarantee
  - Minimizing bandwidth costs

	For the Year Ended December 31,						
	2022		2023		2024		
	RMB	%	RMB	%	RMB	US$	%
	(in thousands, except for percentages)						
Cost of revenues:							
Revenue-sharing costs	9,115,351	50.5%	9,507,483	55.6%	10,803,944	1,480,134	59.8%
Content costs	3,496,871	19.4%	3,195,620	18.7%	2,729,520	373,943	15.1%
Server and bandwidth costs	1,752,878	9.7%	1,477,116	8.7%	1,643,678	225,183	9.1%
IP derivatives and others	3,684,772	20.4%	2,905,903	17.0%	2,880,420	394,616	16.0%
Total cost of revenues	**18,049,872**	**100.0%**	**17,086,122**	**100.0%**	**18,057,562**	**2,473,876**	**100.0%**

Server and bandwidth costs for Bilibili in 2024: **164 million RMB**



Large-scale Measurement

❖ Request-level

- Time: April—Jun 2023
- **4.86 billion** requests: request timestamp, stream ID, edge node that served the request, …

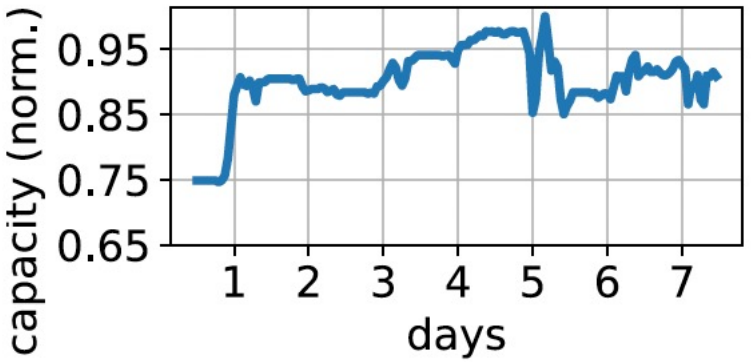
❖ Stream-level

- Time: July 2023
- Peak number of concurrent live streams: **52.13k**

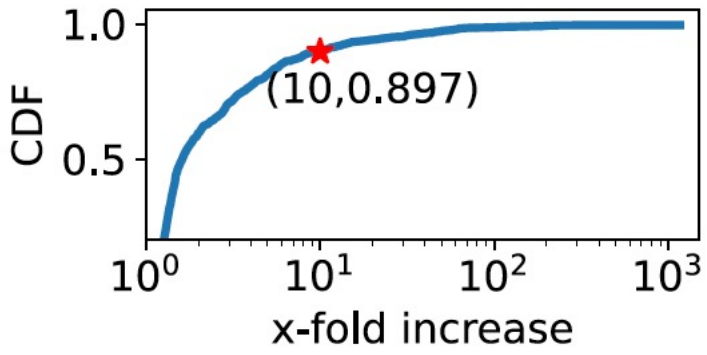
❖ Node-level

- Time: July 2023
- Dynamic bandwidth capacity (which may fluctuate), and bandwidth usage at 5-minute granularity, …

Motivation



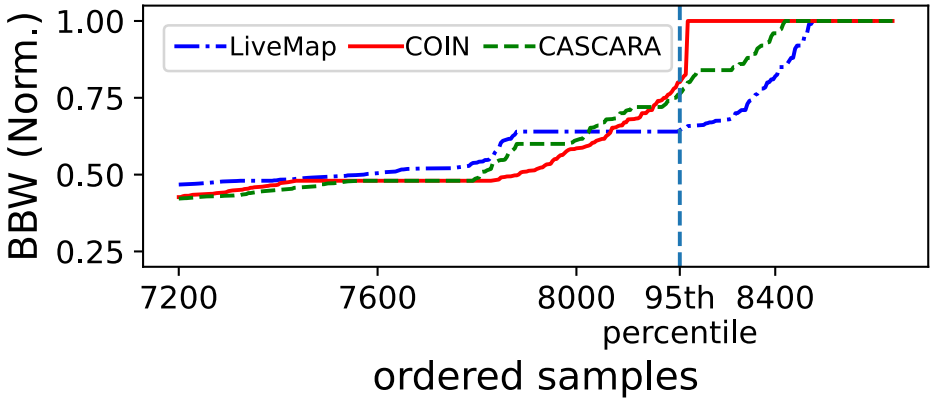
Dynamic Supply



Unpredictable Demand Surges

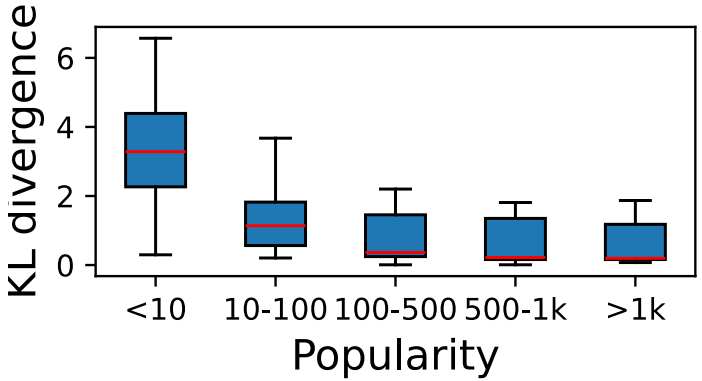
Regions	R1	R2	R3	R4	R5	R6	R7
Ratio	2.75	**0.40**	2.71	2.51	**0.91**	1.62	1.47

Issue1: **Dynamic inter-region S-D imbalance forces frequent cross-region scheduling, leading to inflated access latency.**

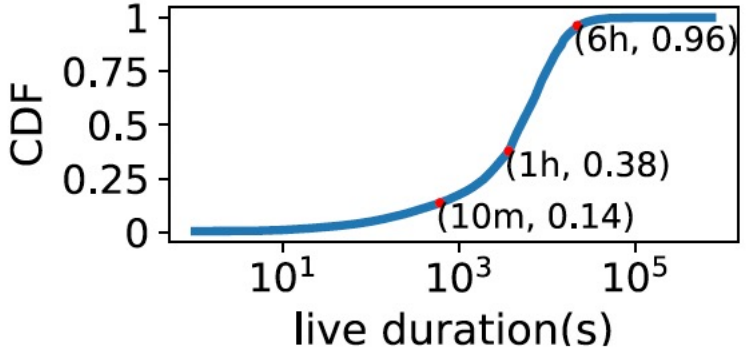


Issue2: **Dynamic S-D fluctuations cause mismatches between planned and actual bandwidth, inflating bandwidth cost.**

Motivation



The KL Divergence between Per-stream and Overall Viewer Distributions.



Live duration distribution.

Issue3: **Stream popularity exhibits significant spatio-temporal heterogeneity, requiring fine-grained spatio-temporal scheduling.**

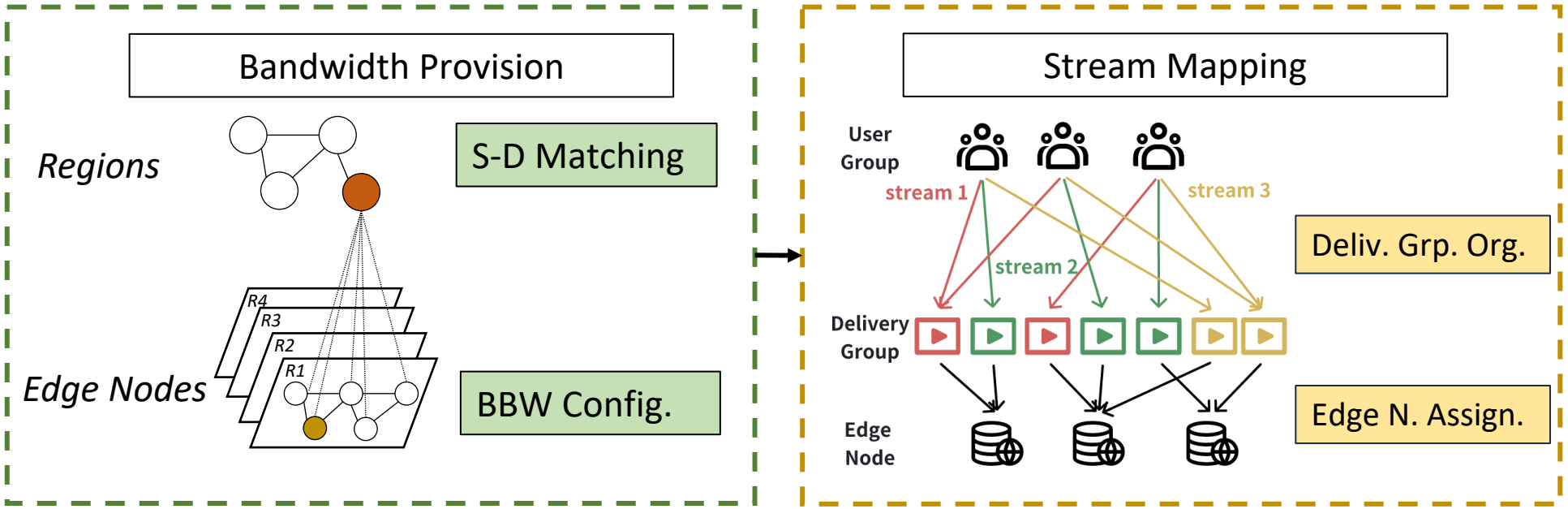
Problem and system design

Challenges:

- Dynamic supply-demand imbalances
- Spatio-temporal skewed popularity

Research goal:

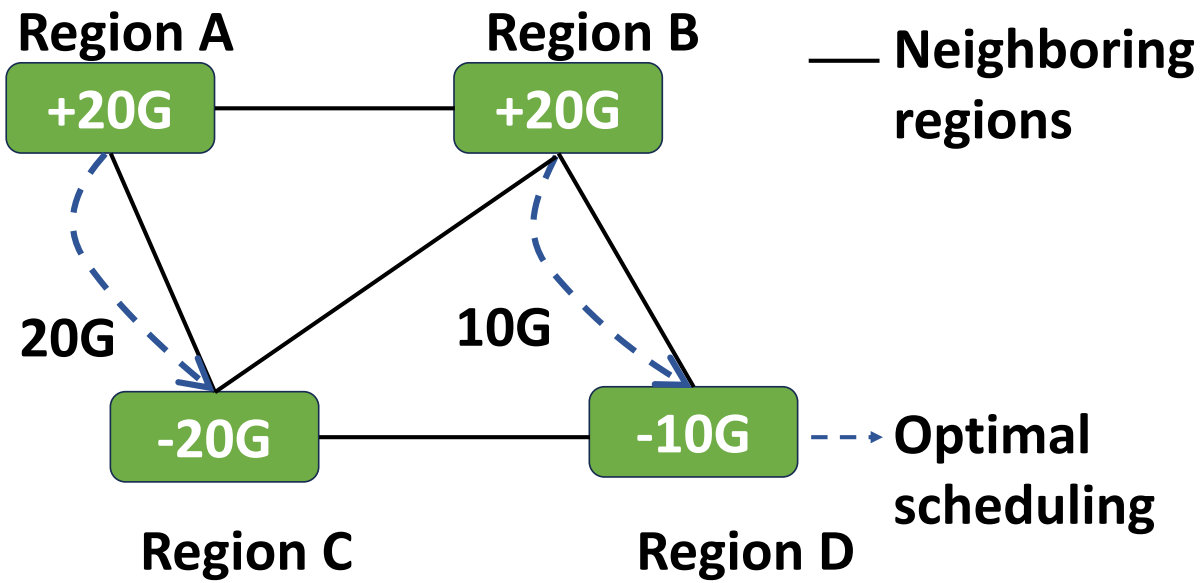
1 minimizing bandwidth cost and 2 maintaining low access latency



Dual-level Coordination **Adaptive Mapping**

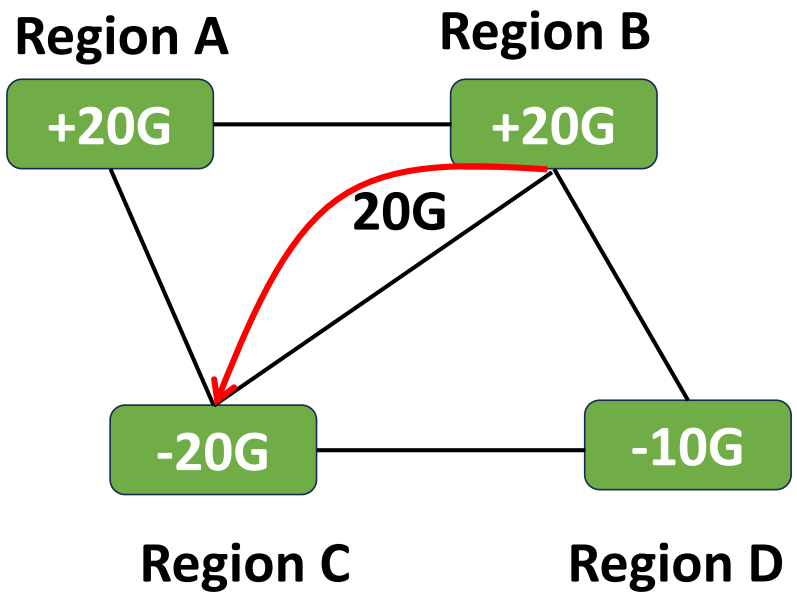
Dual-level Coordination: Regional

Objective： maximize demand served within candidate low-latency regions.



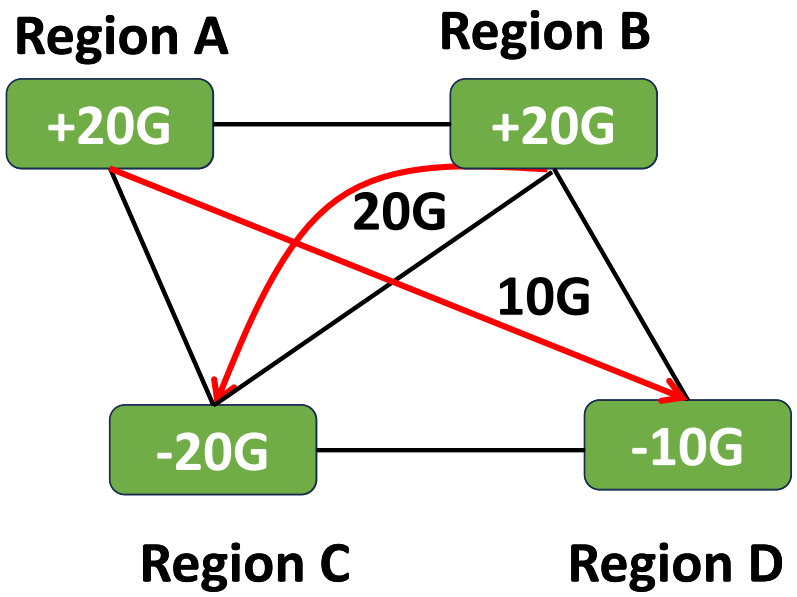
Dual-level Coordination: Regional

Objective： maximize demand served within candidate low-latency regions.



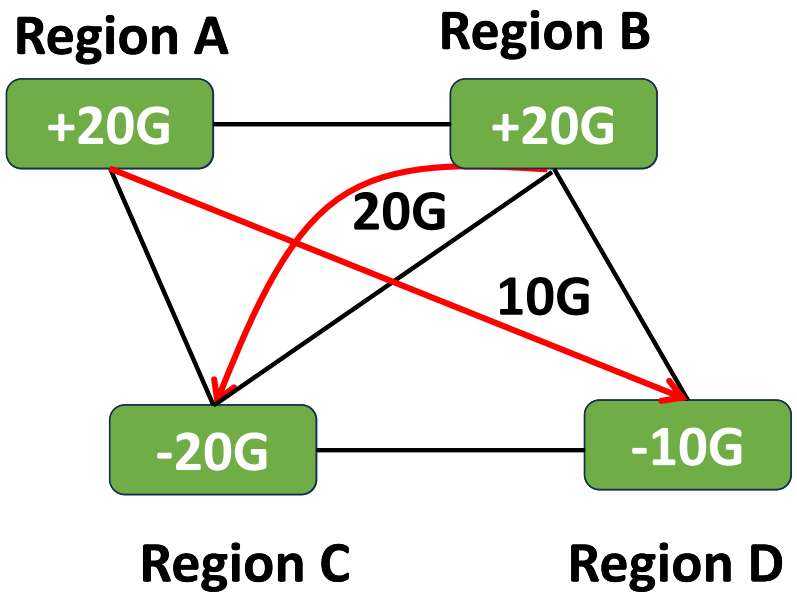
Dual-level Coordination: Regional

Objective： maximize demand served within candidate low-latency regions.



Dual-level Coordination: Regional

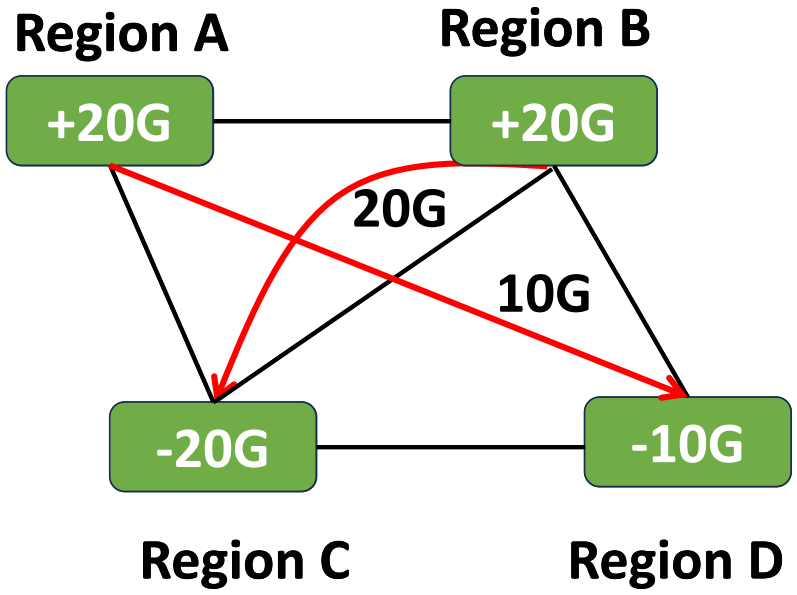
Objective： maximize demand served within candidate low-latency regions.

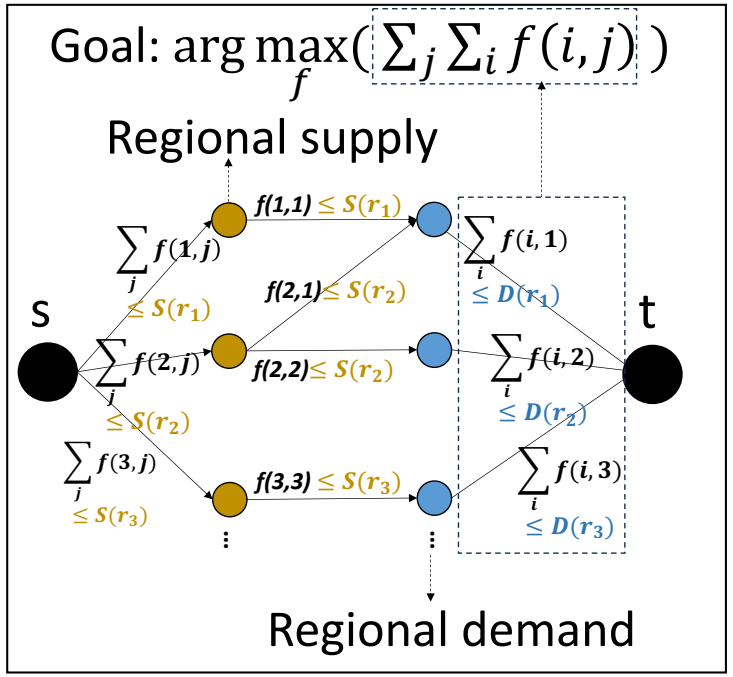


Local heuristics: forcing later regions to use far-away nodes

Dual-level Coordination: Regional

Objective： maximize demand served within candidate low-latency regions.



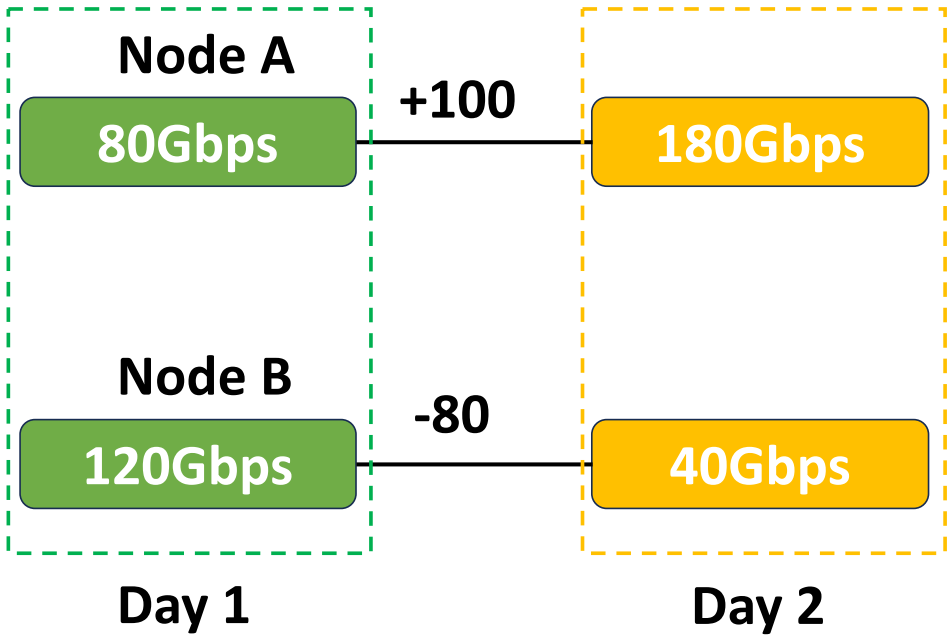


Local heuristics: forcing later regions to use far-away nodes

Global optimization via maximum flow

Dual-level Coordination: Node-level

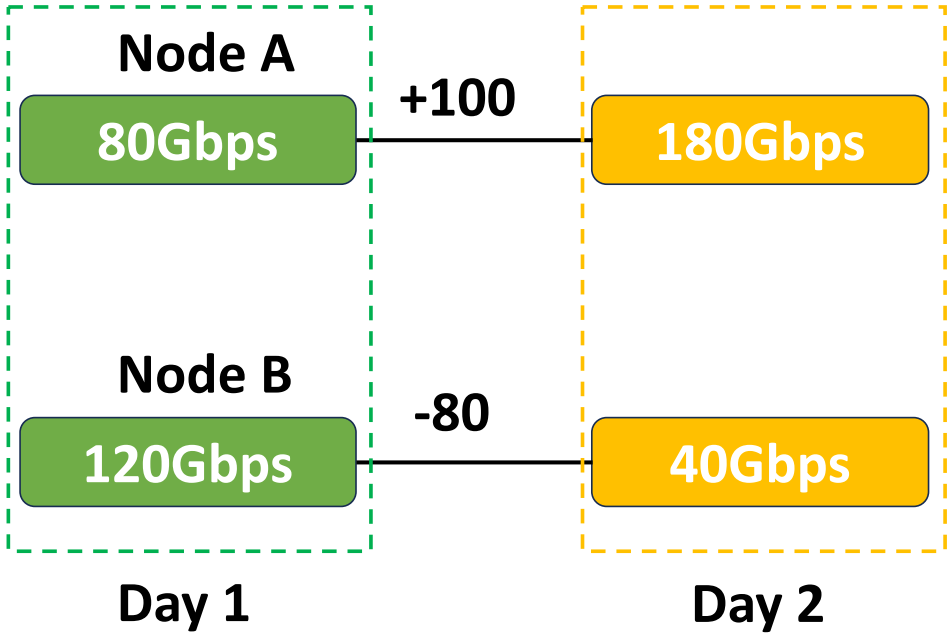
Objective： minimize bandwidth cost within minimum billable bandwidth updates.



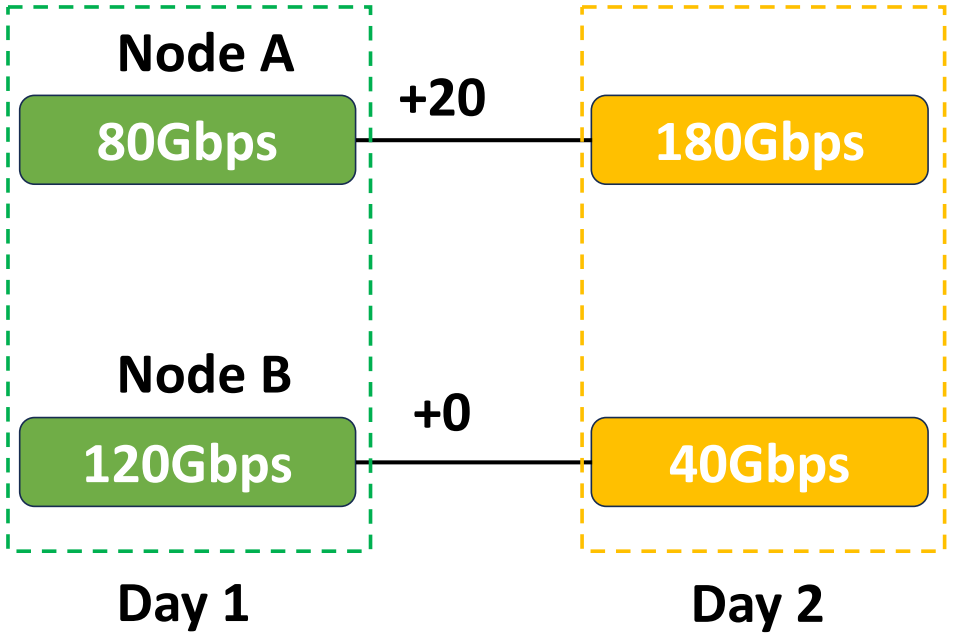
Direct replanning: drastically different allocation

Dual-level Coordination: Node-level

Objective： minimize bandwidth cost within minimum billable bandwidth updates.



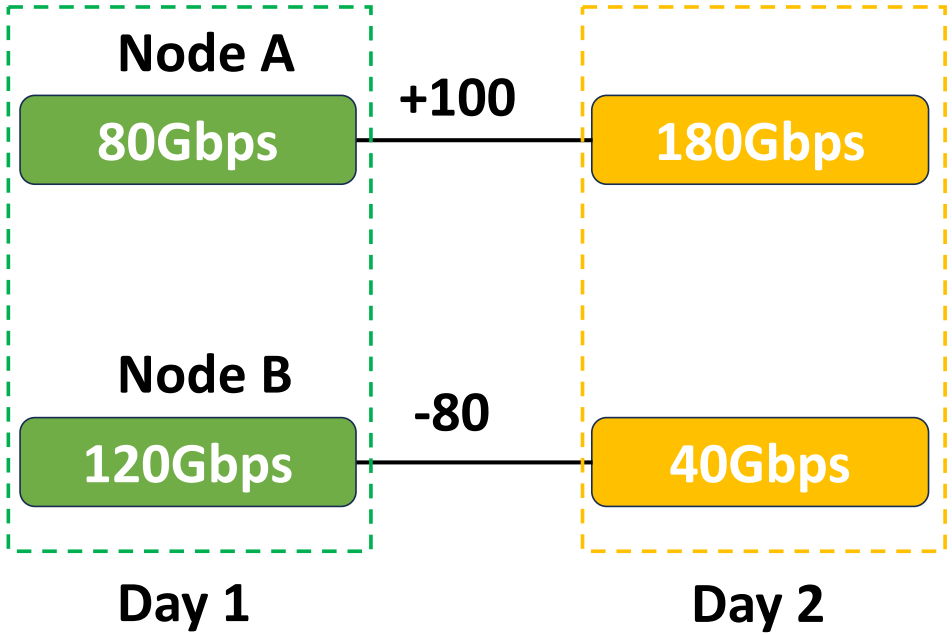
Direct replanning: drastically different allocation.



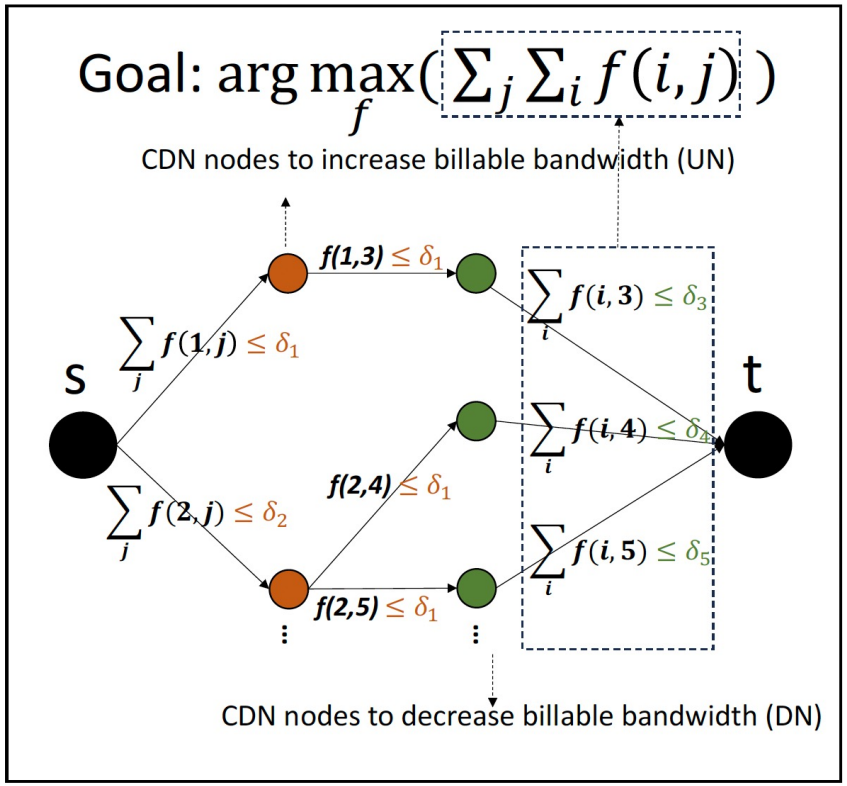
Coordinated replanning: minimize day-to-day billable bandwidth changes.

Dual-level Coordination: Node-level

Objective： minimize bandwidth cost within minimum billable bandwidth updates.



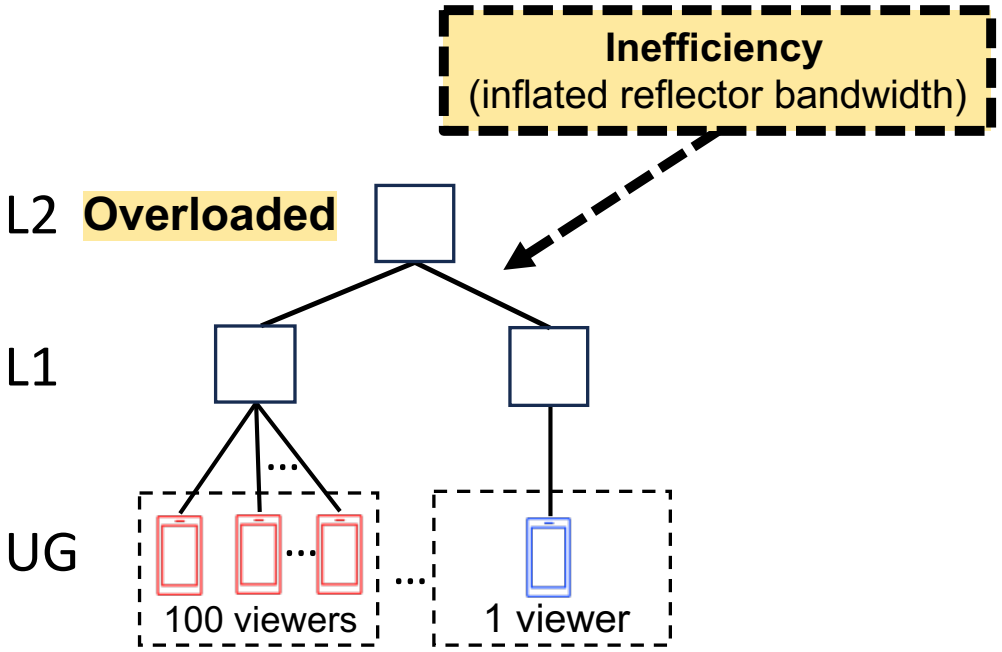
Direct replanning: drastically different allocation.



Coordinated replanning through maximum flow

Adaptive Stream Mapping

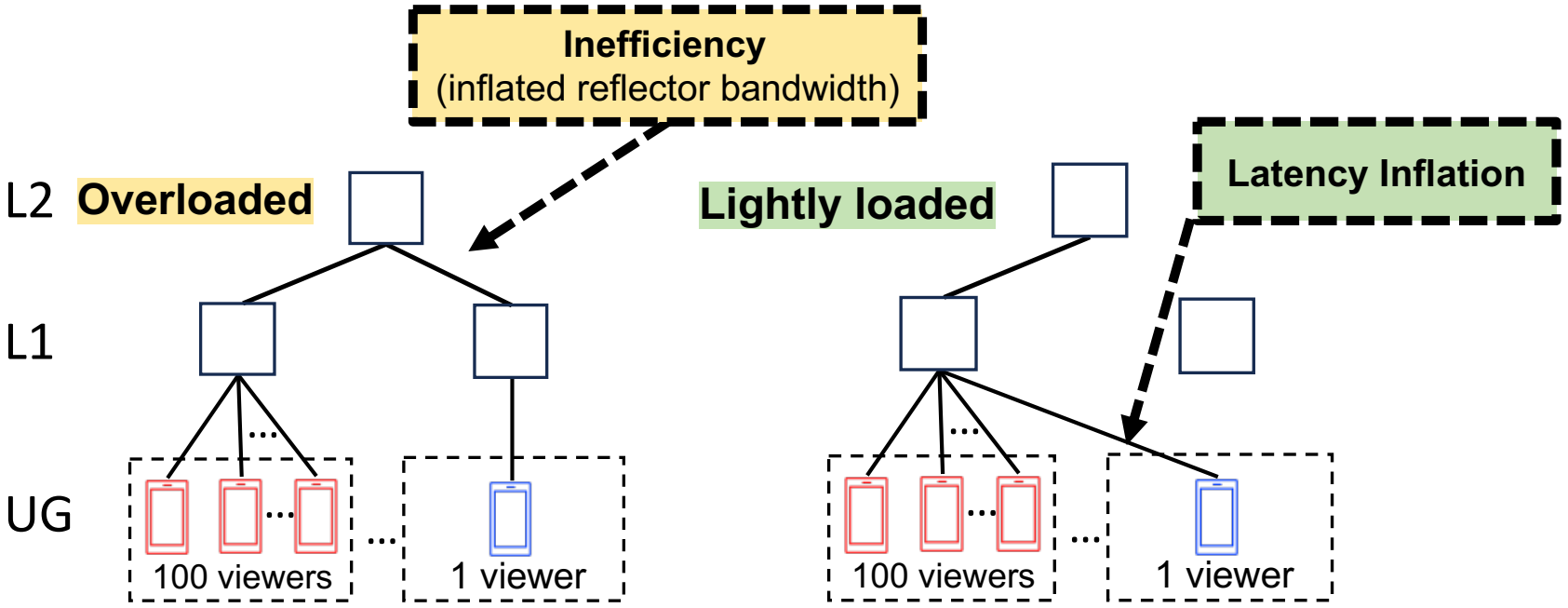
Objective: Dynamic balance between proximity and efficiency



**Proximity-first →
high reflector overhead.**

Adaptive Stream Mapping

Objective: Dynamic balance between proximity and efficiency

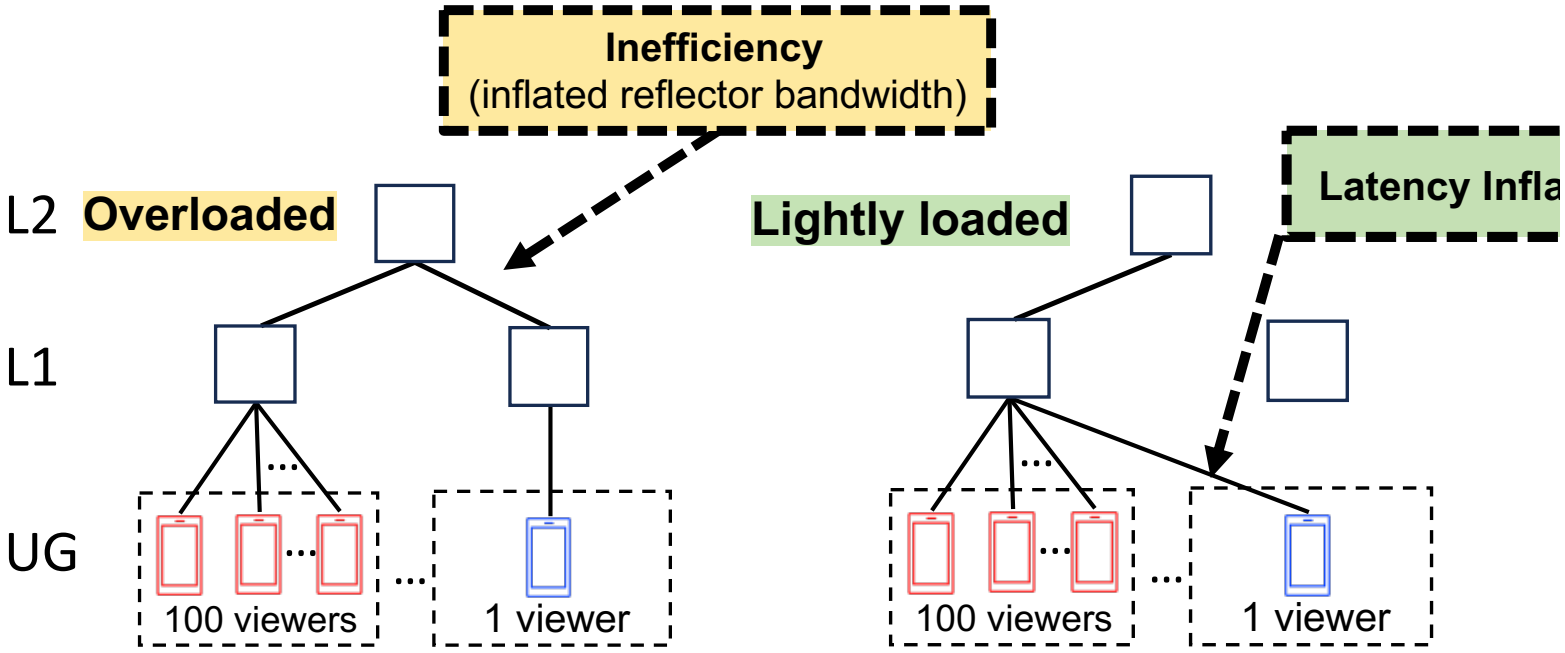


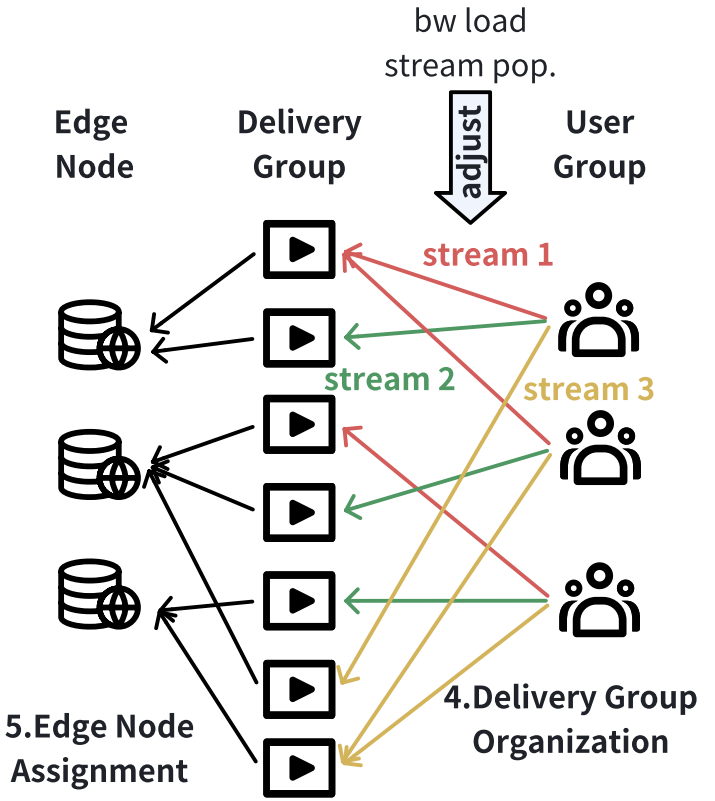
Proximity-first → high reflector overhead.

Efficiency-first → high latency.

Adaptive Stream Mapping

Objective: Dynamic balance between proximity and efficiency





Proximity-first → high reflector overhead.

Efficiency-first → high latency.

Organizes user groups into delivery groups (dynamically).

Trace-Driven Simulation

❖ Baselines

- Offline planning: **COIN** [Infocom '23]
- Offline planning with online updates: **CASCARA** [NSDI '21]
- Cost-first Mapping: **Entact** [NSDI '10]

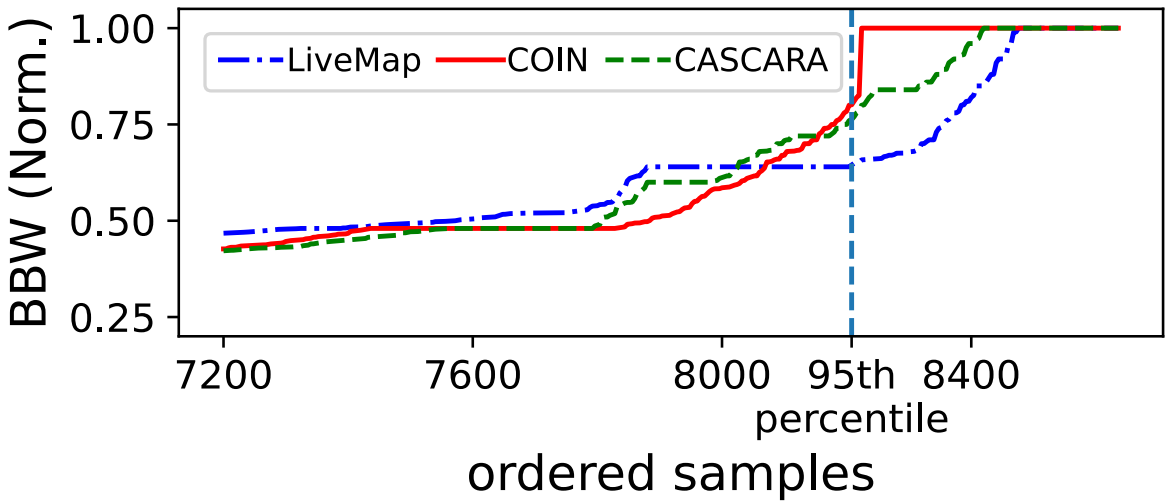
❖ Network Traces

- Collected from Bilibili CDN over three months (April—June 2023).
- Includes:
  - Node-level info: capacity, unit price, and billing model.
  - Request-level data: start time, duration, stream ID, and user origin.
  - Provides network measurements: RTT between each user group and edge node.
- 4.86 billion requests replayed in trace-driven simulations.

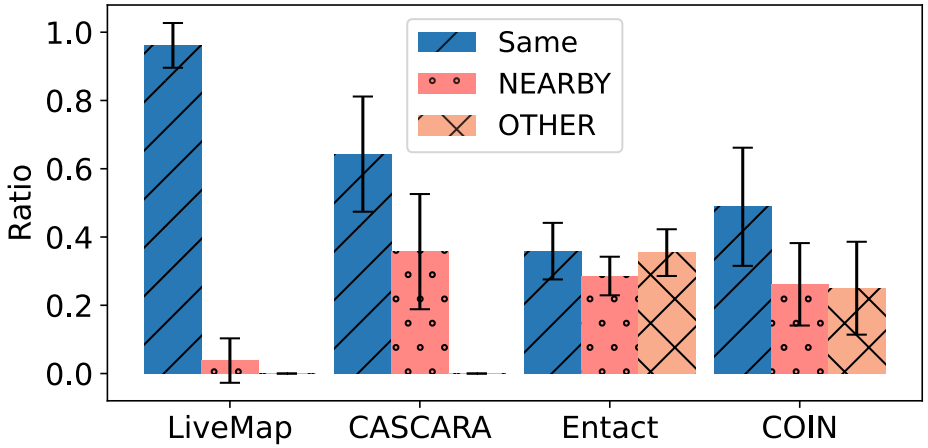
❖ Metrics

- Bandwidth cost
- Access latency

LiveMap Performance Validation



Reduced billable bandwidth



Reduce cross-region scheduling

- Reduces node billable bandwidth through **dynamic bandwidth provisioning**, achieving **20.1%–30.7%** bandwidth cost reduction.
- Reduces access latency by **20.7%–29.9%** by combining **cross-region bandwidth coordination** and stream popularity–aware mapping.

Large-Scale Evaluation in Production System

❖ Deployment

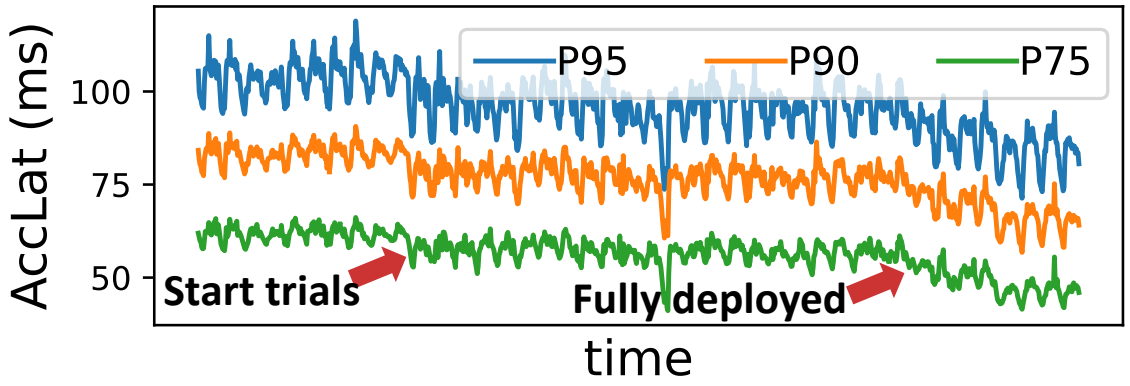
- Deployed in the Bilibili CDN, serving **tens of millions of daily users**.
- 2 deploying stages: Only BP, and LiveMap (BP+SM).

❖ Real-world results:

- Cost savings: **41.2%▼**
- Rebuffering Rate: **12.7%▼**; Access Latency: **20.3% ▼**

Table 5. Live streaming service overview.

	#streams	#viewers	Traffic Volumn
12 p.m.	17.85k	258.99k	0.78 Tbps
8 p.m.	41.78k	468.45k	1.90 Tbps



Takeaway

- **Discovery**: Modern live CDNs face challenges from dynamic demand, unstable edge capacity, and heterogeneous stream popularity.
- **Solution**:
  - LiveMap introduces dual-level bandwidth provisioning and popularity–aware mapping to tame **regional supply–demand imbalance** and stream-level popularity skewness.
- **Impact**:
  - LiveMap has been deployed in the Bilibili CDN since **March 2024**, serving **tens of millions of daily users**.
  - Large-scale deployment shows that LiveMap achieves over 40% cost saving and up to 20% latency reduction.

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

中国科学院大学
University of Chinese Academy of Sciences

Thanks for listening / Q&A

tianyu21b@ict.ac.cn