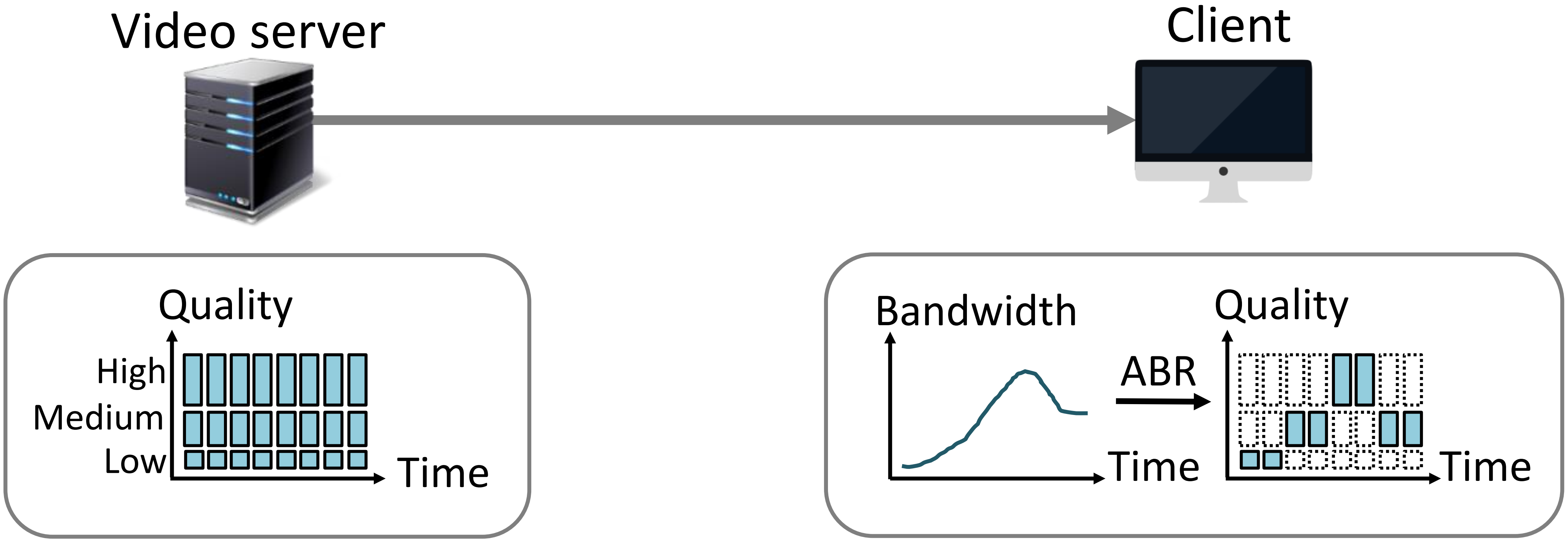
Neural Adaptive Content-aware Internet Video Delivery

Hyunho Yeo, Youngmok Jung, Jaehong Kim,
Jinwoo Shin, Dongsu Han

KAIST

Observation on Current Video Ecosystem

Adaptive streaming has been widely deployed
(a primary tool for improving user QoE)



Traditional Approaches

Optimizing ABR algorithms
Pensieve [SIGCOMM 17], MPC [SIGCOMM 15]

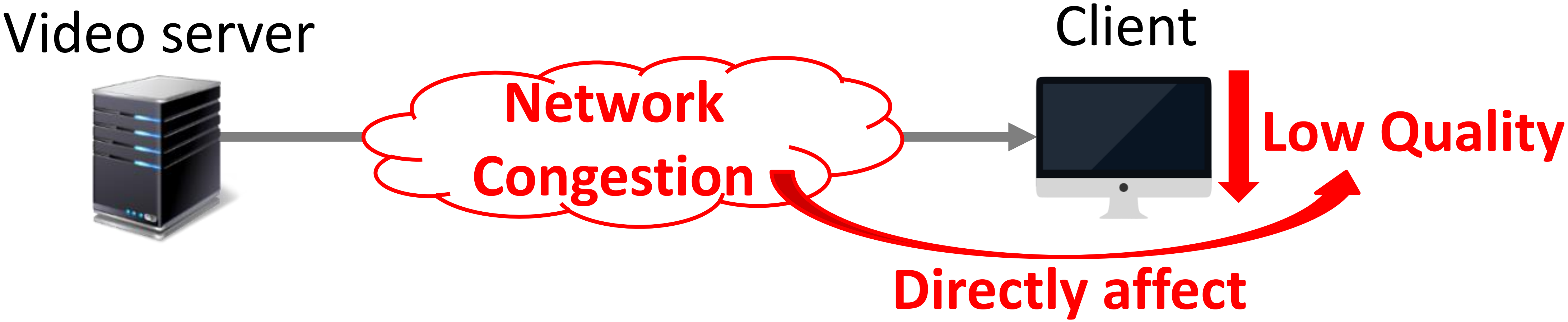
Choosing better servers, CDNs
Content Multihoming [SIGCOMM 12], VDN [SIGCOMM 15]

Leveraging centralized control plan
Video Control Plane [SIGCOMM 12], Pythease [NSDI 17]

Goal: Find how to best utilize the network resource

Limitation of Current Video Delivery

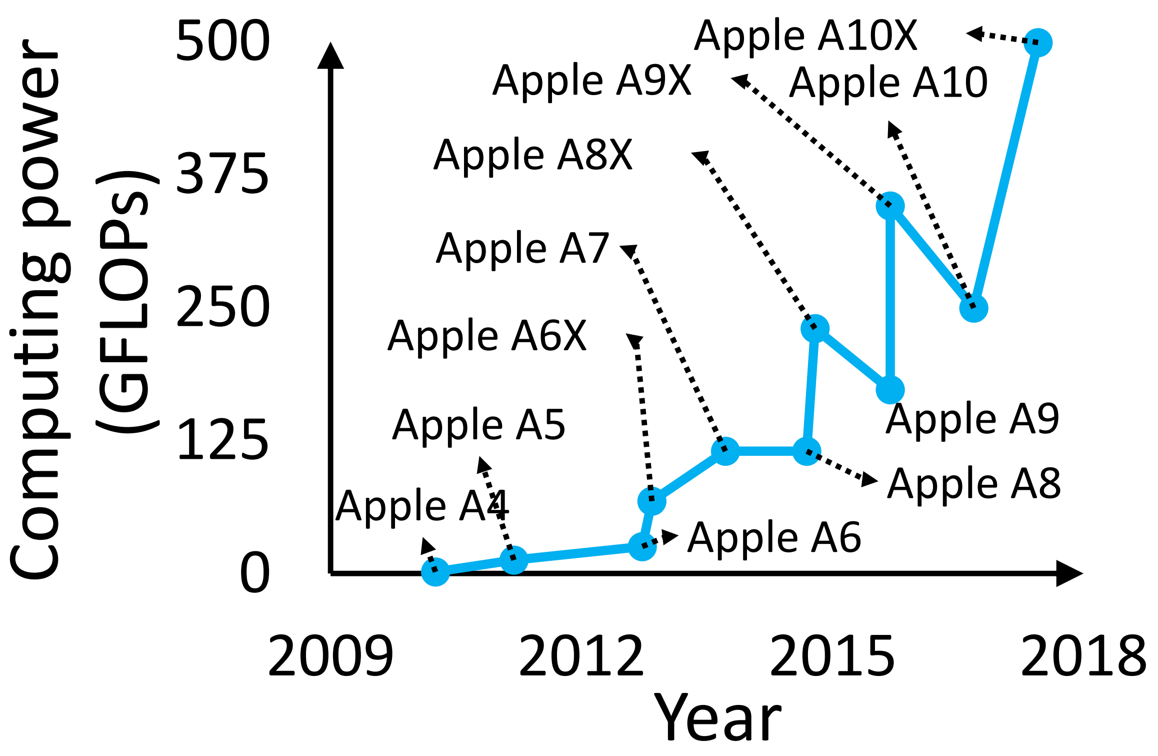
Video quality heavily depends on available bandwidth



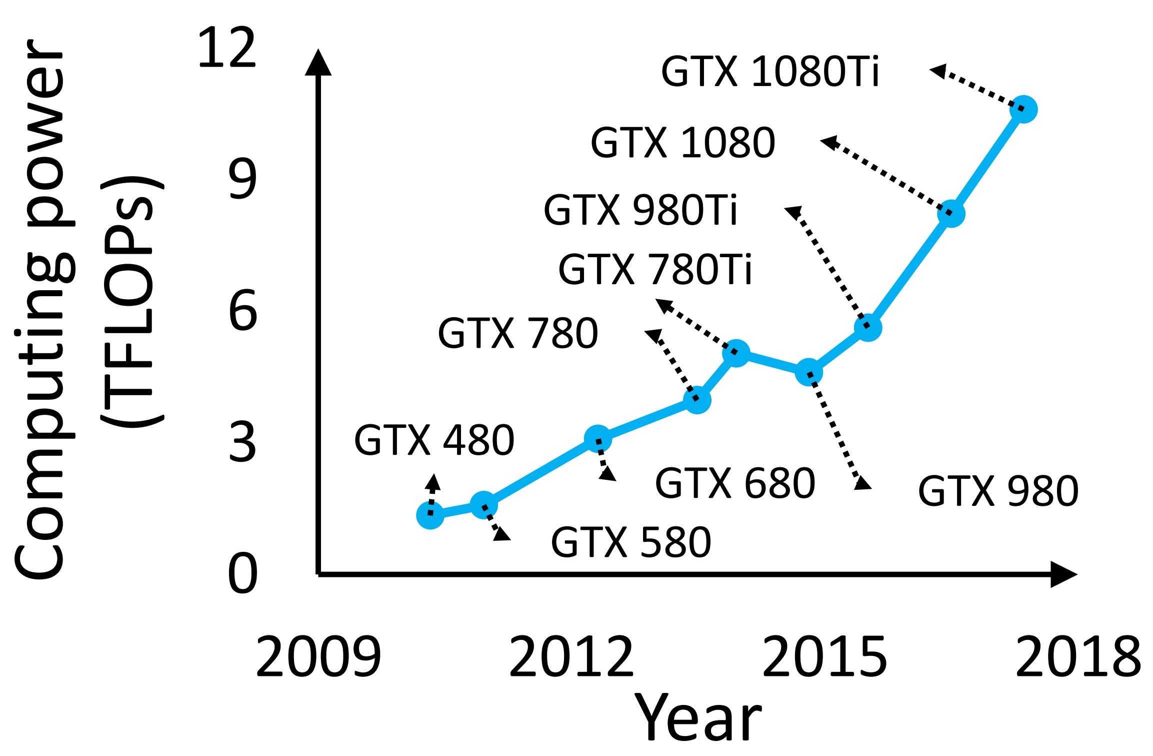
Limitation of Current Video Delivery

Client computing power is scarcely utilized other than for decoding

Mobile GPU

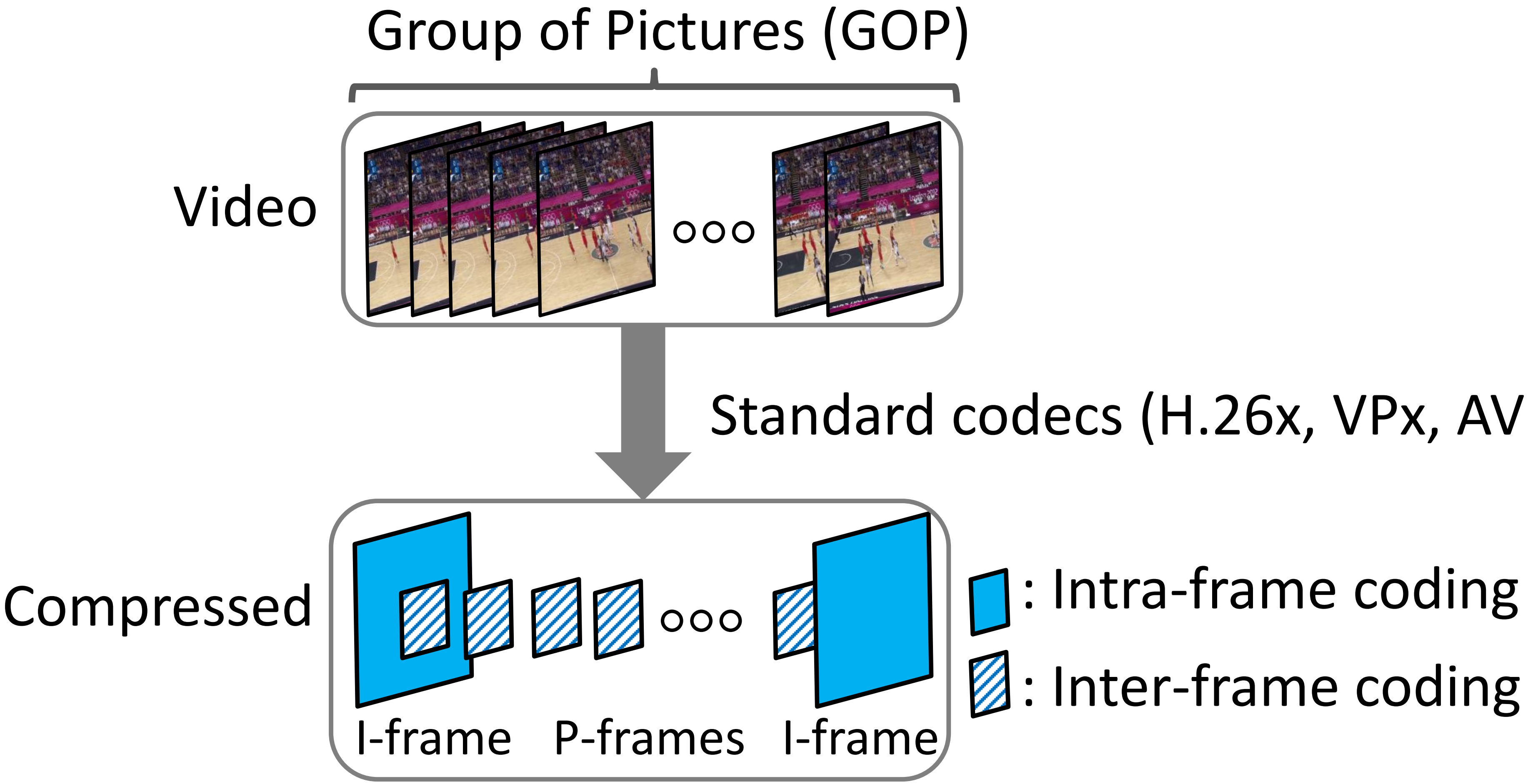


Desktop GPU



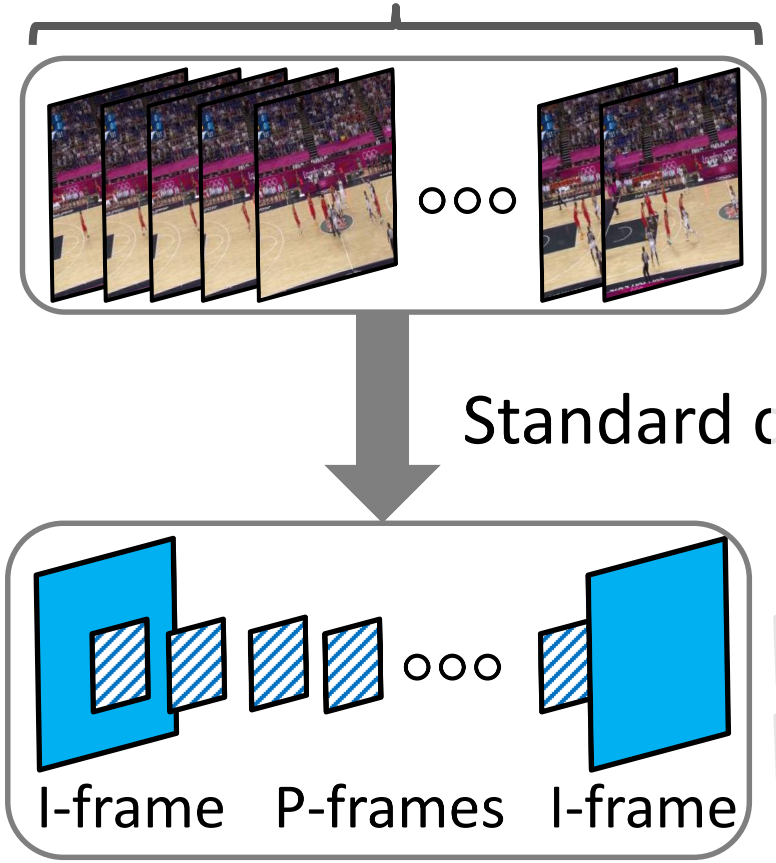
Observation on Current Video Ecosystem

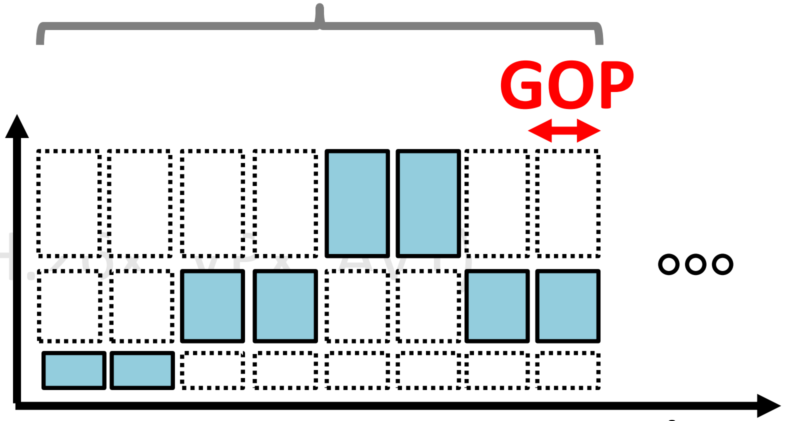
Standard codecs efficiently reduce redundancy inside GOP



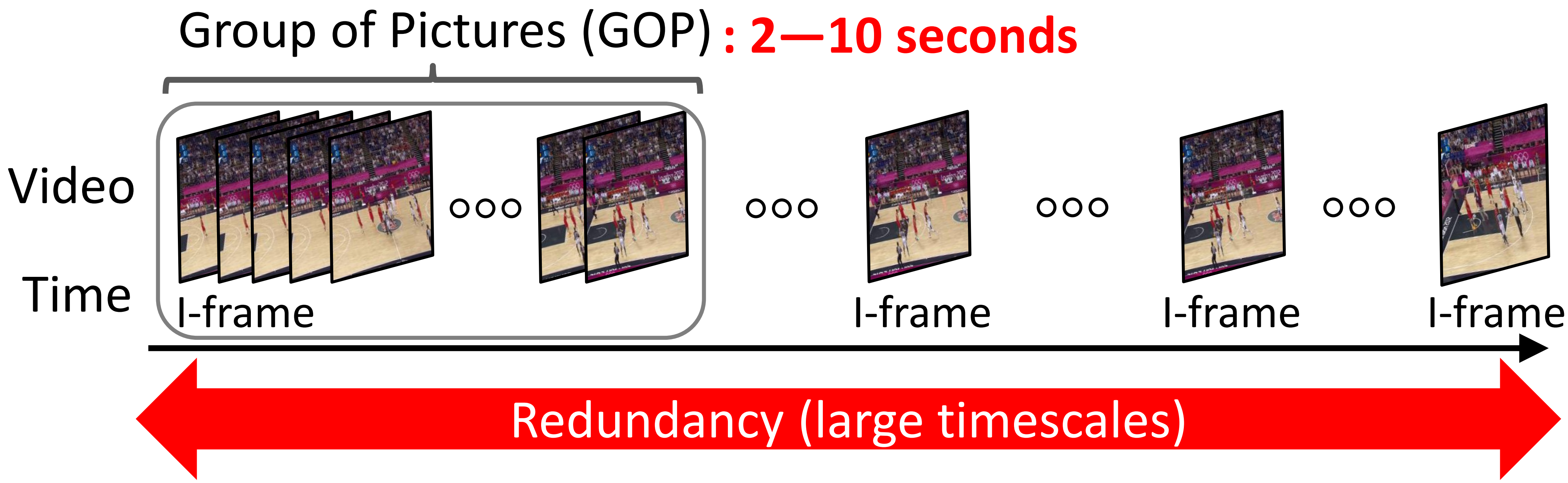
Observation on Current Video Ecosystem

Standard codecs efficiently reduce redundancy inside GOP





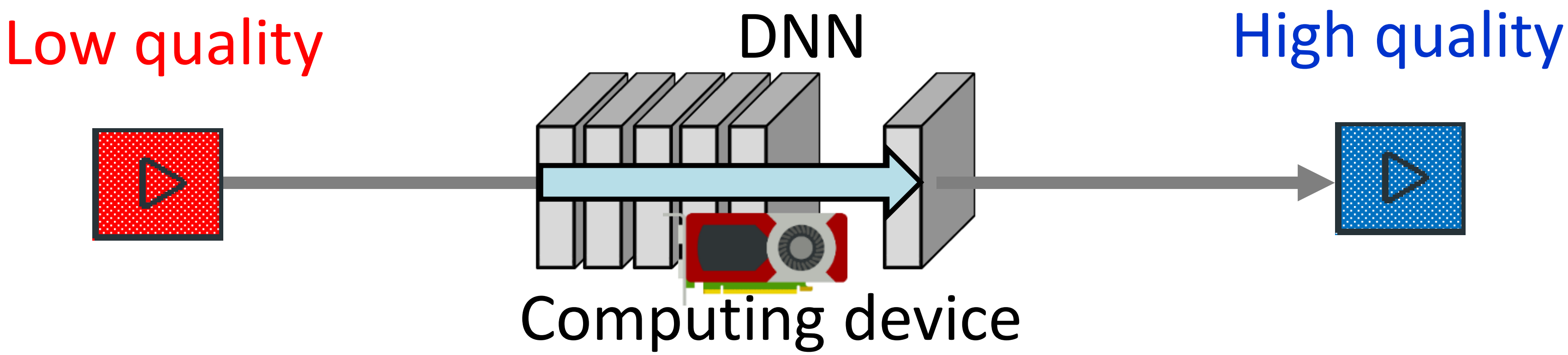
Limitation of Current Video Delivery



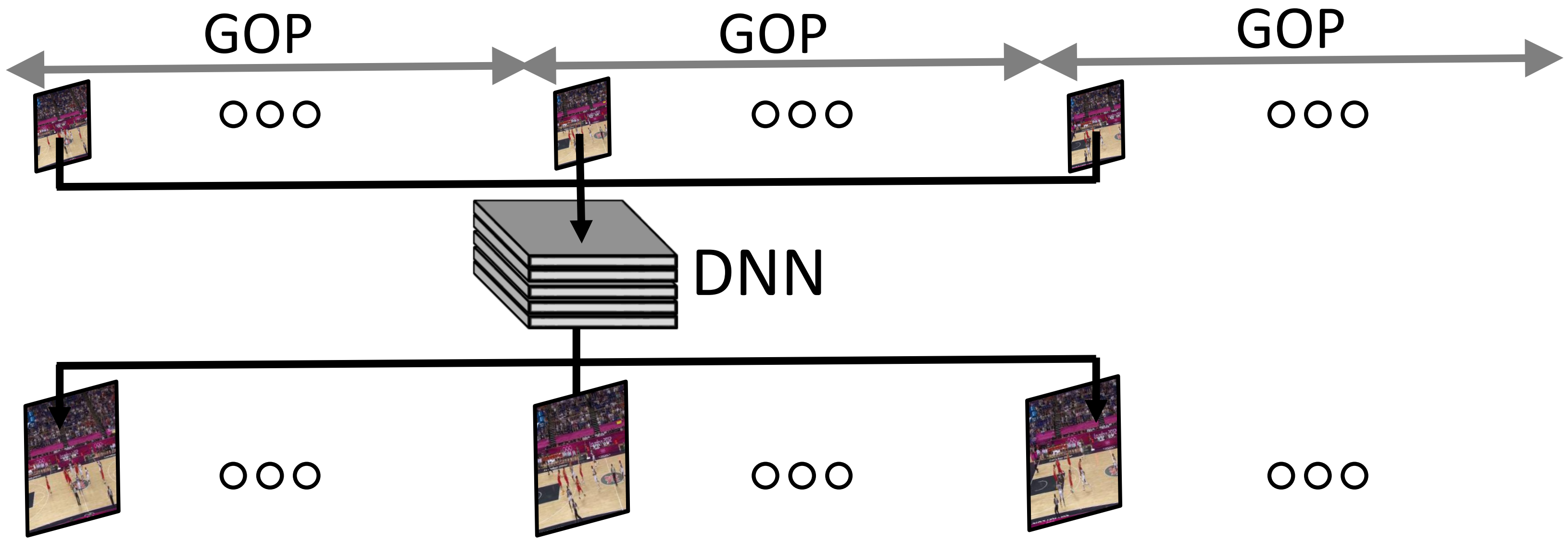
Standard codecs lack any mechanisms for exploiting redundancy that occurs at large timescales

Key Observations on Deep Neural Network

1. Utilizes computing resource to enhance video quality

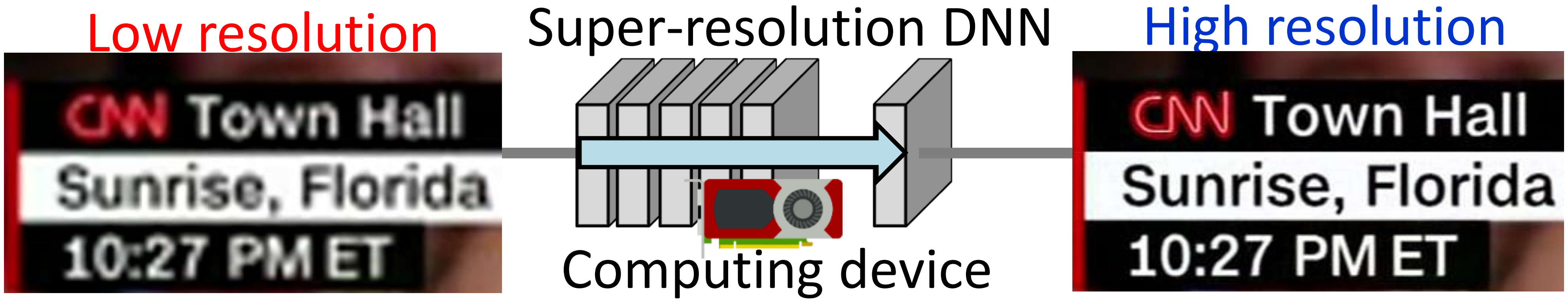


2. Trained and operate in large timescales (video)

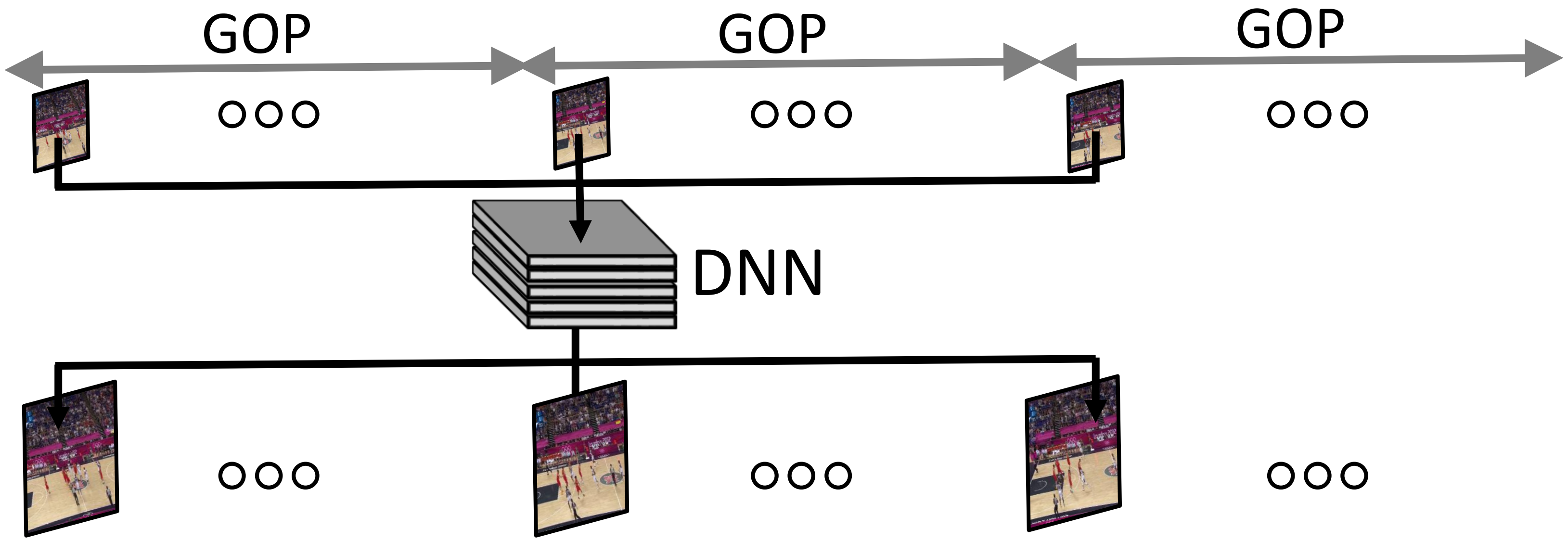


Key Observations on Deep Neural Network

1. Utilizes computing resource to enhance video quality

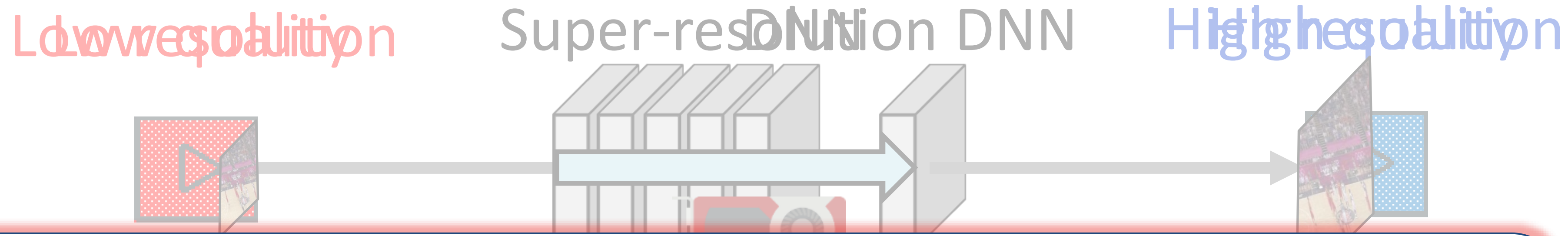


2. Trained and operate in large timescales (video)



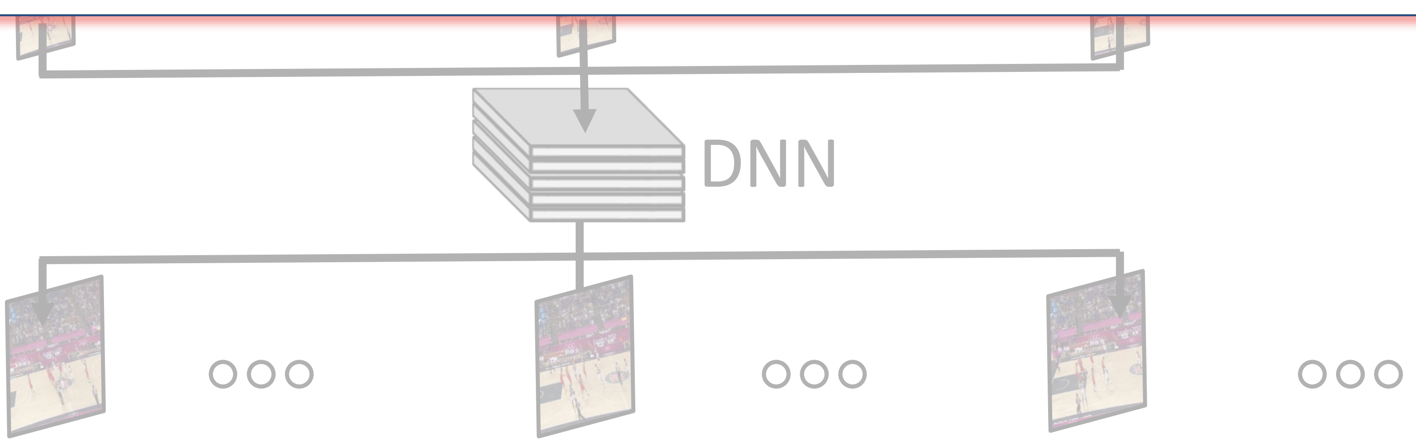
Key Observations on Deep Neural Network

1. Utilizes computing resource to enhance video quality



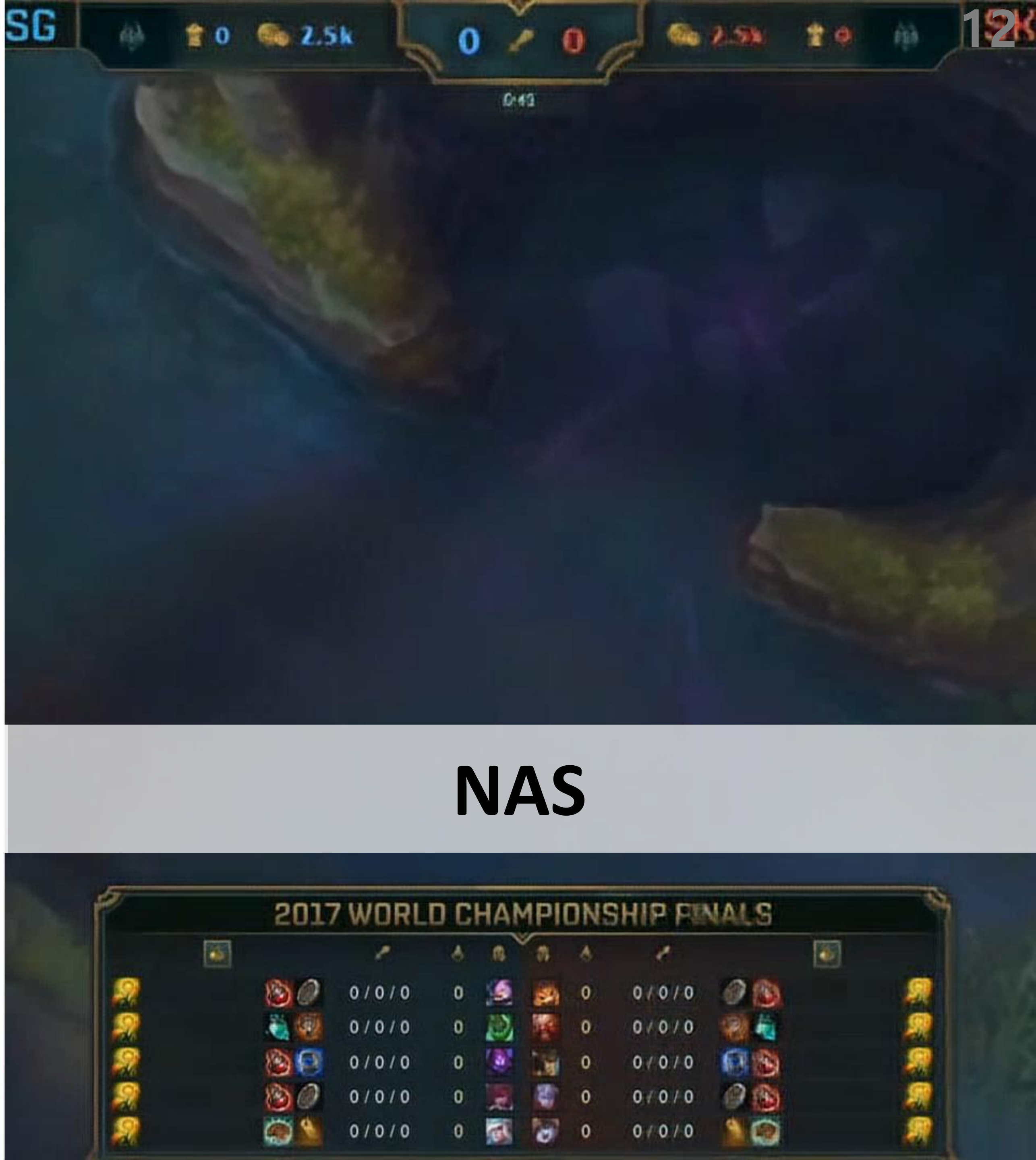
Can we overcome the current limitations via DNN?

How much QoE improvement can we achieve?



Existing Approach

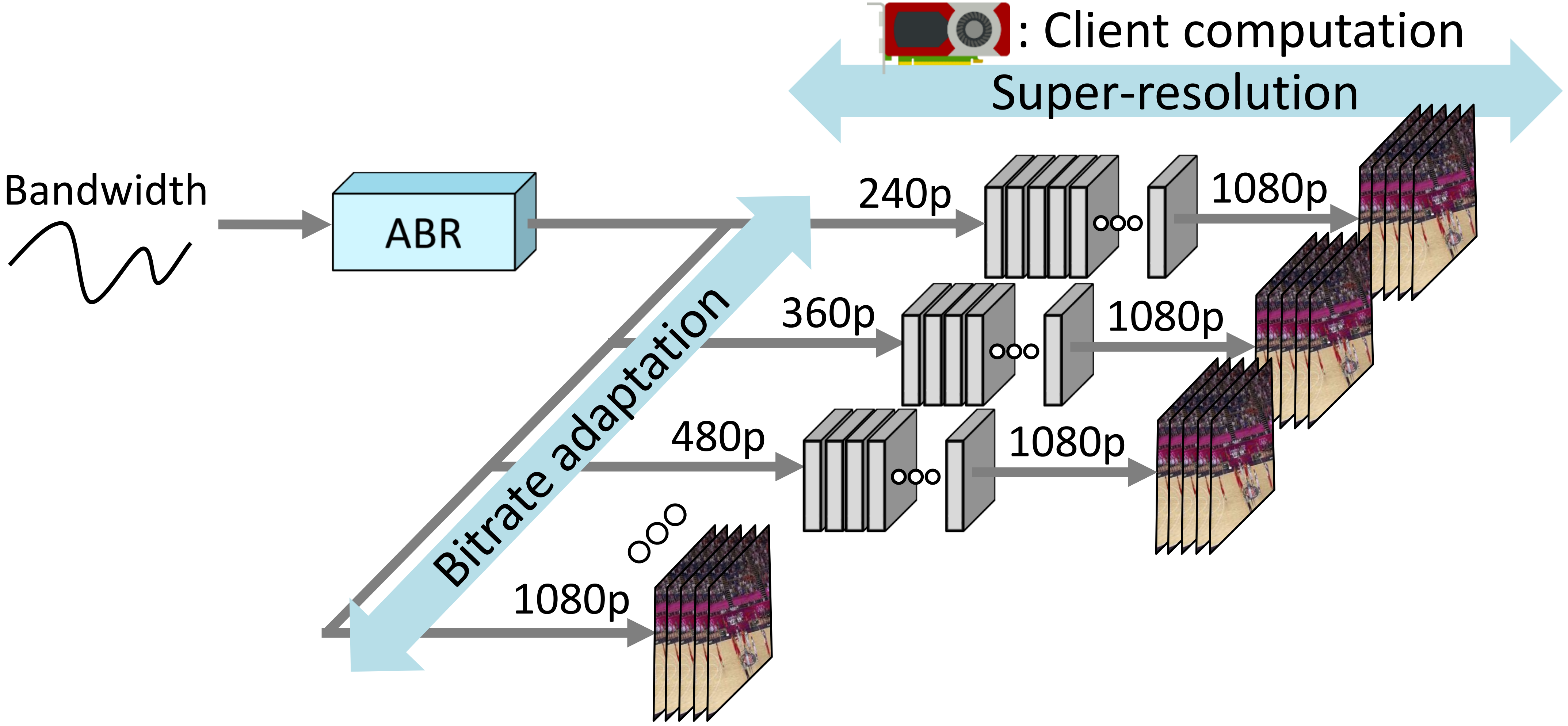
(Pensieve – SIGCOMM 17)



NAS

NAS: DNN-based Video Delivery

Apply super-resolution DNN on top of bitrate adaptation



NAS: Design Scope

1. Content: Video on demand (VOD)

Example

YouTube NETFLIX hulu amazon instant video

2. Computing device: Desktop-class GPUs

Example

GTX 1050 Ti (Entry-level)

○○○

Titan Xp (High-end)

Price

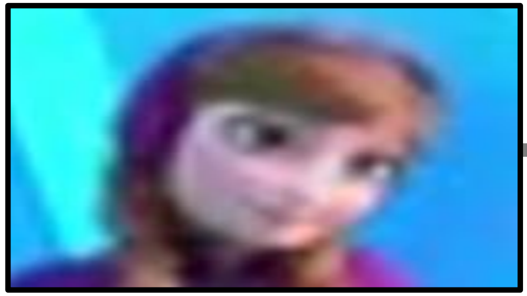
$139

$1,200

NAS: Two Initial Challenges

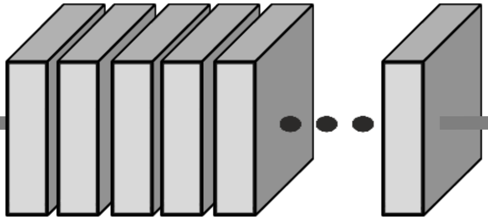
1. DNN accuracy is unreliable for new content ⇨ **Guarantee performance**

New content

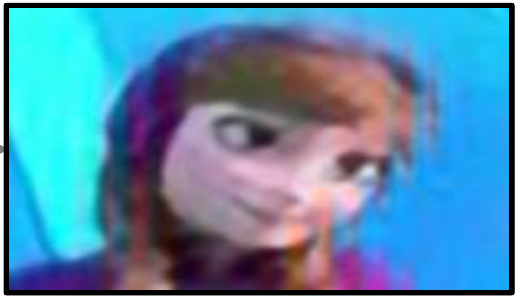


SSIM = 0.86

Generic super-resolution[1,2,3]

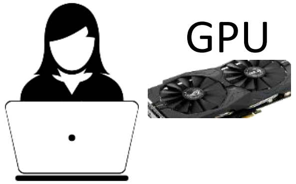


↓Quality



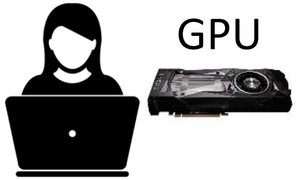
SSIM = 0.84

2. Client must process DNN at real-time,
   but computing power varies across space and time,



GPU

x5 slower!



GPU

⇨ **Require adaptation to computing power**

Client A: Entry-level GPU Client B: High-end GPU

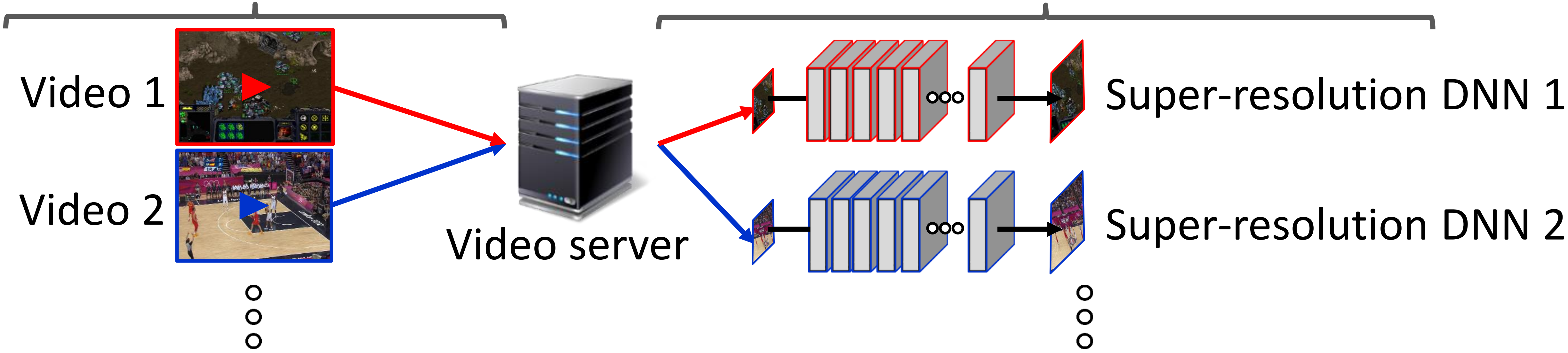
1: SRCNN-ECCV14, 2:VDSR-CVPR 16, 3:EDSR-CVPRW 17

Key Design (1): Content-aware DNN

Challenge: Providing reliable DNN quality

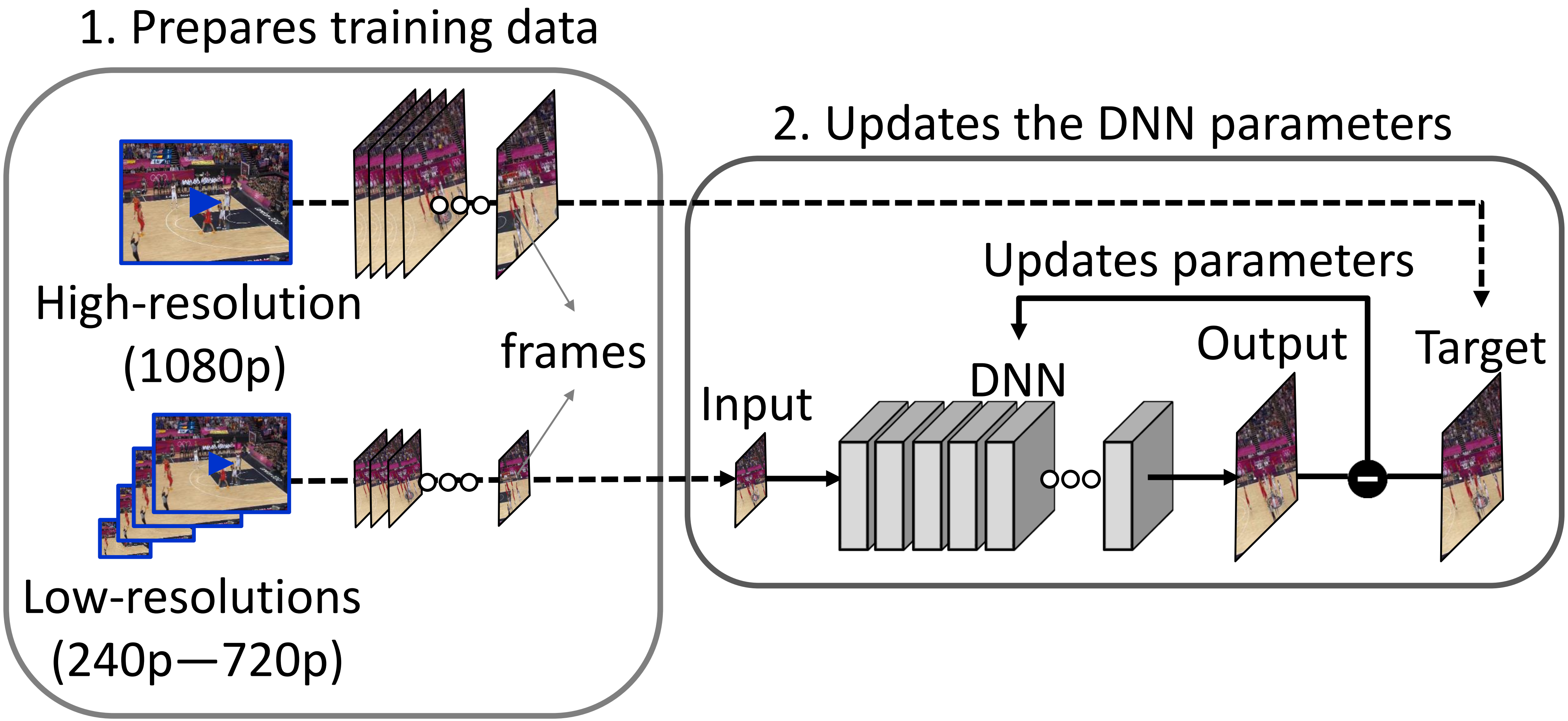
1. New video admission

2. Generates a content-aware DNN per-video

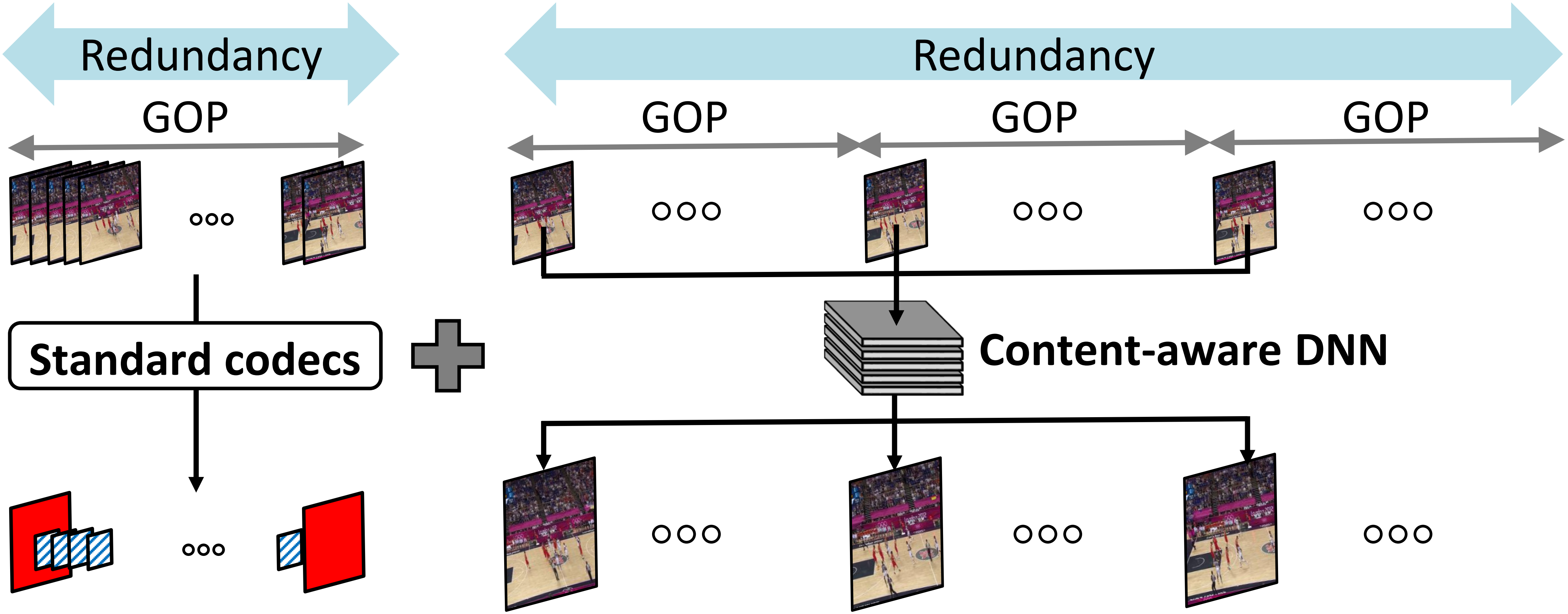


Content-aware DNN delivers the reliable training accuracy instead of the unpredictable testing accuracy.

Training a content-aware super-resolution

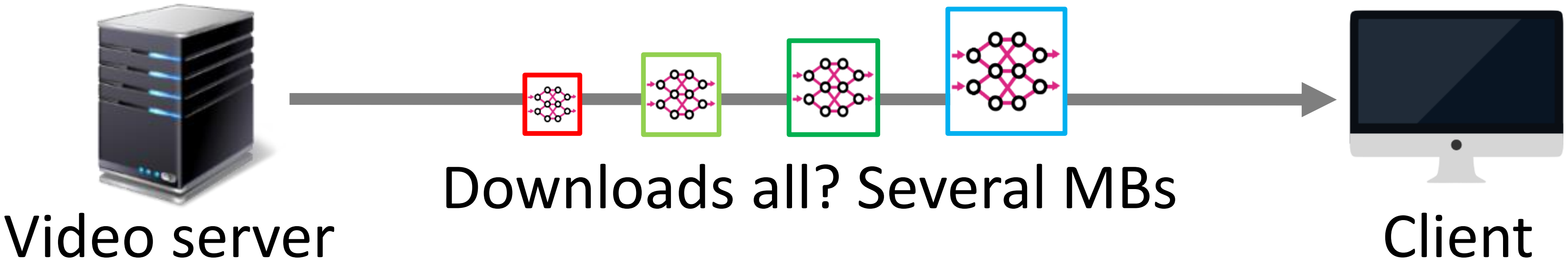


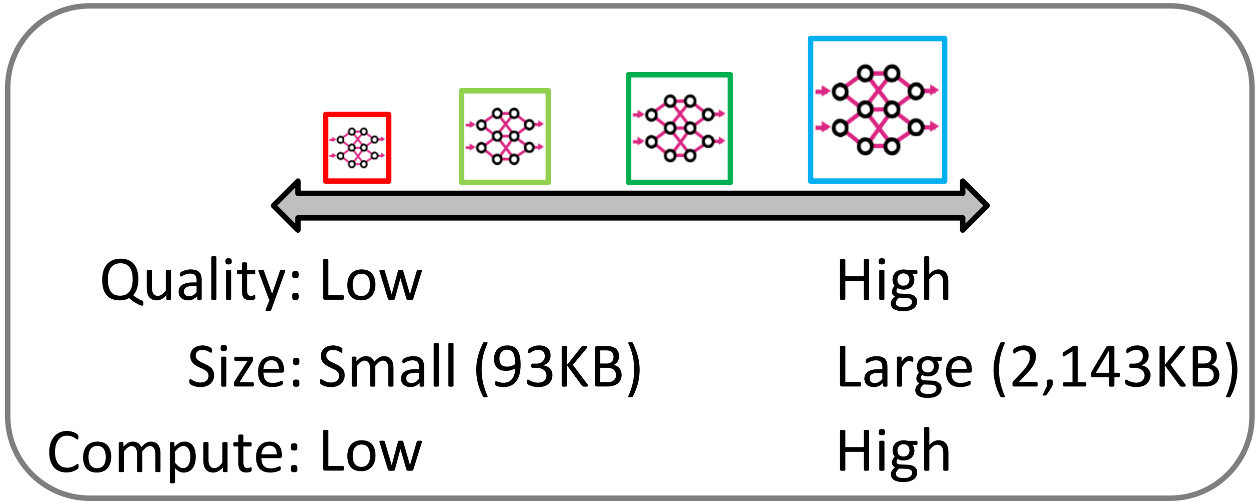
Implication on Video Encoding



Key Design (2): Multiple Quality DNNs

Challenge: Enabling real-time super-resolution on heterogeneous clients

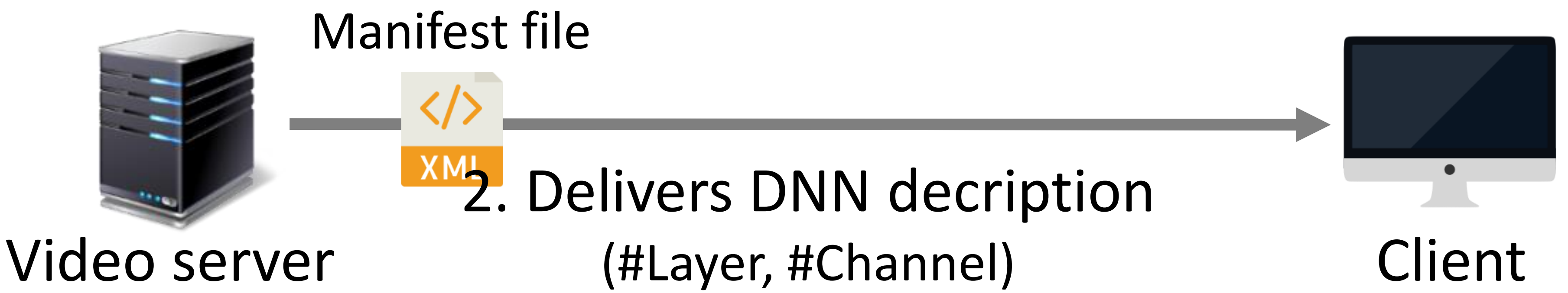


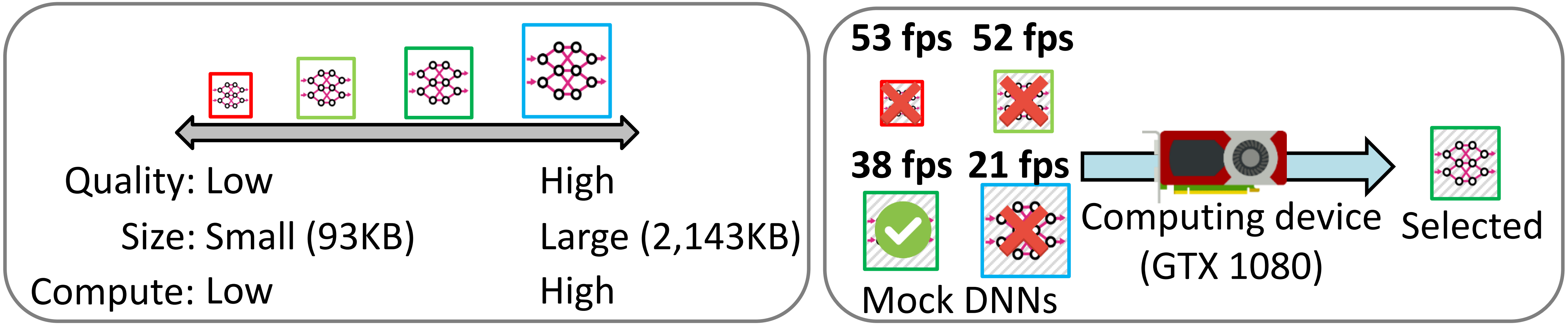


1. Provides multiple quality DNN options

Key Design (2): Multiple Quality DNNs

Challenge: Enabling real-time super-resolution on heterogeneous clients





1. Provides multiple quality DNN options

3. Test-runs and selects the highest-quality running at real-time

NAS: Two Additional Challenges

NAS streams video with a content-aware DNN, but …

1. Takes long time to download and utilize a DNN → **Incremental benefit**

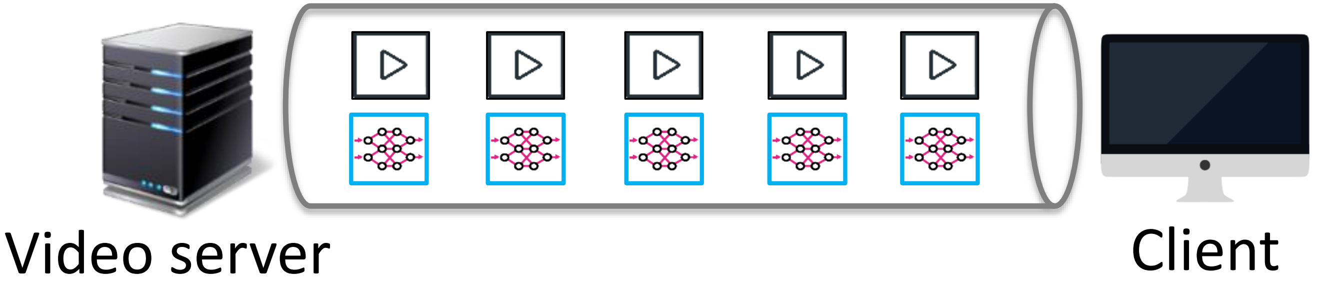
Example

Ultra-high (2,145KB) | 360p video (400Kbps)

1 x = 21 seconds x

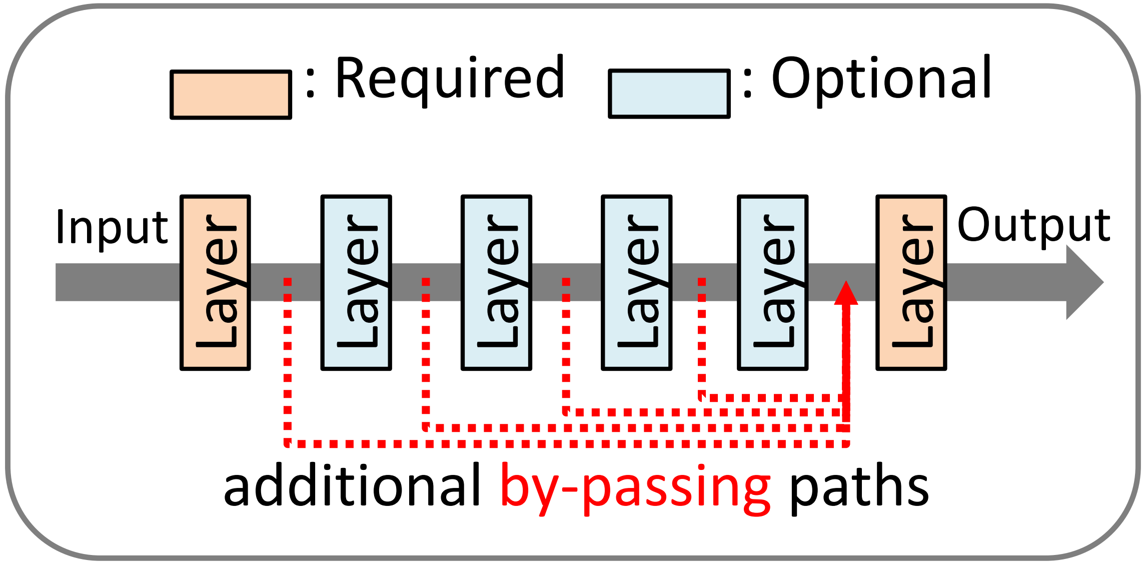
2. A DNN competes bandwidth with video → **Integrate with ABR**

Example

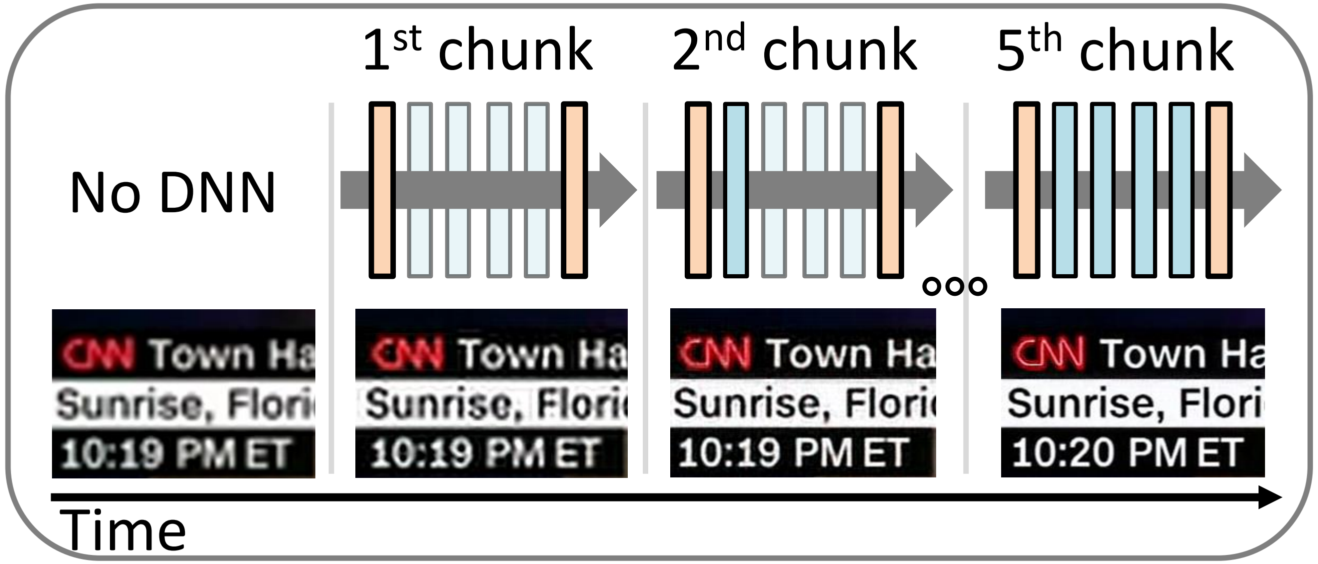


Key Design (3): Scalable DNN

Challenge: Takes a long time to utilize a DNN



1. Implement a scalable DNN
(+ divide into similar-size chunks)

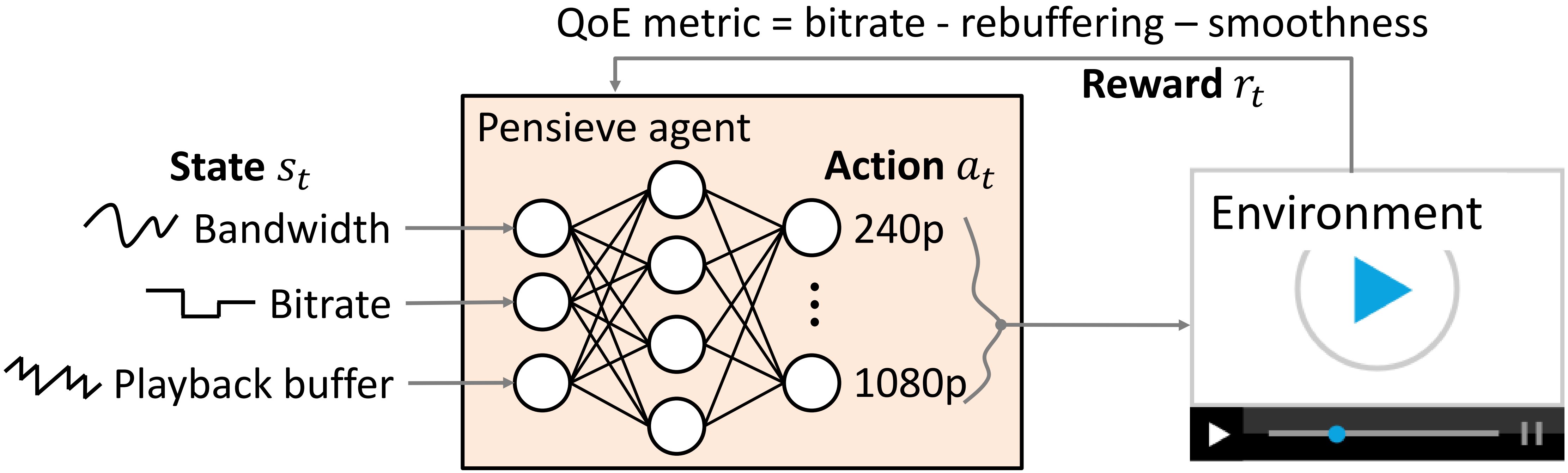


2. Download/Apply a partial DNN

Key Design (4): Integrated ABR

Challenge: How to decide when to download a DNN

- Extends a reinforcement-learning based ABR (Pensieve [SIGCOMM17])

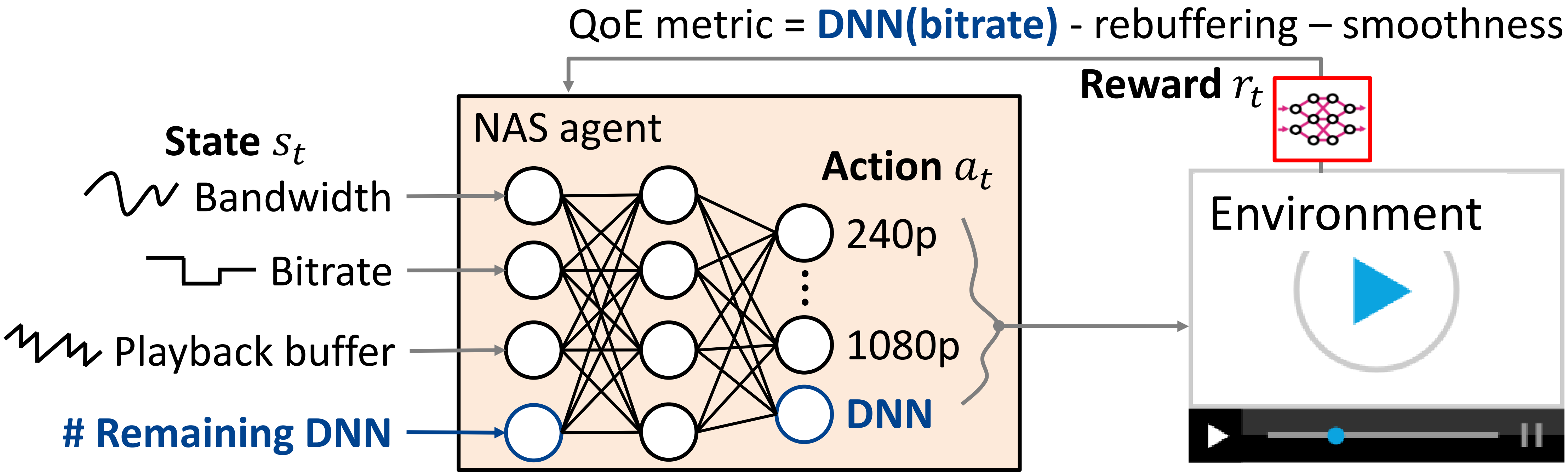


Goal: Maximize the total QoE over an entire video

Key Design (4): Integrated ABR

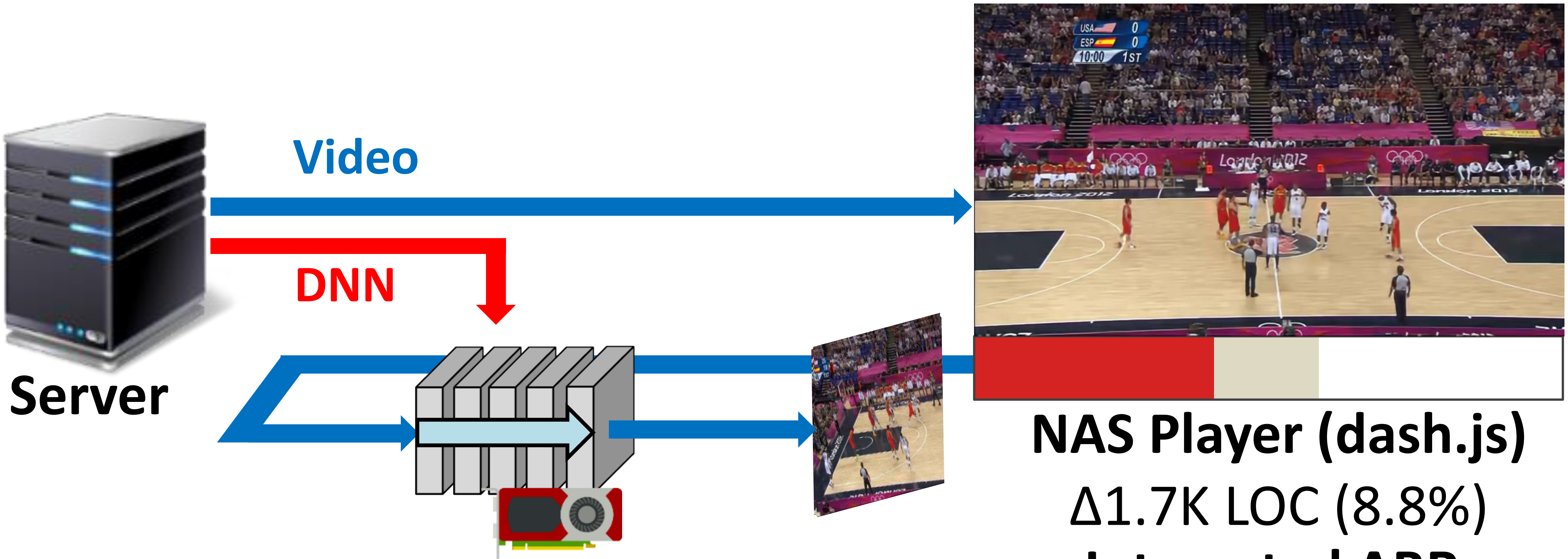
Challenge: How to decide when to download a DNN

- Extends a reinforcement-learning based ABR (Pensieve [SIGCOMM17])



Goal: Maximize the total QoE reflecting DNN-based quality enhancement

Putting All Together: Implementation



DNN Processor (PyTorch)
6.3K LOC

NAS Player (dash.js)
Δ1.7K LOC (8.8%)
Integrated ABR
5.5K LOC

Evaluation

1) How much benefit does NAS deliver?

2) What are the cost and benefit of NAS ?

3) Does NAS effectively adapt to heterogeneous clients?

NAS vs. Existing Video Delivery : QoE

- **17.8 hours real-world network traces**: collected from 3G network and broadband (average bandwidth: 1.31Mbps)
- **27 YouTube videos**: 5-24 minutes, encoded at {400, 800, 1200, 2400, 4800}kbps
- **Computing device:** NVIDIA Titan Xp, **DNN quality:** Ultra-high
- **Video player**: Chromium browser, **Video server**: Apache server

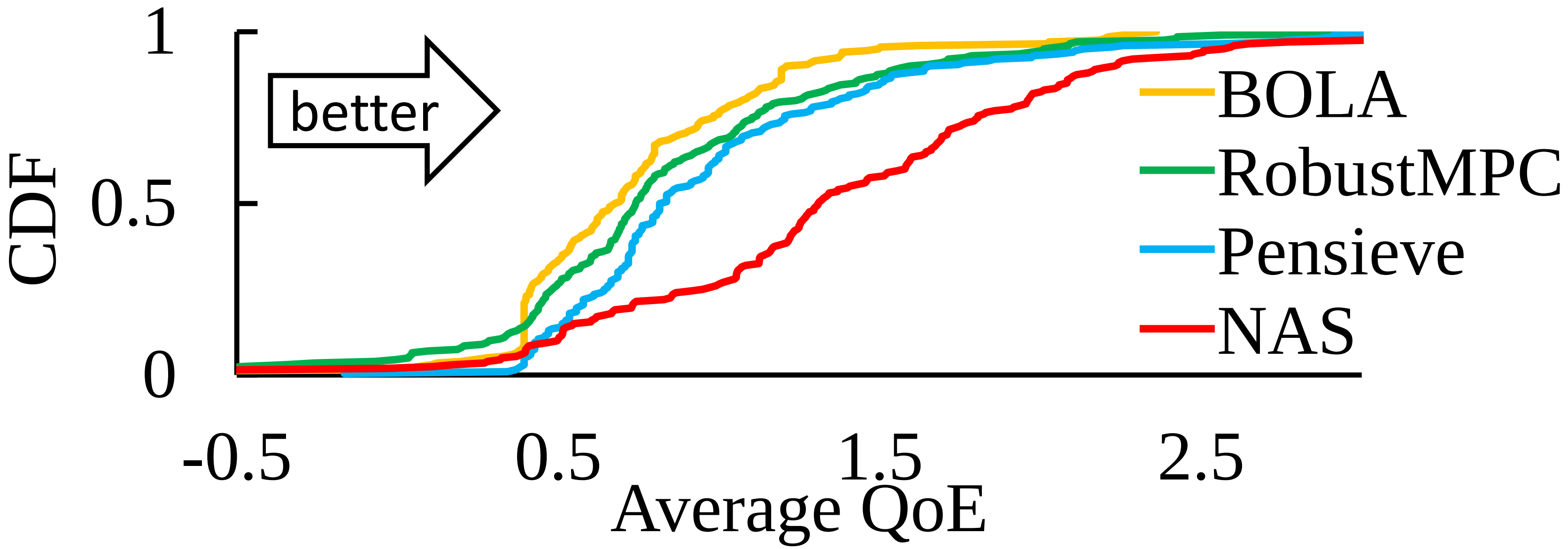
Existing Approach

(Pensieve – SIGCOMM 17)

NAS

NAS vs. Existing Video Delivery : QoE

- **17.8 hours real-world network traces**: collected from 3G network and broadband (average bandwidth: 1.31Mbps)
- **27 YouTube videos**: 5-24 minutes, encoded at {400, 800, 1200, 2400, 4800}kbps
- **Computing device:** NVIDIA Titan Xp, **DNN quality:** Ultra-high
- **Video player**: Chromium browser, **Video server**: Apache server



NAS improves user QoE by 43.08% compared to Pensieve and 92.28% compared to BOLA using same amount of bandwidth.

NAS vs. Existing Video Delivery : Cost

Pensieve CDN

= 0.085$/GB

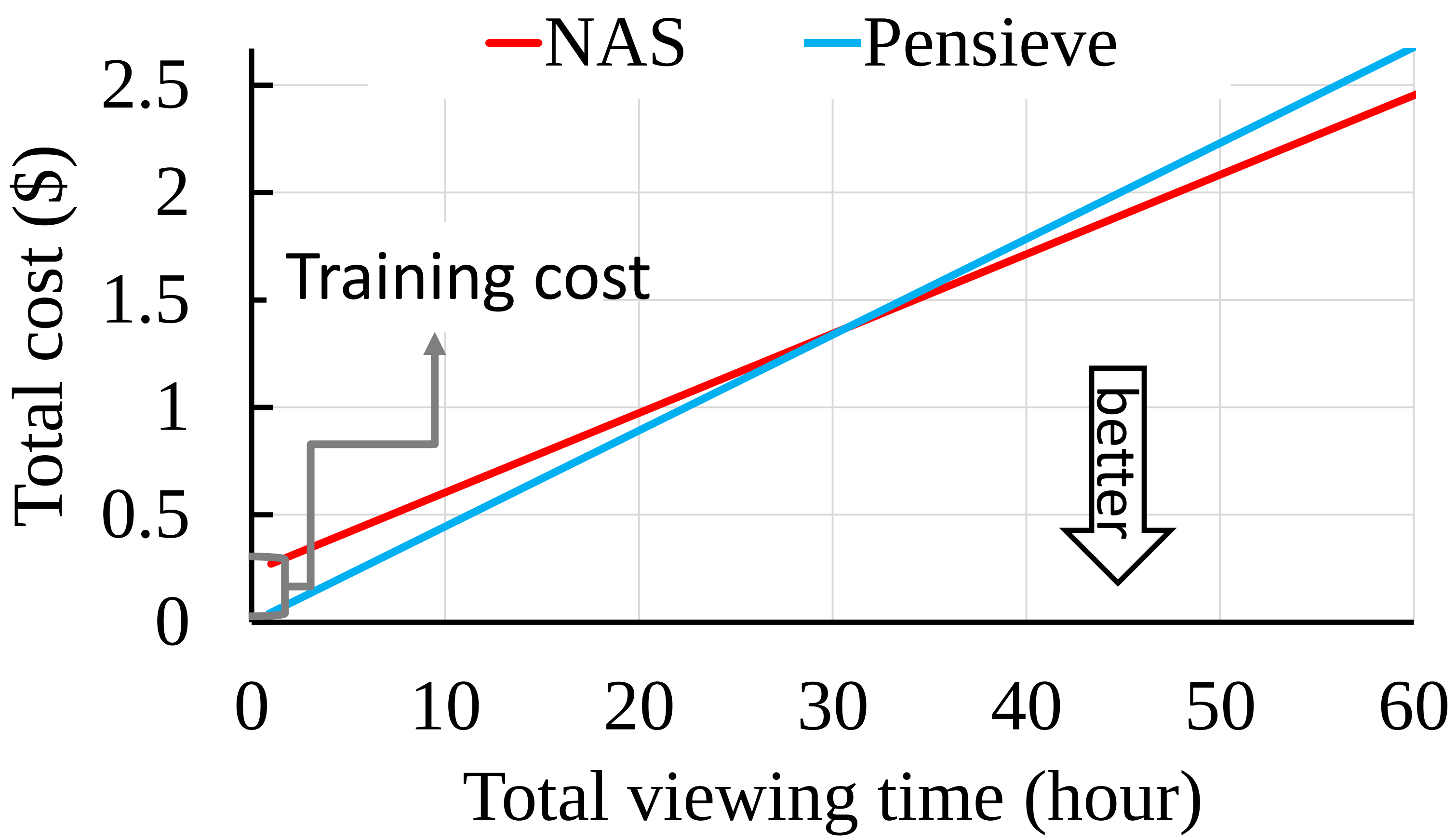
NAS CDN

= 0.085$/GB

: ↓**17.13% bandwidth** for same quality

10 mins ✖ 1.4$/hour

= 0.23$/minute of video

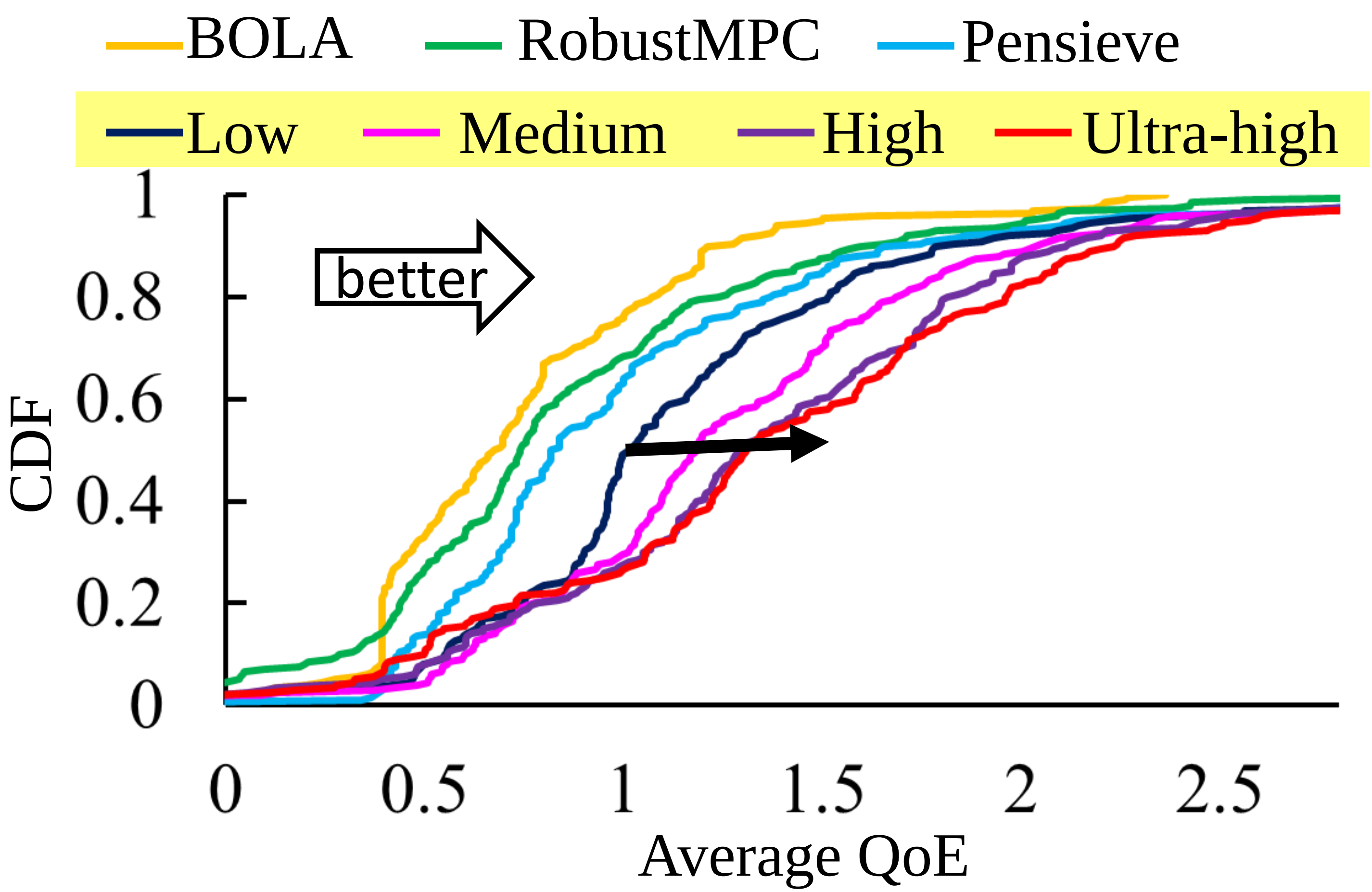


When the total viewing reaches 30 hours (per minute of video), NAS CDN recoups the initial training cost.

Heterogeneous Clients

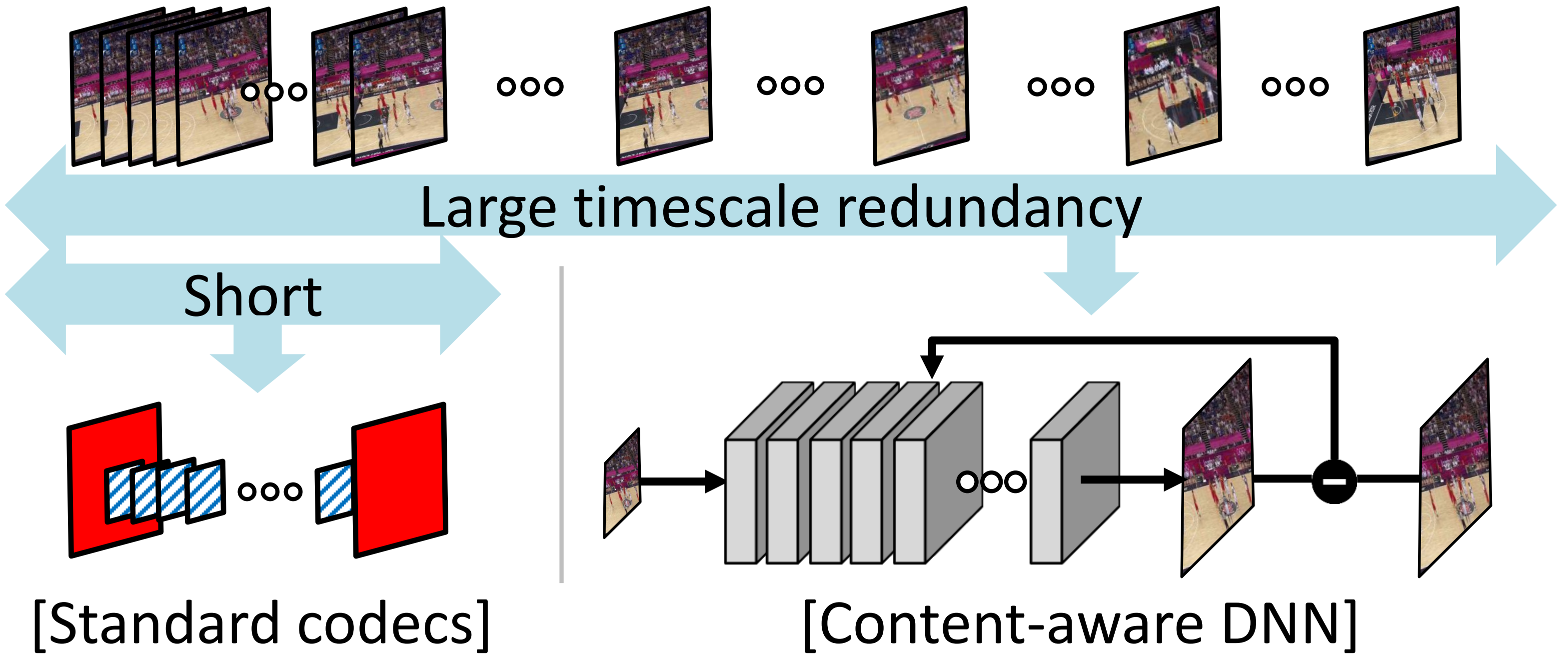
Each GPU processes at real-time
(> 30fps for all resolutions)

DNN quality	GPU model (Price)
Low	GTX 1050 Ti ($139)
Medium	GTX 1060 ($249)
High	GTX 1070 Ti ($449) GTX 1080 ($559)
Ultra-high	GTX 1080 Ti ($669) Titan Xp ($1,200)



NAS adapts to heterogeneous devices,
and a device with higher computing power receives greater benefit.

Conclusion



- NAS presents a new type of QoE maximization & encoding via DNN
- NAS accommodates four key designs: Content-aware DNN, Multiple quality DNNs, Scalable DNN, Integrated ABR.
- NAS can improve user QoE or reduce the video delivery cost for CDN.