How will Deep Learning Change Internet Video Delivery?

Hyunho Yeo

Ph.D. Candidate@KAIST

- Work on DL-based video delivery
- Publish papers at top-tier system/network conferences

Neural Adaptive Content-aware Internet Video Delivery
Hyunho Yeo, Youngmok Jung, Jaehong Kim, Jinwoo Shin, **Dongsu Han**
USENIX OSDI 2018

How will Deep Learning Change Internet Video Delivery?
Hyunho Yeo, Sunghyun Do, **Dongsu Han**
ACM HotNets 2017

Overview

"How will Deep Learning Change Internet Video Delivery?"

1. Observation/Limitation of Current Video Delivery

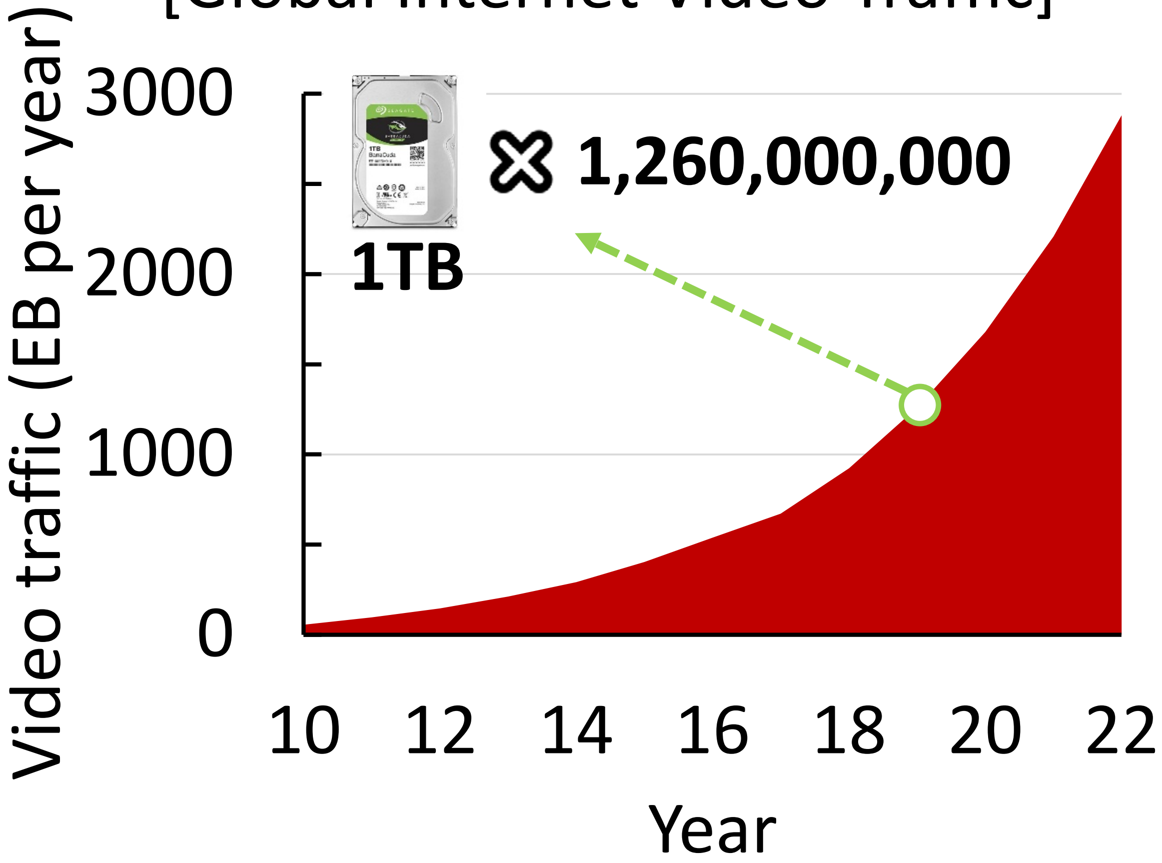
2. Recent research: DL-based adaptive streaming [OSDI 18]

3. Vision of DL-based Video Delivery

Era of Internet Video Delivery

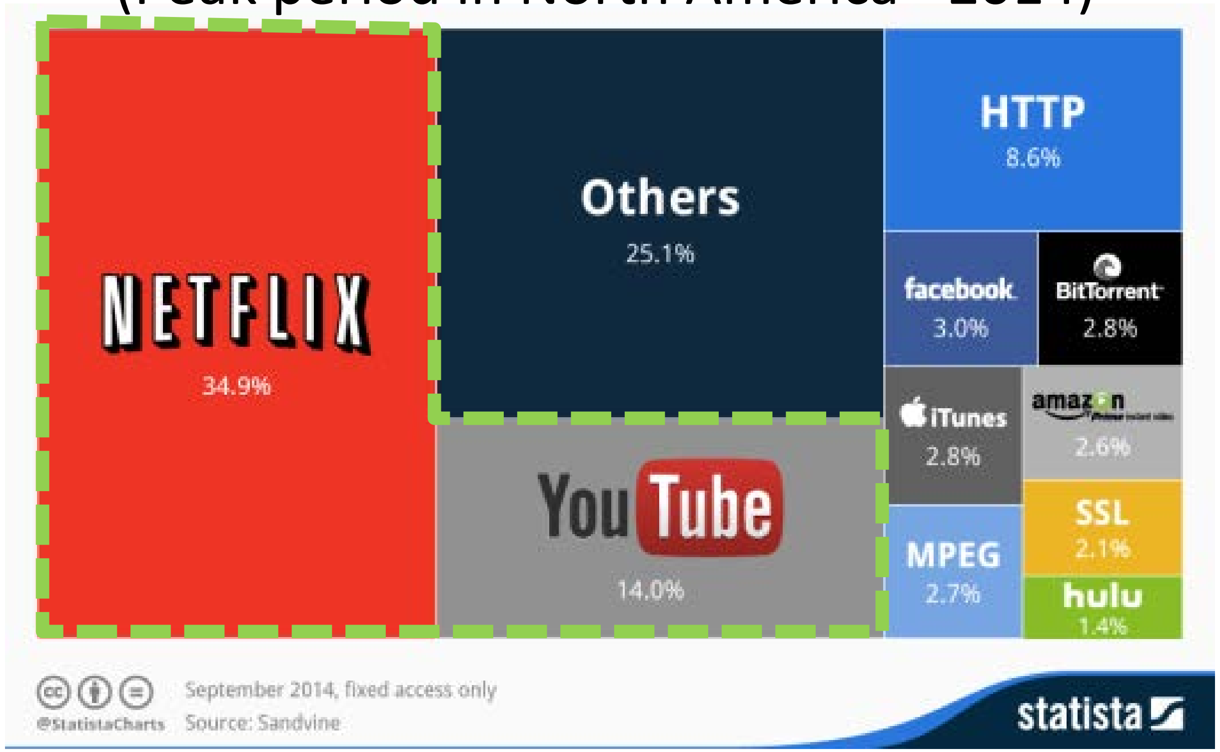
Internet video traffic has *exponentially* grown over last decade!

[Global Internet Video Traffic][1]



[Percentage of downstream traffic][2]

(Peak period in North America - 2014)

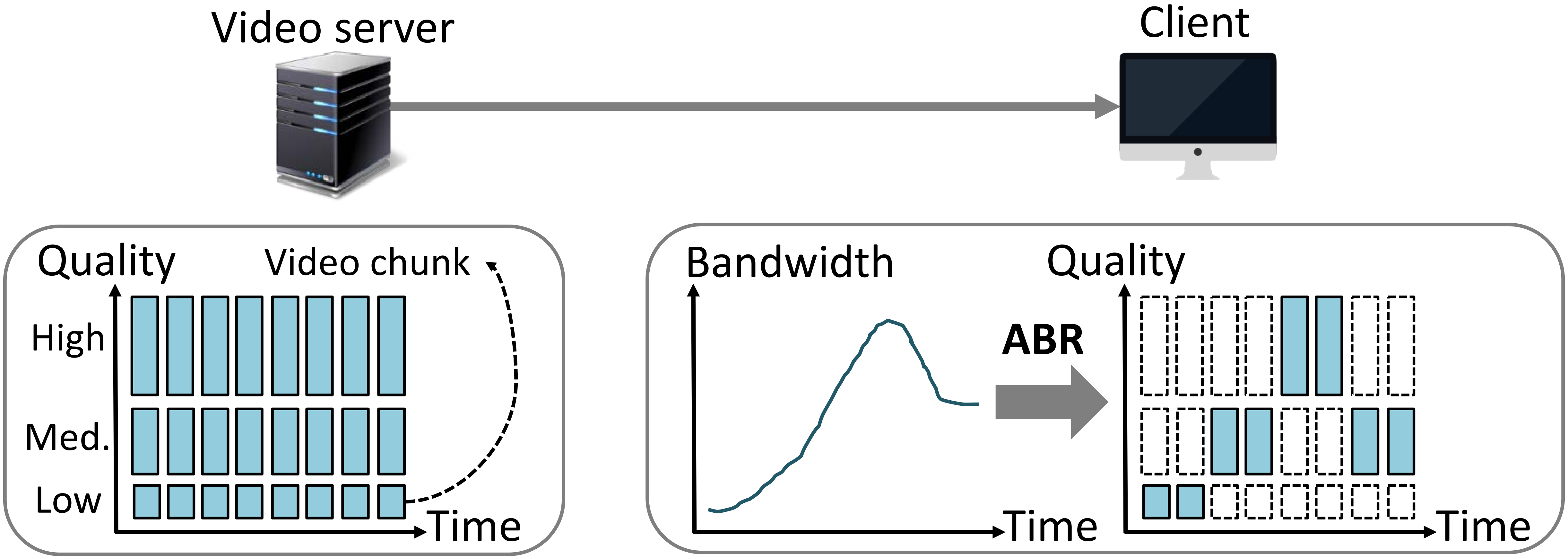


1: CISCO Visual Networking Index, 16 data was interpolated

2: https://digitalbusinessblog.wordpress.com/2014/11/25/who-are-the-biggest-bandwidth-hogs/

Observation on Current Video Ecosystem

To handle bandwidth heterogeneity over space and time,
Adaptive streaming has been widely deployed



Traditional Approaches

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

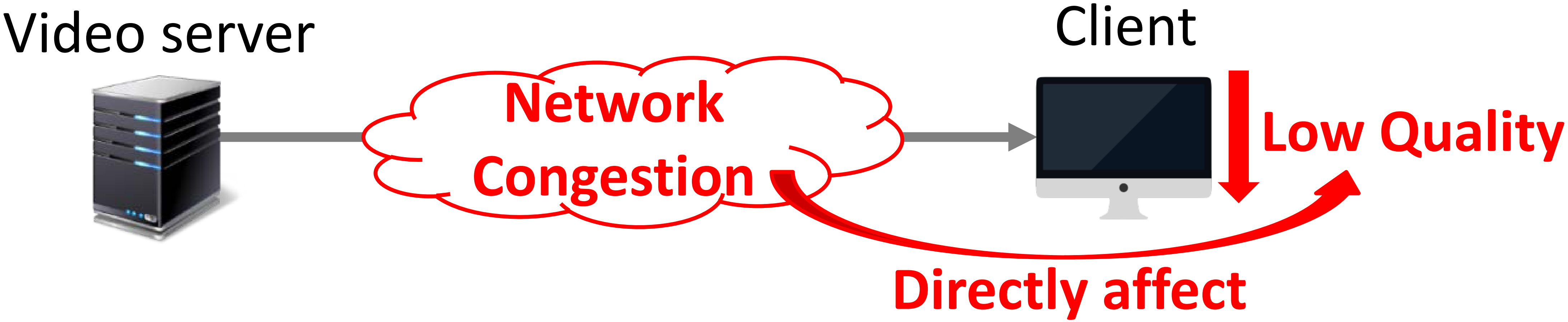
Choosing better servers, CDNs
Content Multihoming [SIGCOMM 12], VDN [SIGCOMM 15]

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

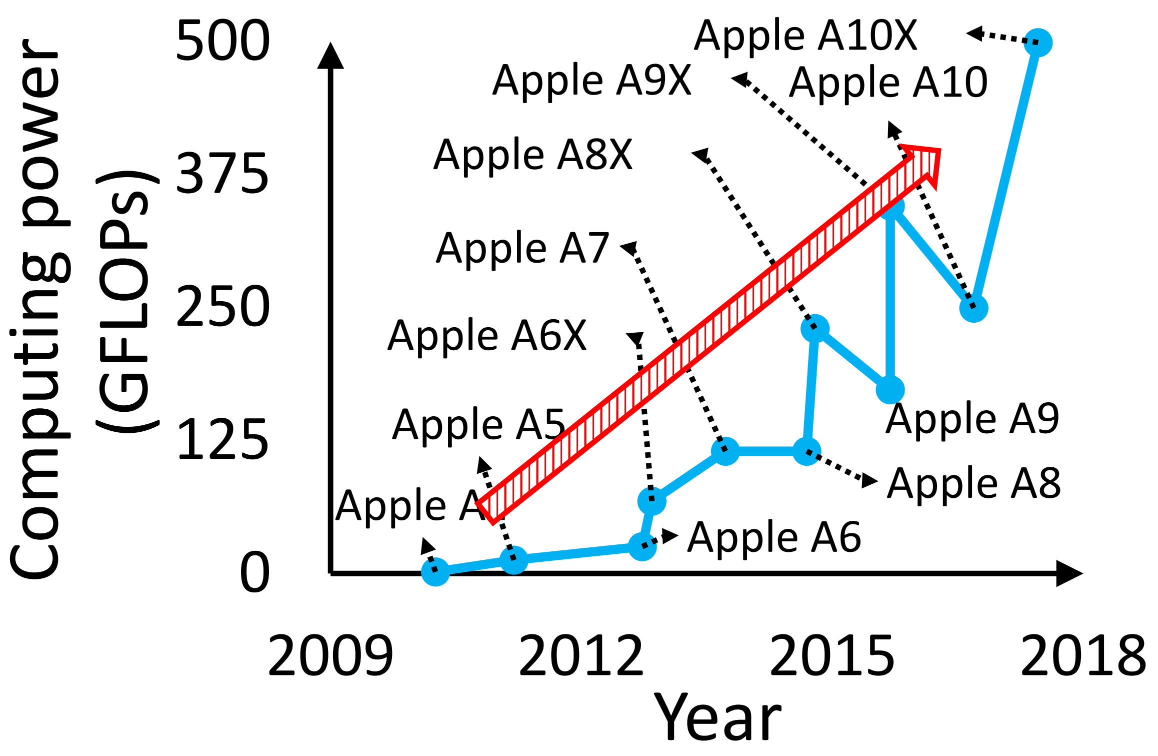
Video quality heavily depends on available bandwidth



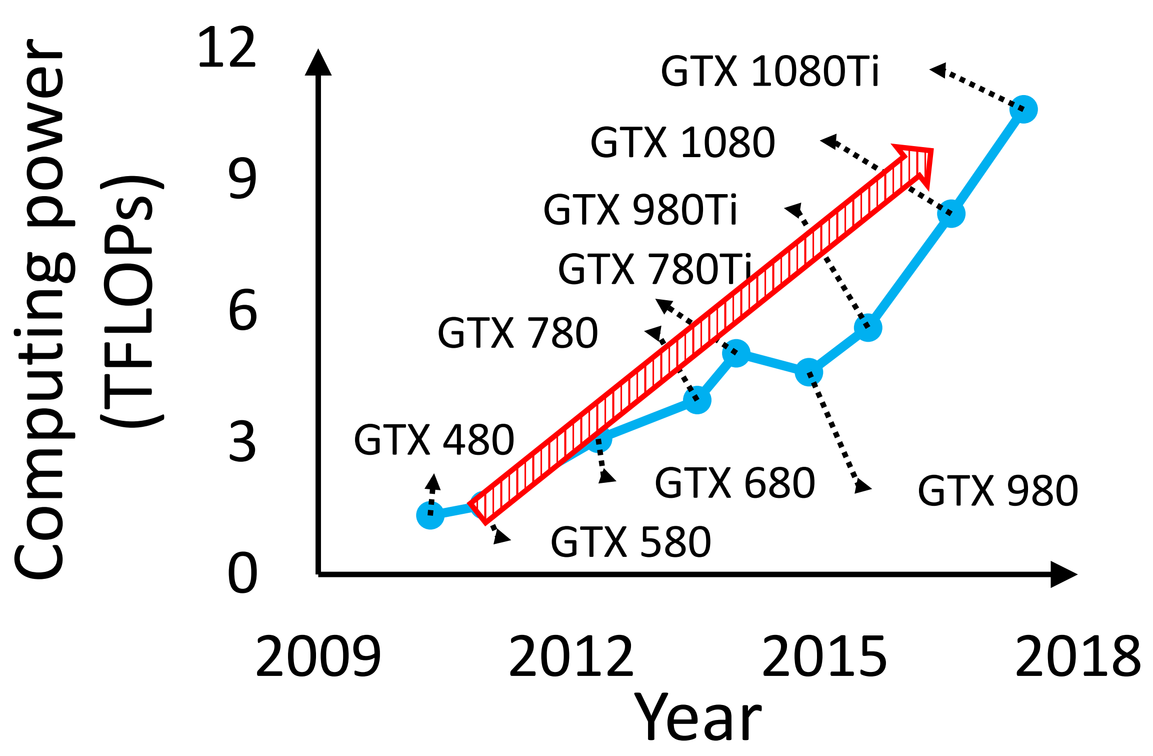
Limitation of Current Video Delivery

Client computing power is scarcely utilized other than for decoding

Mobile GPU

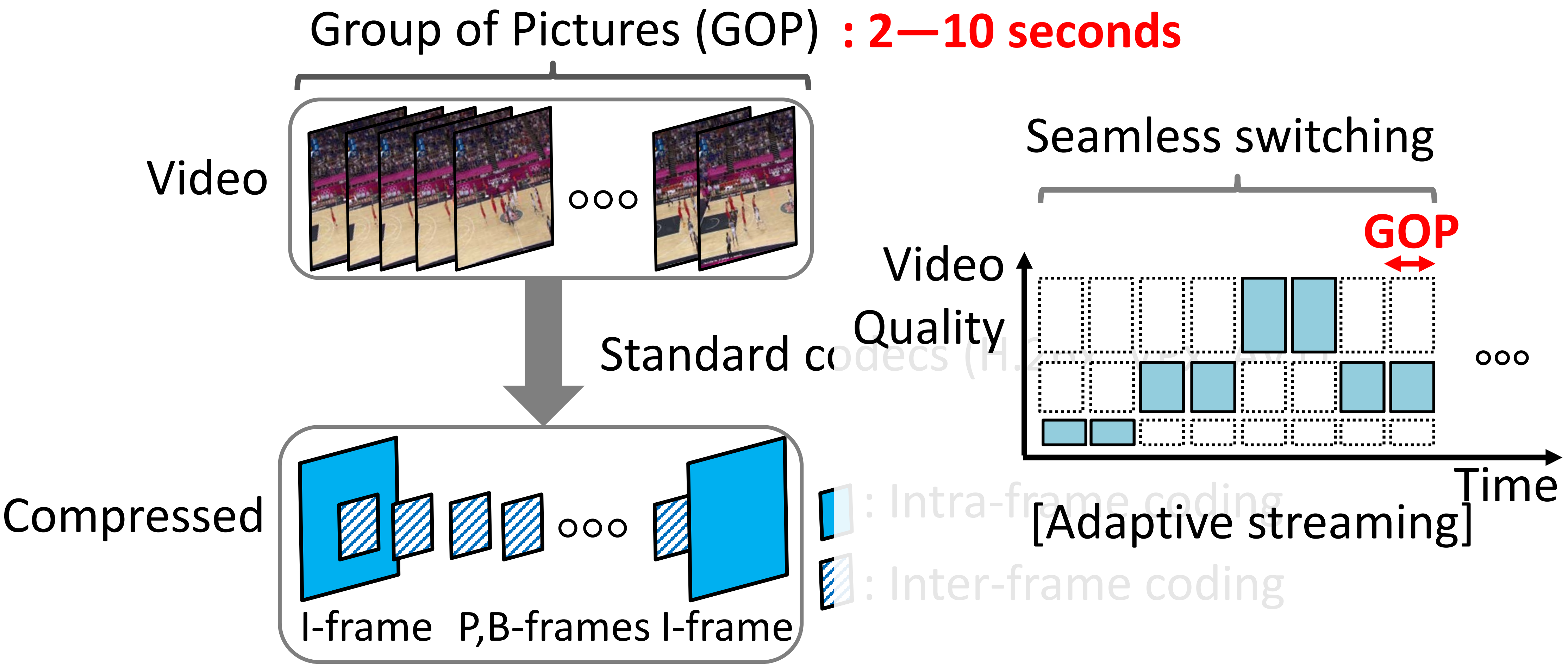


Desktop GPU

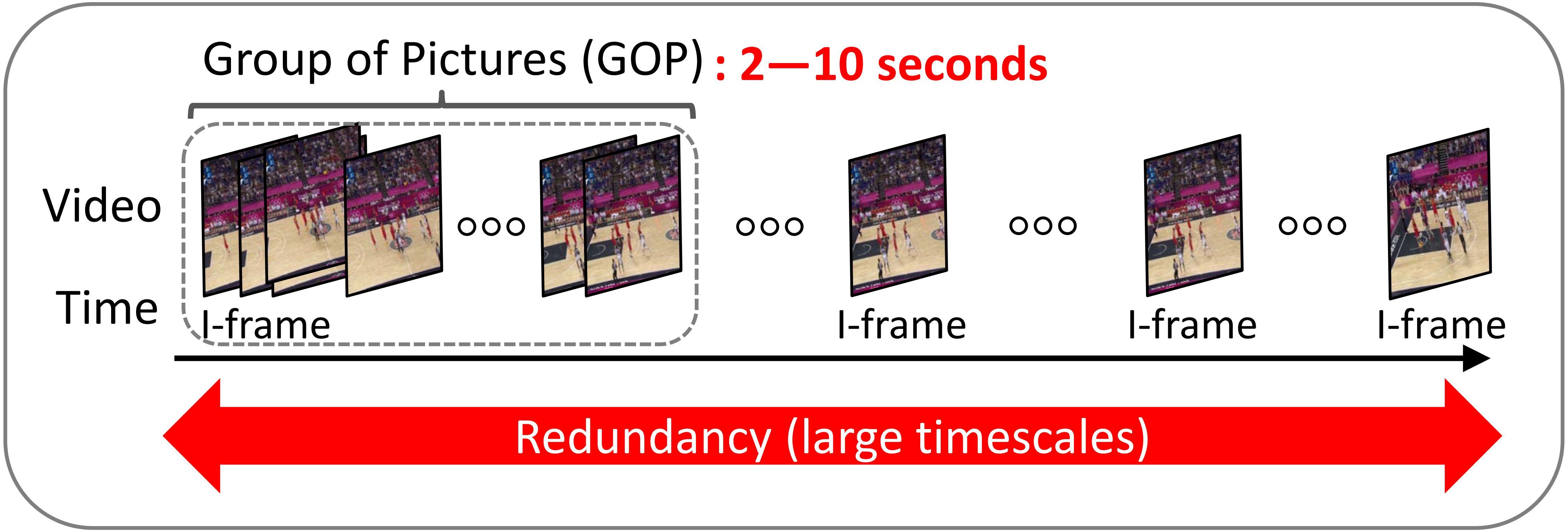


Observation on Current Video Ecosystem

Standard codecs efficiently reduce redundancy *only* inside GOP



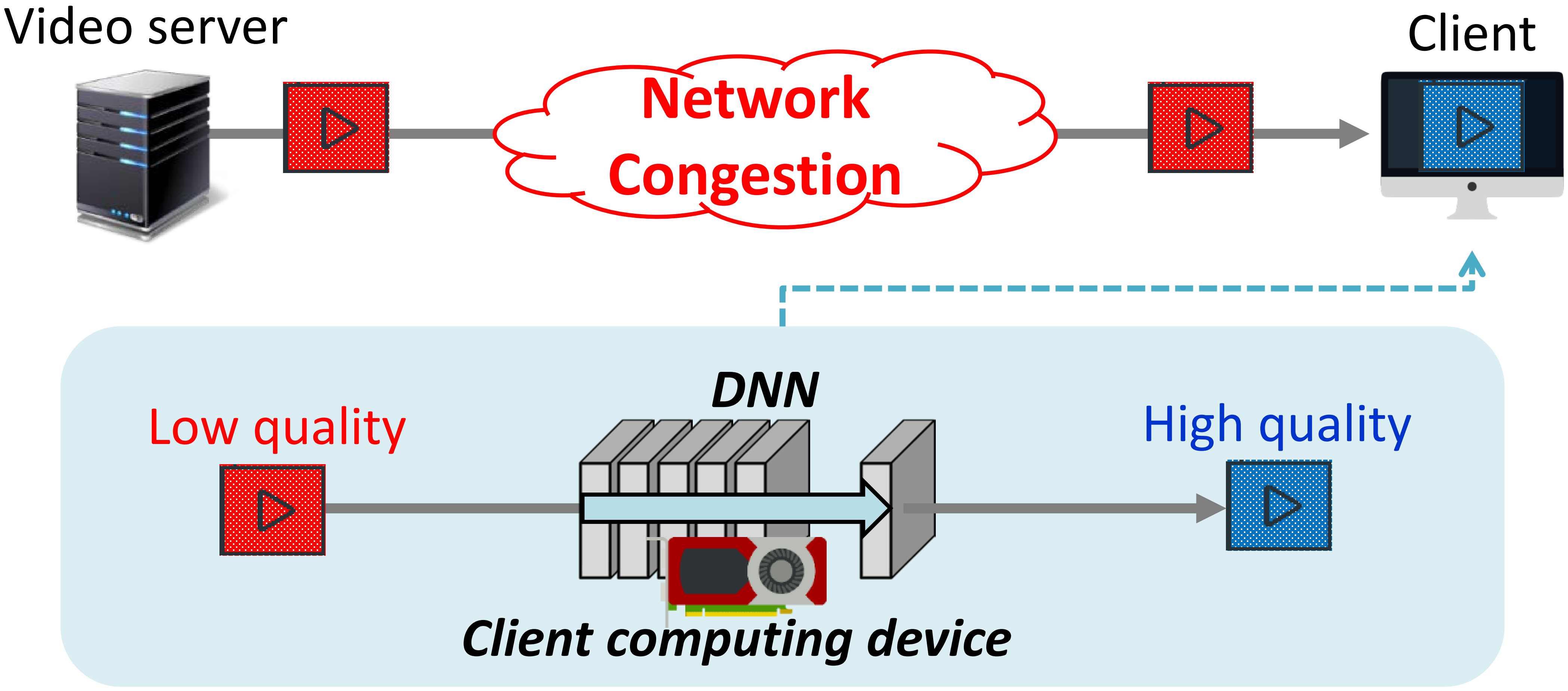
Limitation of Current Video Delivery



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

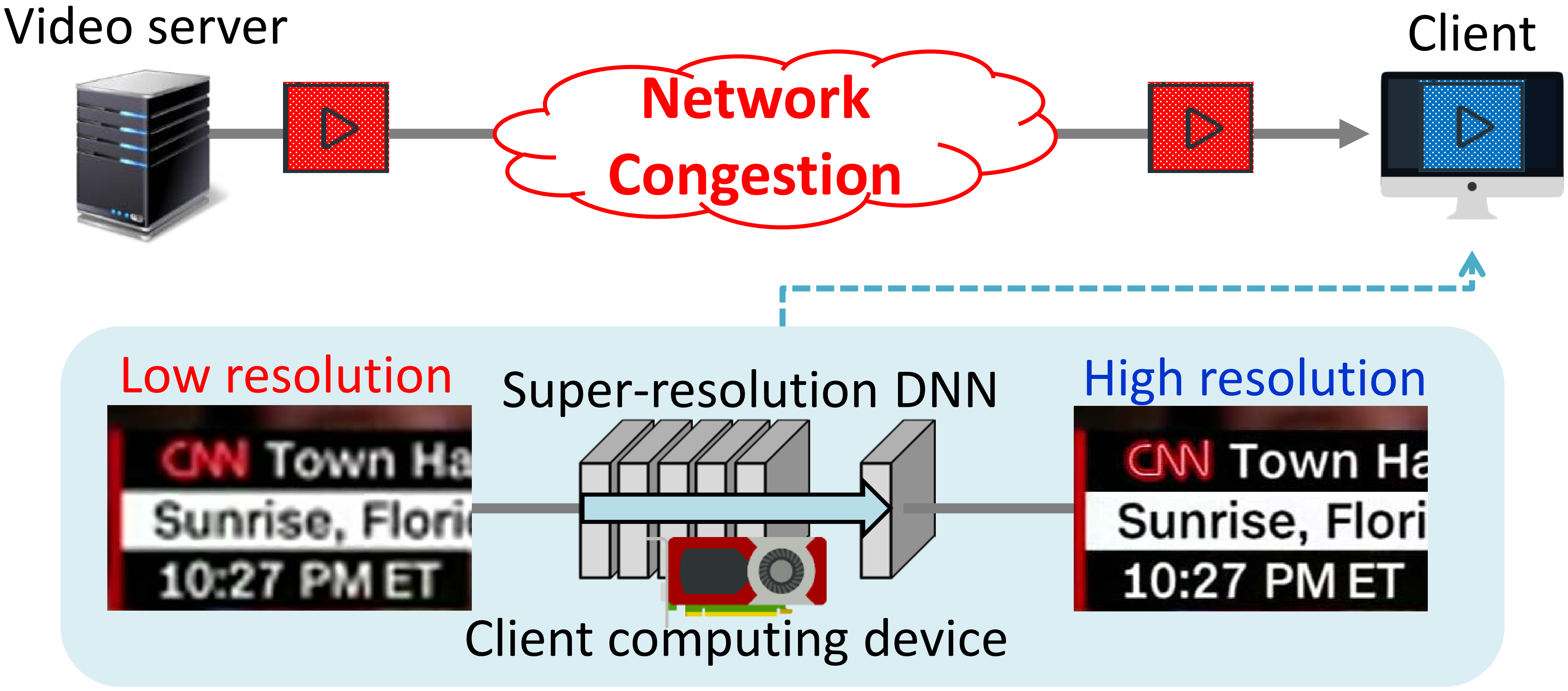
What Deep Neural Network (DNN) Can Do?

1. Utilize **client computation** to enhance video quality



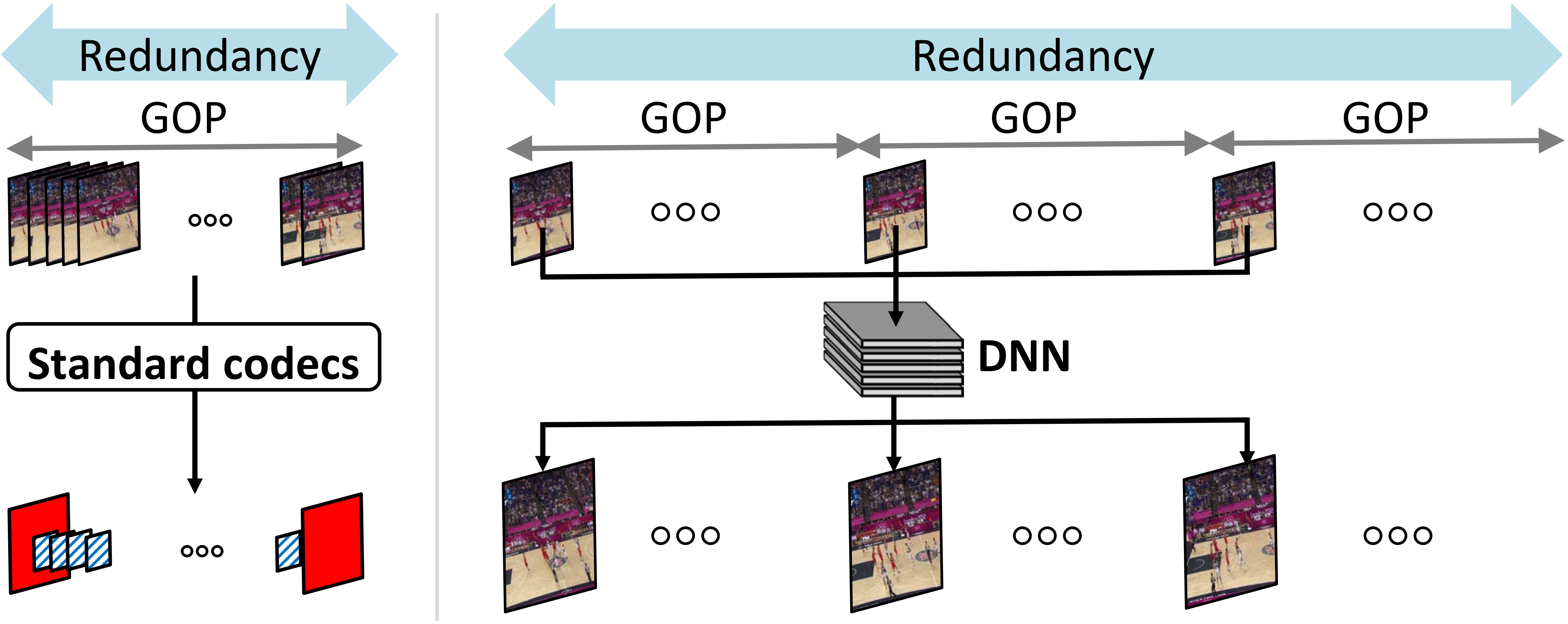
What Deep Neural Network (DNN) Can Do?

1. Utilize **client computation** to enhance video quality



What Deep Neural Network (DNN) Can Do?

2. Trained and operate in **large timescales** (video)



What Deep Learning (DL) can Do?

Can we overcome the current limitations via DNN?

How much quality improvement can we achieve?

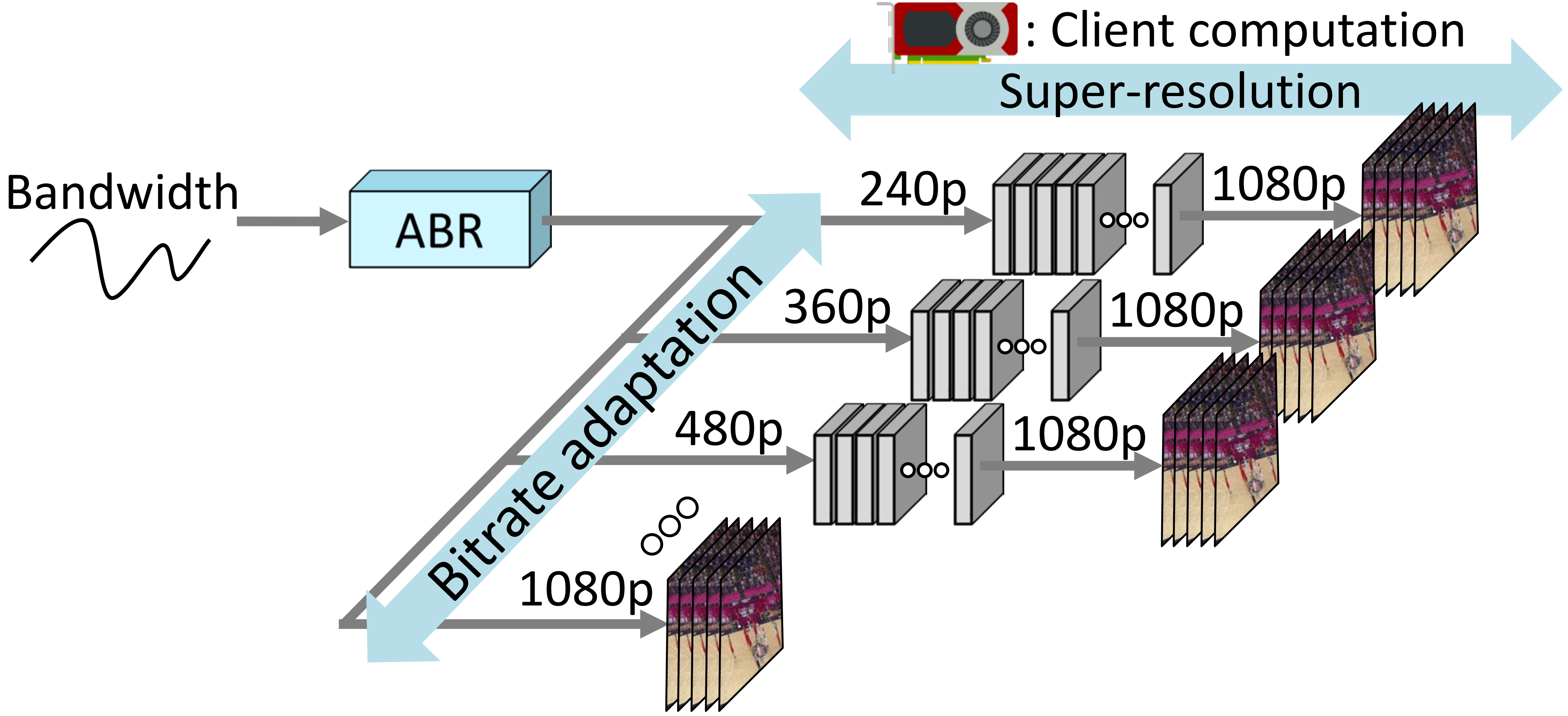
To answer these, let's move to our recent research, NAS [OSDI18]

Neural Adaptive Content-aware Internet Video Delivery

Hyunho Yeo Youngmok Jung Jaehong Kim Jinwoo Shin Dongsu Han
KAIST

NAS: DL-based Adaptive Streaming

Apply super-resolution DNN on top of bitrate adaptation




0 SSG
2.5k
0
Pensieve
2017 WORLD CHA
acer


0 SSG
0
2.5k
0
NEXT DRAKE
MOUNTAIN
NAS
SSG CuVee
SSG Ambition
SSG Crown
SSG Ruler
SSG CoreJJ
2017 WORLD CHAN
0/0/0
acer
RULER

NAS: System Target

1. Content: Video on demand (VOD)

Example

YouTube NETFLIX hulu amazon instant video

2. Computing device: **NVIDIA** GTX 10 series

NVIDIA GEFORCE GTX

Example

GTX 1050 Ti (Entry-level) | Titan Xp (High-end)

○○○

Price

$139 → $1,200

NAS: Two Initial Challenges

⚠️ NAS utilizes DNN and client computation, but …

NAS: Two Initial Challenges

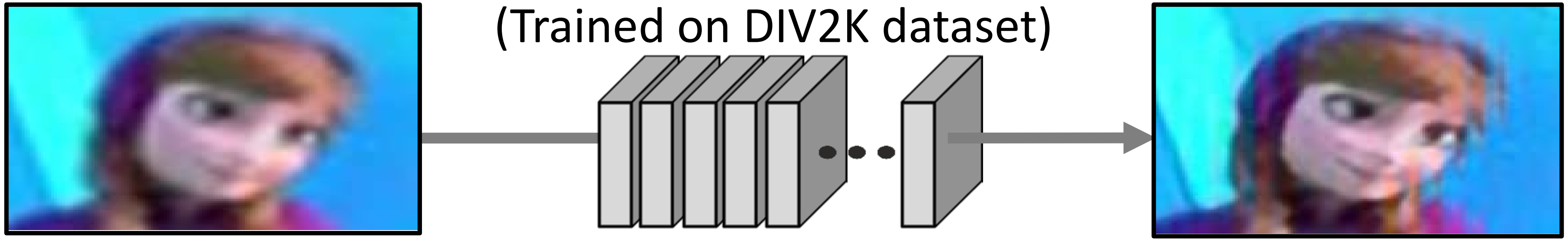
⚠ NAS utilizes **DNN** and client computation, but …

1. DNN testing accuracy is unreliable for unseen/new content
   - Even worse, degradation can occur (below figure)

Unseen content

EDSR [CVPRW 17]
(Trained on DIV2K dataset)

↓Quality



SSIM = 0.86

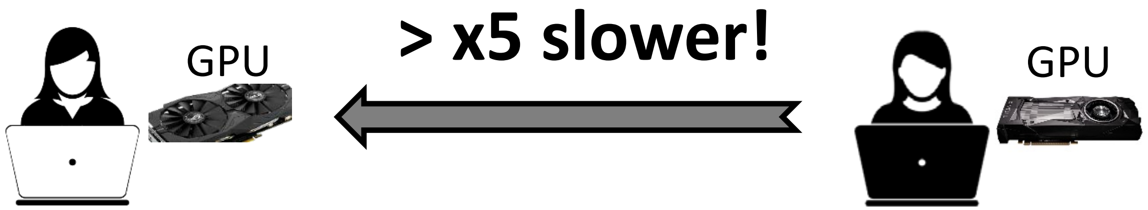
SSIM = 0.84

For the real-world deployment, DNN accuracy should be **guaranteed**

NAS: Two Initial Challenges

⚠ NAS utilizes DNN and **client computation**, but …

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



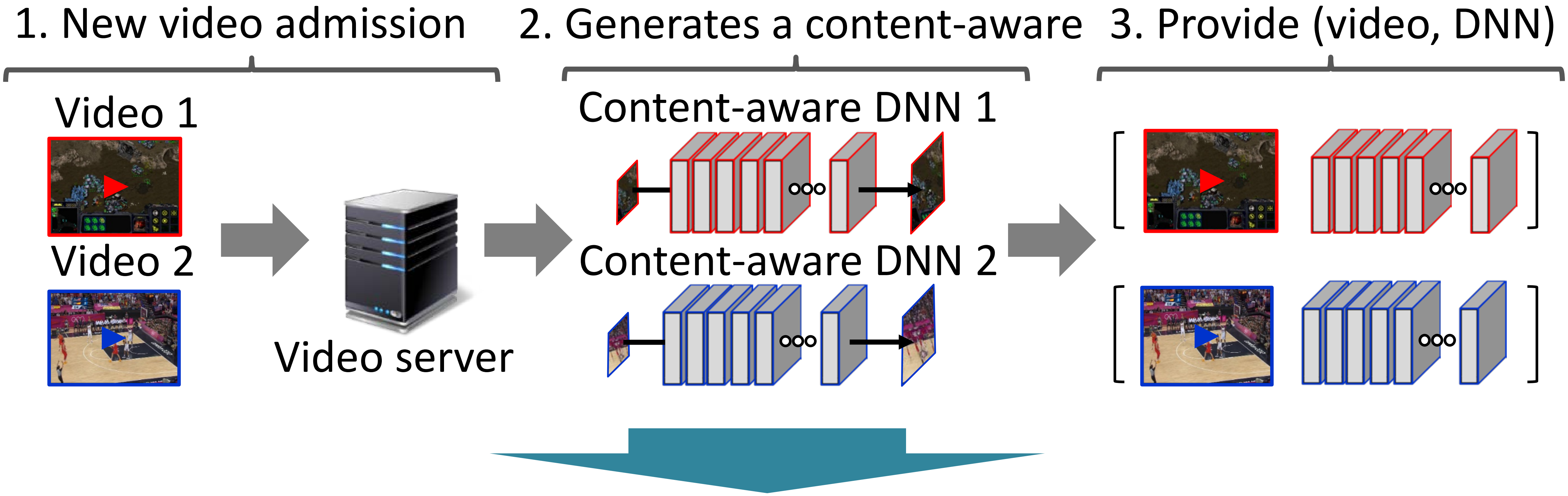
Client A: Entry-level GPU
(1.98 TFLOPS – 1050 Ti)

Client B: High-end GPU
(10.79 TFLOPS – Titan Xp)

Adaptation to computing power is required

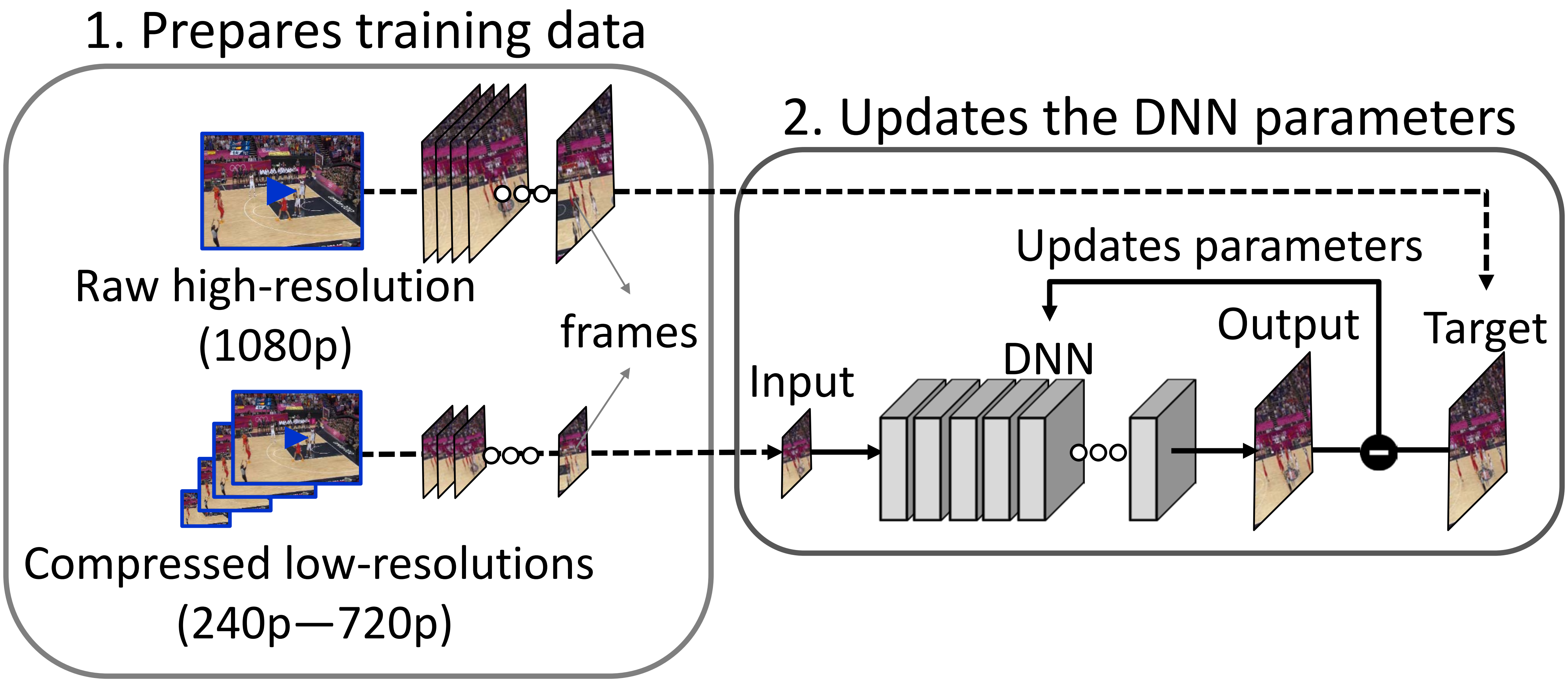
Key Design 1: Content-aware DNN

Challenge: Providing reliable DNN quality

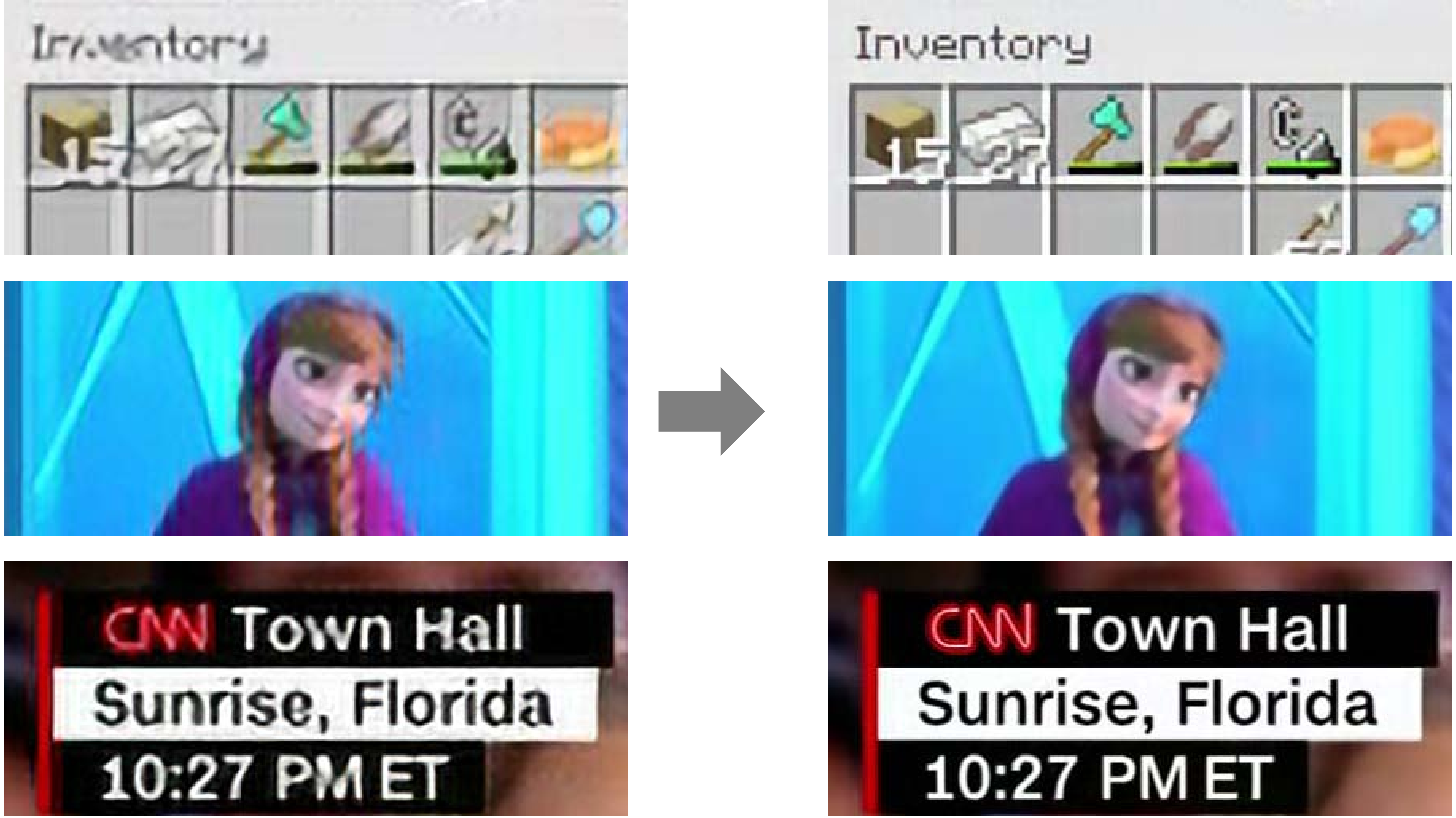


Content-aware DNN delivers the reliable (over-fitted) **training accuracy** instead of the unpredictable **testing accuracy**.

Training a content-aware super-resolution



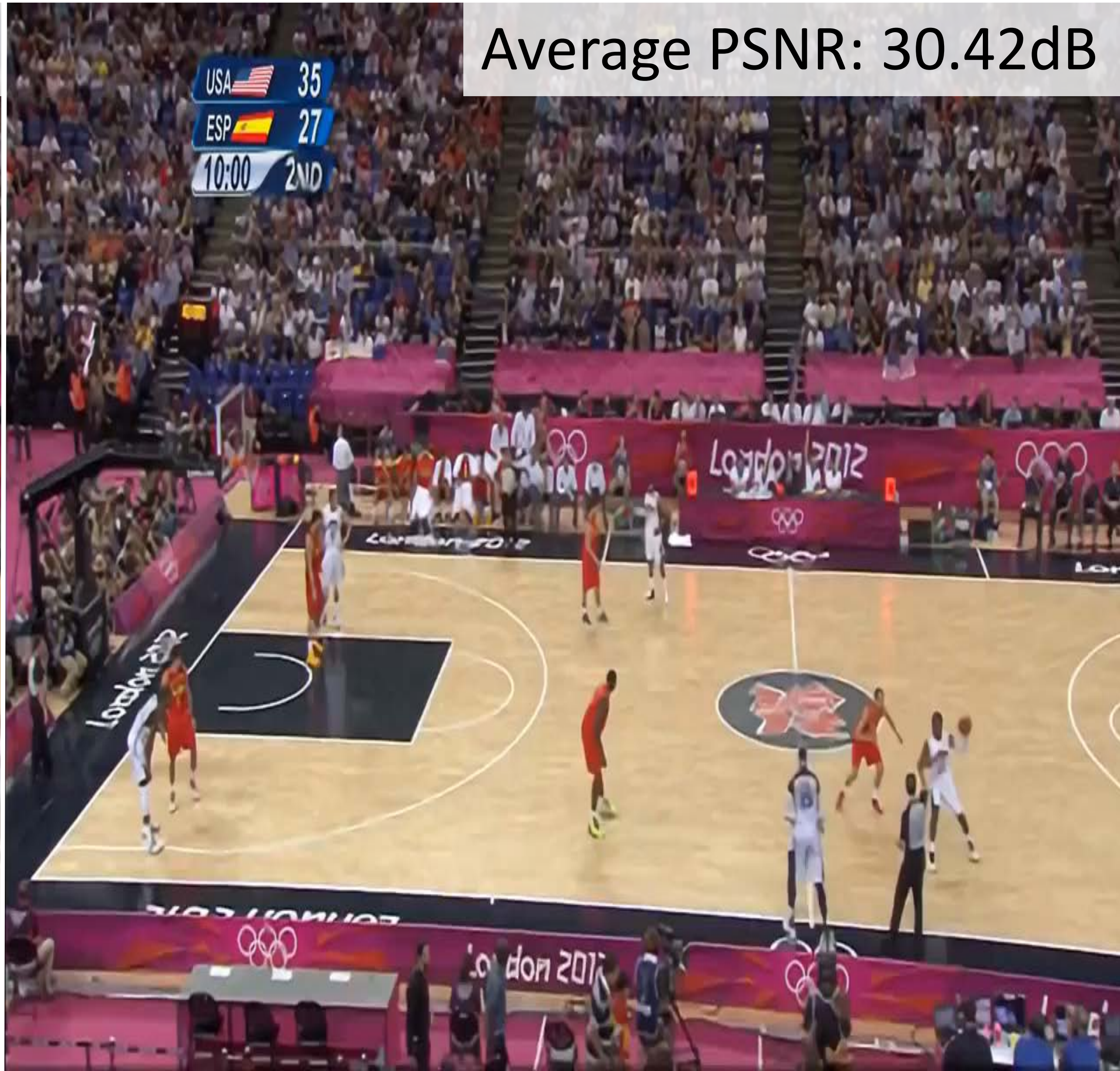
Content-agnostic vs. Content-aware



“PSNR 2~4 dB gain over content-agnostic”

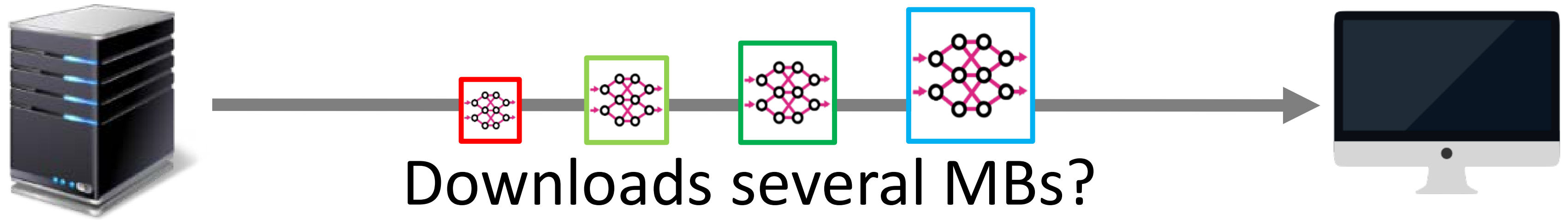
Bicubic vs. Content-aware DNN

Bicubic vs. Content-aware DNN



Key Design 2: Multiple Quality DNNs

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

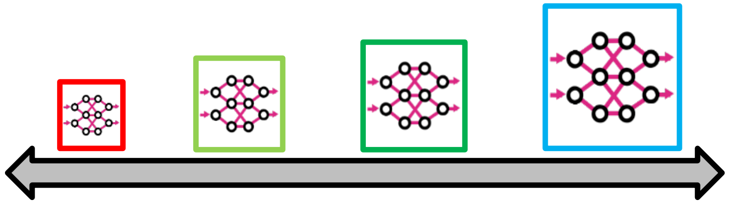


Video server

Downloads several MBs?

→ Delay video streaming

Client



Quality:	Low	High
Size:	Small (93KB)	Large (2,143KB)
Compute:	Low	High

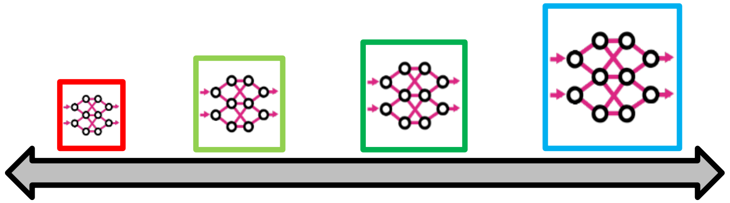
1. Provides multiple quality DNN options

Key Design 2: Multiple Quality DNNs

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

2. Delivers **DNN description**

(#Layer, #Channel)



Quality: Low — High

Size: Small (93KB) — Large (2,143KB)

Compute: Low — High

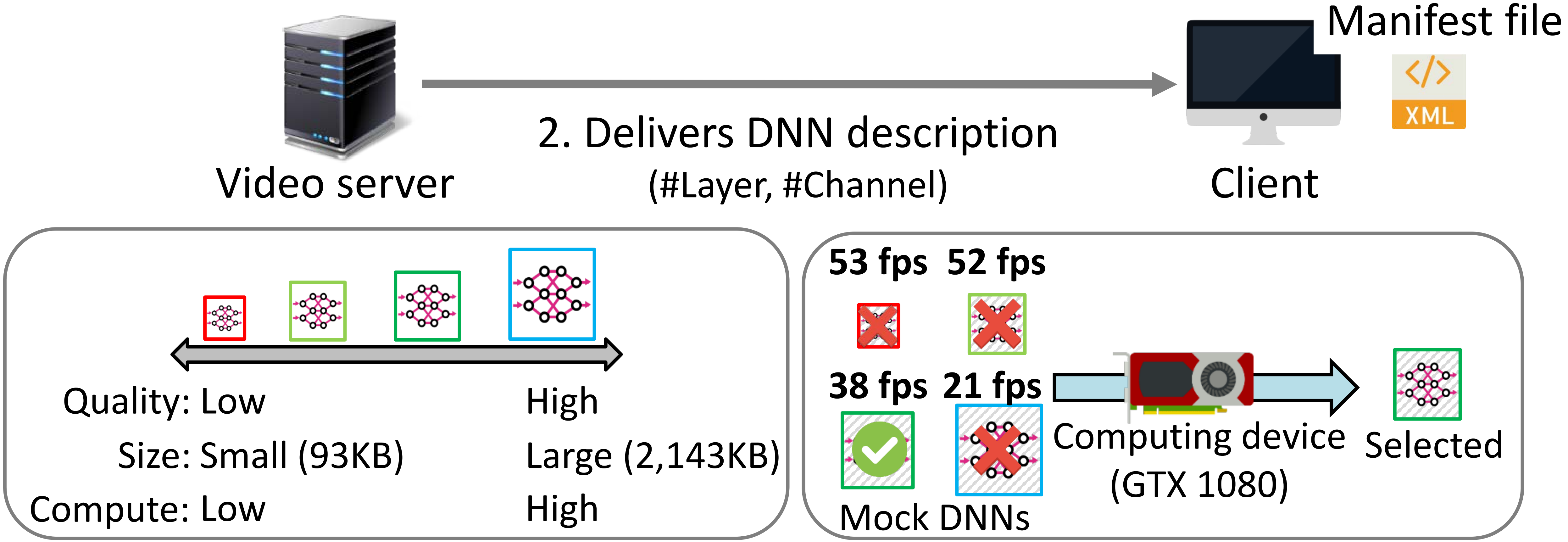
1. Provides multiple quality DNN options

MPD (Media Presentation Description)

- Period
  - Adaptation Set (Video)
    - 1080p 4.8Mb/s
    - 720p 2.4Mb/s
    - 480p 1.2Mb/s
    - ooo
  - **Adaptation Set (DNN) (#layer, #filter)**
    - Low-240p (20, 9)
    - Med.-240p (20, 21)
    - High-240p (20, 32)
    - ooo

Key Design 2: Multiple Quality DNNs

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

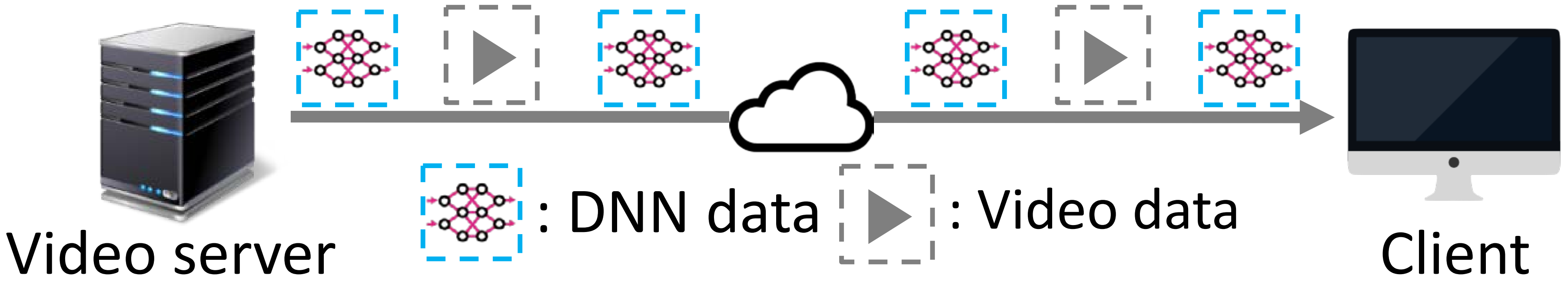


1. Provides multiple quality DNN options

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

NAS: Two Additional Challenges

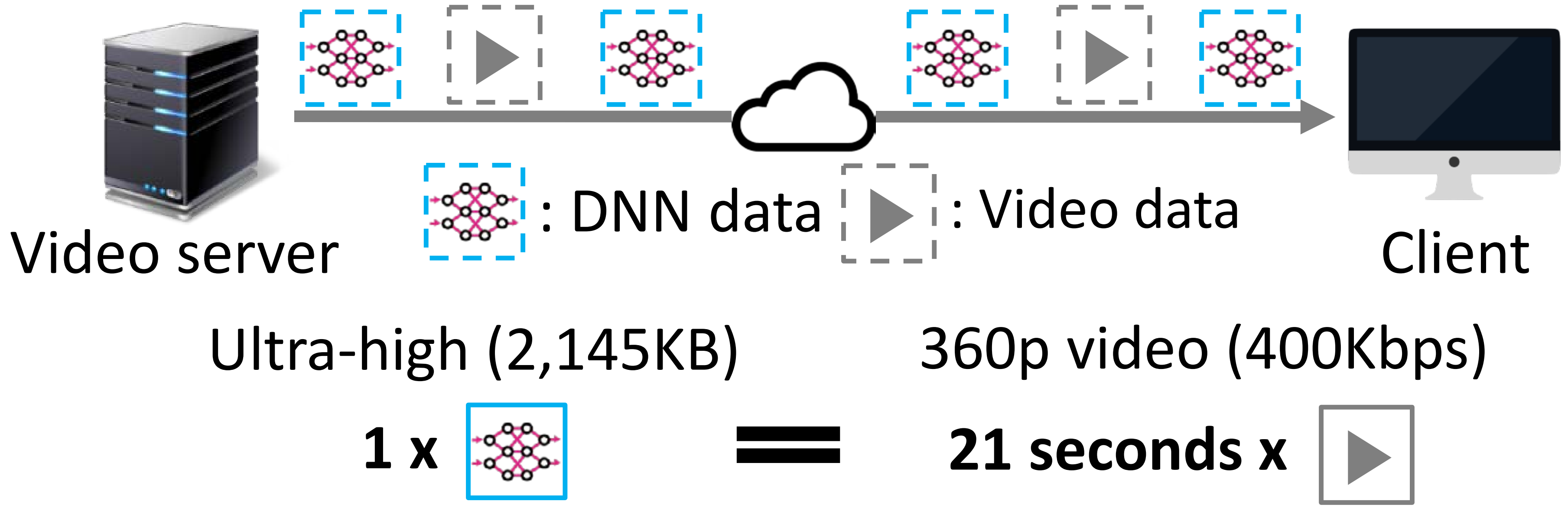
NAS streams video with a content-aware DNN, but …



NAS: Two Additional Challenges

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

1. Takes long time to download and utilize a DNN

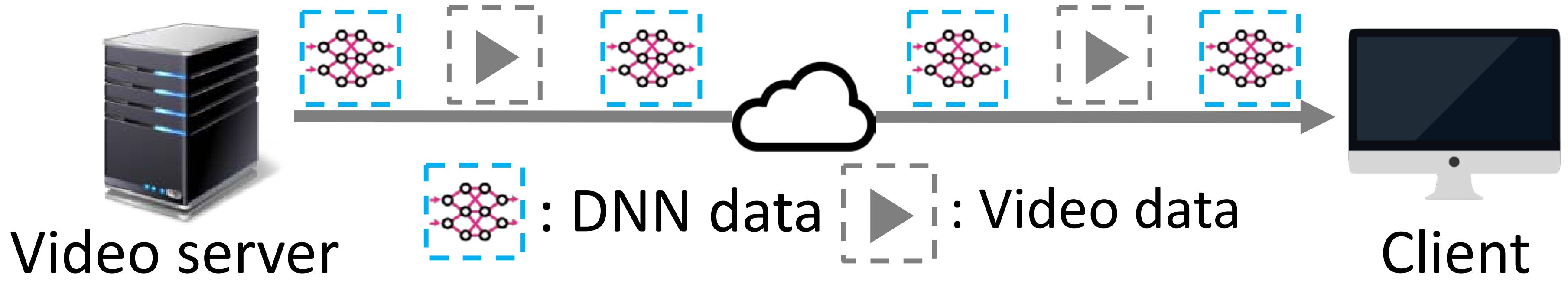


Need to provide incremental benefit during downloading a DNN

NAS: Two Additional Challenges

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

2. A DNN competes bandwidth with video



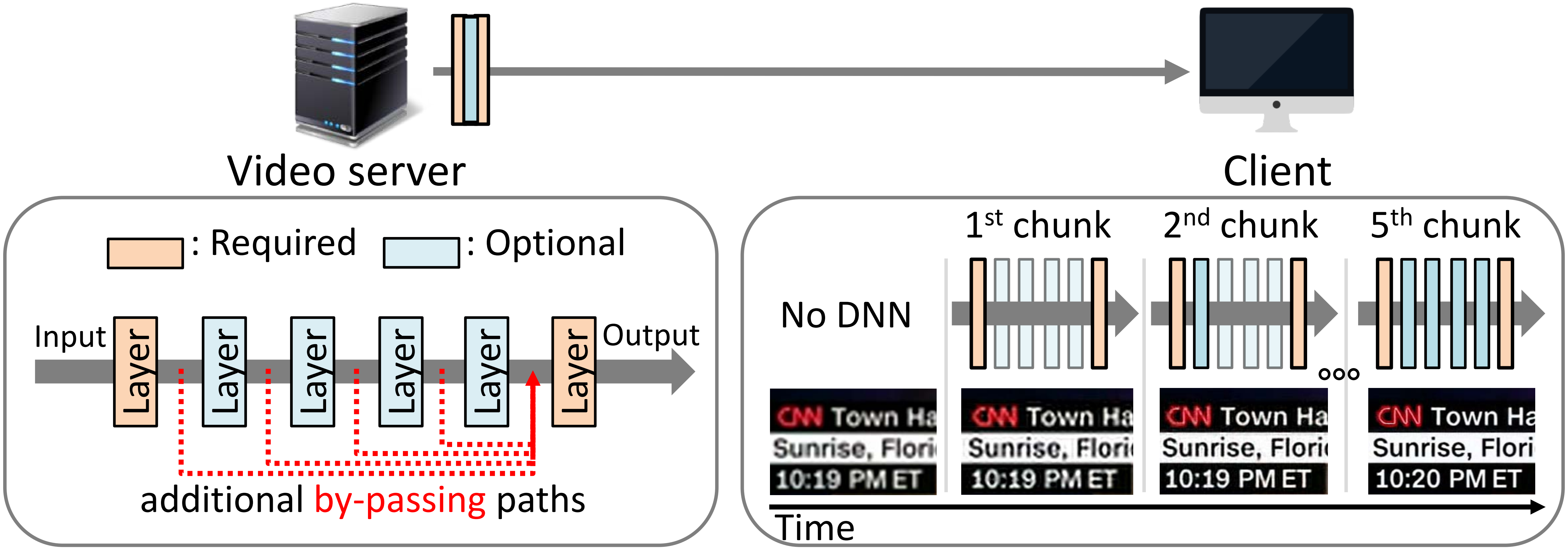
(-) Aggressive download: rebuffering, low video quality

(-) Conservative download: low DNN benefit

Need to *carefully* decide how/when to download a DNN model

Key Design 3: Scalable DNN

Challenge: Takes a long time to utilize a DNN

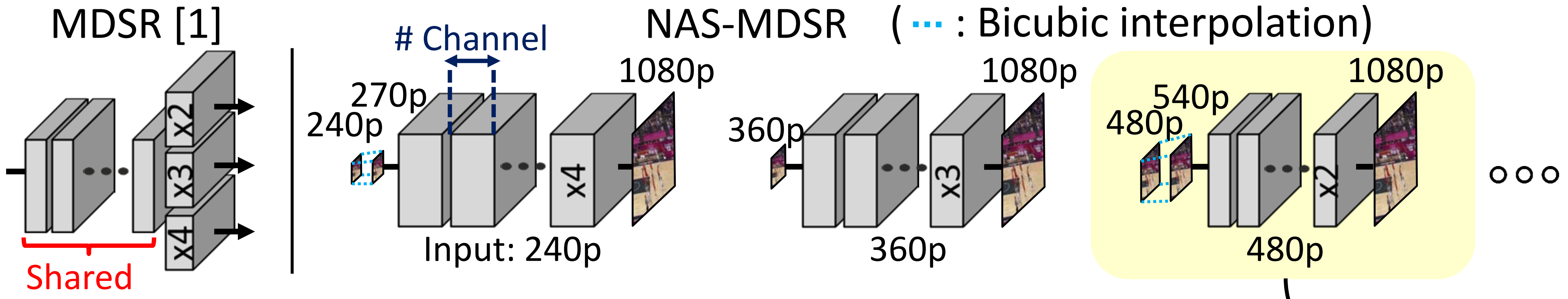


1. Implement a scalable DNN
(+ divide into similar-size chunks)

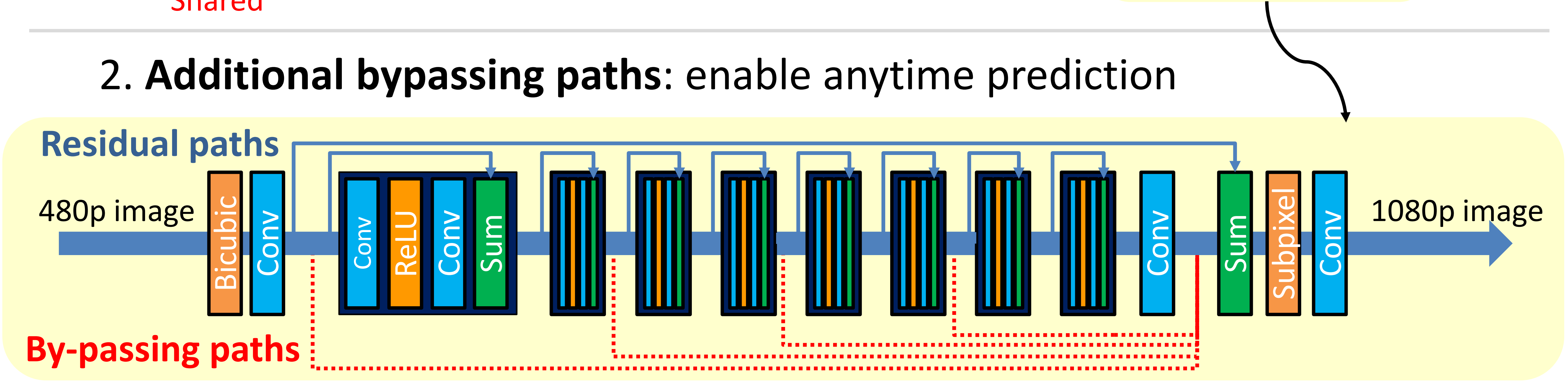
2. Download/Apply a partial DNN

NAS DNN Architecture

1. **Per-resolution DNN**: enable real-time processing



2. **Additional bypassing paths**: enable anytime prediction



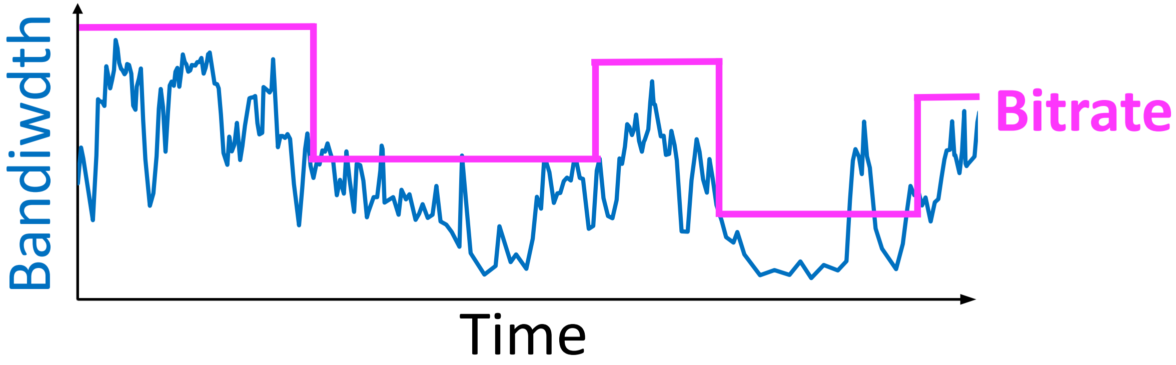
[1] Enhanced Deep Residual Networks for Single Image Super-Resolution, CVPRW, 2017

Key Design 4: Integrated ABR

Challenge: How to decide when to download a DNN

ABR already handles unpredictable bandwidth variations

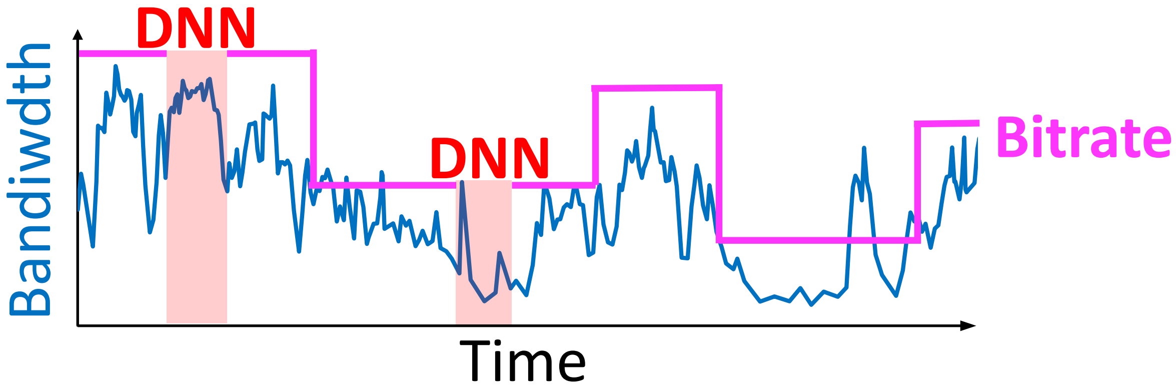
Bandwidth → **ABR** → Bitrate



→ **Integrate** DNN download decisions with existing **RL-based ABR** (Pensieve [1])

Bandwidth → **Integrated ABR** → Bitrate *or* DNN

But, non-trivial!

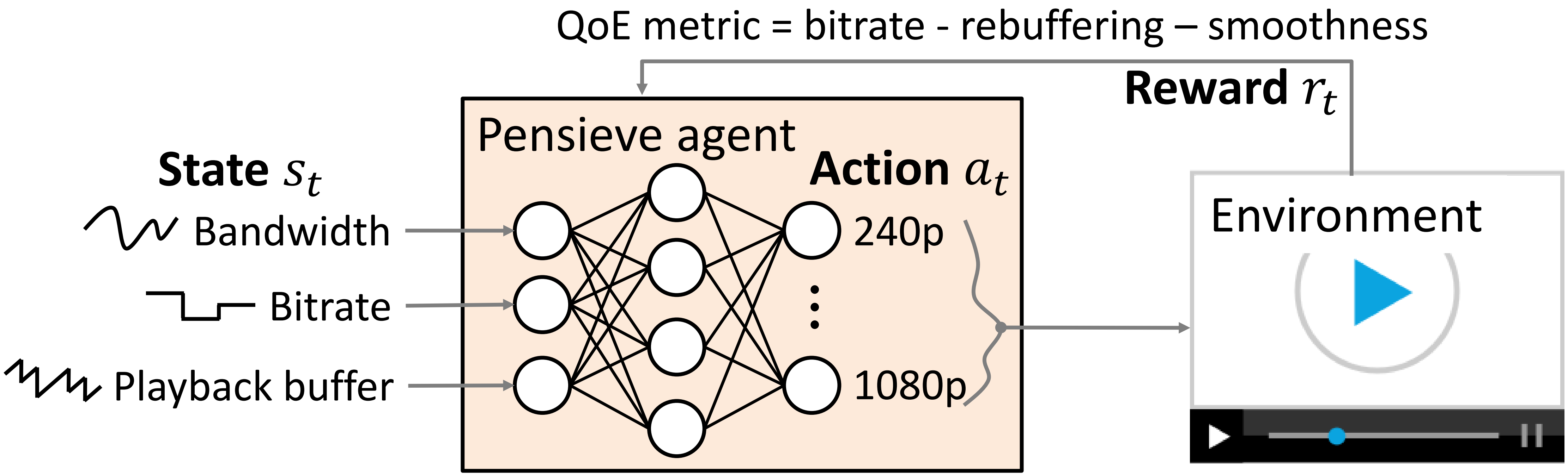


[1]: Mao, Hongzi, Ravi Netravali, and Mohammad Alizadeh. "Neural adaptive video streaming with pensieve.", SIGCOMM, 2017.
[2]: Upper right figure is from the slide of"Neural adaptive video streaming with pensieve.",

Key Design 4: Integrated ABR

Challenge: How to decide when to download a DNN

- **Integrate** DNN download decisions with existing **RL-based** ABR (Pensieve) [1]



Goal: Maximize the total QoE over an entire video

[1]: Mao, Hongzi, Ravi Netravali, and Mohammad Alizadeh. "Neural adaptive video streaming with pensieve.“, SIGCOMM, 2017.

Key Design 4: Integrated ABR

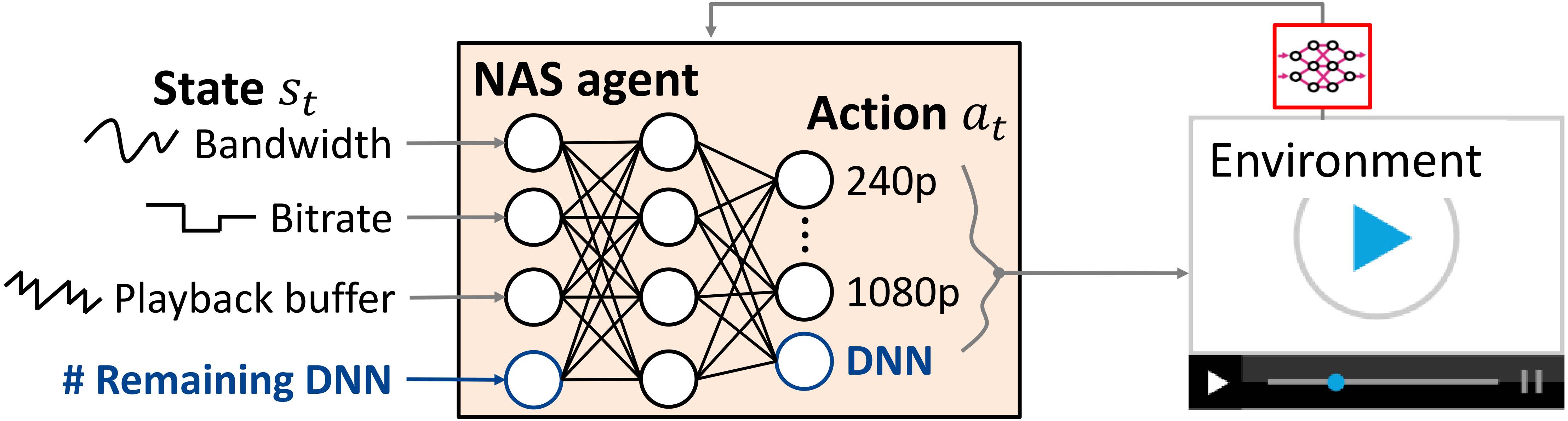
Challenge: How to decide when to download a DNN

- **Integrate** DNN download decisions with existing **RL-based** ABR (Pensieve) [1]

Reward r_t

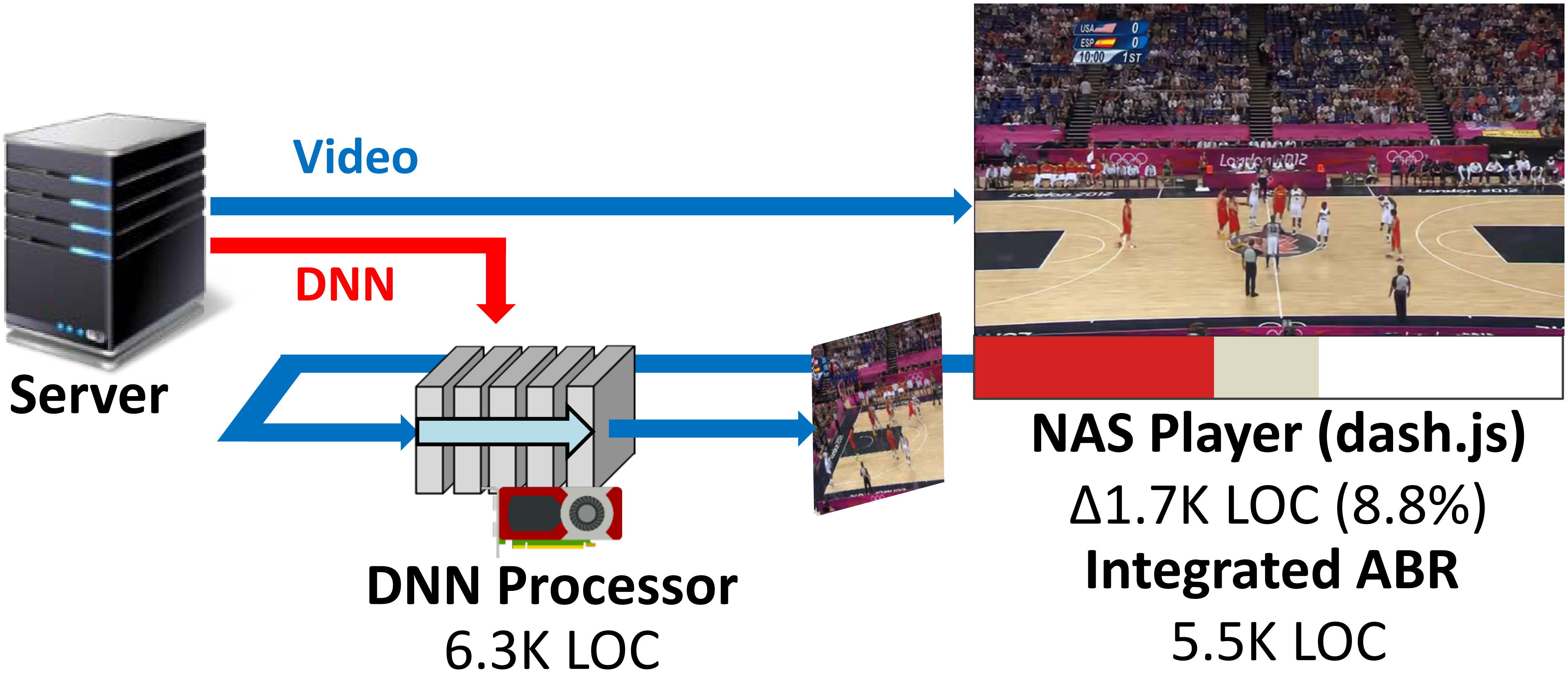
QoE metric = bitrate – rebuffering – smoothness

QoE metric = **DNN(bitrate)** – rebuffering – smoothness



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

Putting All Together: Implementation



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

QoE Metric

Quantify user experience of video streaming

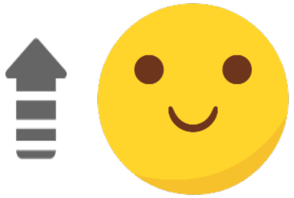
1080p

vs.

240p

Higher quality

Less rebuffering



Better experience

+(Quality) −(rebuffering) −(smoothness)

Generalized QoE model[1,2,3] : $q(R_n) - \mu T_n - |q(R_n) - q(R_{n-1})|$

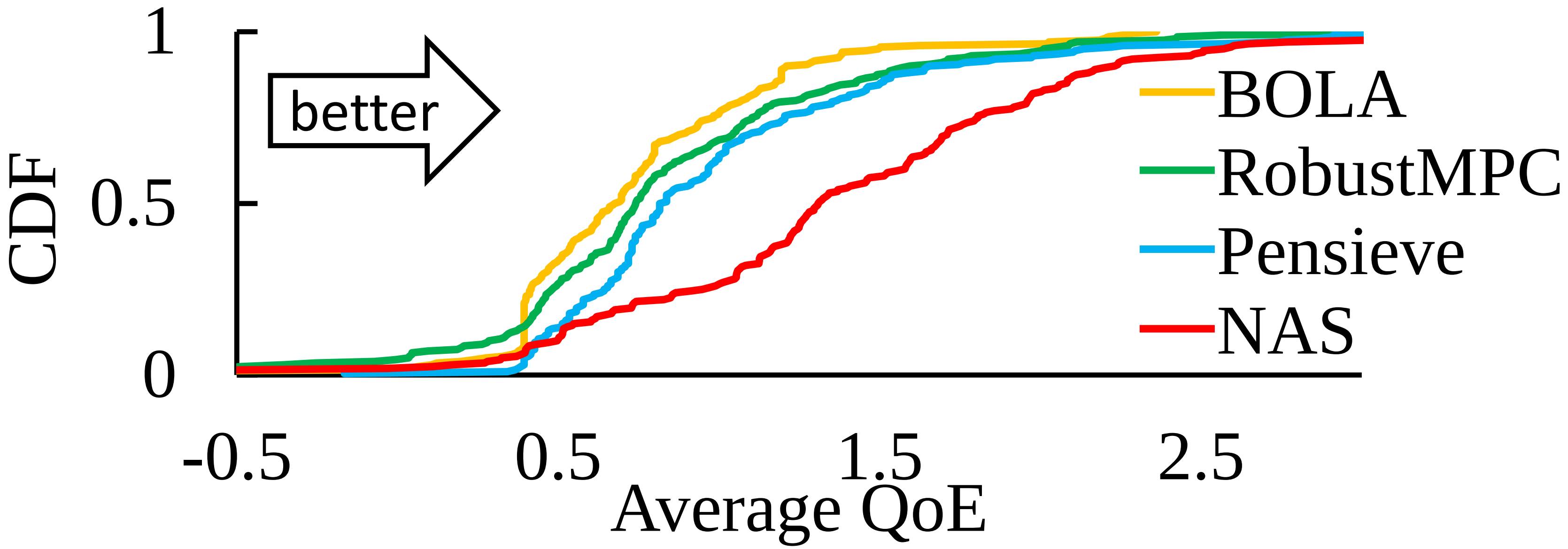
+(Quality) −(rebuffering) −(smoothness)

- $q(R_n)$: Perceptual quality of n^{th} video chunk bitrate R_n
- T_n : Rebuffering time for downloading n^{th} video chunk

1: MPC-SIGCOMM15, Pensieve-SIGCOMM17, Oboe-SIGCOMM18

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

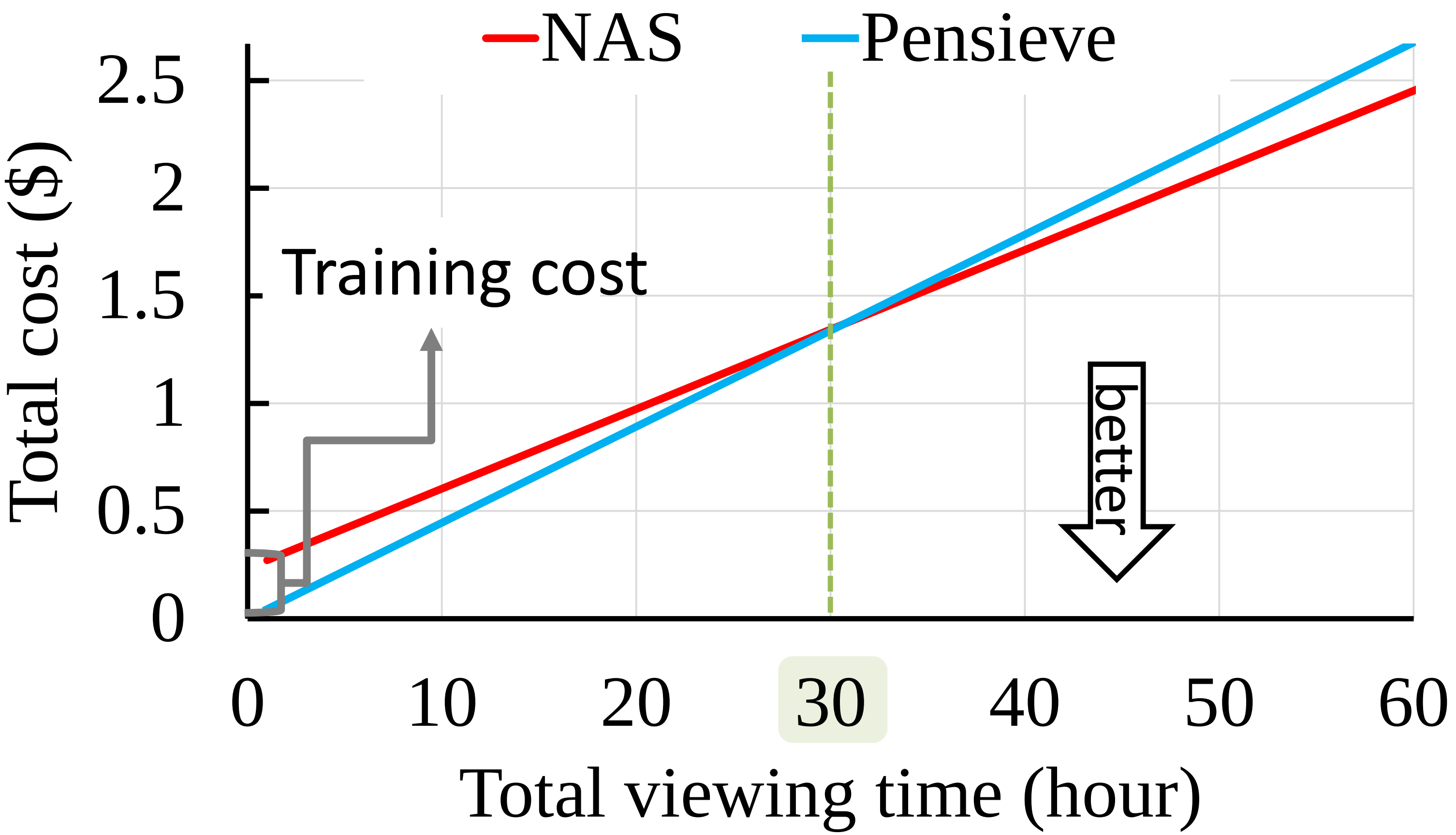
NAS CDN

= 0.085$/GB

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

10 mins ✖ 1.4$/hour

= 0.23$/minute of video

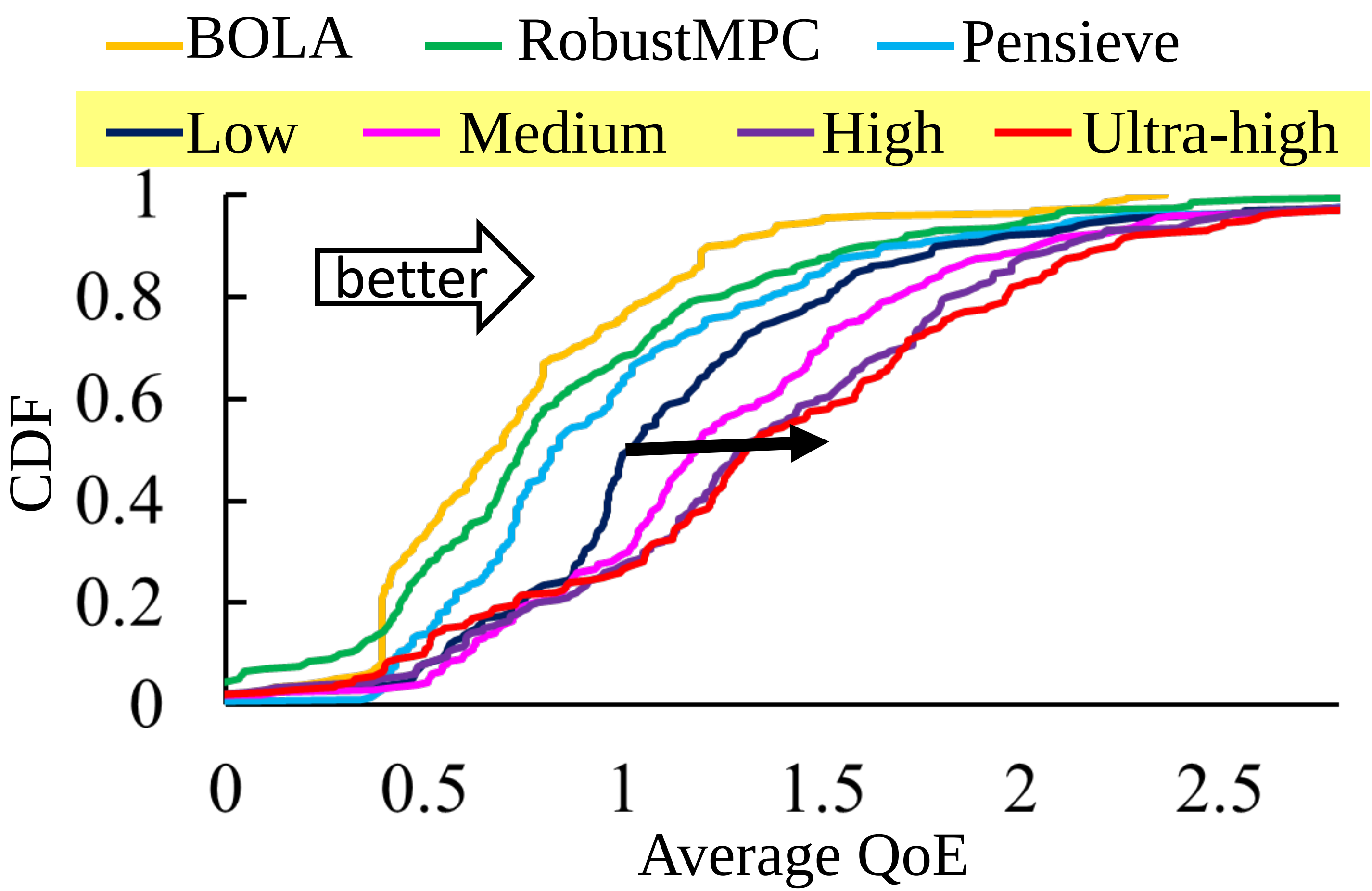


When the total viewing reaches 30 hours (per minute of video), NAS CDN recoups the initial training cost.

Heterogeneous Clients

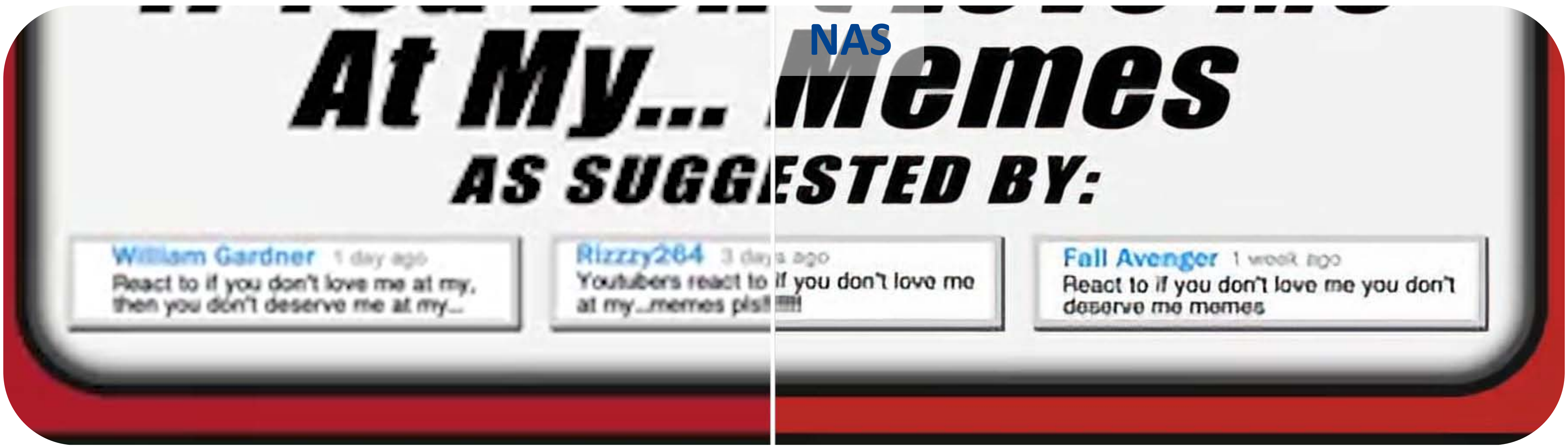
Each GPU processes at real-time
(> 30fps for all resolutions)

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.

NAS: DL-based Adaptive Streaming

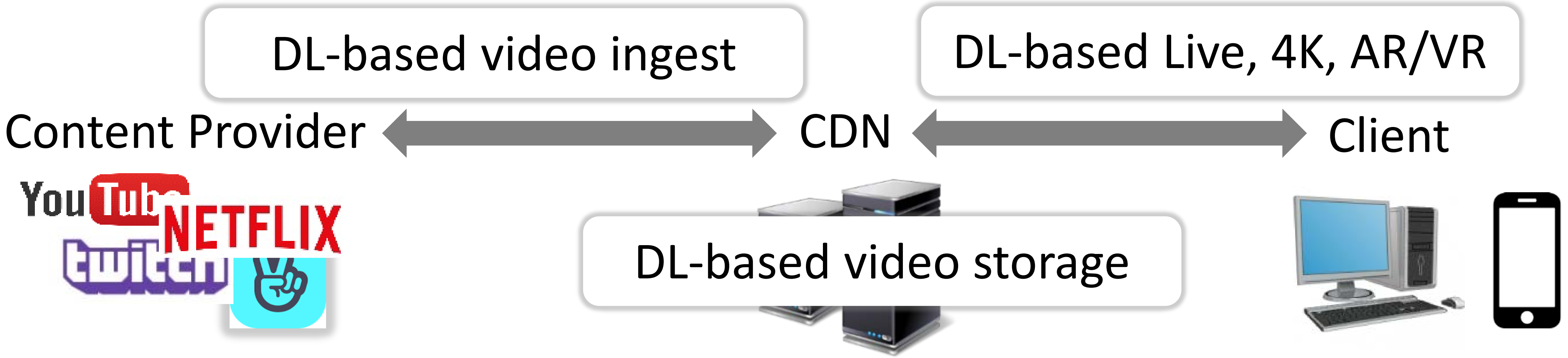


- NAS shows that applying *DNN* on *video content* utilizing *client computation* can significantly enhance user QoE.
- NAS accommodates *four key designs*: Content-aware DNN, Multiple quality DNNs, Scalable DNN, Integrated ABR.

What's Next?

NAS = Adaptive streaming + VoD contents + Desktop-class GPUs

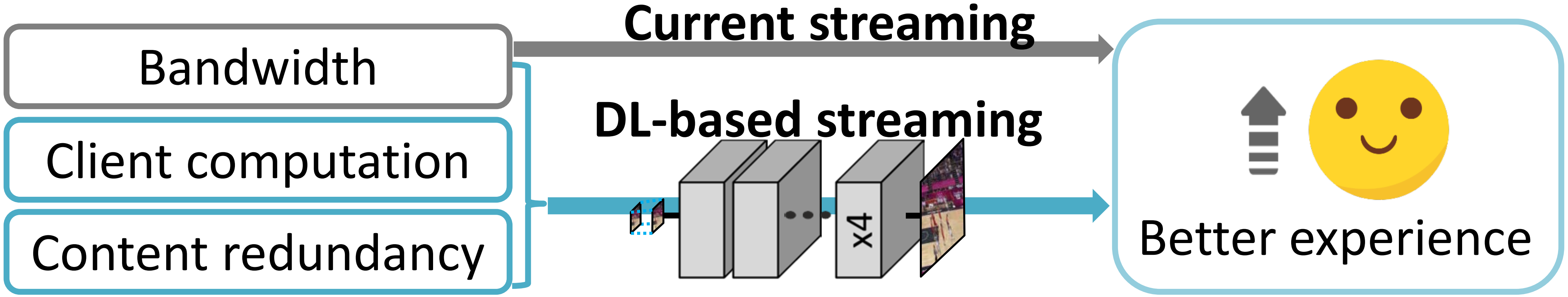
- Integrate DL with **various parts** in video delivery infrastructure
- Apply DL on **diverse video applications** (e.g., Live/4K/AR/VR)
- Deploy DL-based streaming on **commercial mobile devices**



Conclusion

"How will Deep Learning Change Internet Video Delivery?"

- The advance of deep learning presents **unseen opportunities**



- Rethinking the video delivery infrastructure is required to take advantage of the new opportunities

Neural Adaptive Content-aware Internet Video Delivery

Hyunho Yeo   Youngmok Jung   Jaehong Kim   Jinwoo Shin   Dongsu Han
KAIST

: **First step** toward this direction

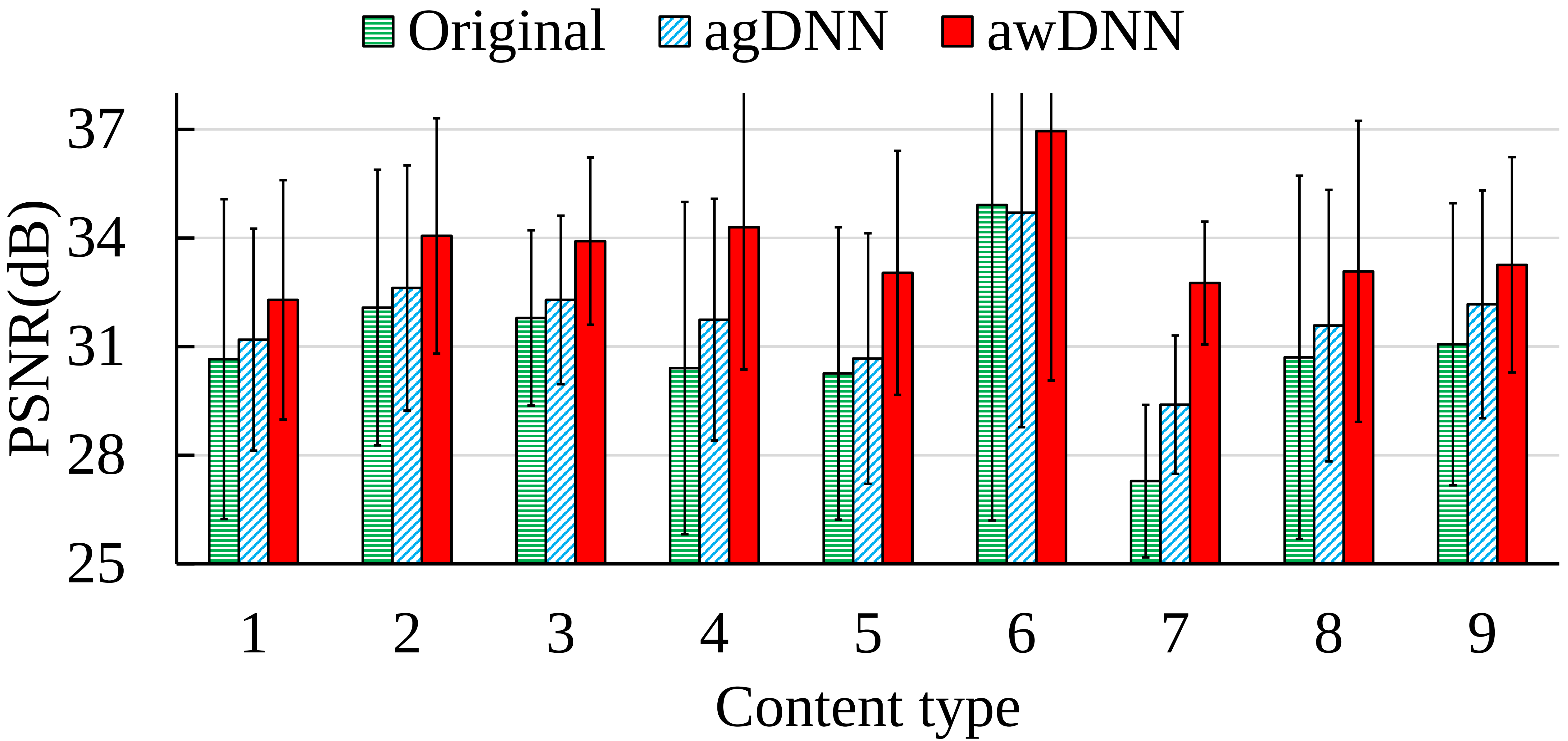
Thank you

OSDI conference @ Carlsbad, CA, USA

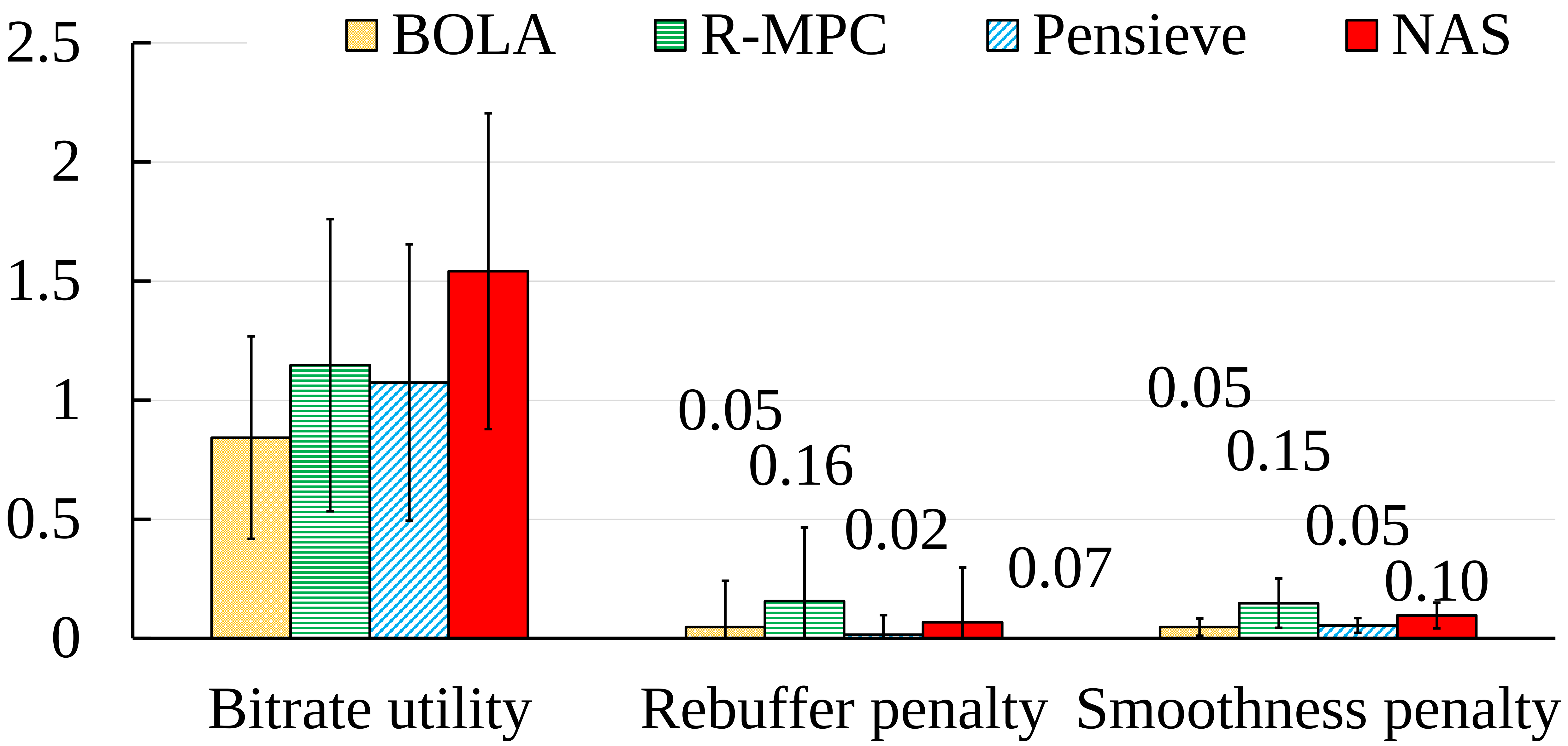
- **Personal homepage**
  http://ina.kaist.ac.kr/~hyunho/
- **Lab homepage** INA
  http://ina.kaist.ac.kr/
- **Project homepage**
  http://ina.kaist.ac.kr/~nas/

KAIST

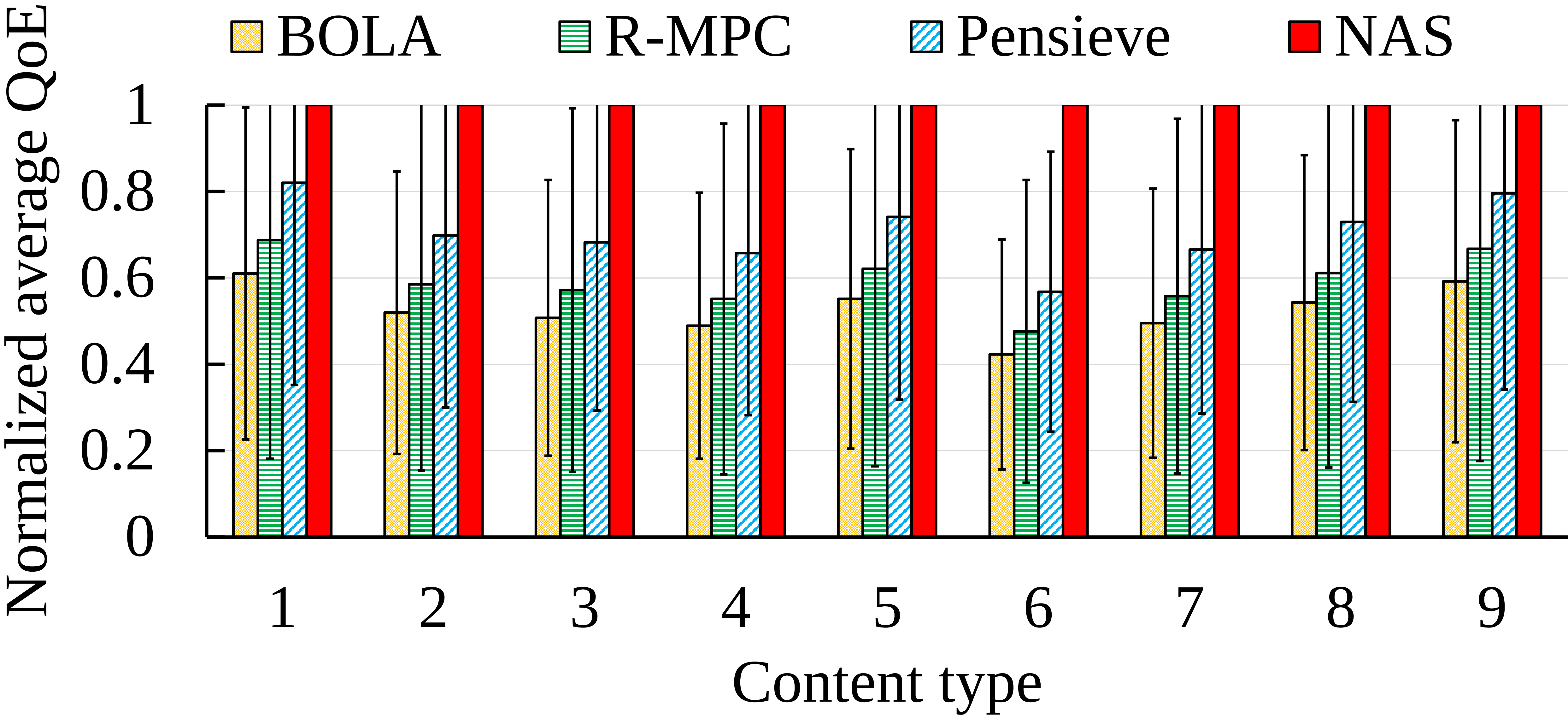
Content-agnostic vs. Content-aware



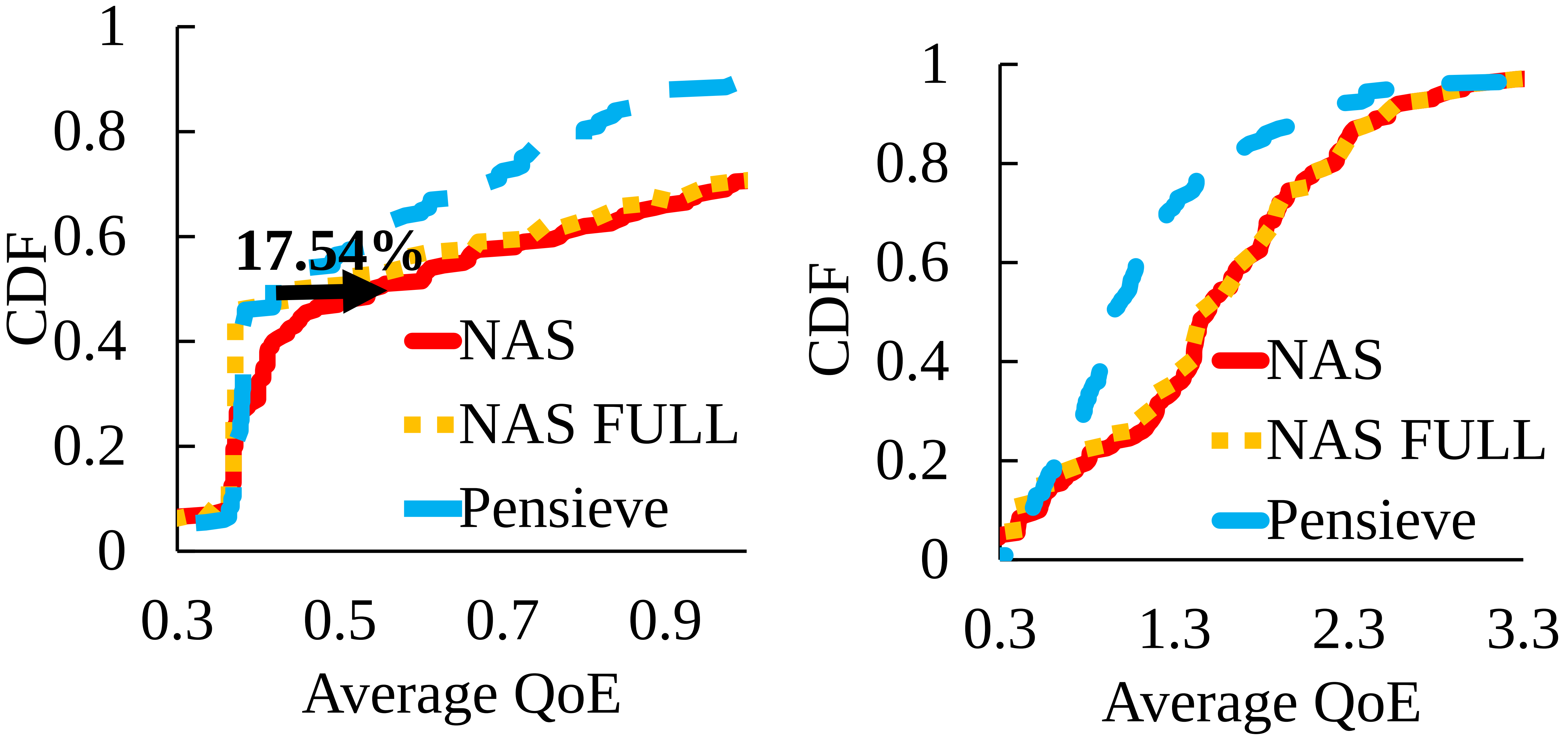
QoE breakdown



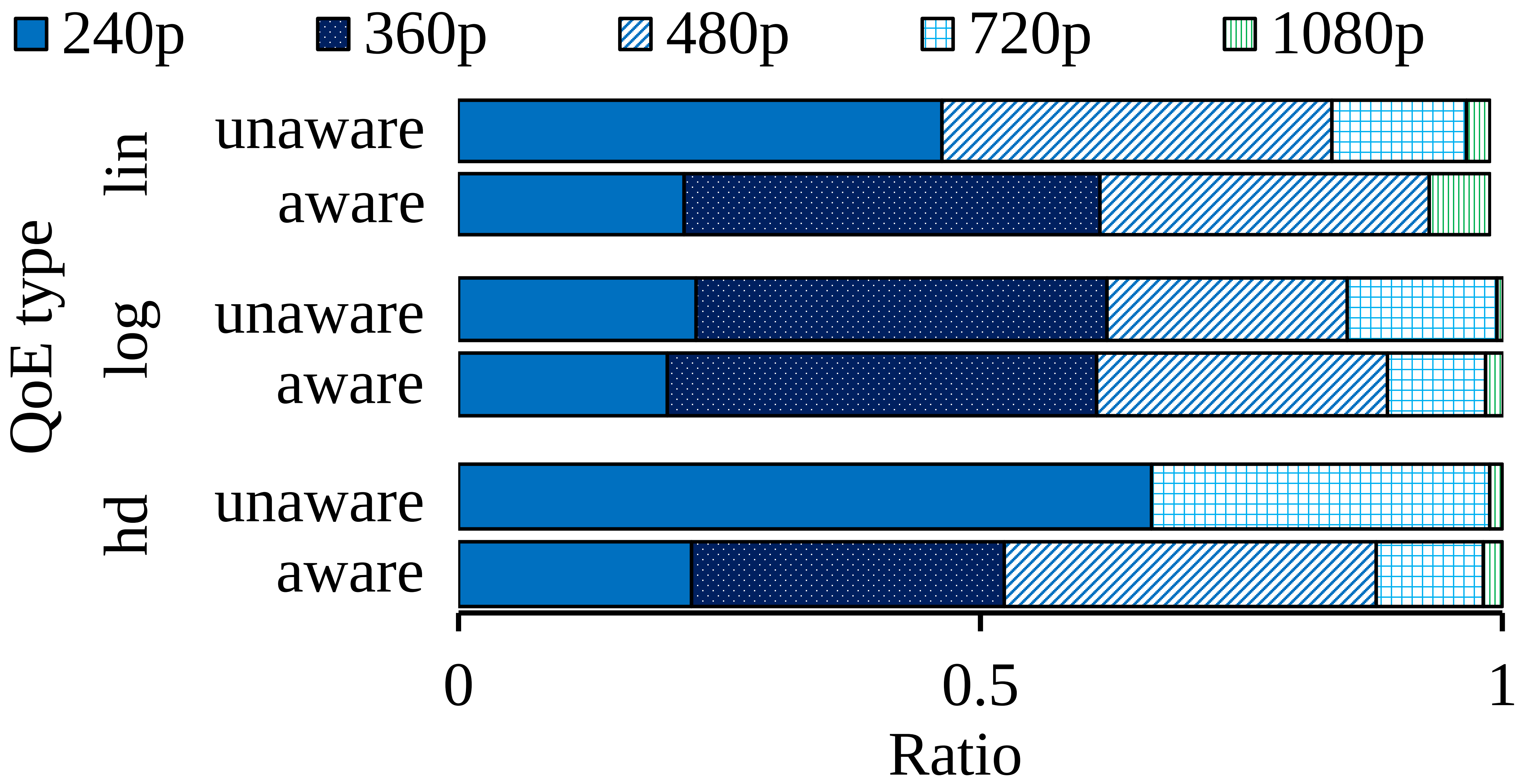
Average QoE over Content Types



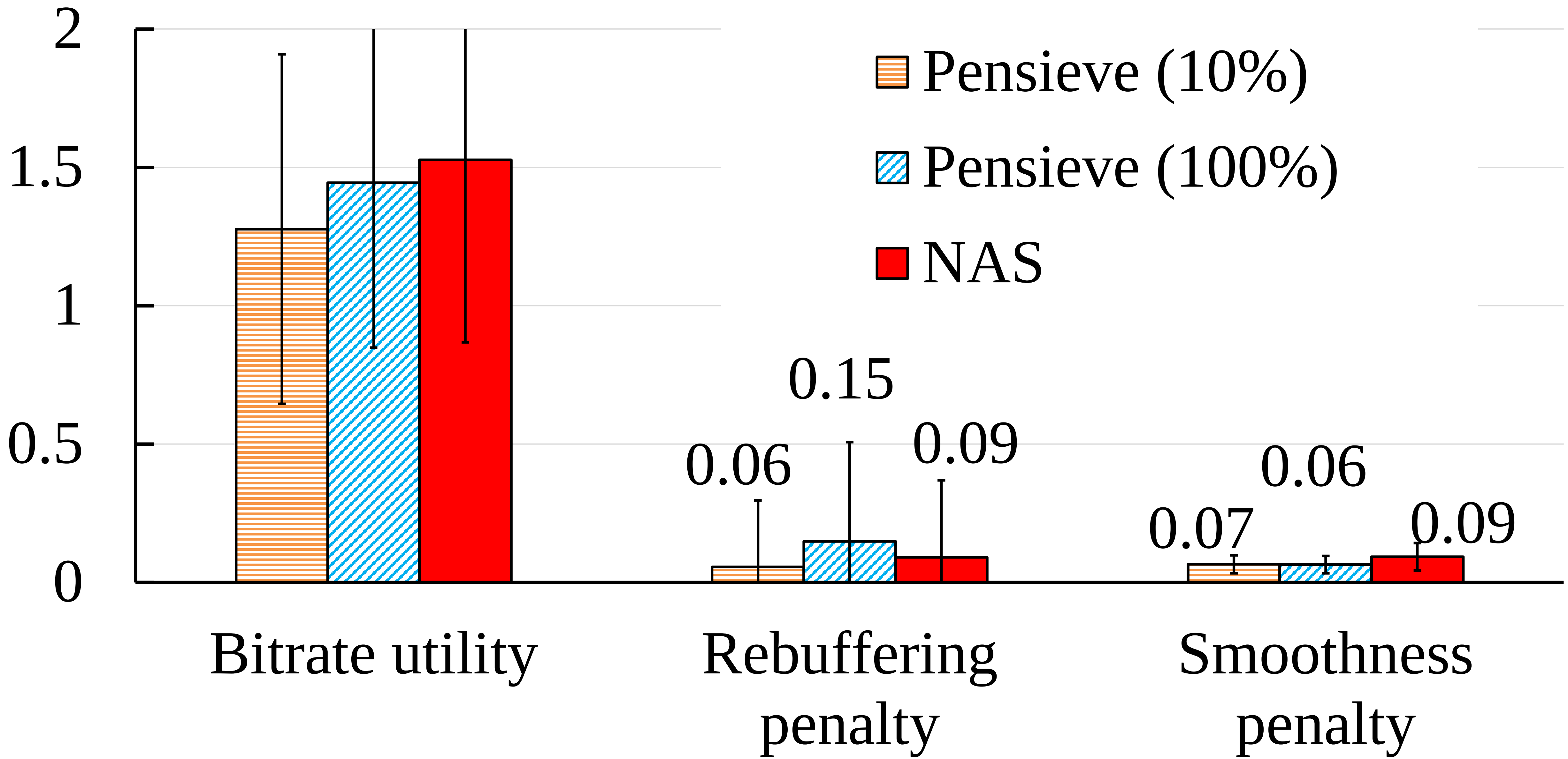
Scalable DNN



Integrated ABR (Quality-awareness)



Integrated ABR (DNN downloads)



Case Study: Timeline