OASIS: Collaborative Neural-Enhanced Mobile Video Streaming

Shuowei Jin, Ruiyang Zhu, Ahmad Hassan[†], Xiao Zhu, Xumiao Zhang,
Z. Morley Mao, Feng Qian[†], Zhi-Li Zhang[‡]

University of Michigan, [†]University of Southern California, [‡]University of Minnesota

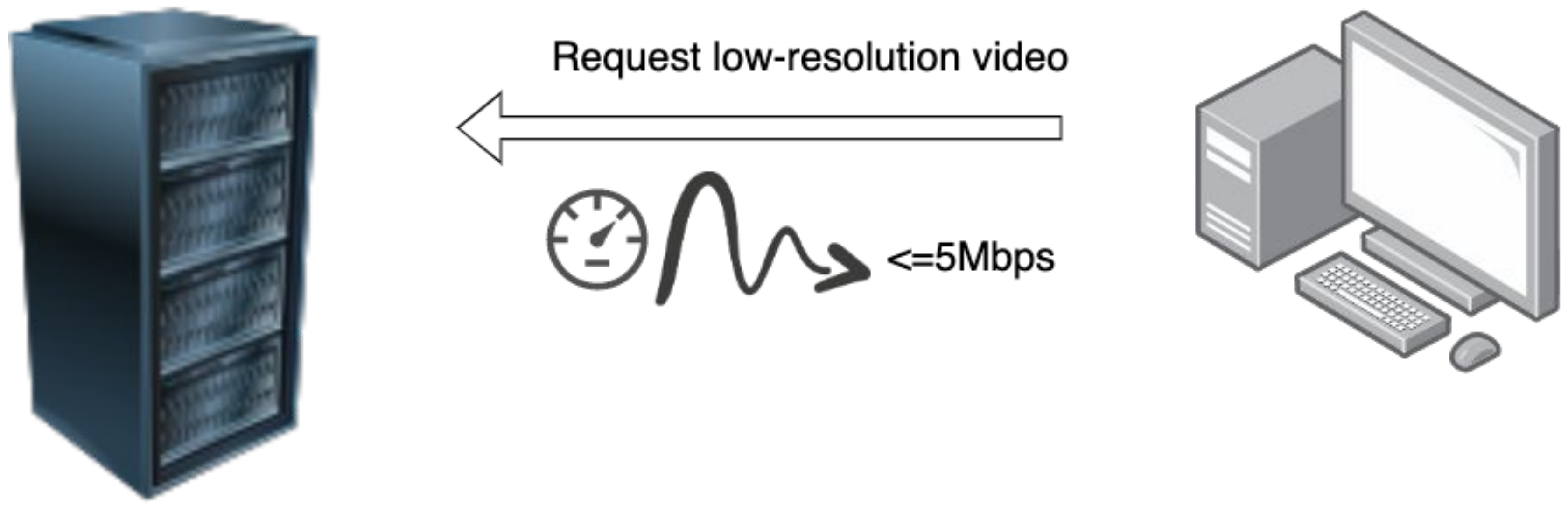
jinsw@umich.edu

Presenter: Jinwei Zhao, University of Victoria, Canada

Outline

- **Background and Motivation**
- System Design
- Evaluation Results
- Conclusion

Background: Neural-enhanced Video Streaming[1,2]

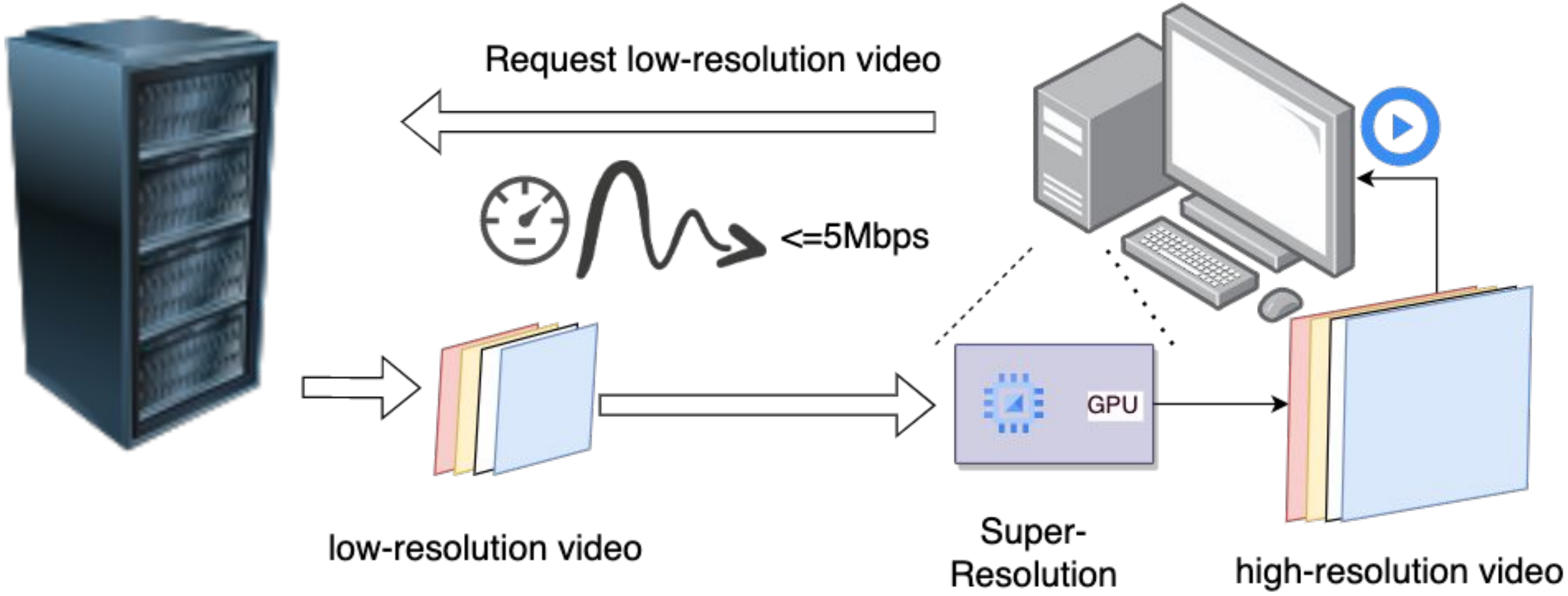


Neural-enhanced Video Streaming Workflow

- Idea: Transfer the burden from network to computation.
- Such system can do high-quality video streaming under low/fluctuating network bandwidth conditions.
- Save data usage while delivering high-quality video.

[1] Neural Adaptive Content-aware Internet Video Delivery. OSDI 2018.
[2] Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning. SIGCOMM’20

Background: Neural-enhanced Video Streaming[1,2]



Neural-enhanced Video Streaming Workflow

- Idea: Transfer the burden from network to computation.
- Such system can do high-quality video streaming under low/fluctuating network bandwidth conditions.
- Save data usage while delivering high-quality video.

[1] Neural Adaptive Content-aware Internet Video Delivery. OSDI 2018.
[2] Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning. SIGCOMM'20

Previous systems focus on the PC with high-performance GPU[1,2].

[1] Neural Adaptive Content-aware Internet Video Delivery. OSDI 2018.
[2] Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning. SIGCOMM'20

Previous systems focus on the PC with high-performance GPU[1,2].

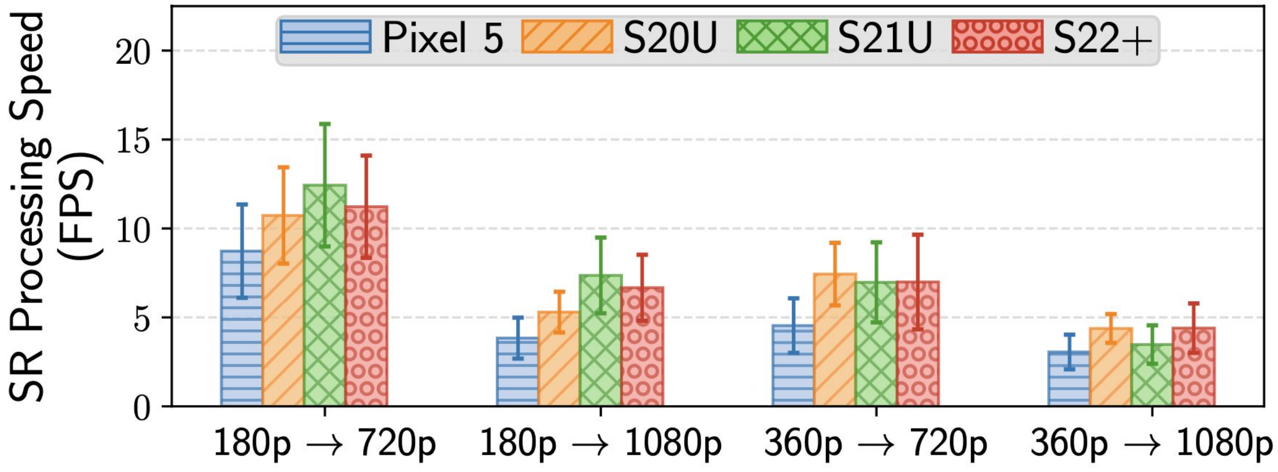
How can we do it on the **Mobile** Side?

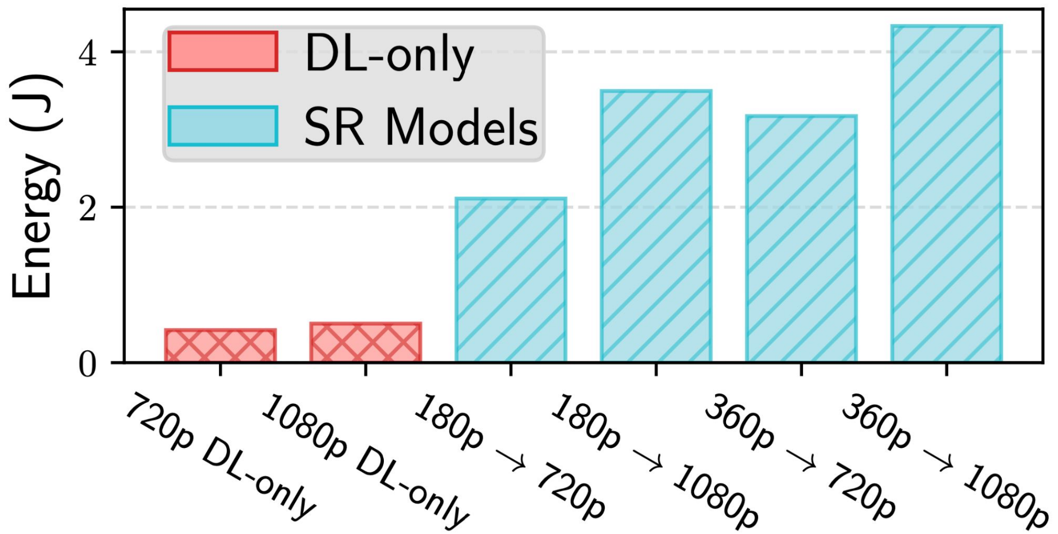
Qualcomm snapdragon

[1] Neural Adaptive Content-aware Internet Video Delivery. OSDI 2018.
[2] Neural-Enhanced Live Streaming: Improving Live Video Ingest via Online Learning. SIGCOMM'20

Motivation Experiment

- Experiment Setting:
  - Measure different super resolution (SR) model processing speed on different devices.
  - Record total energy consumption (Screen energy consumption is subtracted from the result.)
- **Conclusion**:
  - Single mobile device SR processing speed is ***less than 24-30 FPS***, not enough.
  - SR procedure consumes **too much energy** for single device. 30-min SR video streaming leads to 28%-57% battery drain.

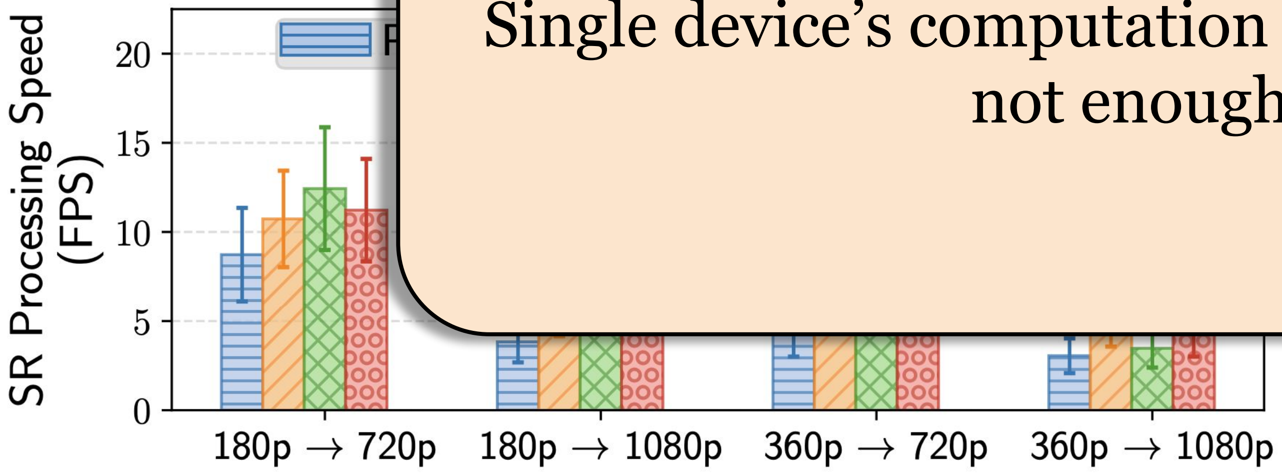


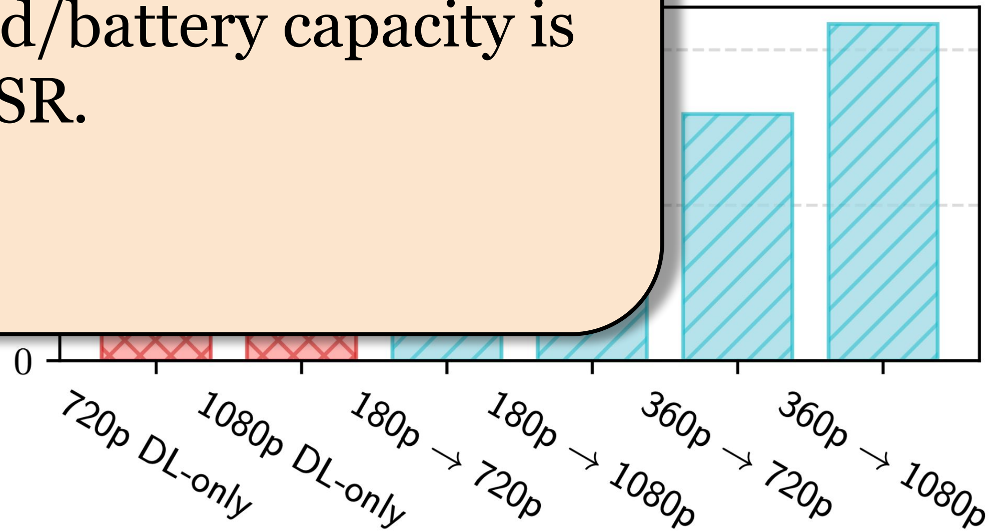


Motivation Experiment

- Experiment Setting:
  - Measure different super resolution (SR) model processing speed on different devices.
  - Record total energy consumption (Screen energy consumption is subtracted from the result.)
- **Conclusion**:
  - Single mobile device SR processing speed is ***less than 24-30 FPS***, not enough.
  - SR pr … video strea…

Single device's computation speed/battery capacity is not enough for SR.





How to perform neural-enhanced video streaming on the Mobile Side?

Solution: Multi-device collaboration.

- Leverage all devices' network and computation resources to perform neural-enhanced video streaming.
- Benefits:
  - **Scalability**: Capable of processing complicated SR models as the system scales up.
  - **Energy Saving**: The heavy computation task is distributed across all devices. Each device will have less energy consumption.

[1] NEMO: Enabling Neural-enhanced Video Streaming on Commodity Mobile Devices. MobiCom 2020.
[2] Basicvsr: The search for essential components in video super-resolution and beyond. CVPR 2021.

Motivation: Incentives to use multi-devices

- It is becoming more usual for users to own numerous mobile devices.
  - 53% of adults in the United States possess a tablet.[1]
  - 33% of American households own three or more smartphones.[2]
- A group of people gather to watch the same video clip from YouTube, Netflix.
  - 50% of male YouTube viewers between the ages of 18 and 34 watch YouTube clips in person with friends.[3]

[1] Statista. https://www.statista.com/statistics/756045/tablet-owners-among-us-adults/.
[2] pewresearch. https://www.pewresearch.org/fact-tank/2017/05/25/a-third-of-americans-live-in-a-household-with-three-or-more-smartphones/
[3] Gen V research, Google. http://www.youtube.com/yt/advertise/medias/pdfs/research-gen-v-men-2.pdf

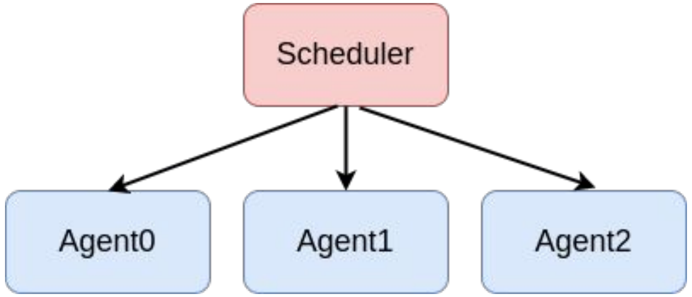
Problem Formulation & Challenges

Problem: A video streaming system enables multiple mobile devices in close proximity to do collaboratively neural-enhanced video streaming on each devices.

Challenges:

ABR in multi-device system:
select bitrate and SR model

Scheduling: heterogeneous network
and computation resources

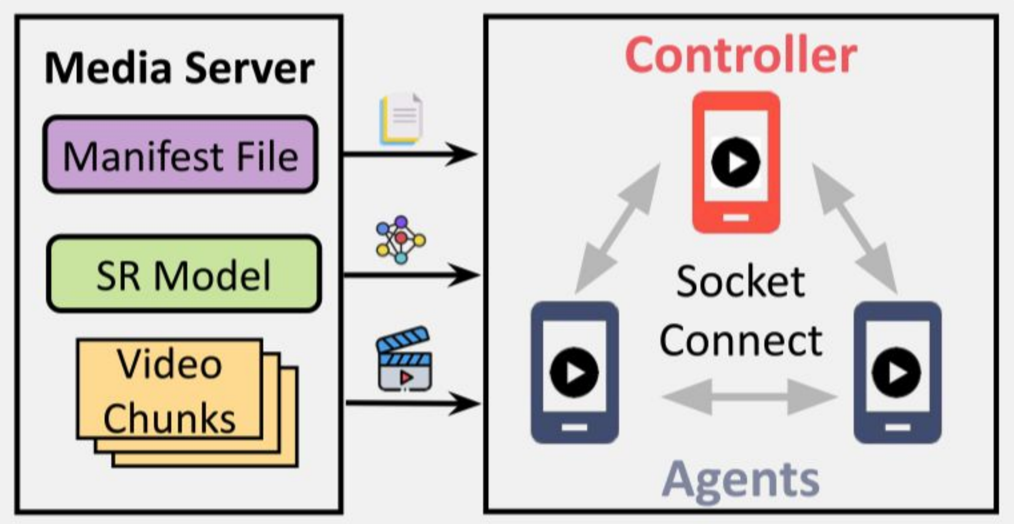


Outline

- Background and Motivation
- **System Design**
- Evaluation Results
- Conclusion

System Workflow

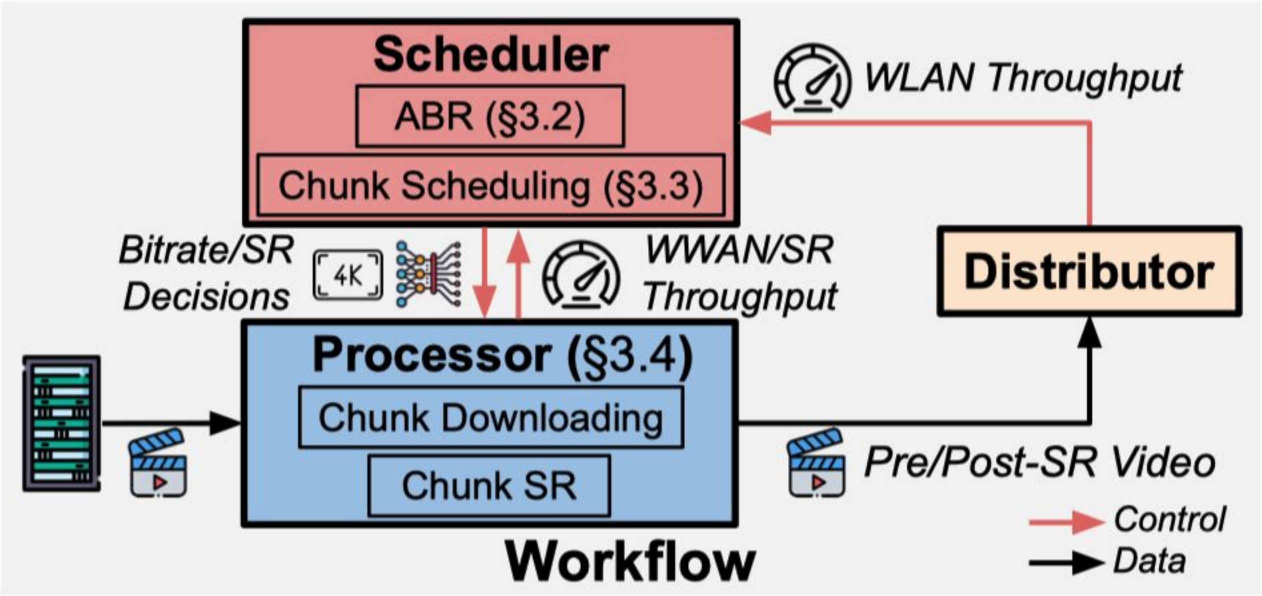
- **Offline Video-content Preparation:**
  - Encode, segment the upload video to multiple-bitrate video chunks, train SR models.
- **Online Video Streaming:**
  - Multiple devices connect via peer-to-peer socket.
  - One device as the controller, the rest as the agents.



Architecture

System Architecture

- **Task Scheduling (Scheduler):**
  - OASIS-ABR: choose download bitrate and SR model.
  - OASIS-SCHED: schedule the chunk downloading, forwarding and SR tasks to each device.
- **Task Processing (Processor):**
  - Execute assigned tasks, measure performance and inform the Scheduler
- **Chunk Distribution (Distributor):**
  - Forward downloaded chunks, broadcast post-SR chunks.



OASIS-ABR

- Goal:
  - Adaptively select the optimal download bitrate and SR model combination based on system parameters: throughput, buffer, SR speed of each device.

OASIS-ABR

- Goal:
  - Adaptively select the optimal download bitrate and SR model combination based on system parameters: throughput, buffer, SR speed of each device.
- Algorithm Steps:
  - System total throughput modeling.
  - ABR decision making.

OASIS-ABR: System Throughput Modeling

- *Insights 1*: Bottleneck determines upper bound throughput.
  - Our system employs a **pipeline design**, streamlining chunk downloading, SR processing, and post-SR chunk distribution.
  - Bottleneck throughput represents the upper bound for total throughput.

OASIS-ABR: System Throughput Modeling

- <u>*Insights 1*</u>: Bottleneck determines upper bound throughput.
  - Our system employs a **pipeline design**, streamlining chunk downloading, SR processing, and post-SR chunk distribution.
  - Bottleneck throughput represents the upper bound for total throughput.
- <u>*Insights 2*</u>: Due to multiple device synchronization, overhead exists.
  - Predicting overhead based on historical overhead values.
  - ***Total_Throughput = Bottleneck_Throughput - Overhead.***

OASIS-ABR: System Throughput Modeling

- *Insights 1*: Bottleneck determines upper bound throughput.
  - Our system employs a **pipeline design**, streamlining chunk downloading, SR processing, and post-SR chunk distribution.
  - Bottleneck throughput represents the upper bound for total throughput.
- *Insights 2*: Due to multiple device synchronization, overhead exists.
  - Predicting overhead based on historical overhead values.
  - ***Total_Throughput = Bottleneck_Throughput - Overhead.***

Predicted **Throughput** -> Estimate the Rebuffer Time -> Predict **QoE**.

OASIS-ABR: Balance exploration and exploitation

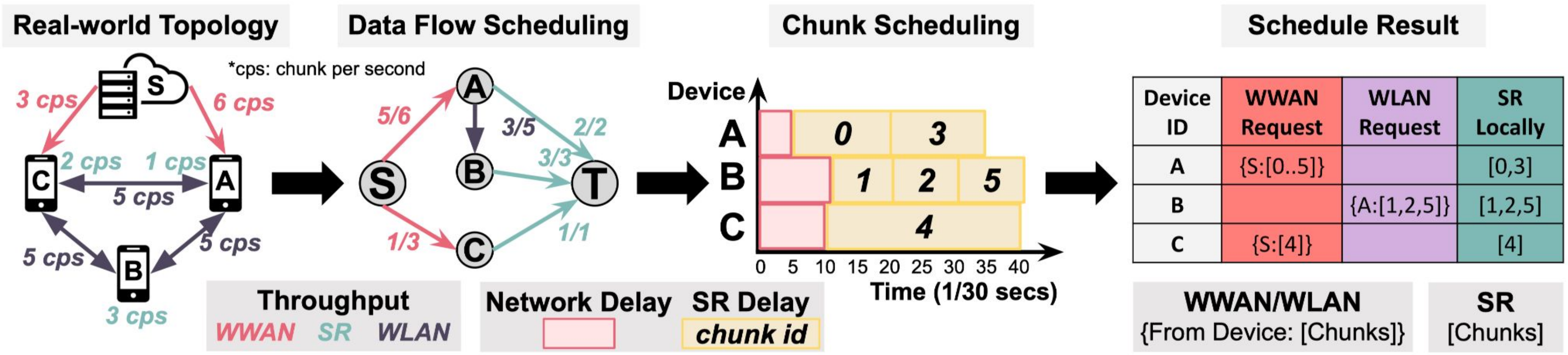
- For each (bitrate, SR Model) **combination**, throughput -> Predicted QoE.
- Rather than selecting highest QoE, we choose highest upper confidence bound (**UCB**) value.
  - **UCB = Predicted QoE + Uncertainty term**.
    - As a combination is explored more (C_i increases), its uncertainty diminishes, promoting the exploration of less investigated combinations.

$$UCB_i = Q\hat{o}E_i + \alpha\sqrt{\frac{\log(\sum C_i + 1)}{C_i + 1}}$$

Number of times the combination has been explored.

OASIS-SCHED: Chunk Scheduling Algorithm

- Output: Tasks for each device: download/forward/SR tasks.
- Goal:
  - Maximize throughput & Prioritizing the completion time of earlier chunks -> Minimize stall time -> Improve QoE.
- Workflow: (two steps)
  - First step (high-level): schedule the data flow across devices.
    - **Key Idea: Find the dataflow to maximally utilize all devices' network and computation resources.**
  - Second (detailed-level): schedule the chunk IDs to devices.
    - **Key Idea: Ensure earlier chunk finishes earlier.**



Outline

- Background and Motivation
- System Design
- **Evaluation Results**
- Conclusion

Experiment Setup

Baselines.

- End-to-end system baselines: MicroCast, MPBond.

Devices: 7 devices in total.

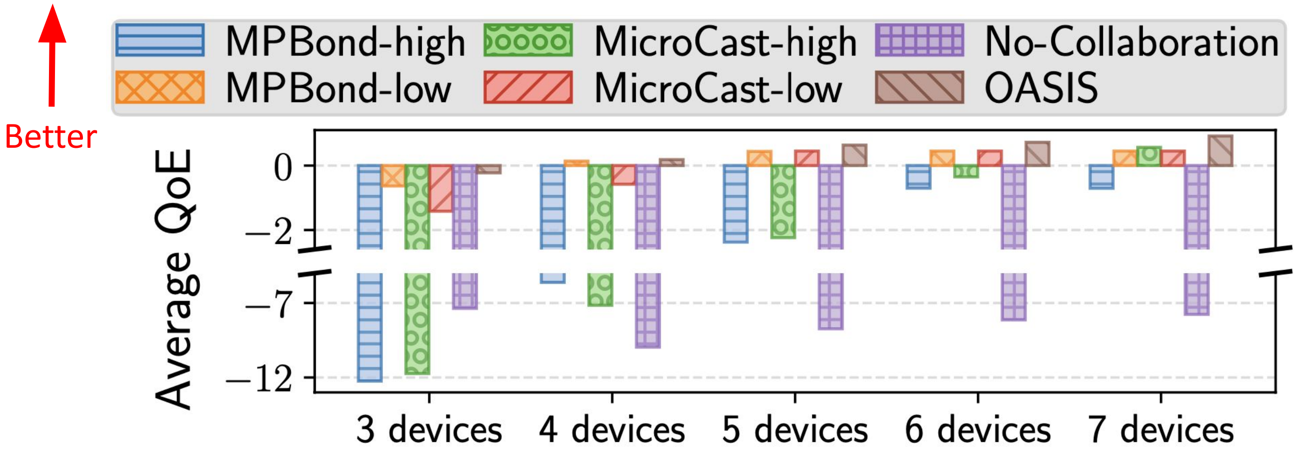
- 2 Pixel5 (PX5), 1 Samsung S10 (S10), 3 Samsung S20 (S20), 1 Samsung S21U (S21U).
- A monsoon power monitor connected to S10, the other one connected to S20.

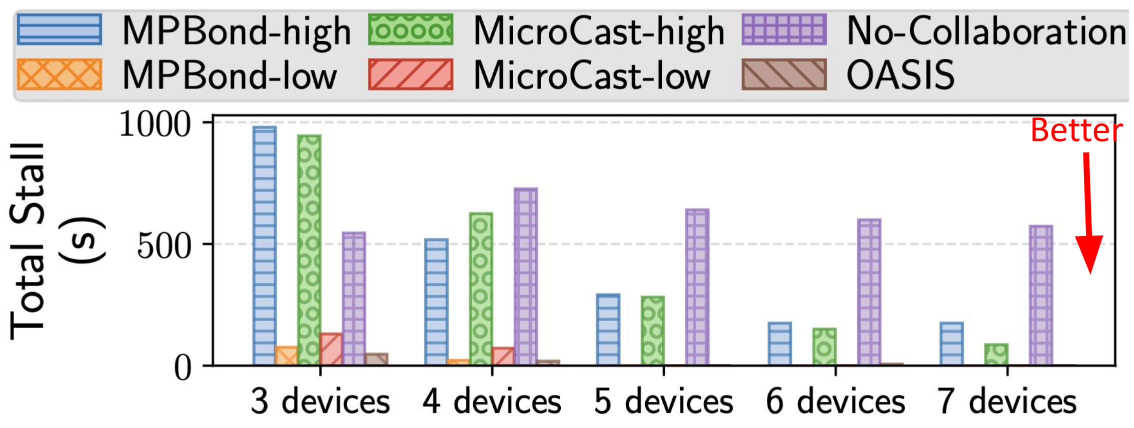
SR Models:

- 180p->720p, 180p->1080p, 360p->720p, 360p->1080p

Evaluation Results

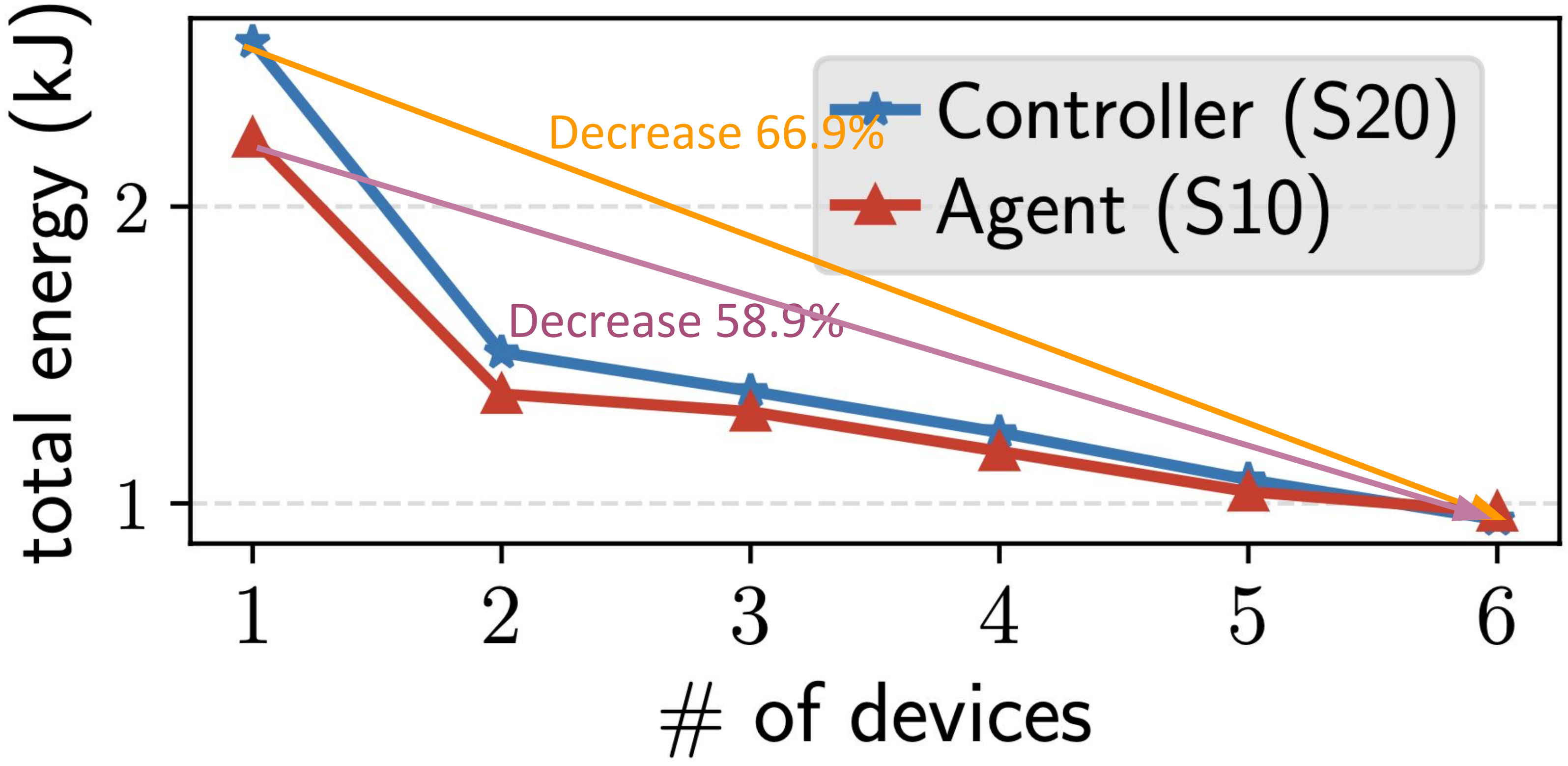
- OASIS outperforms MPBond, MicroCast, No-Collaboration by improving 35%-230% on average QoE.
- OASIS reaches 37% to 100% less stall comparing with the baselines.
- OASIS's average QoE keep increasing when the system scales up.





Energy Experiment

- Adding more device into the system can reduce per device energy usage.
    - Per device energy consumption decreases by 60% when system scales up from 1 device to 6 devices.



Outline

- Background and Motivation
- System Design
- Evaluation Results
- **Conclusion**

Conclusion

- OASIS is the first system to realize **both network-level and computation-level collaboration** to perform neural-enhanced video streaming.
- OASIS proposes a new direction in multi-device collaboration, setting a precedent for future research.

Thank you