NOVA: QoE-driven Optimization of DASH-based Video Delivery in Networks

Vinay Joseph* and Gustavo de Veciana

Department of Electrical and Computer Engineering

The University of Texas at Austin

*joined Qualcomm, San Diego

ASCEND lab

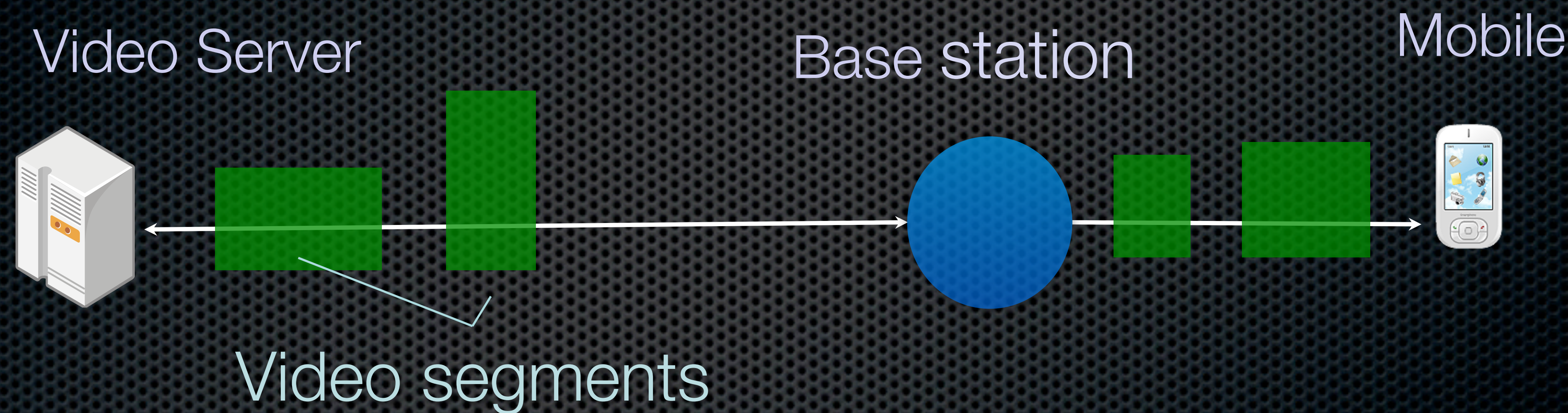
Wireless Networking & Communications Group

Elements of Video Delivery

- Video compression
- Transport and transmission over wireline and wireless networks
- <u>Humans are the “receivers”</u>

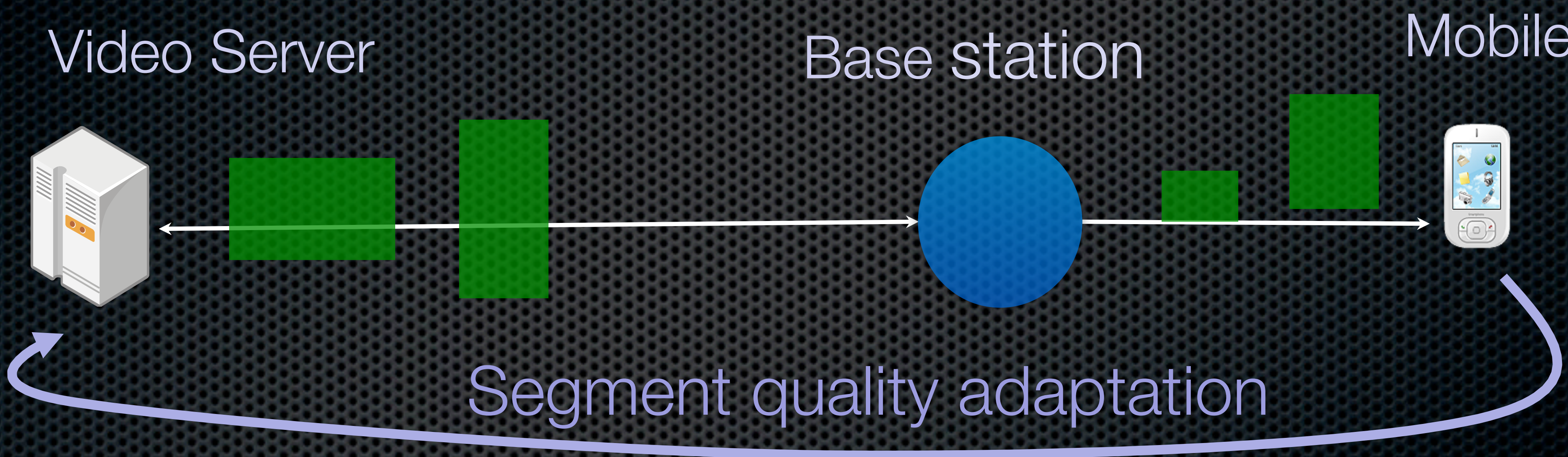
Other important aspects CDNs, caching, transcoding,……we focus on the above.

State of the Art: Dynamic Adaptive Streaming over HTTP (DASH)



- Video stream broken down into segments
- Multiple representations per segment (video quality/size)
- Segment requests/representations are receiver driven with asynchronous decision points

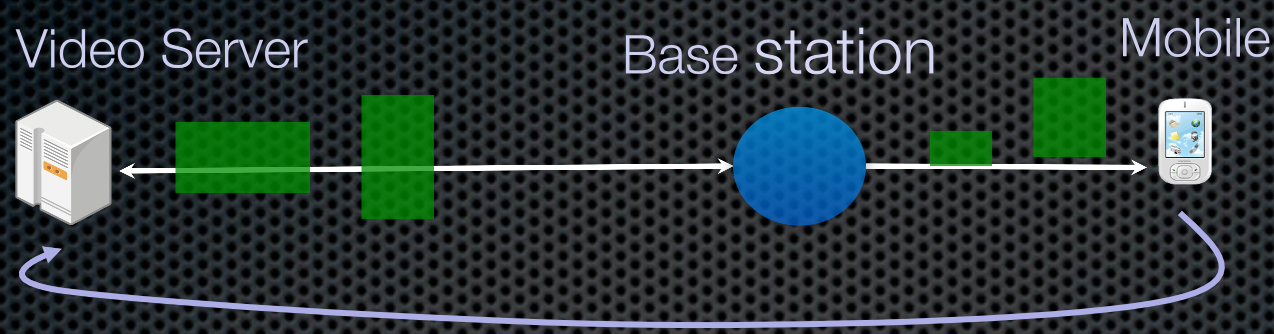
State of the Art*: Dynamic Adaptive Streaming over HTTP (DASH)



- Using TCP as transport protocol for segments
- Adapting choice of segment representation (quality/size) to match estimated throughput

Old problem with lots related and complementary work !!!!!!!

DASH Algos : Some Shortcomings

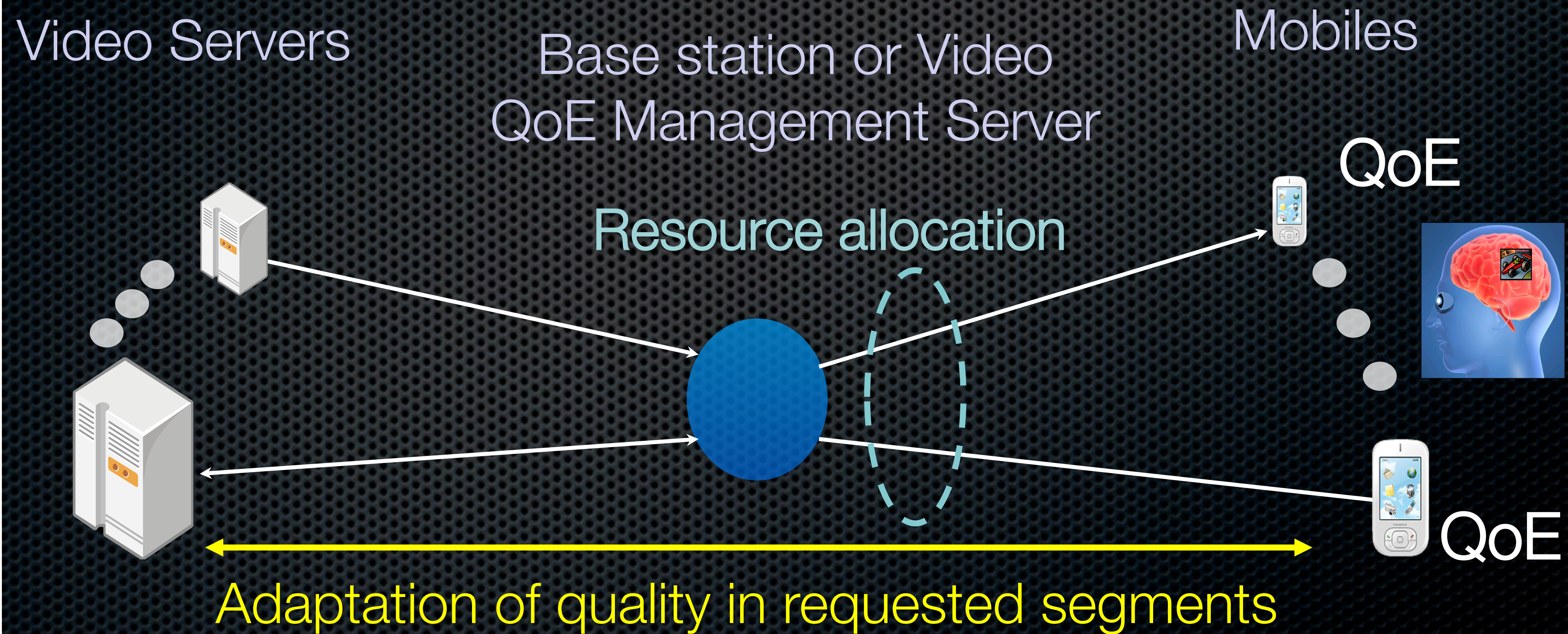


Only indirectly aware of users' Quality of Experience (QoE) i.e., through compression "rate"

Only indirectly optimizing QoE tradeoffs across users sharing (wireless) bottlenecks.

No complementary network resource allocation

Goal: "Optimal" but Practical Joint Multi-User Network Resource Allocation and Quality Adaptation



Talk Trajectory

- Humans are the "receivers"
  - Video Quality and Quality of Experience (QoE)

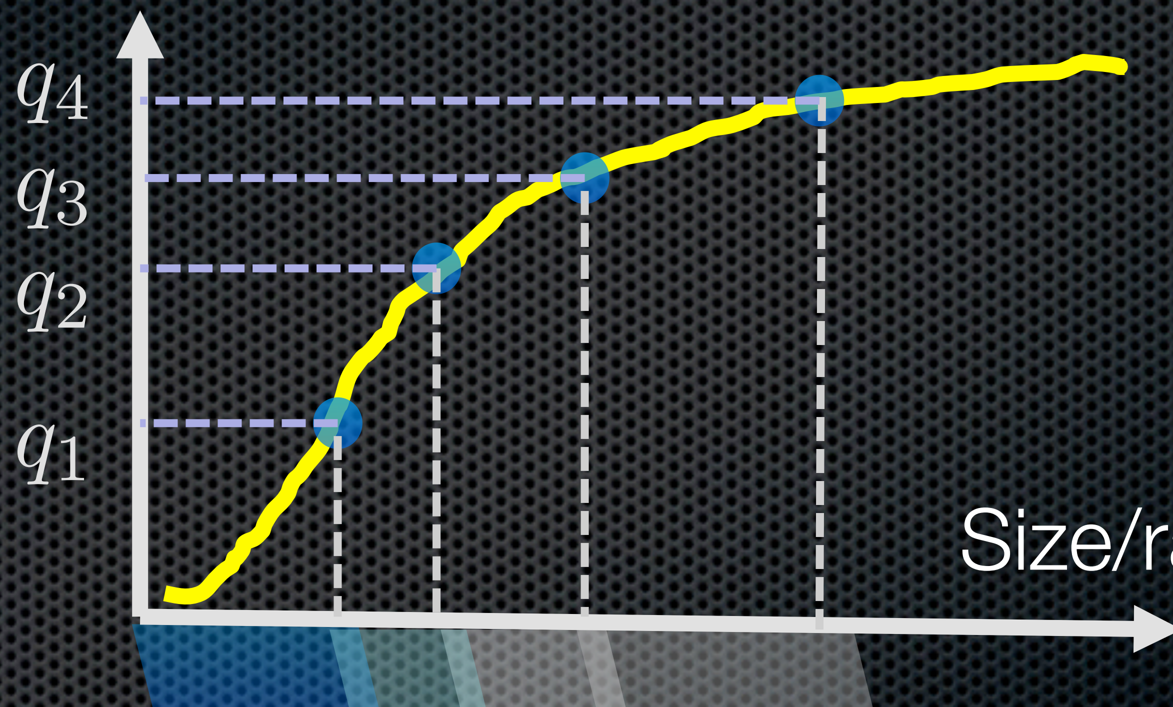
- New Class of Network Utility Maximization Problem
  - Algorithms which optimize QoE of delivered video

- Performance Evaluation and Comparisons

Quality vs Segment Size Tradeoffs

Objective Metrics Tracking Subjective Quality

Segment video Quality

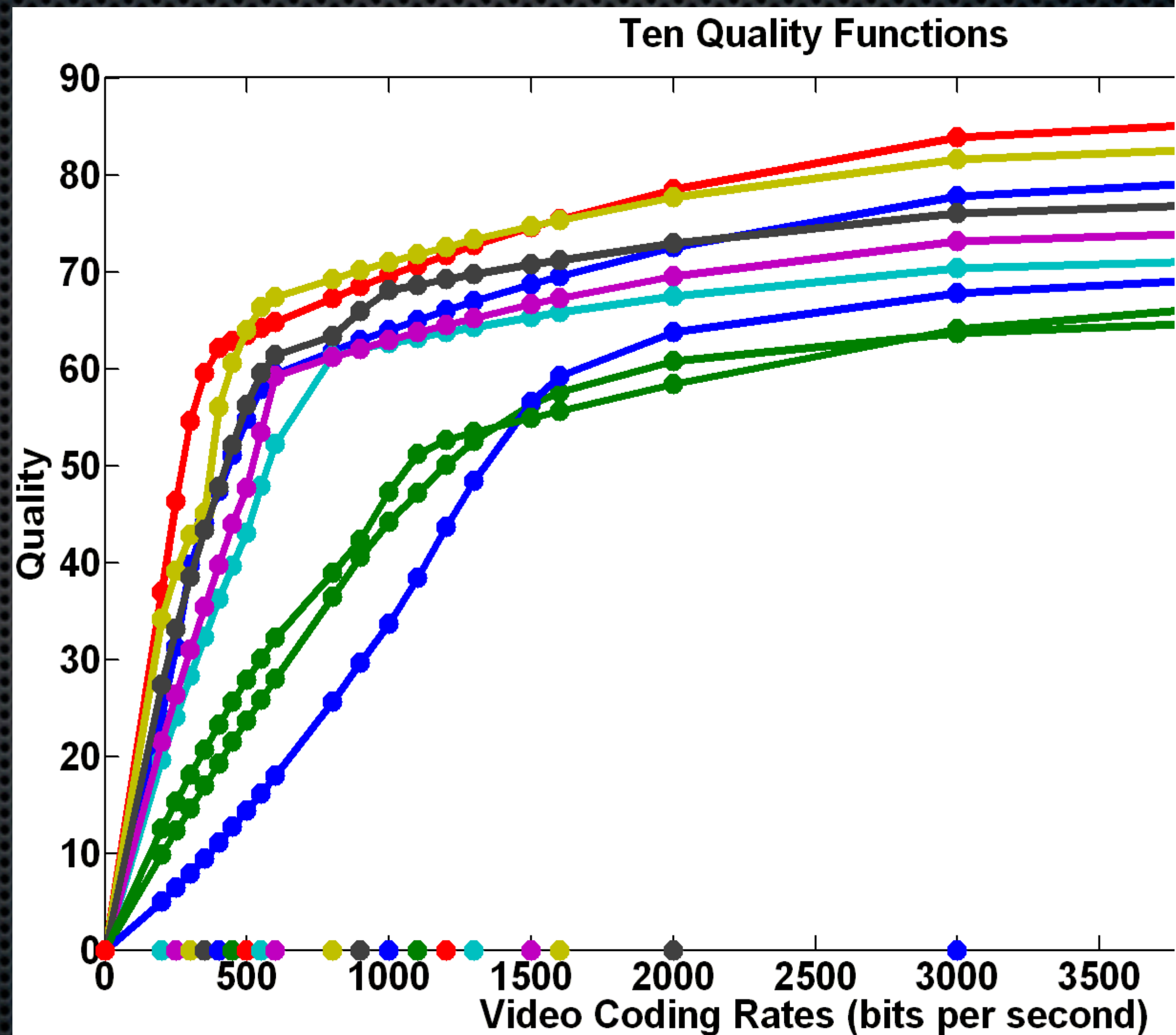


Scalable or Adaptive Video Coding

10+ years of image/video quality research: computable "utility functions" -> key abstraction to drive resource allocation

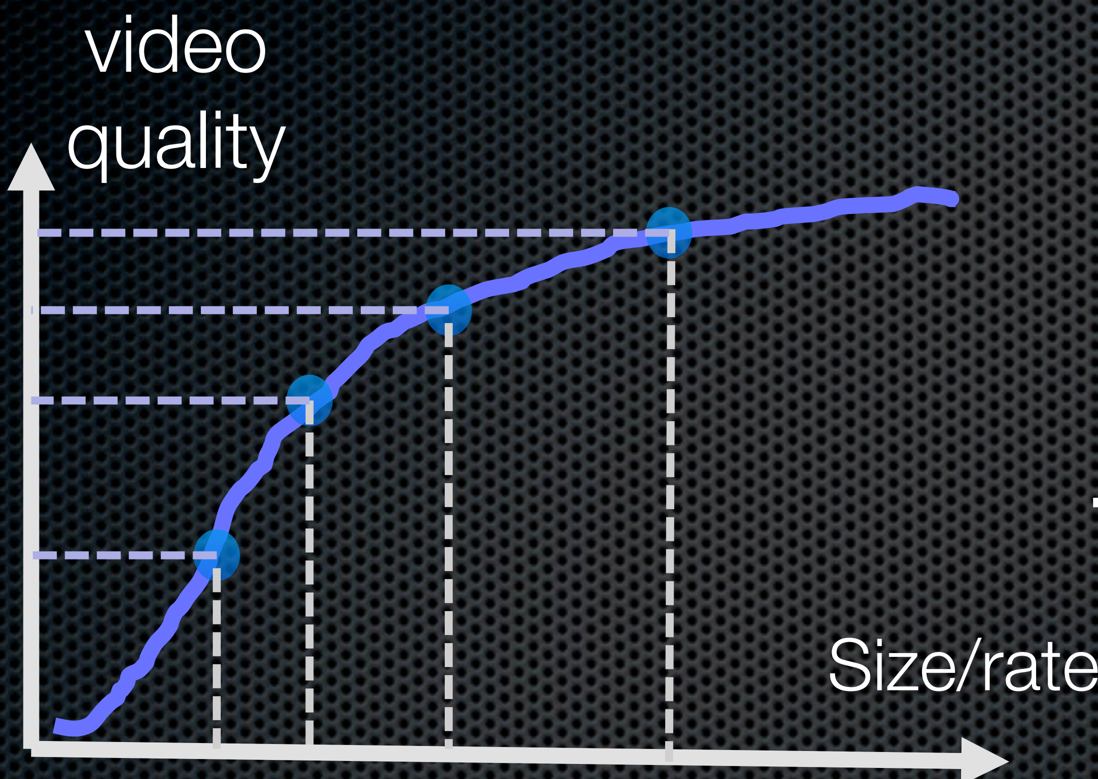
Quality vs Segment Size Tradeoffs:
Heterogeneous and Temporally Variable

- device dependent
- content dependent
- and time-varying,
  - i.e., across segments



Seshadrinathan et al. LIVE Video Quality Database 2010

Optimizing Video Delivery for Humans' Quality of Experience



+ aversion to variability in quality

Perceptual aspects of video quality.

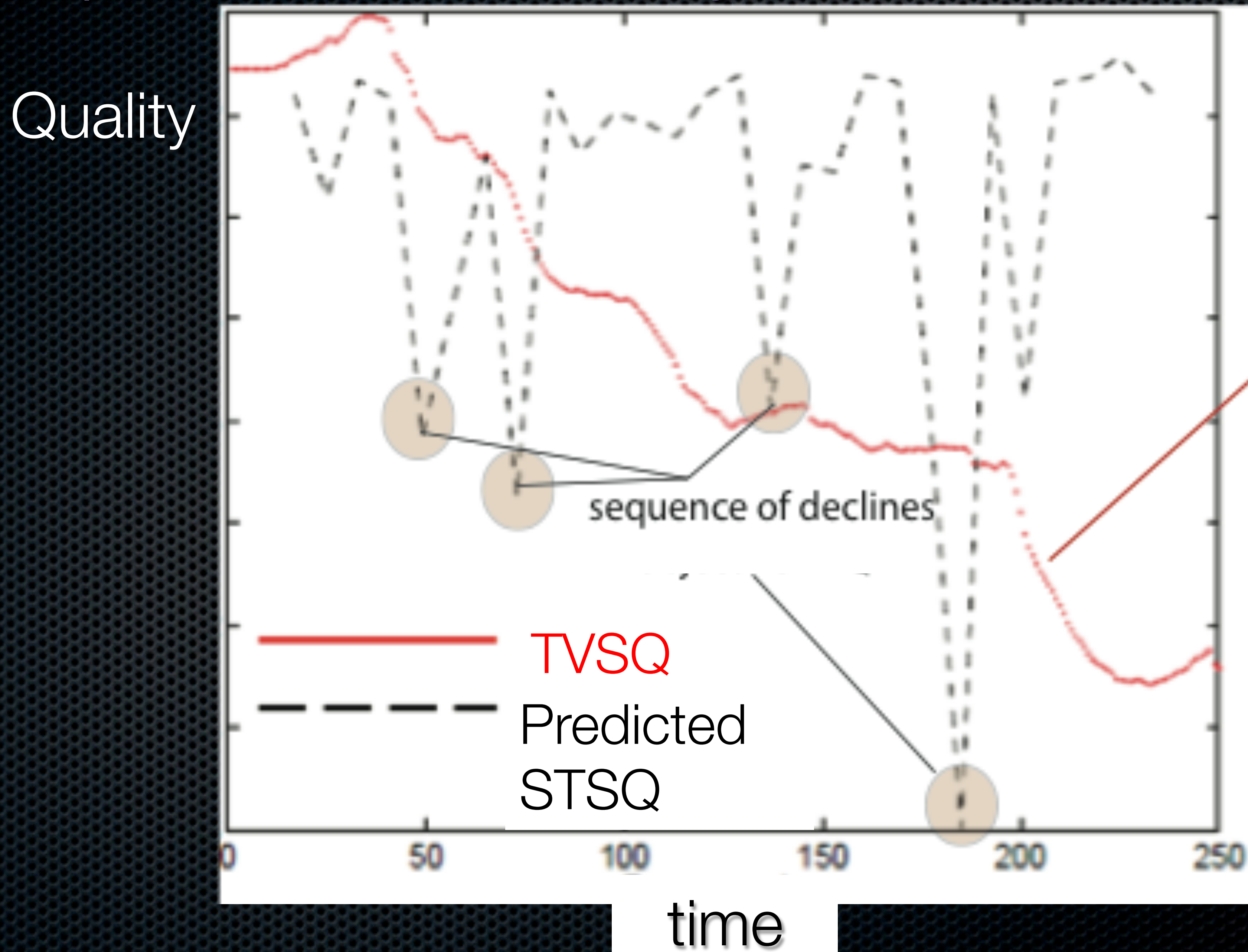
Behavioral aspects of video quality, e.g., memory

STSQ: Short Term Subjective Quality

TVSQ: Time-varying Subjective Quality

VQ vs Quality of Experience (QoE)

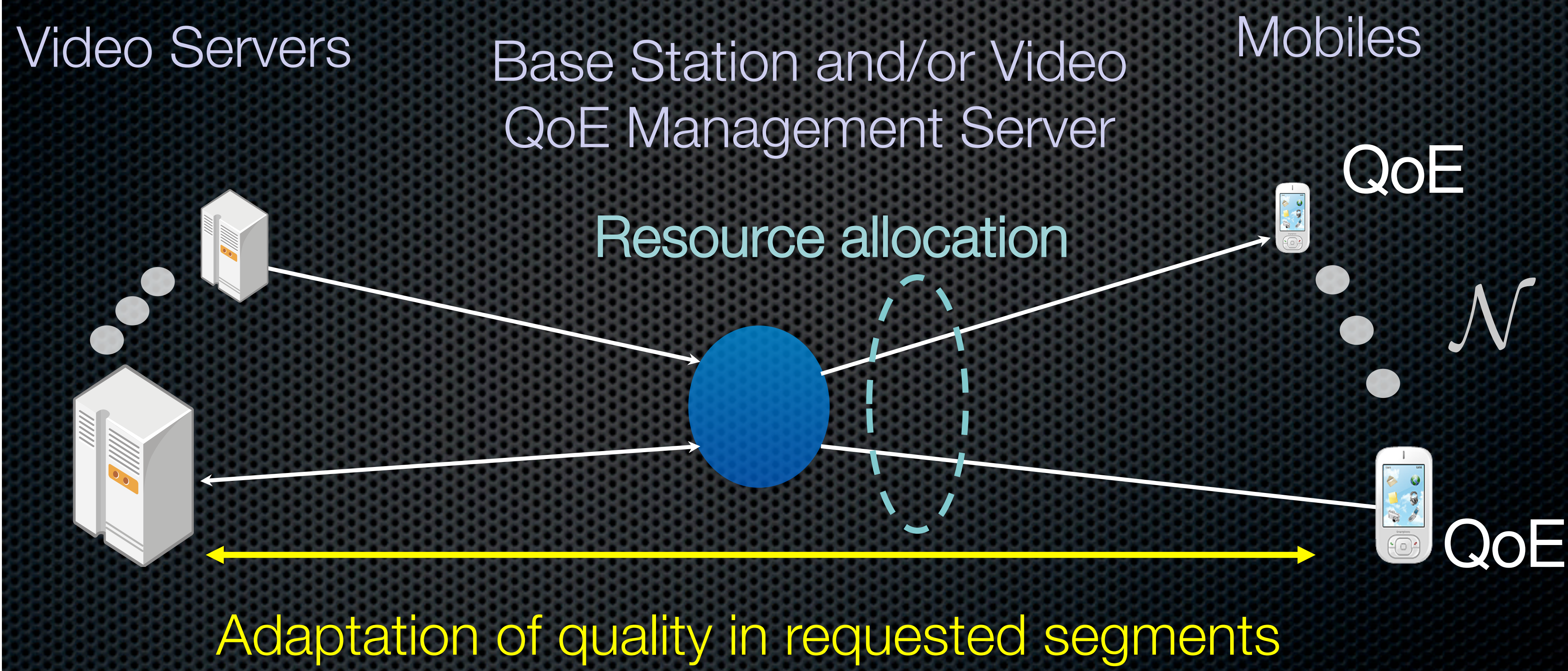
Temporal Dimension - Hysteresis



Right Metric(s) to Capture Video Quality and Drive Resource Allocation

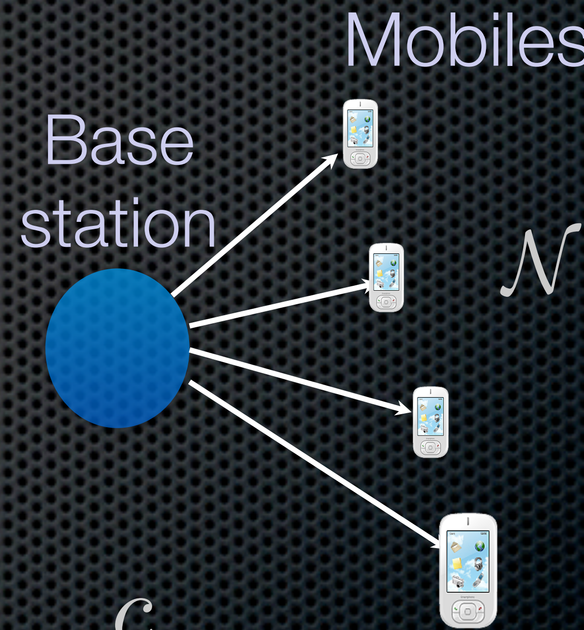
- Universal all encompassing metric?
  - Temporal variability in video quality
  - Rebuffering: dynamics & startup time
- Tractable metric
  - Approximate quality-size tradeoffs
  - Capture aversion to quality variability
  - Prioritize controlling rebuffering
  - Enable user specific QoE preferences/tradeoffs

Model and Theory …

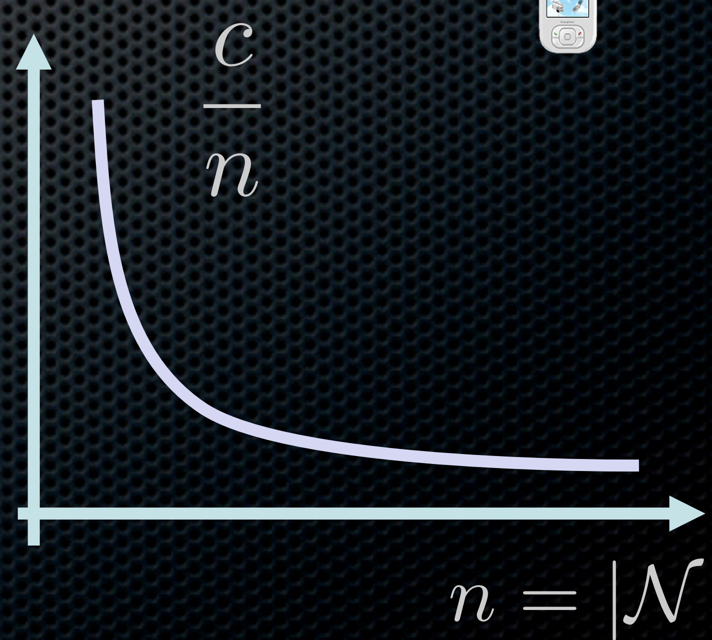


Heterogeneity and Variability in Users' Wireless/Network Capacity

- path loss, shadowing
- fast fading, interference
- mobility, load variability

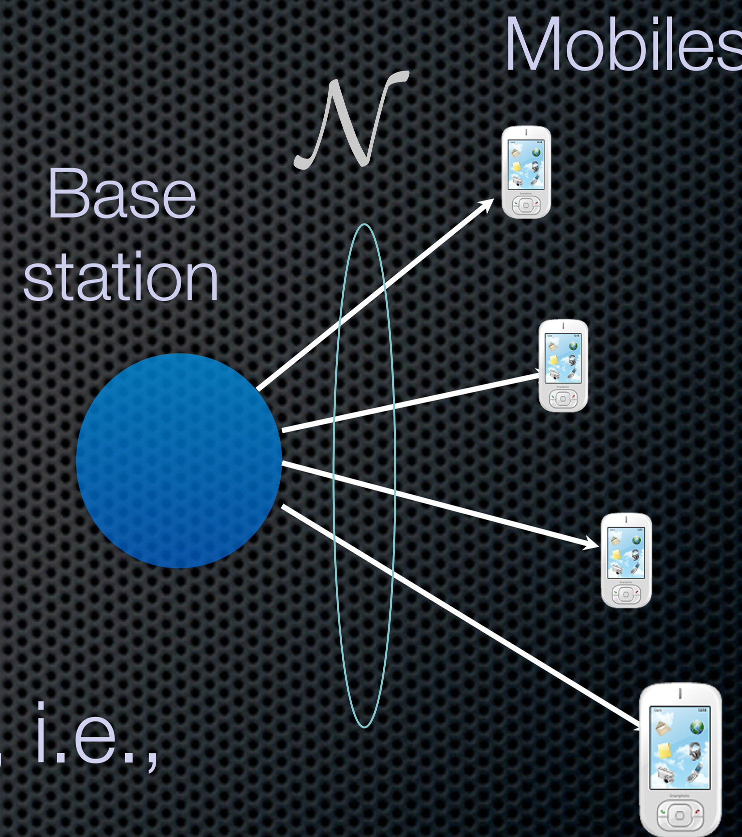


4G densification increases system capacity as well as per user capacity variability



Users' Rate Allocations in time "slot" k

$$\mathbf{r}_k = (r_{ik})_{i \in \mathcal{N}}$$



Feasible Allocations (current)

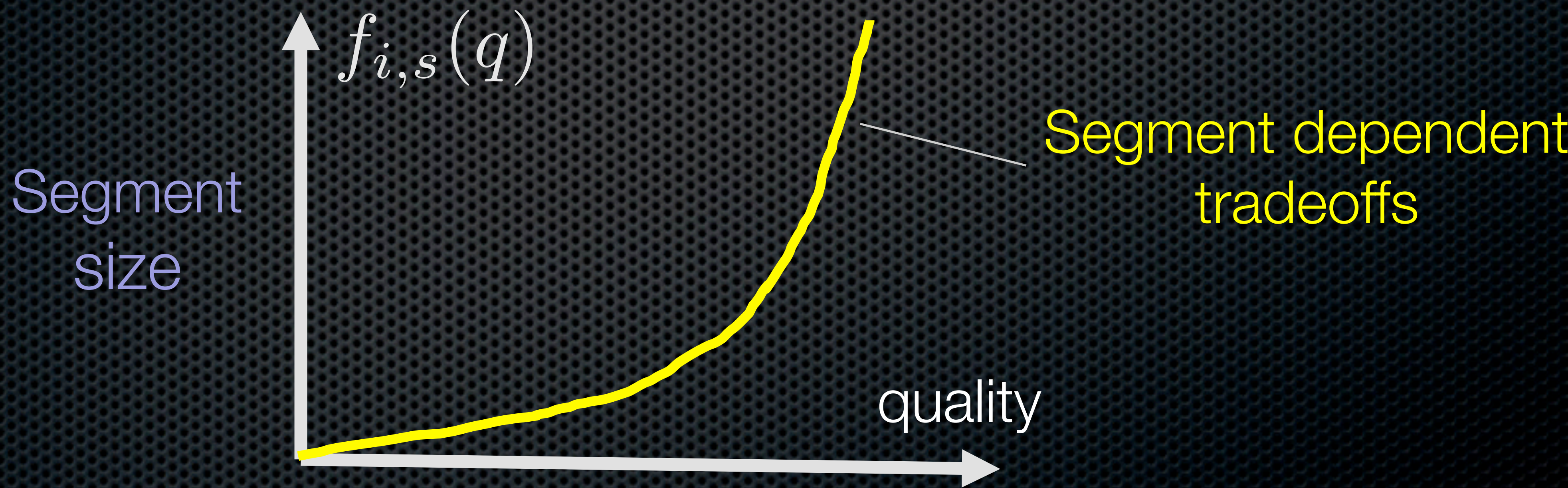
$$c_k(\mathbf{r}_k) \leq 0$$

where c_k is a convex function, i.e.,

Quality Adaptation: Each user i

- sequentially downloads video segments indexed by s each corresponding to viewing time τ_{seg}

- size (in bits) of segment s is an increasing convex function $f_{i,s}$ of the selected quality $q_i(s)$



Utility Maximization: Resource Allocation and Quality Adaptation

$$\max \sum_{i \in \mathcal{N}} U_i^E(\ m_i\ )$$

$$m_i = \frac{1}{S} \sum_{s=1}^{S} q_i(s)$$

Fairness/priority across users' allocations

Mean quality seen by user i

Optimizing over feasible

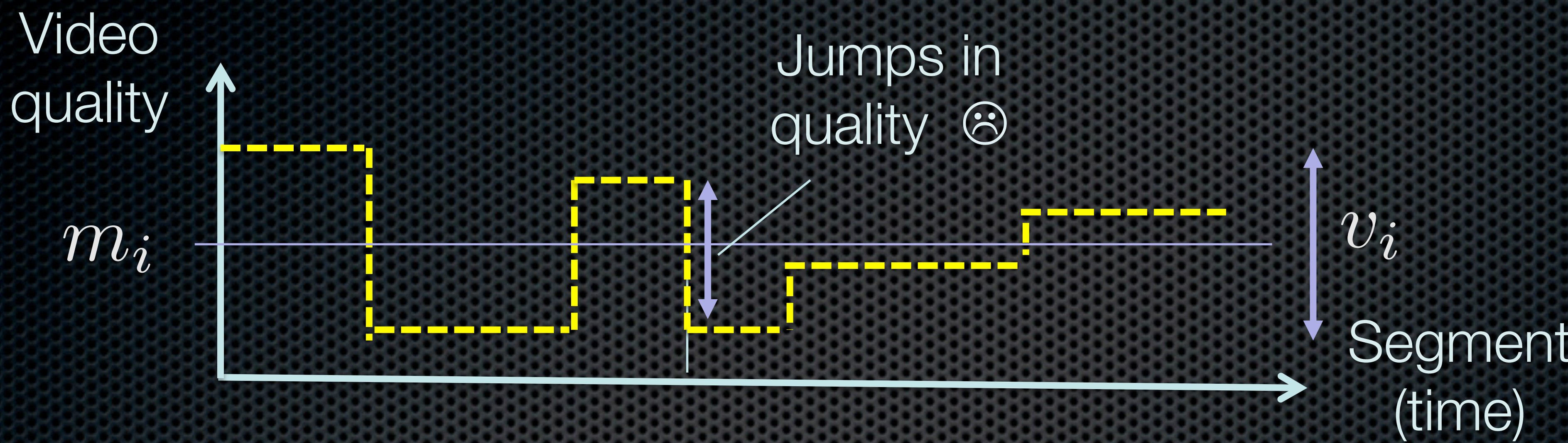
rate allocations per slot

&

quality choices per segment

Note: optimizing over temporal variations of both wireless capacity and quality-rate tradeoffs!

QoE Proxy Metrics: Temporal Dimension



High mean video quality is good m_i

“Variability” in video quality is bad v_i

$$\mathrm{QoE}_i = m_i - \boxed{U_i^V(v_i)}$$

Penalty for variability in segment quality

Mean-variability tradeoff

Extending Utility Maximization Framework:
Utilities which are sensitive to variability

Proxy for User i's QoE

$$\max \sum_{i \in \mathcal{N}} U_i^E\left(\boxed{m_i - U_i^V(v_i)}\right)$$

Fairness across
users' allocated QoE

Penalty function for variability
in a user's quality choices

Optimizing over
feasible

Rate
allocations
per slot

&

Quality
choices
per segment

Extending Utility Maximization Framework: Additional Constraints

$$\max \sum_{i \in \mathcal{N}} U_i^E(\, m_i - U_i^V(v_i)\,)$$

Optimizing over feasible

rate allocations per slot

&

quality choices per segment

Constraint on % rebuffering for each user.

Constraint on average cost/unit time for each user.

Offline Joint Resource and Quality Adaptation

$$\max \sum_{i \in \mathcal{N}} \cancel{U_i^E}(\ m_i - U_i^V(v_i)\ )$$

Lets simplify for this talk ☺

$$c_k(\mathbf{r}_k) \leq 0, \ \forall k$$

Feasible rate allocation per time slot time varying capacity/quality-rate

$$\beta_i \leq \bar{\beta}_i, \quad \forall i \in \mathcal{N}$$

Constraint on % rebuffering β_i for each user

$$p_i \leq \bar{p}_i, \quad \forall i \in \mathcal{N}$$

Constraint on average cost per viewing time p_i for each user

Our Online Solution

NOVA: Network Optimization for Video Adaptation Algorithm

1. A Simple distributed online algorithm
2. Strong optimality guarantees

Online Distributed Algorithm

1. Learns (estimates) key parameters associated with
   - mean and variability in quality
   - variability in system and (Lagrange multipliers) associated with rebuffering/cost constraints

2. Uses those parameters to perform
   - resource allocation in network each slot
   - segment quality adaptation at clients as segments complete

Online Algorithm: Learning Parameters

Client i keeps track of

$m_{i,s}$ = mean quality up to segment s

$v_{i,s}$ = variance in quality up to segment s

Easy segment driven updates!

$b_{i,k}$ = Lagrange multiplier associated with rebuffering constraint at slot k (large -> playback buffer is low)

$d_{i,k}$ = Lagrange multiplier associated with cost constraint at slot k (large-> cost is getting too high)

Online Algorithm: Learning Parameters

Client i keeps track of virtual playback time queue

$$b_{i,k+1} = \max[b_{i,k} - \epsilon(\tau_{seg}), 0]$$

Upon segment transfer completion

$$b_{i,k+1} = b_{i,k} + \epsilon\left(\frac{\tau_{slot}}{(1+\bar{\beta}_i)}\right)$$

Upon slot completion

- Updated asynchronously!
- Large virtual playback time queue means segment delivery is not keeping up!

Online Algorithm: Resource Allocation

Beginning of each slot k base station/network allocates rate based on

$$\max_{\mathbf{r}} \quad \sum_{i \in \mathcal{N}} b_{i,k} r_i$$

Higher weight to users with large virtual playback queues

$$\text{s.t.} \quad c_k(\mathbf{r}) \leq 0, \qquad \mathbf{r} \geq 0$$

Current capacity constraint

- N variable convex optimization, linear program if capacity constraint are linear
- This is simply weighted proportional fair scheduling!

Online Algorithm: Quality Adaptation

Upon completion of segment s on slot k client i selects quality for segment $s+1$ based on

Video quality

Penalize variability

$$\max_{q \geq 0} q - (U_i^V)'(v_{i,k})(q - m_{i,s})^2 -$$

$$- \frac{b_{i,k}}{(1+\bar{\beta}_i)\tau_{seg}} f_{i,s+1}(q) - \frac{p_i^d d_{i,k}}{\bar{p}_i} f_{i,s+1}(q)$$

Penalize rebuffering

Penalize cost

Simple scalar convex optimization!

What can be rigorously shown?

Theorem*: Assuming stationary variations for quality-size tradeoffs and network capacity our online algorithm is asymptotically optimal !

Asymptotic optimality? Over long periods of time performance of

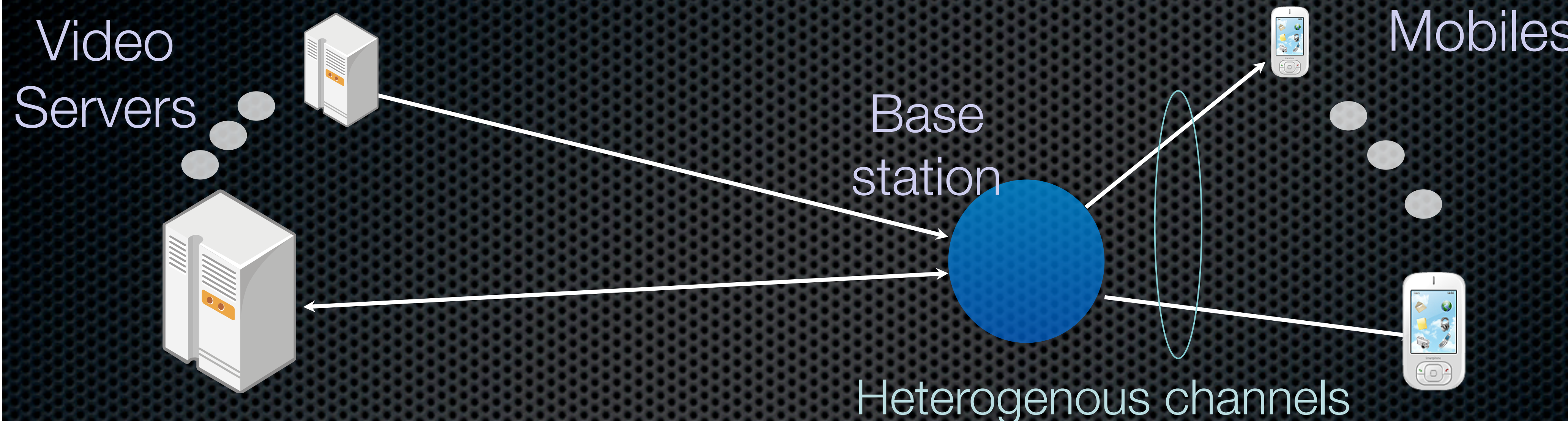
online algo. = optimal offline algo

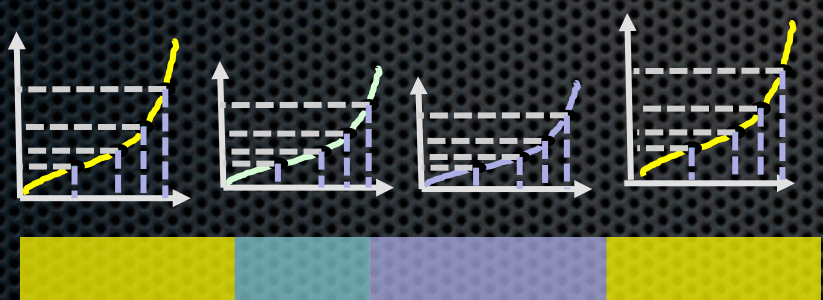
*Simplified statement. This result is quite challenging, role of temporal variations on utility, role of asynchrony, role of playback buffer.

Simulation Setup

Segment =1sec

Slot =10ms

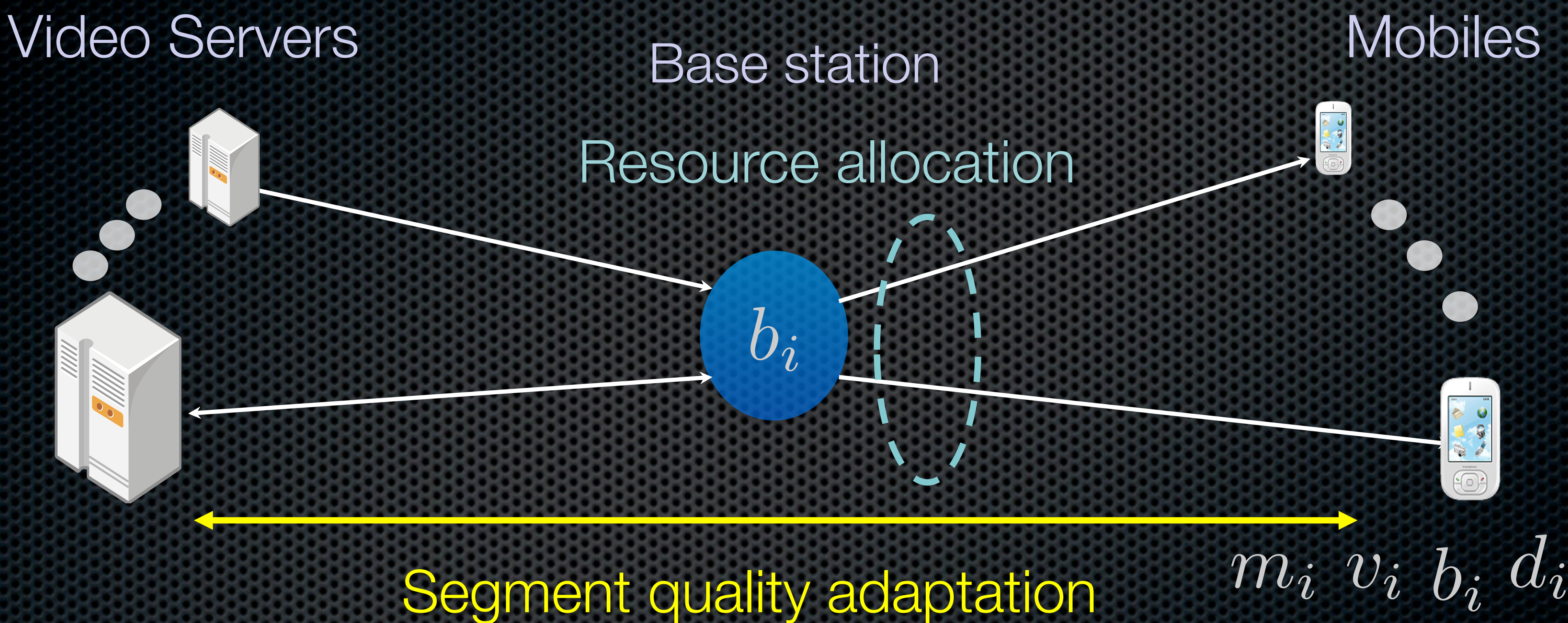




Real Video Segments :
6 compression rates
0.1,0.2, 0.3,0.6,0.9,1.5 Mbps

Slowly varying wireless capacity:
correlated samples from peak
rate distribution for HSDPA system

Simulation Setup: DASH framework

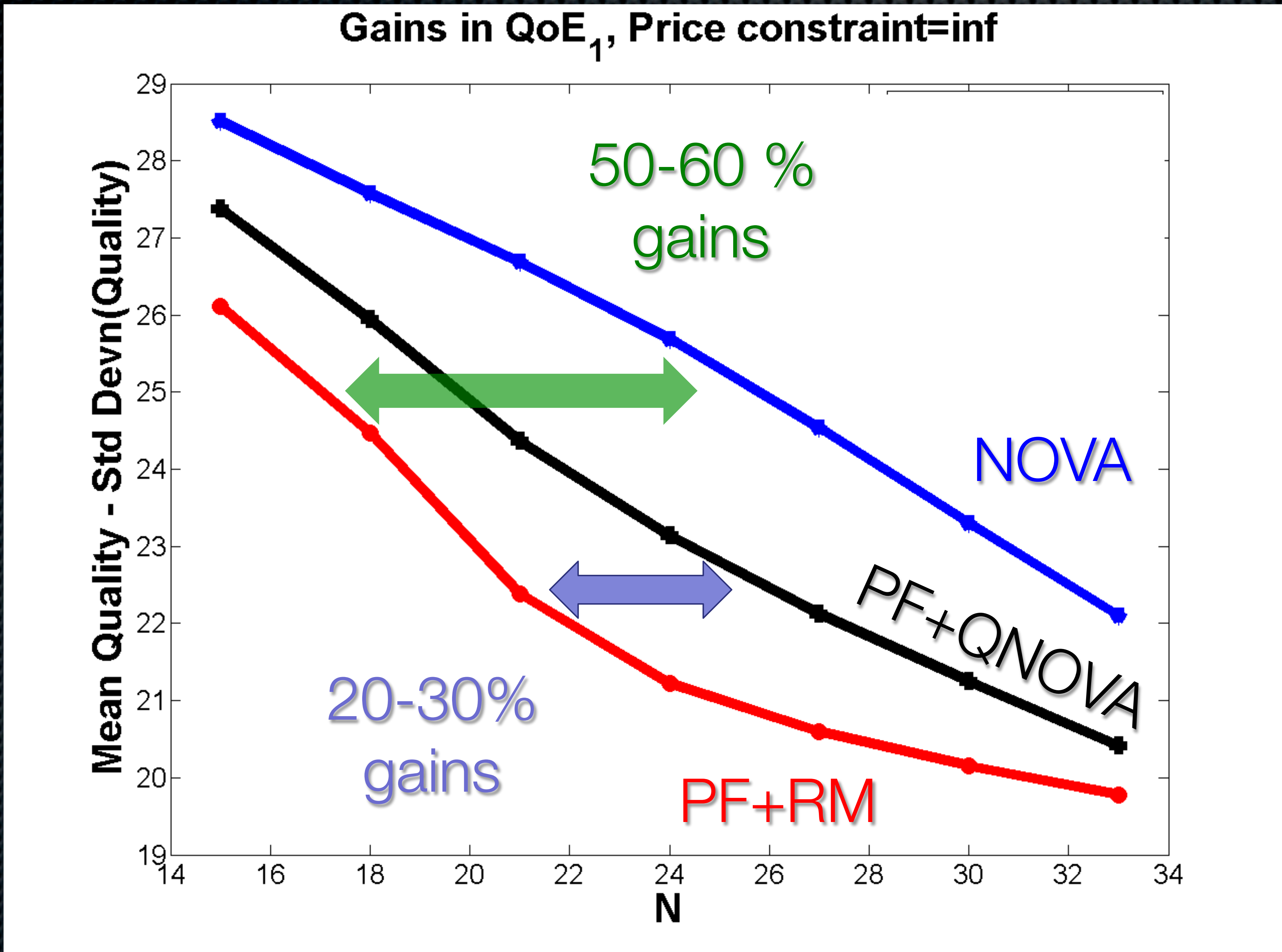


NOVA: our Joint Resource Allocation and Quality Adaptation

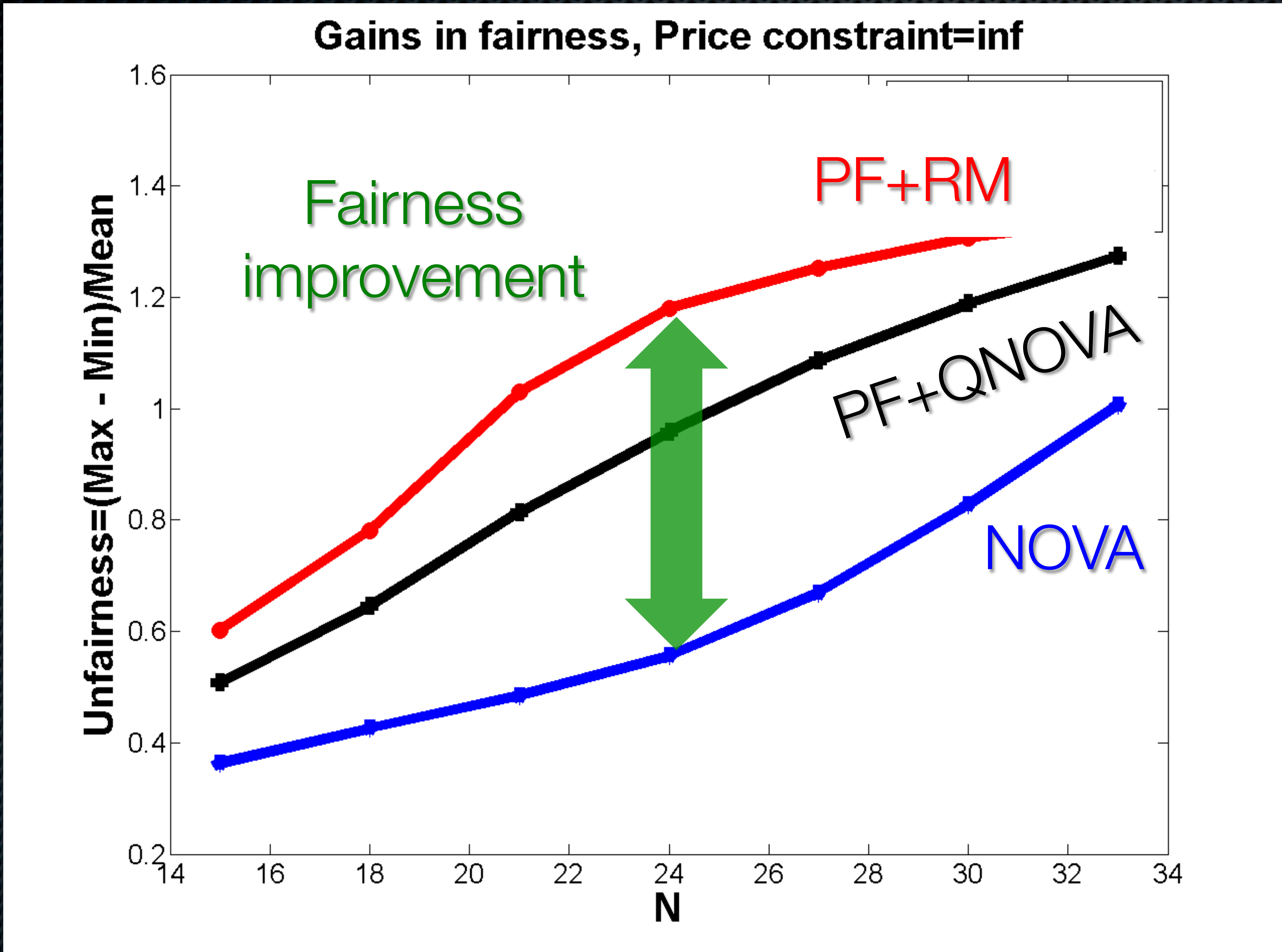
PF+QNOVA: Proportionally Fair Allocation + our Quality Adaptation

PF + RM: Proportionally Fair Allocation + Greedy Rate Matching

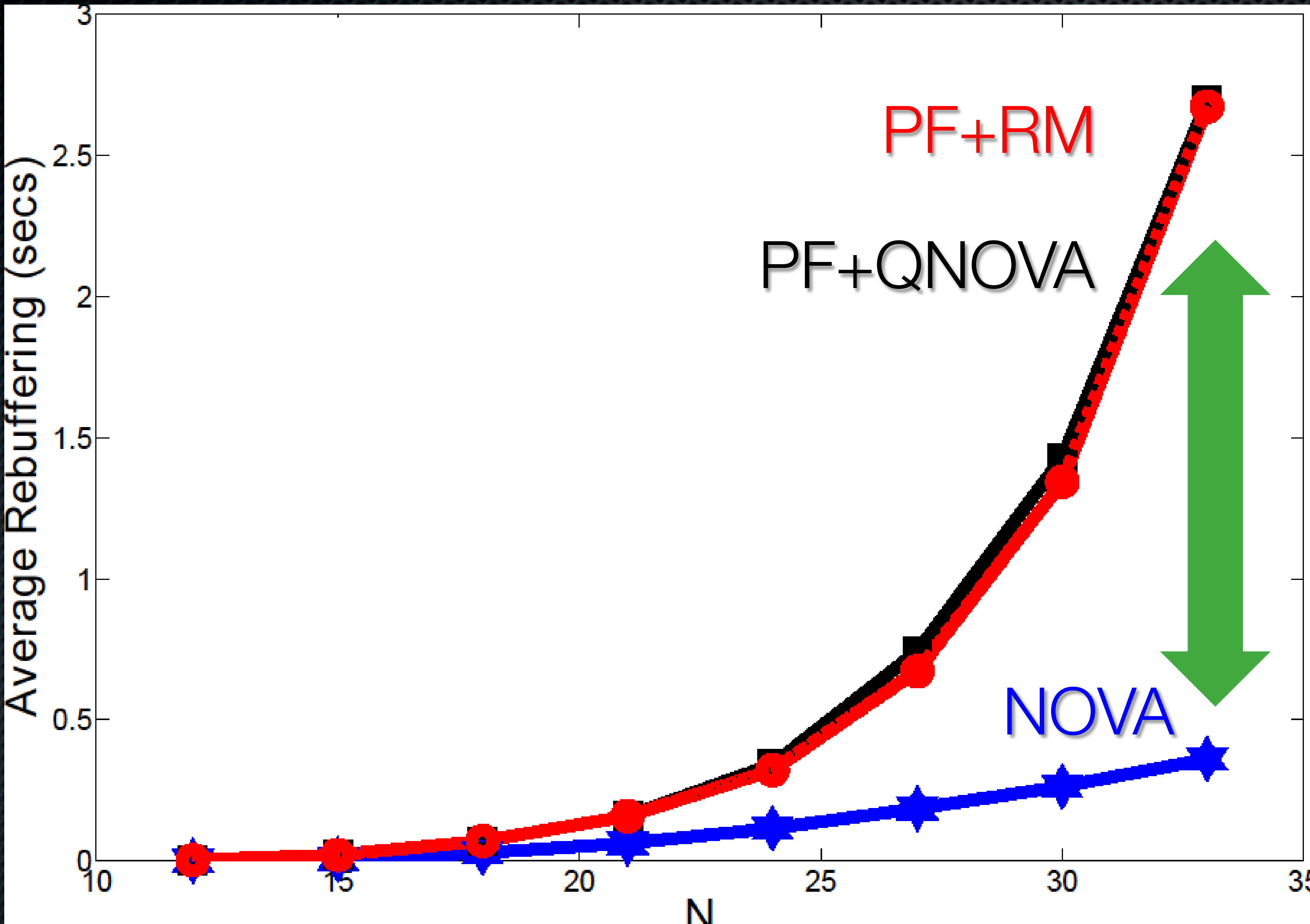
Improved Video `Capacity'



Improved Fairness



Improved Rebuffering



Take Aways

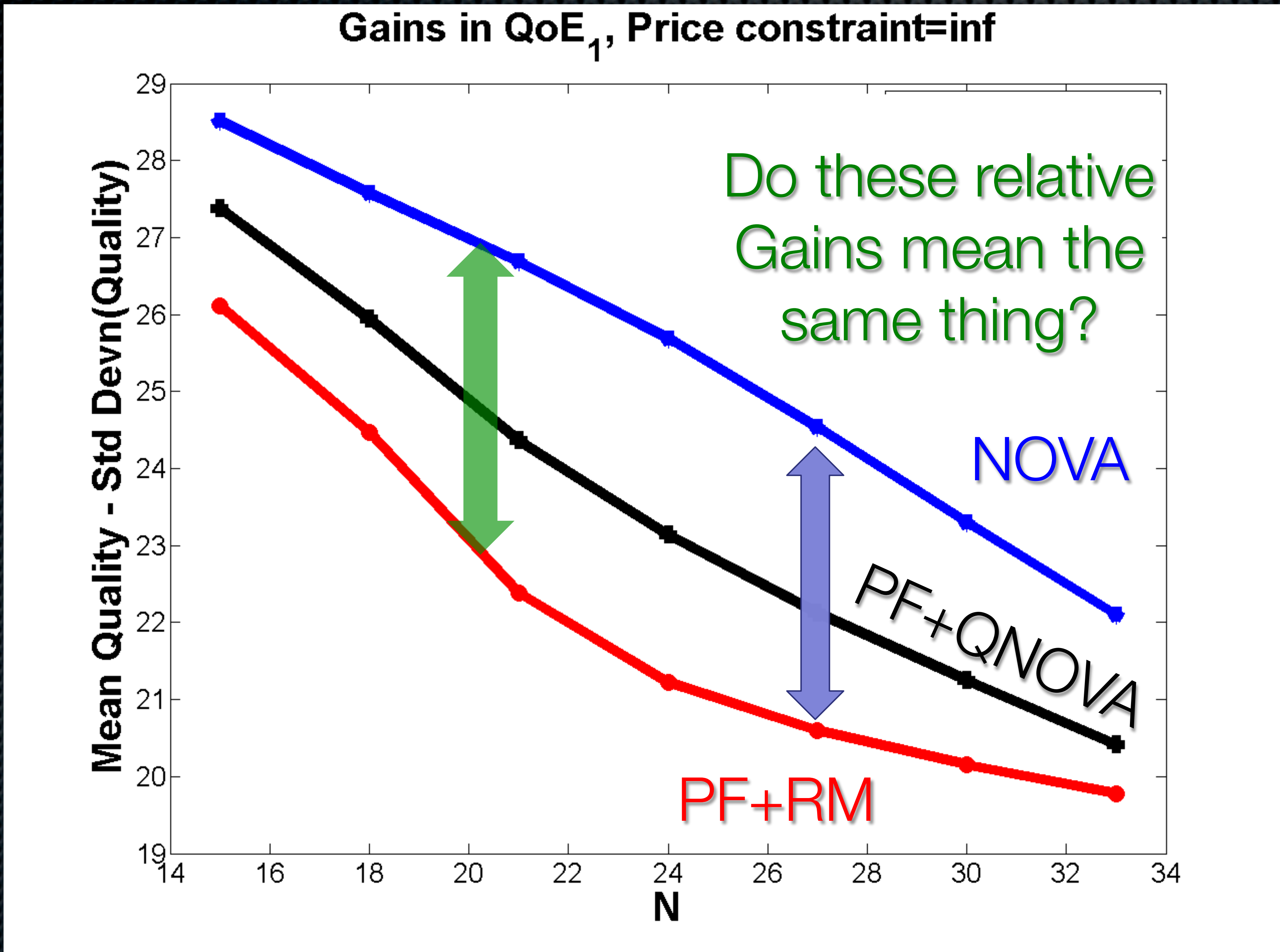
- Distributed, online theoretically “optimal” and practical algorithm.
- Asynchronous nature suits DASH framework.
- NOVA delivers 50-90% capacity gains* over baseline and quality adaptation (only) delivers 25-40%
- Substantial improvement in fairness over baseline
- Opportunity to build delivery infrastructure that is tailored to user, content or system provider preferences.

- Studied a new “buffered” network utility maximization where users are sensitive to “variability” make asynchronous decentralized choices.

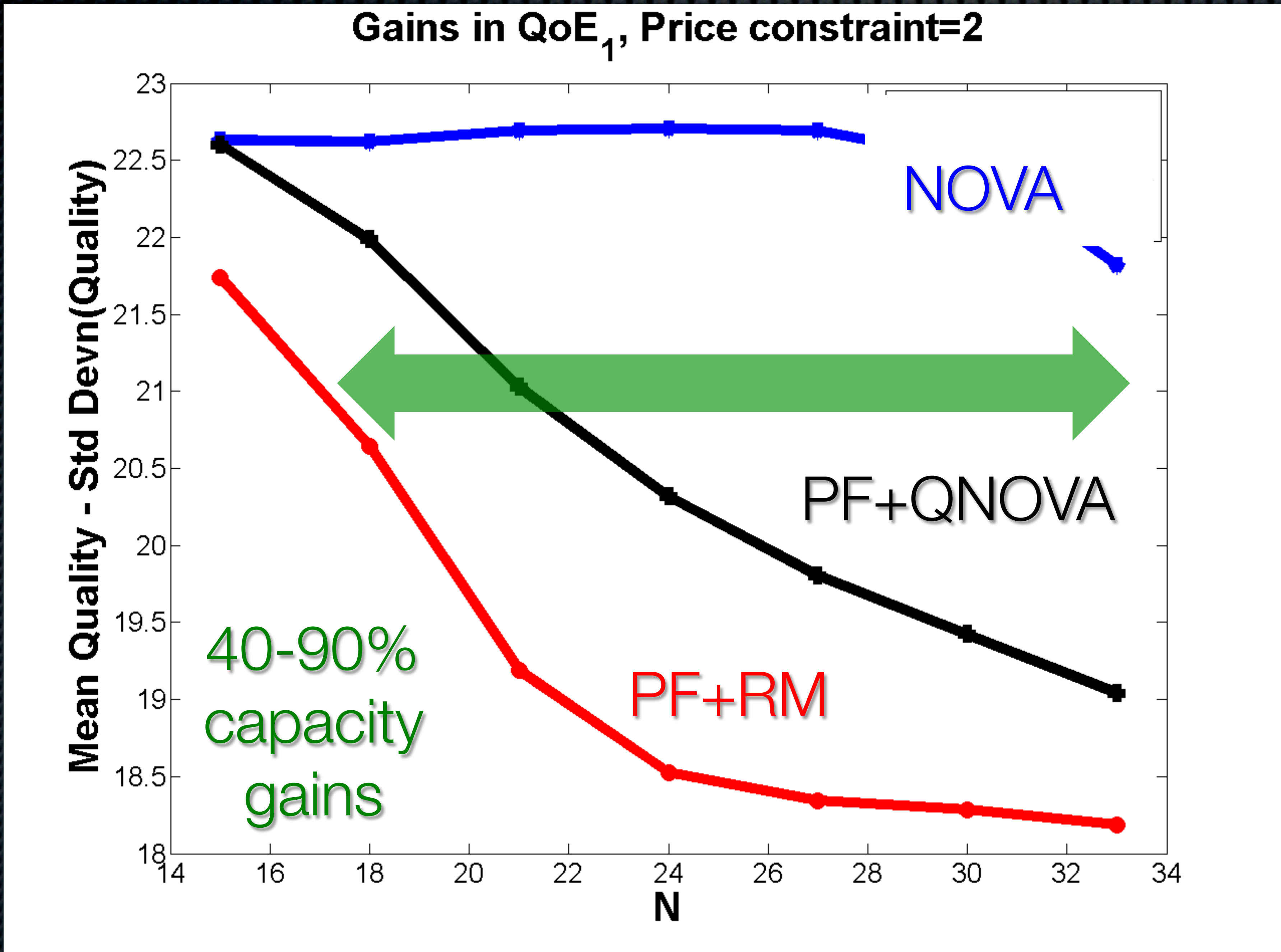
Practical Issues

- Can incorporate best effort data users in resource allocation
- Improvements are robust to “precision” Q-R tradeoffs
  - Can compress tradeoffs using parametric models/PSNR
- Can address legacy issues, e.g., no resource management
- Progressive download vs real-time streaming algorithms?
  - Just limit the client side buffering reduce benefits
- Startup behavior is tuned for aggressive at start

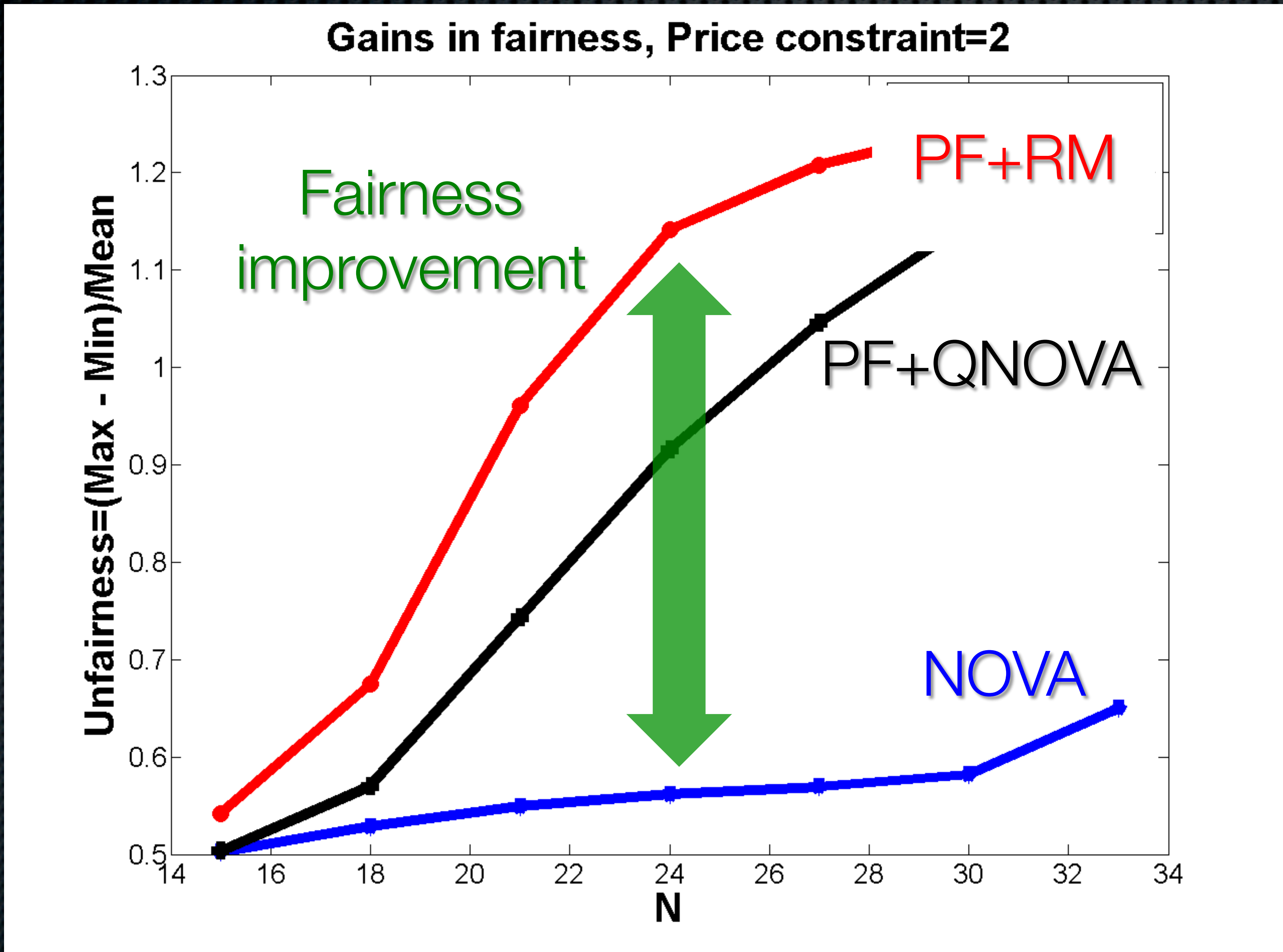
Improved Video `Quality' ?



Improved Video `Capacity'



Improved Fairness



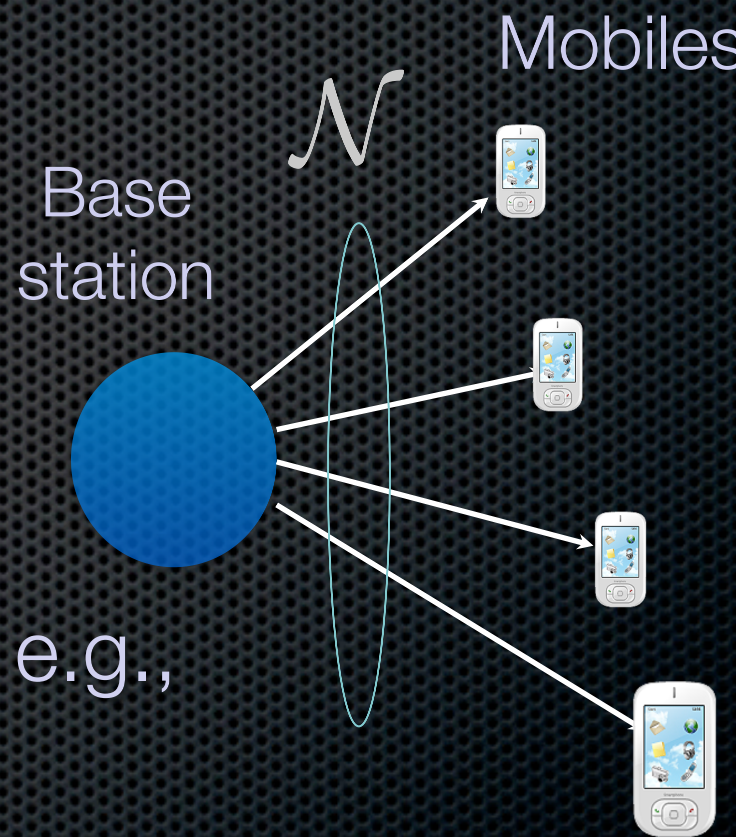
Need Framework That Addresses

- Tradeoffs on mean vs variability in video quality
- Addresses primacy of rebuffering vs video QoE
- Fairness (or prioritization) in allocating QoE across users.
- Accounts for average cost/sec to maintain video QoE
- Can support heterogeneous/content/device dependent user preferences

We delivered theoretically optimal & practical algorithm to achieve these goals in DASH framework

Users' Rate Allocations in time "slot" k

$$\mathbf{r}_k = (r_{ik})_{i \in \mathcal{N}}$$



Feasible Allocations (current)

$$c_k(\mathbf{r}_k) \leq 0$$

where c_k is a convex function, e.g.,

$$c_k(\mathbf{r}_k) = \sum_{i \in \mathcal{N}} \frac{r_{ik}}{p_{ik}} - 1 \leq 0$$

peak rate to user i in slot k