Eliminating Adverse Control Plane Interactions in Independent Network Systems

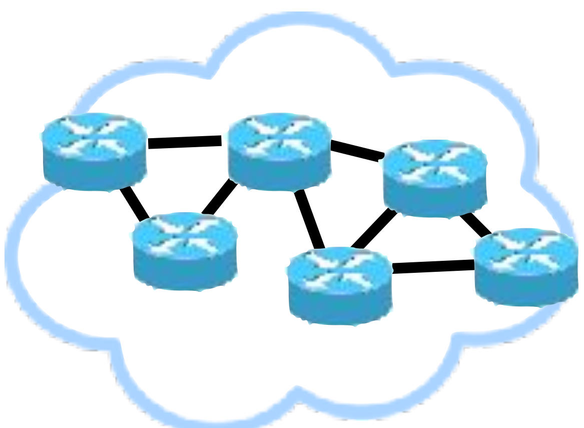
Matthew K. Mukerjee

Computer Science PhD Thesis Defense

May 1st, 2018

Carnegie Mellon University

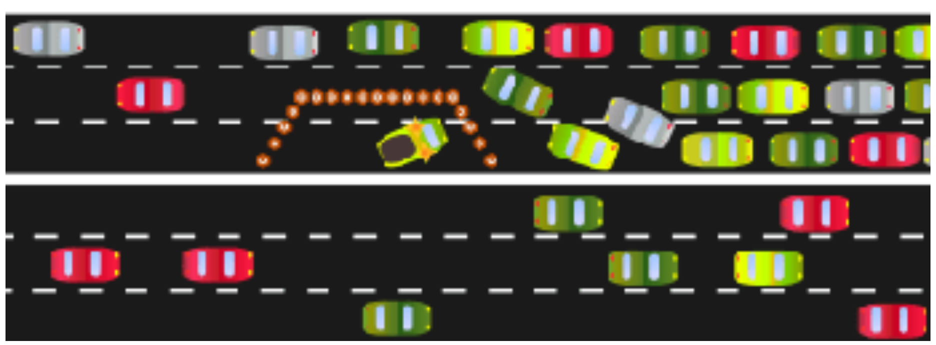
Network Control



Routing

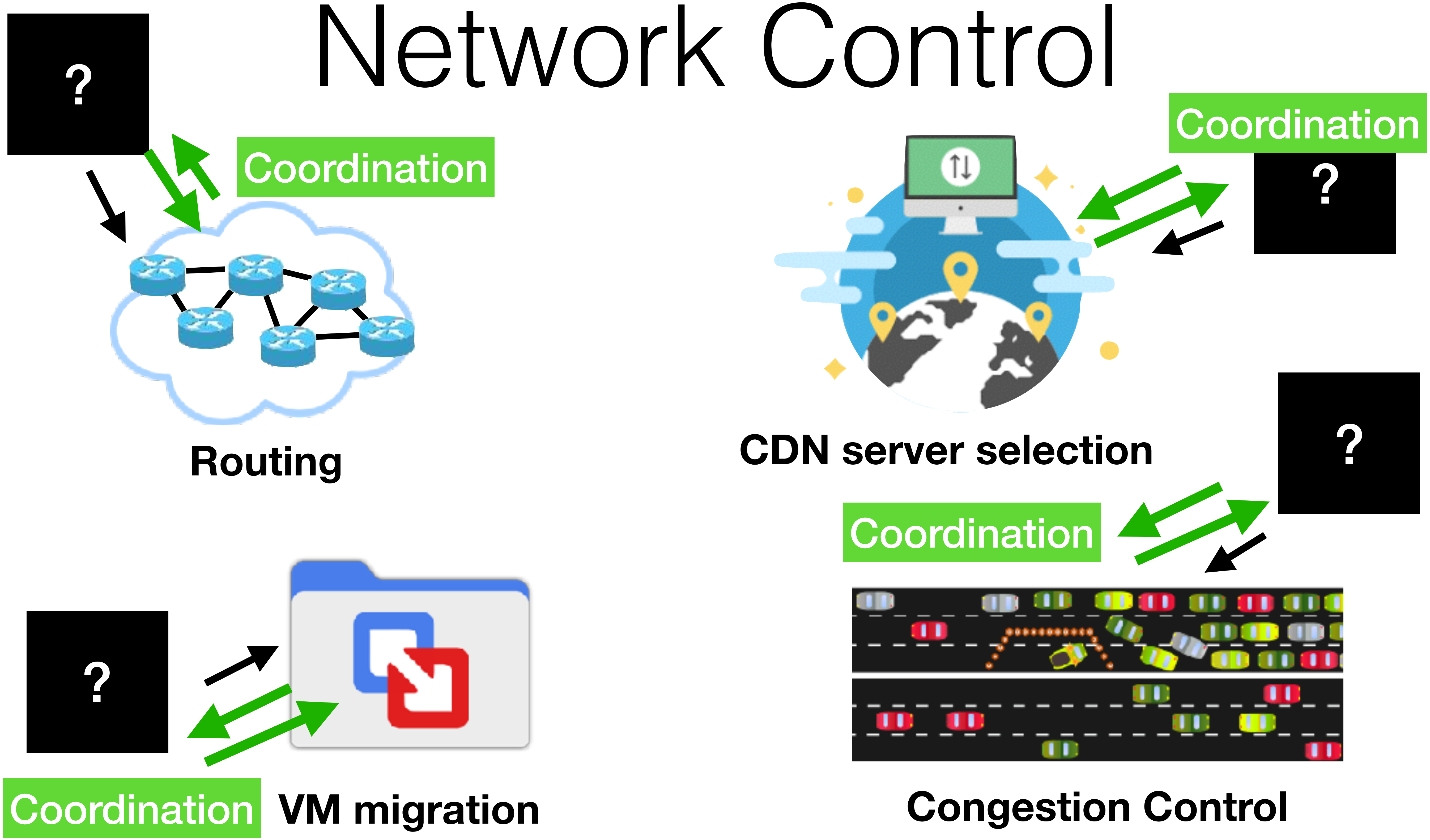
CDN server selection

VM migration



Congestion Control

Network Control



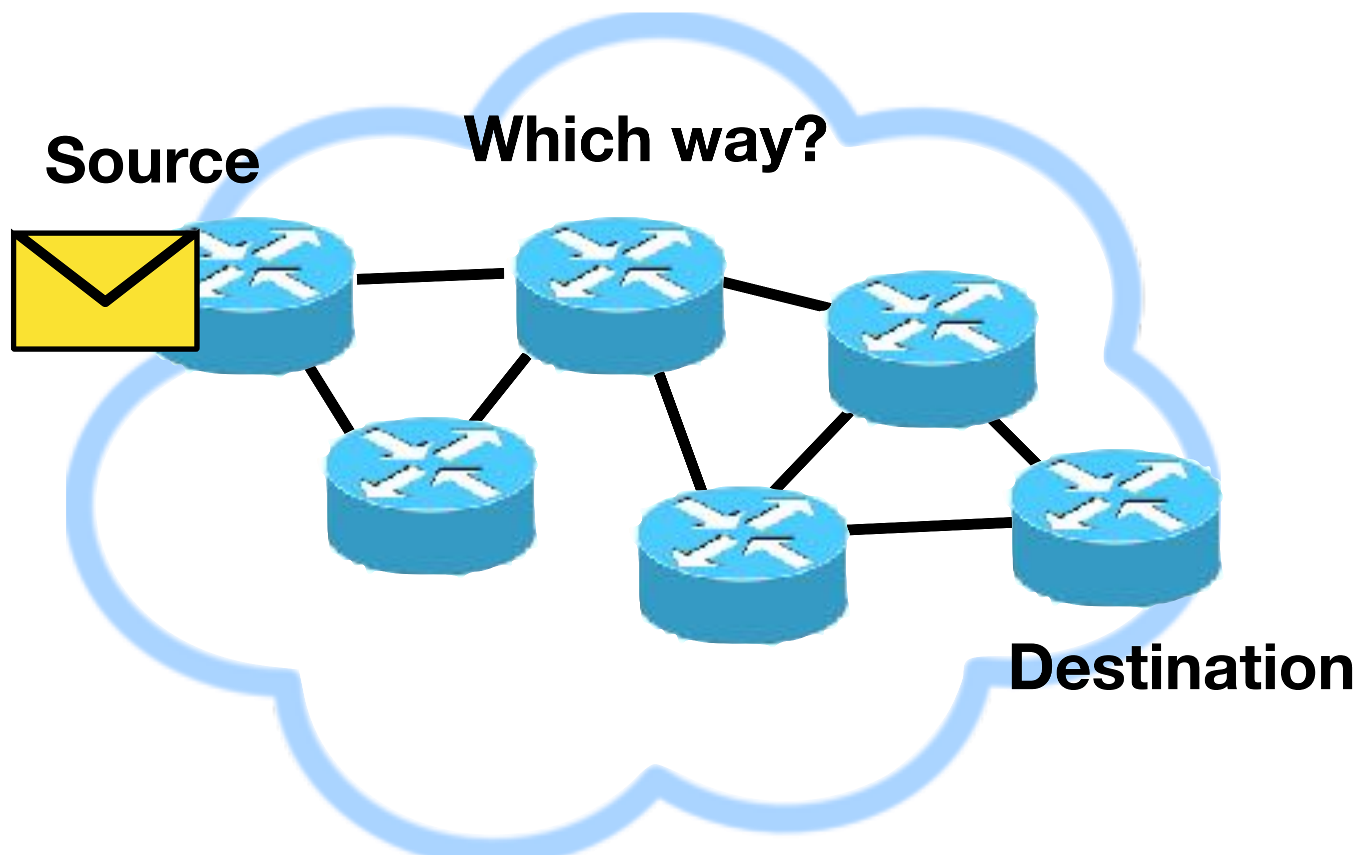
Control Plane

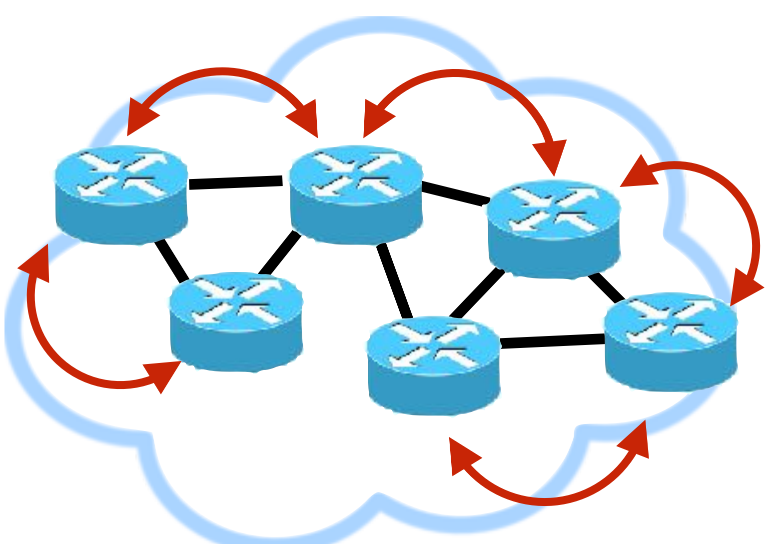
Routing:
"figure out" best path
(periodically computed)

Control

Data Plane

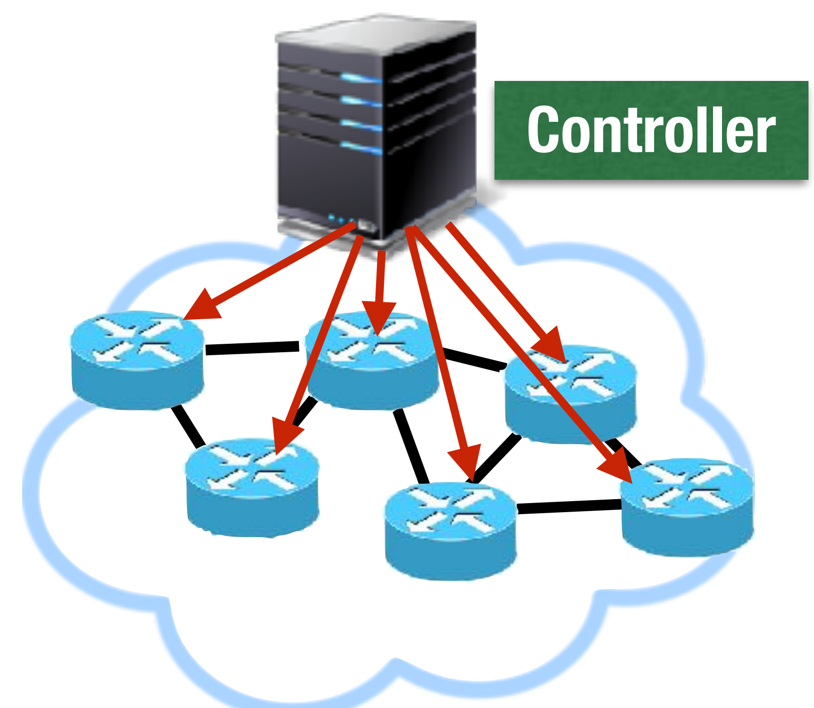
Forwarding:
data transmission
(done per-packet)





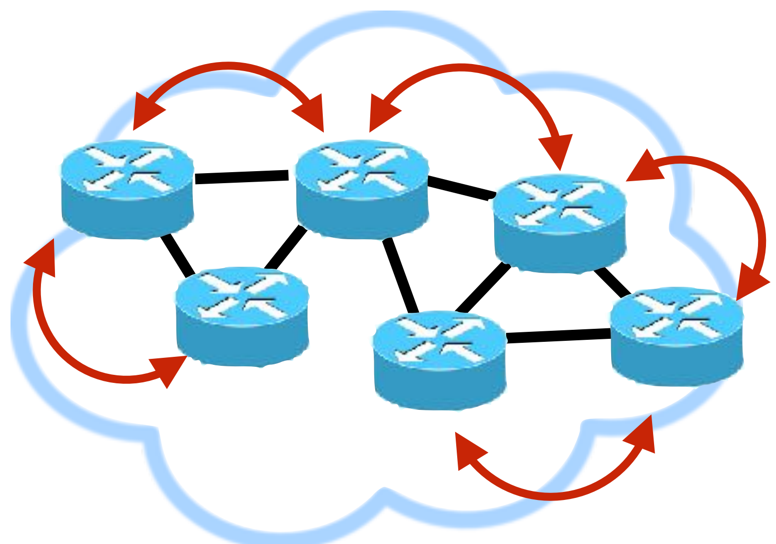
Distributed

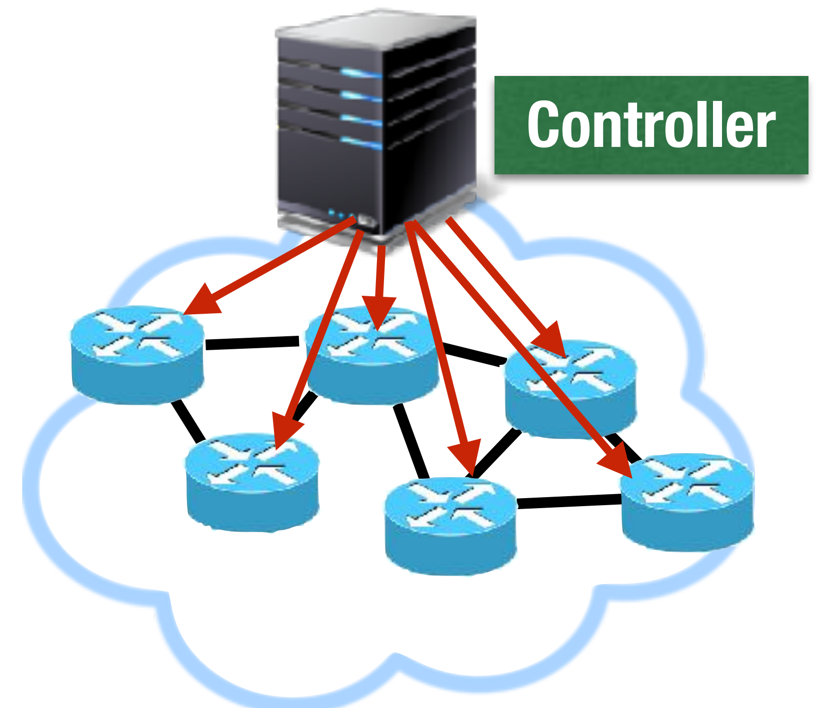
OSPF



Centralized

SDN





Distributed

OSPF

Quick failure response

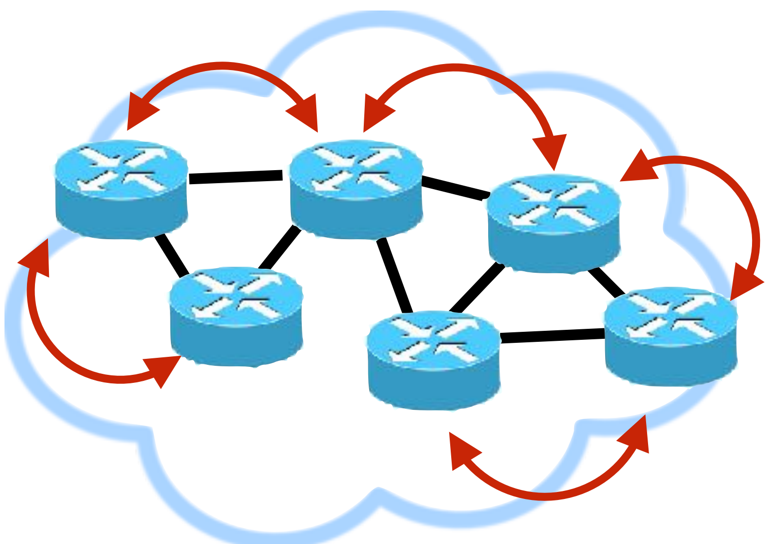
Bad at performance optimization

Centralized

SDN

Good at performance optimization

Slow failure response

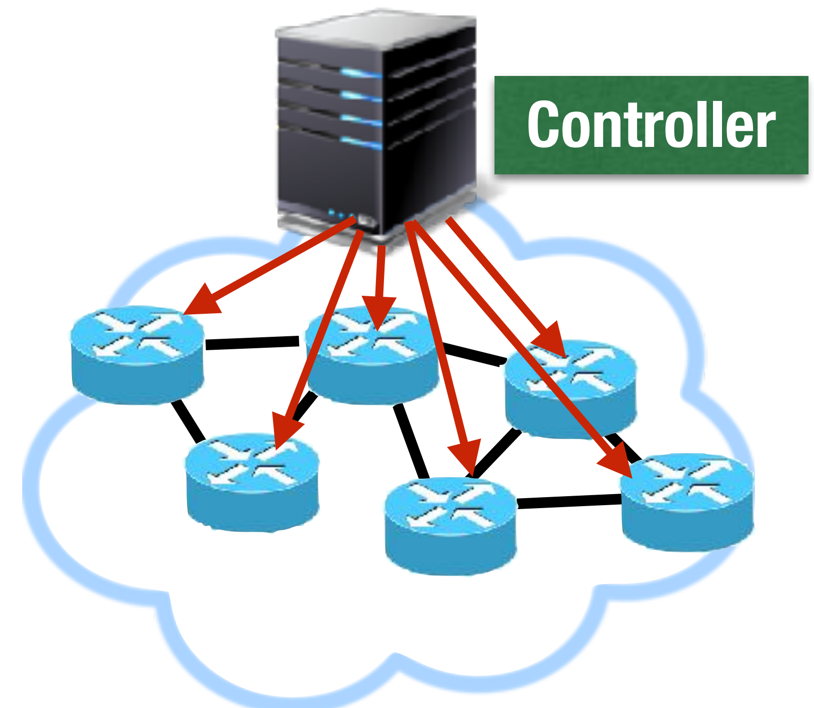


Distributed

OSPF

Quick failure response

Bad at performance optimization



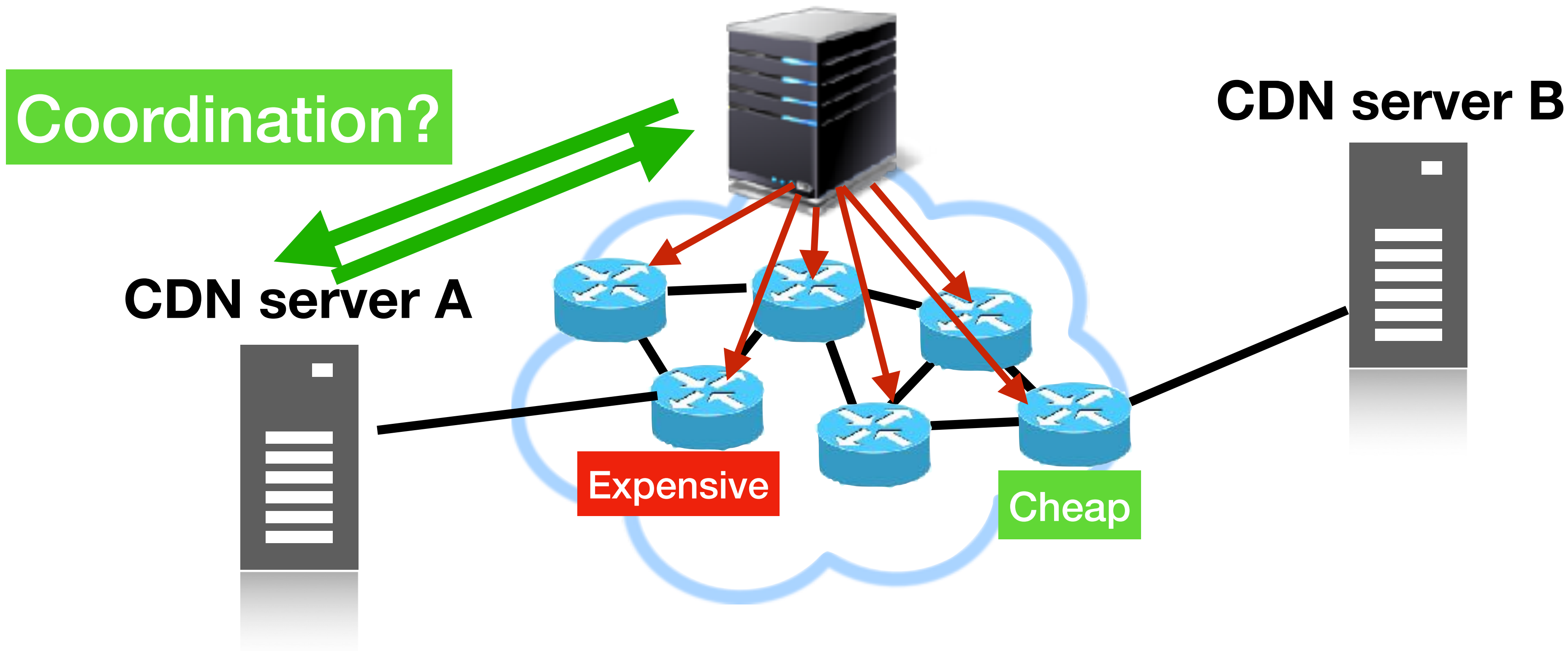
Centralized

SDN

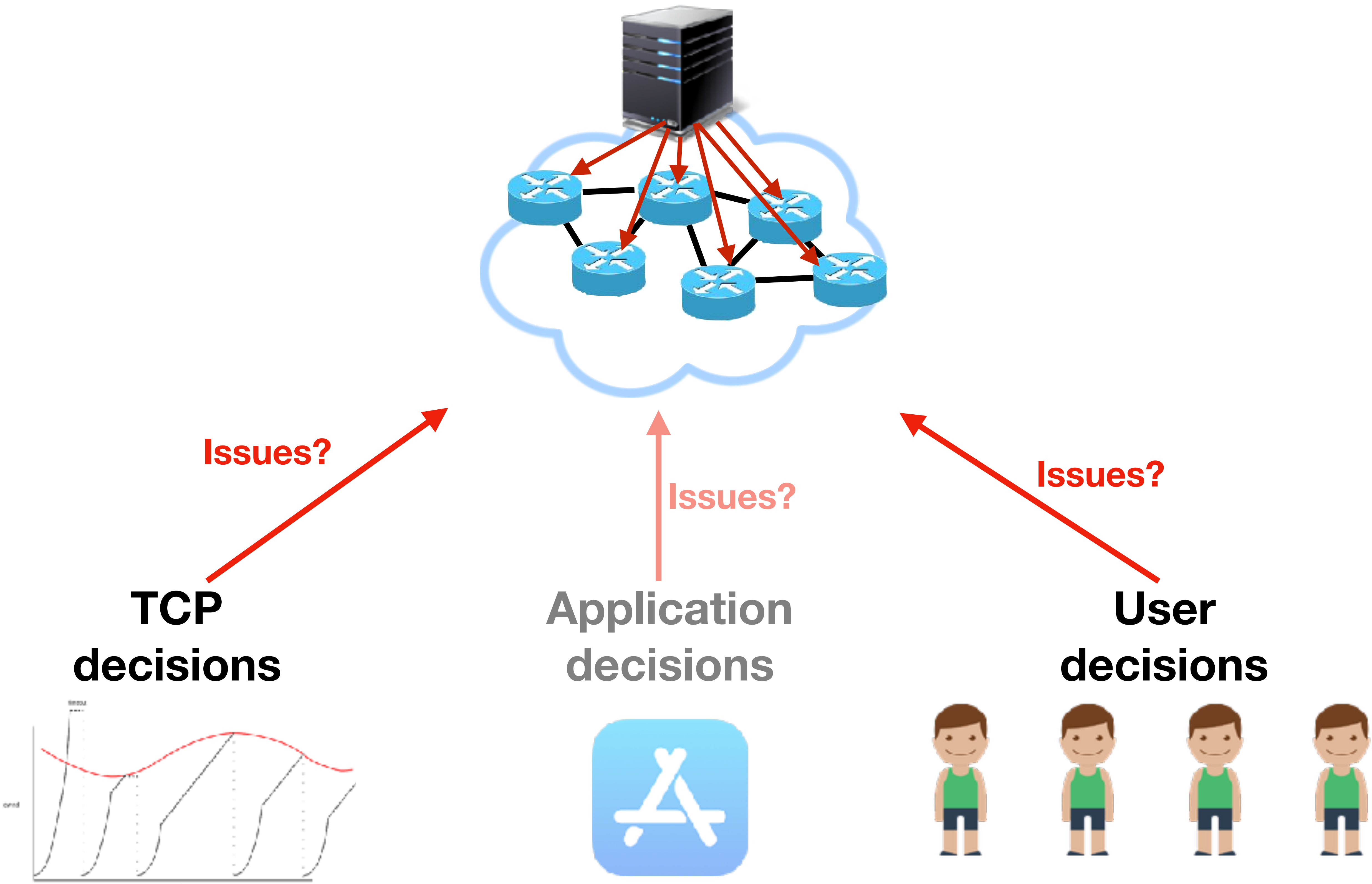
Good at performance optimization

Slow failure response

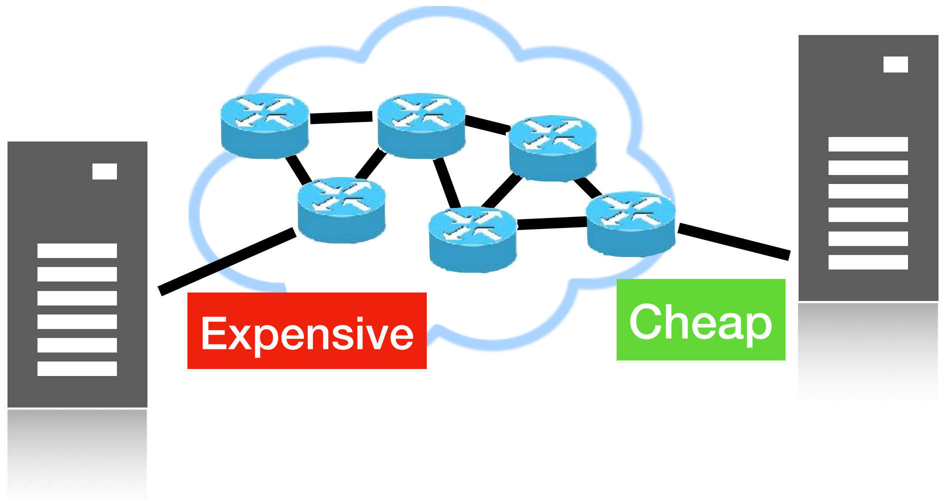
App TE + ISP TE



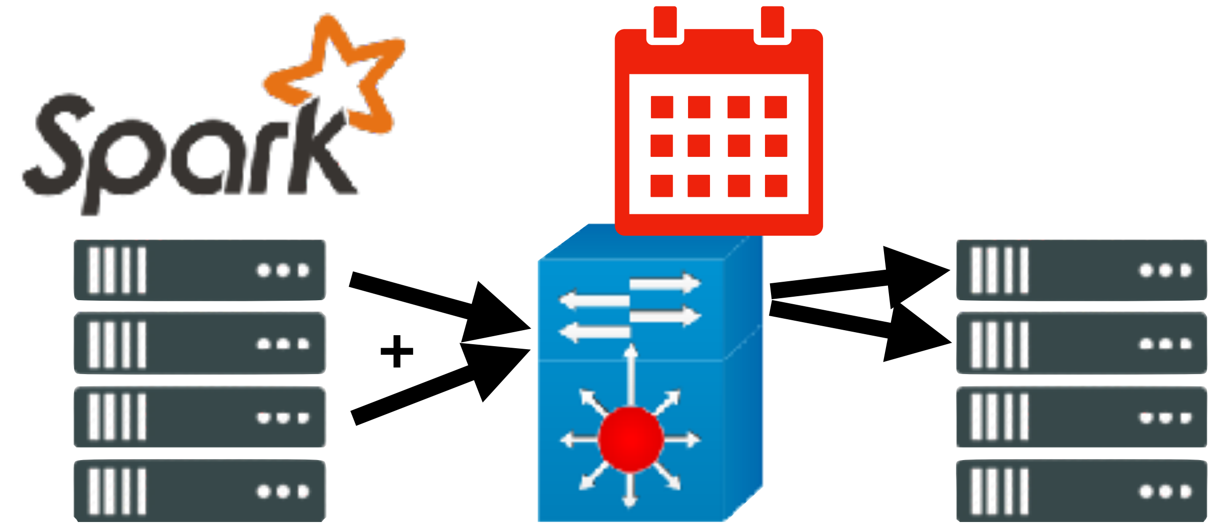
Bad CDN server selection
→ ISP paying for costly routes


Issues?
Issues?
Issues?
TCP decisions
Application decisions
User decisions

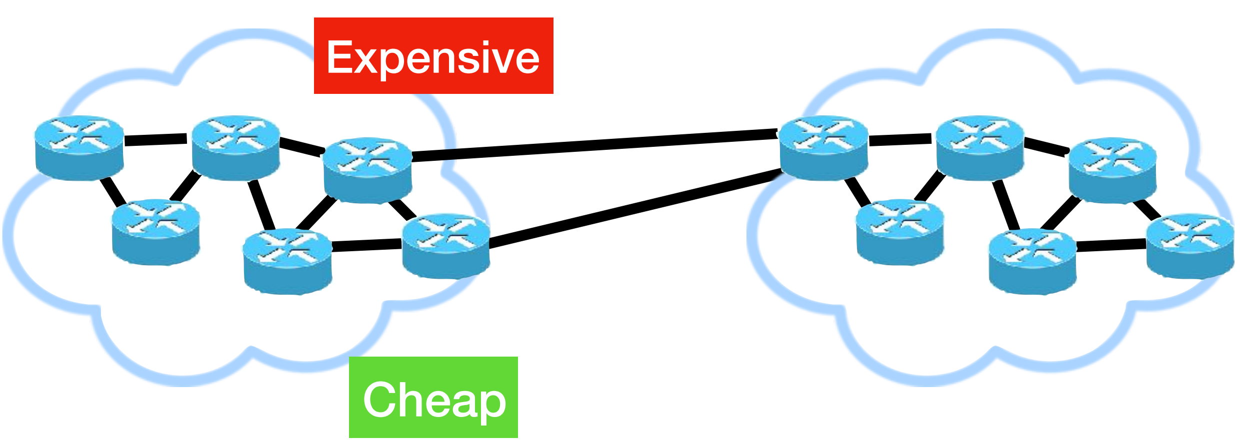
Categorizing Control Coordination



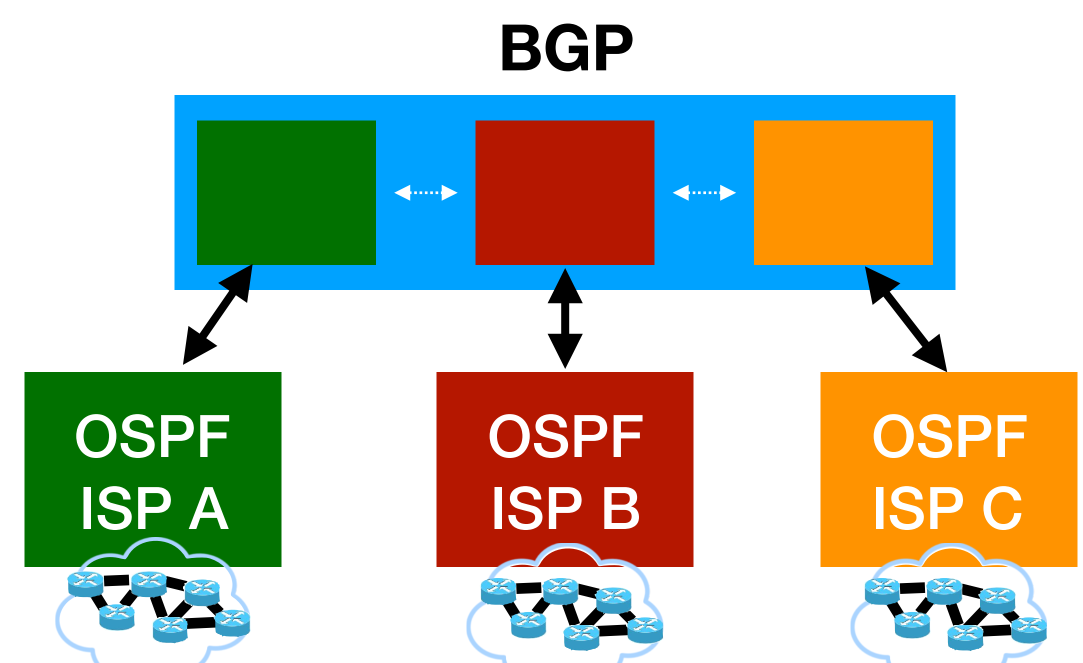
App TE + ISP TE



Coflow (App + DC scheduling)

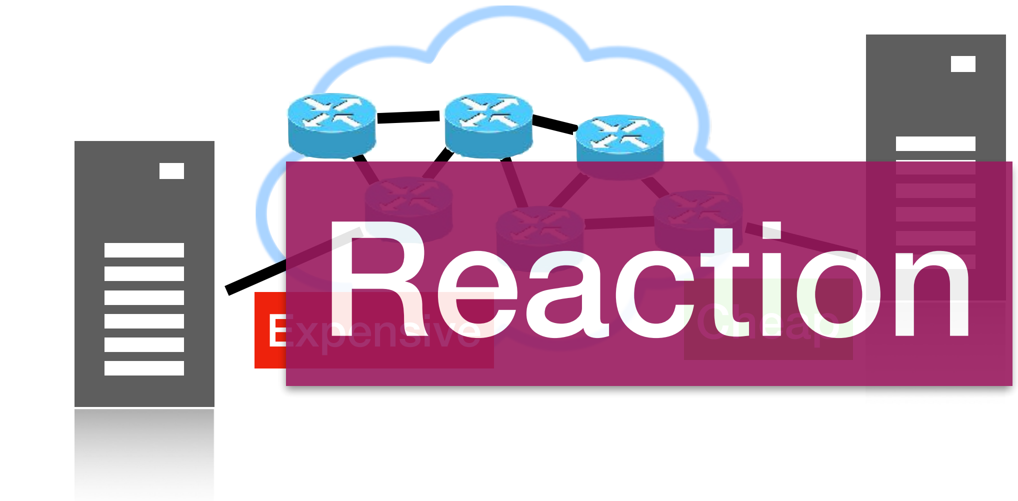


BGP + BGP

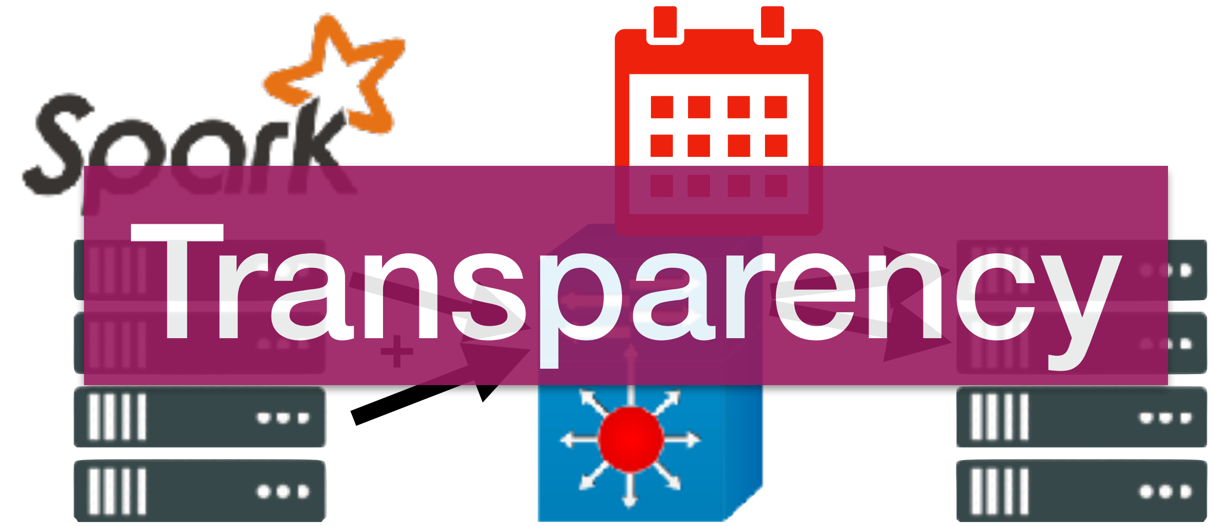


Internet-scale Routing (BGP + OSPF)

Categorizing Control Coordination

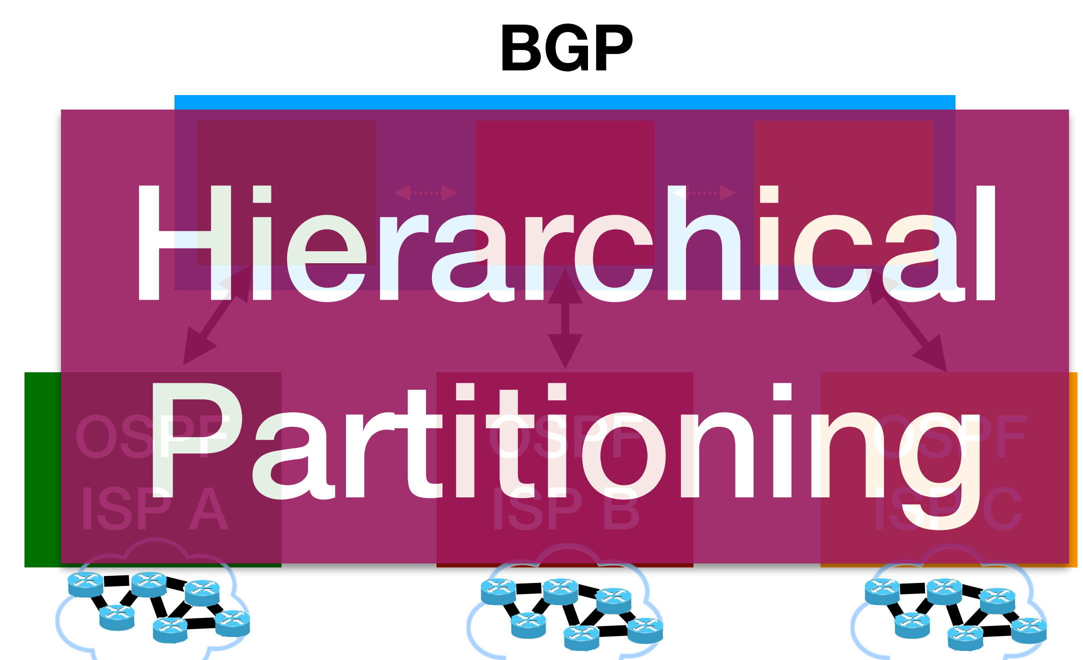


App TE + ISP TE



Coflow (App + DC scheduling)

BGP + BGP



Internet-scale Routing (BGP + OSPF)

Control Coordination

Scenario: **Layering**	Scenario: **Admin**	Scenario: **Scalability**
Etalon	VDX	VDN

App TE + ISP TE **Reaction**	Coflow **Transparency**
BGP + BGP **Priority Ranking**	Internet-scale Routing **Hierarchical Partitioning**

Control Coordination

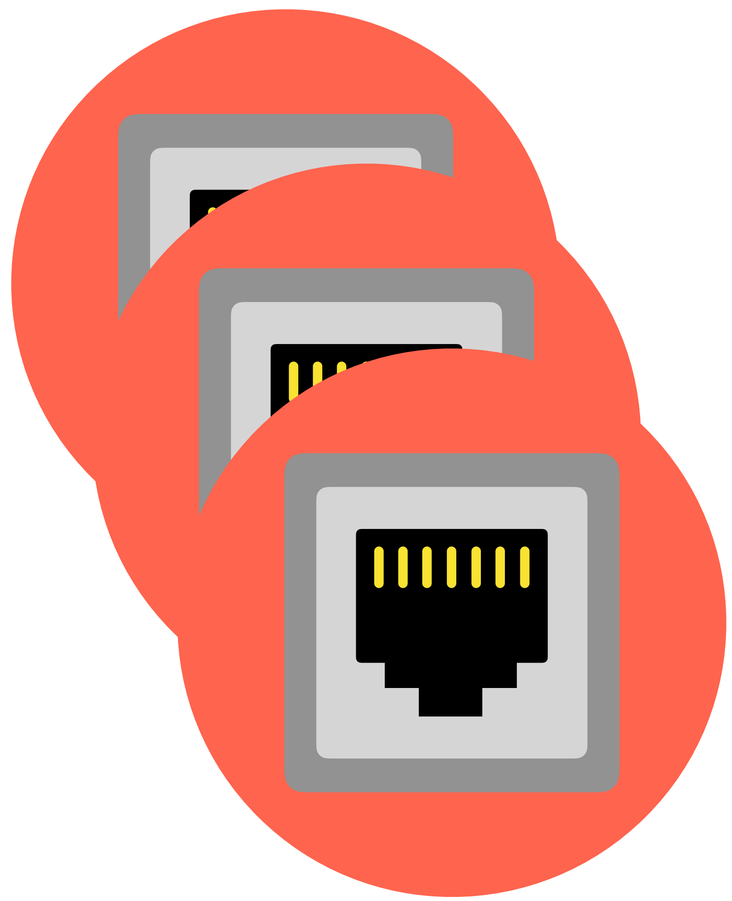
Scenario: Layering | **Scenario: Admin** | **Scenario: Scalability**

Etalon | VDX | VDN

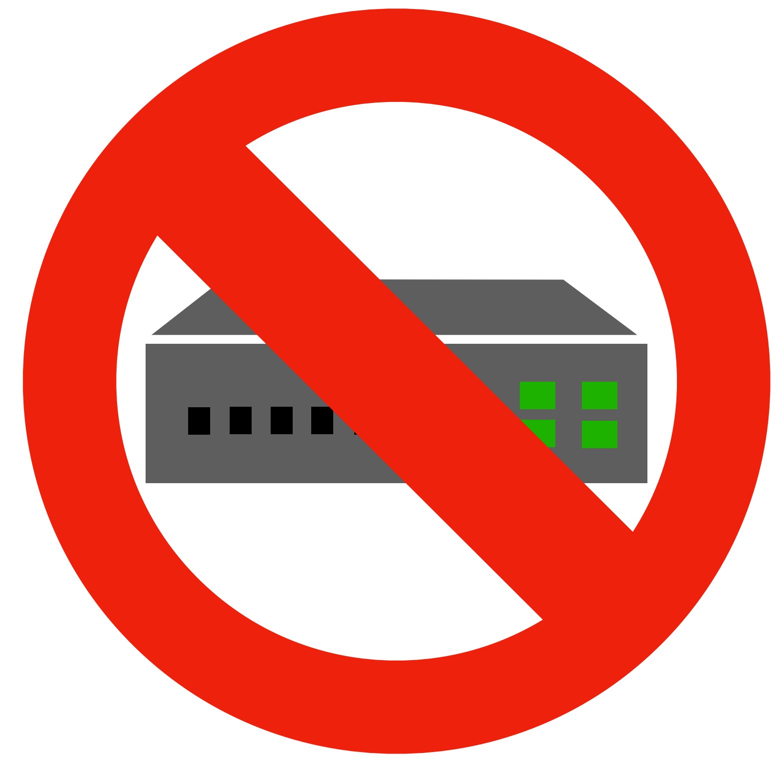
App TE + ISP TE **Reaction**	Coflow **Transparency**
BGP + BGP **Priority Ranking**	Internet-scale Routing **Hierarchical Partitioning**

Difficult to scale datacenters with demand

+



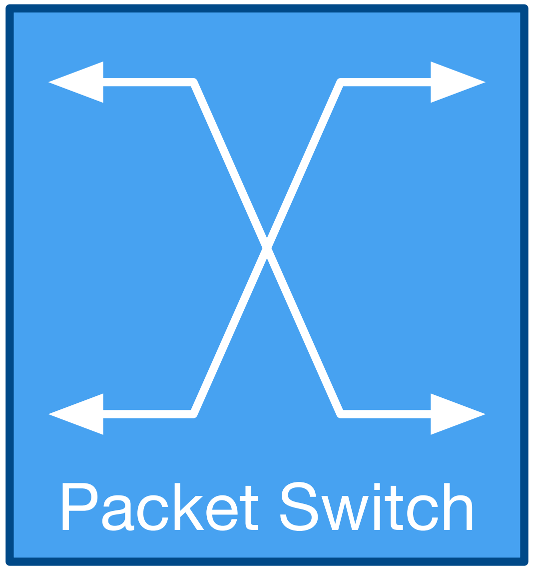
≠



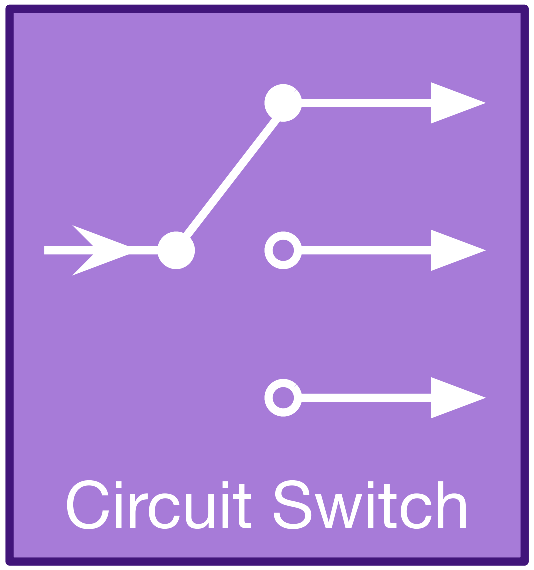
Higher Bandwidth **Higher Port Count** **CMOS limits…**

P-FatTree: A multi-channel datacenter network topology. HotNets 2016.

Use circuits to build bigger + faster networks!



+



Reconfigurable Datacenter Networks (RDCNs)

Circuit Switch Design

How do you physical build it?

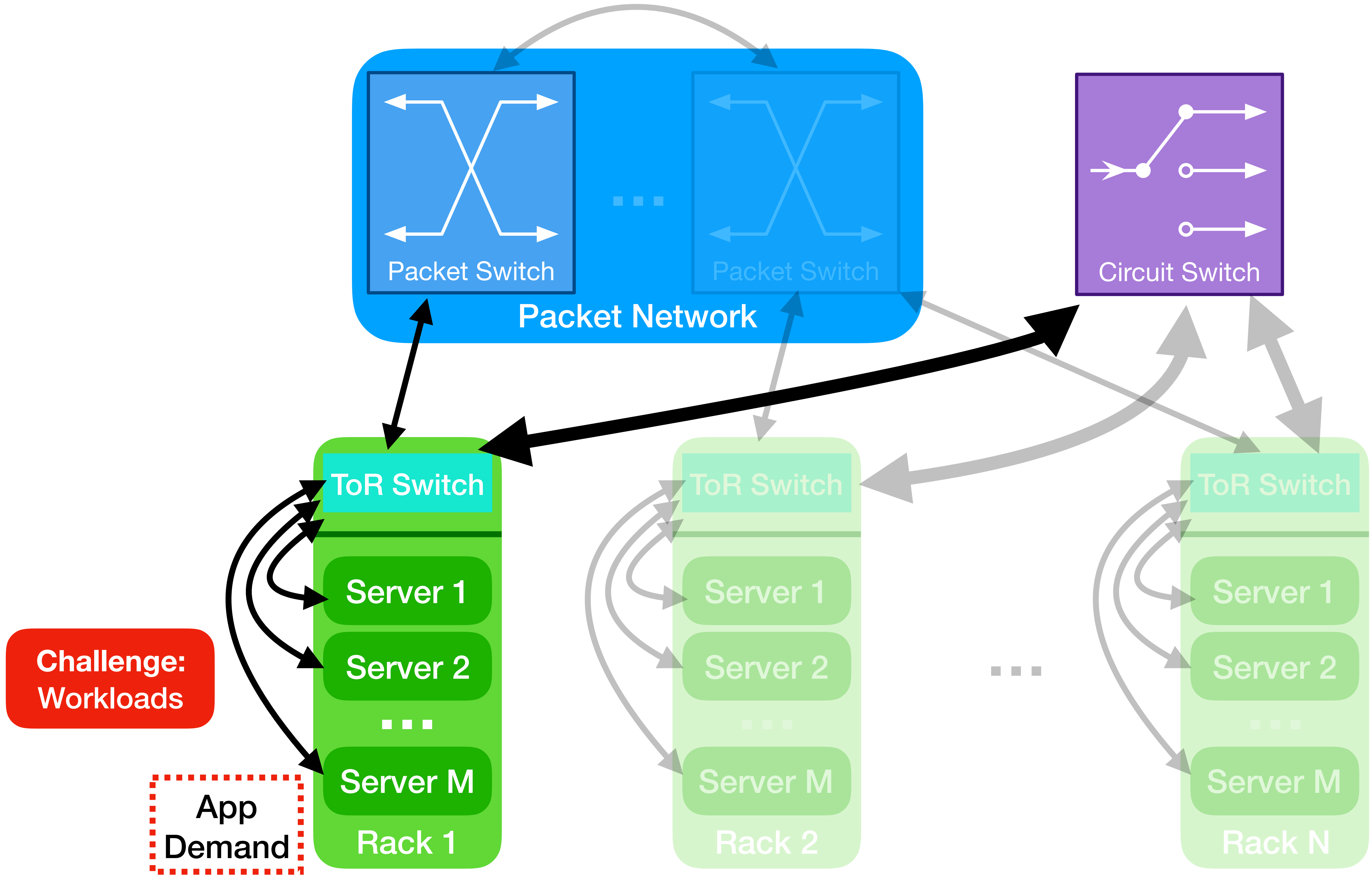
Network Scheduling

How do you make use of it?

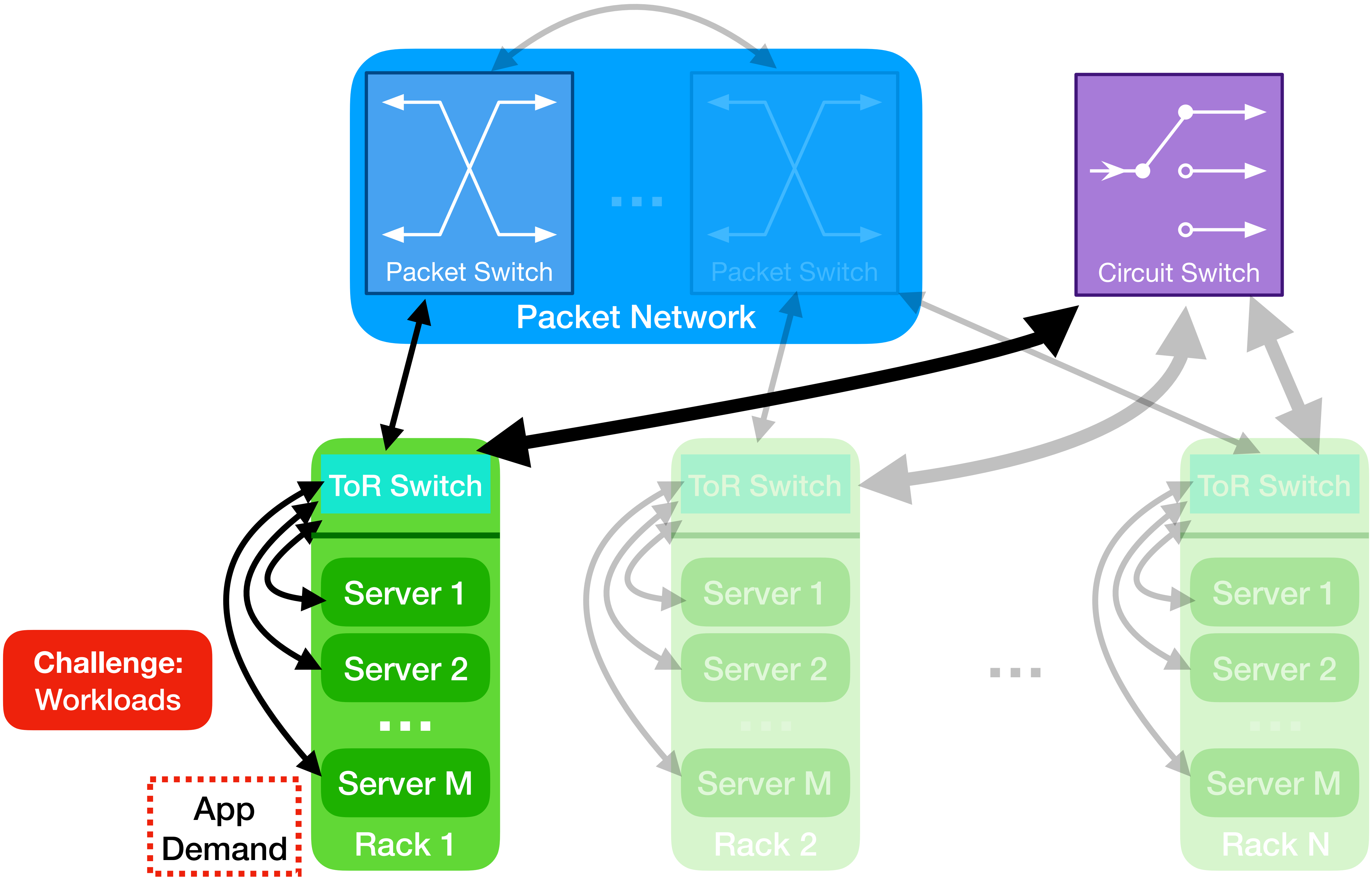
End-to-End Challenges

What existing things break?

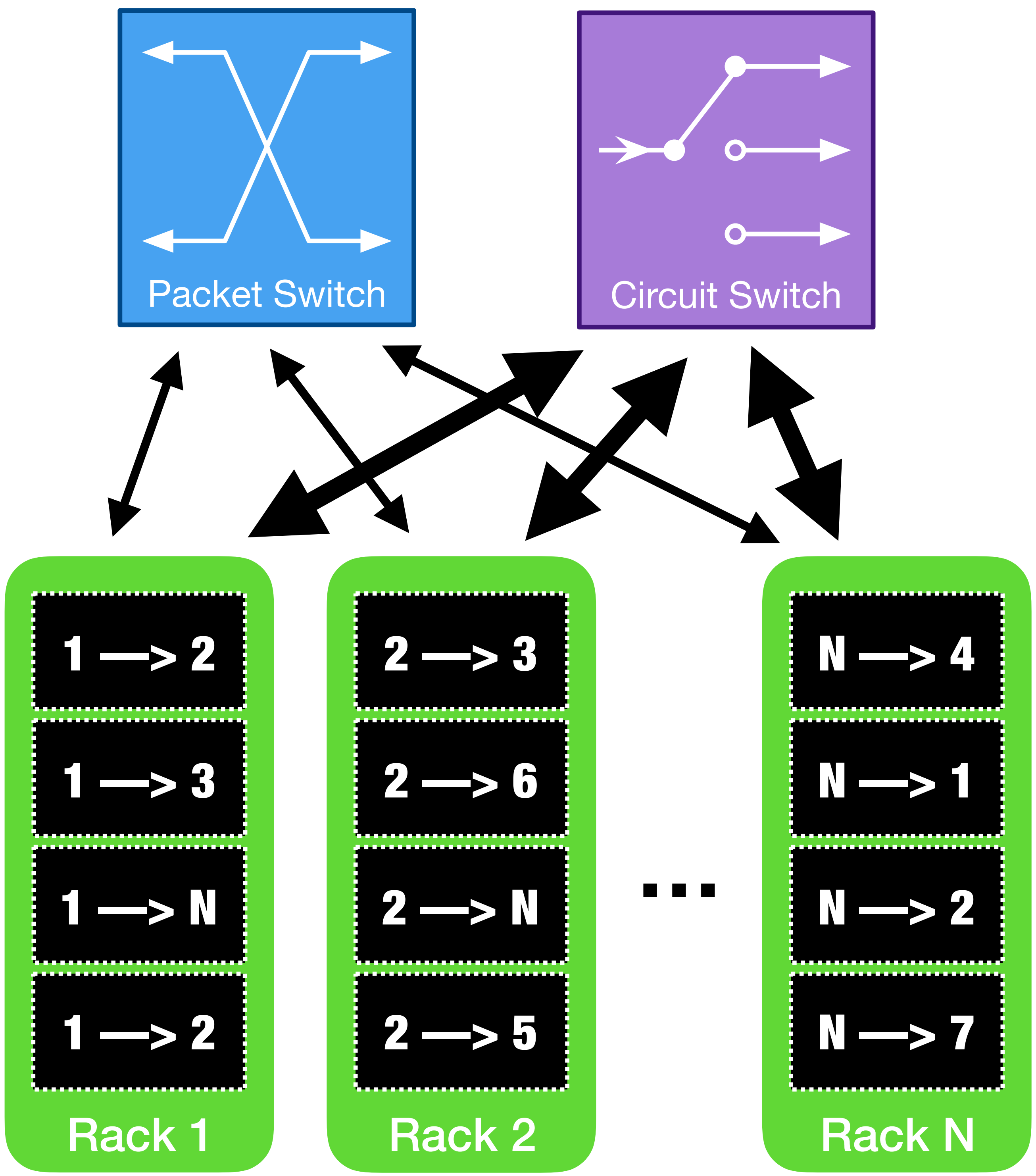
RDCN switch design



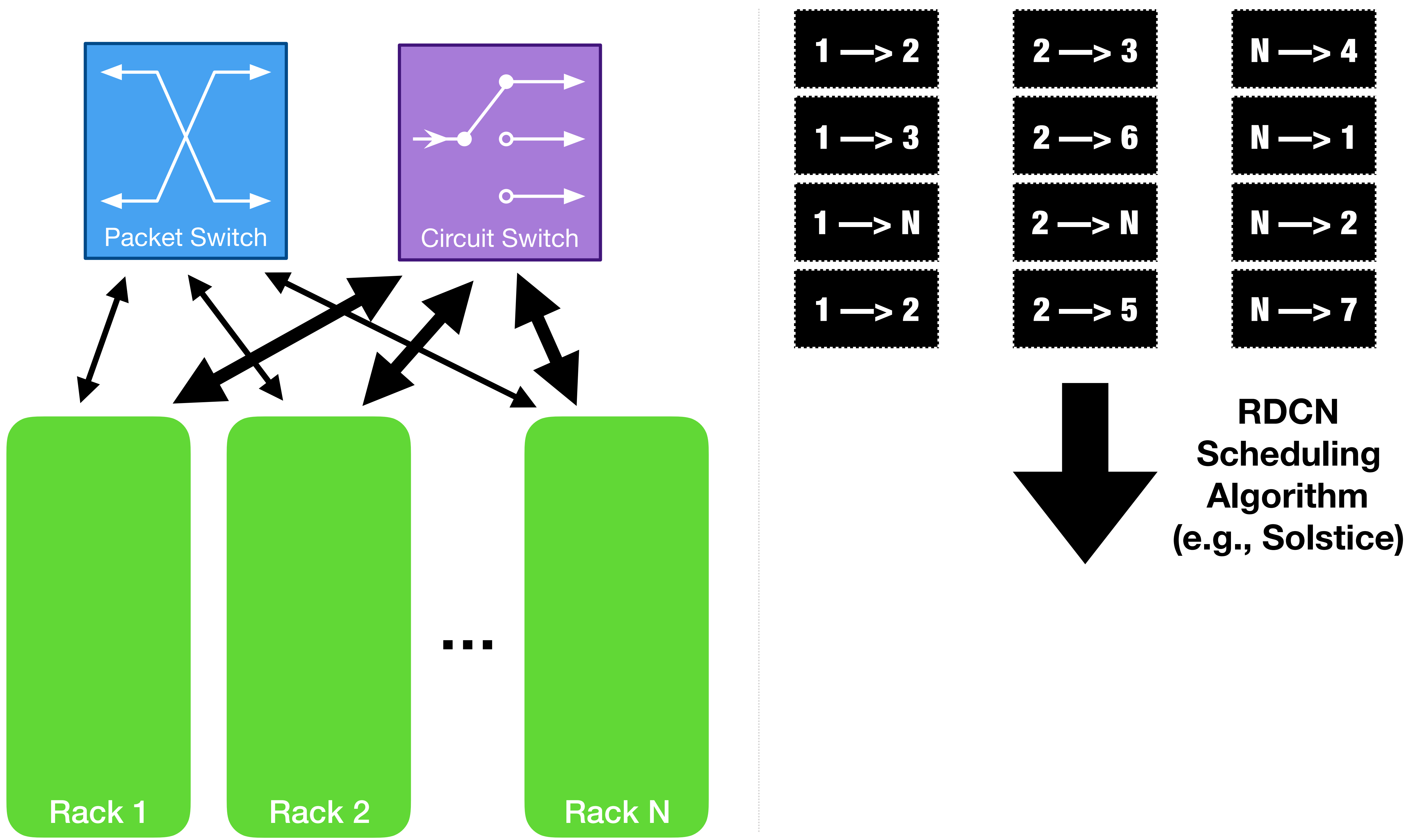
RDCN switch design



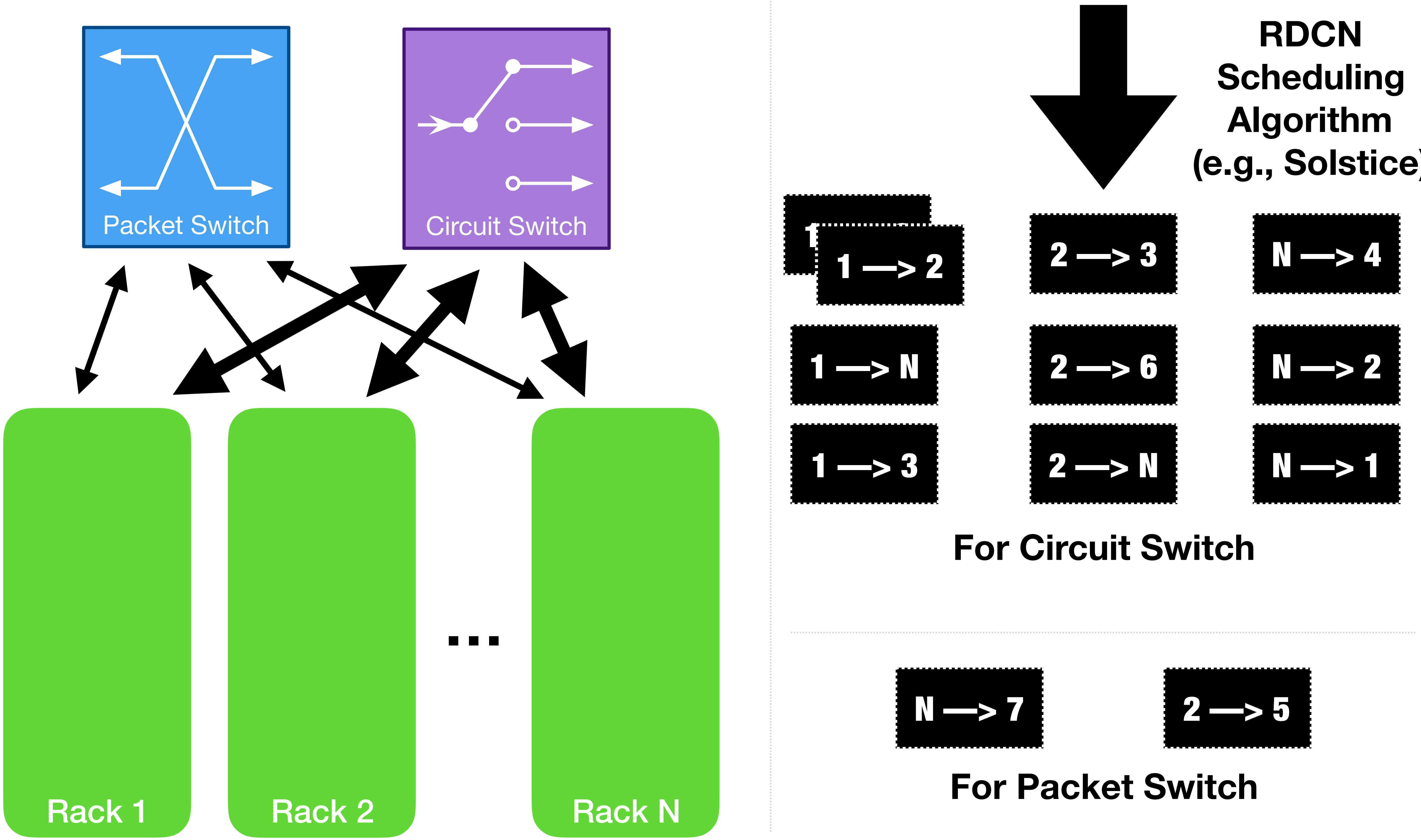
RDCN scheduling



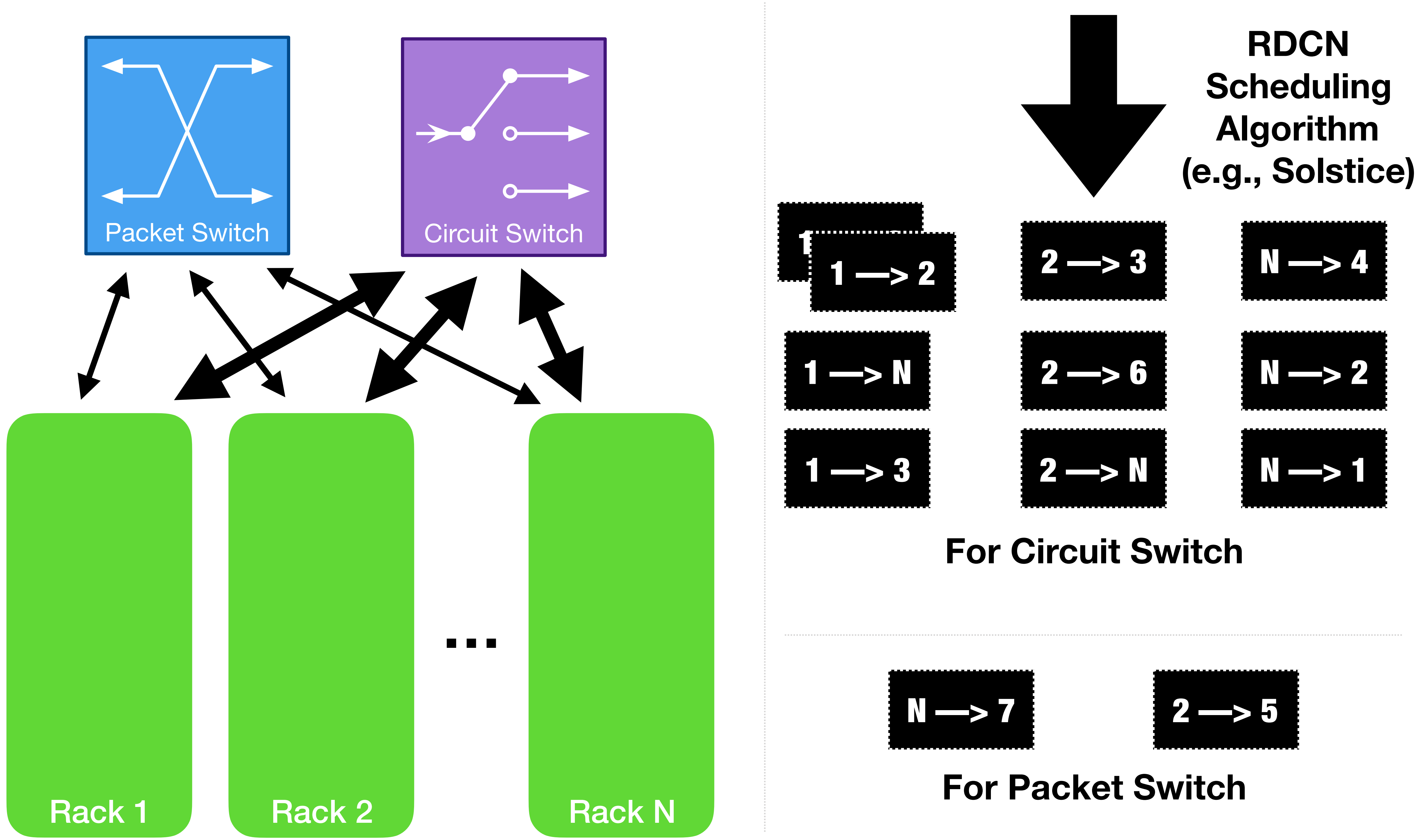
RDCN scheduling



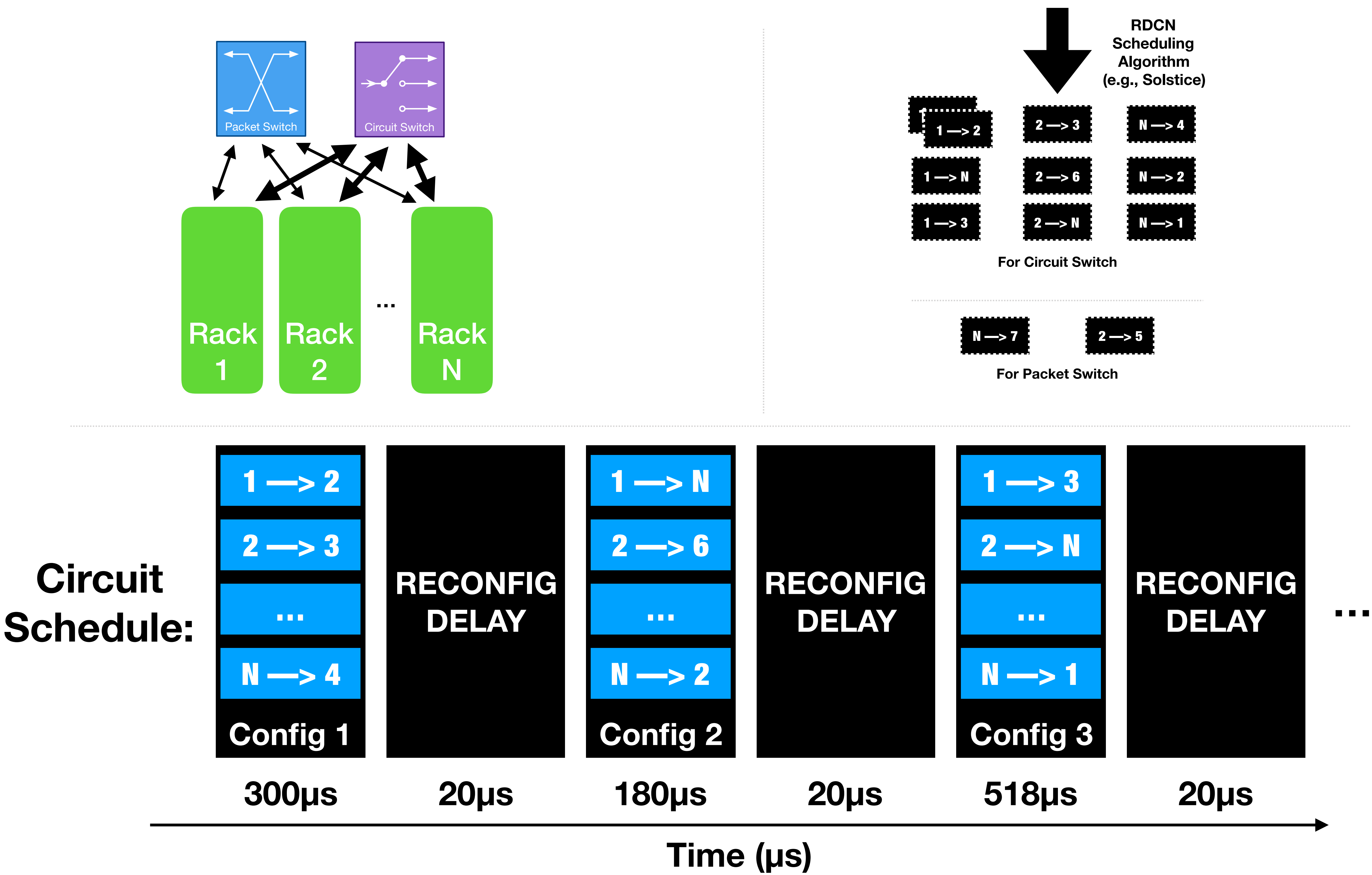
RDCN scheduling



RDCN scheduling



RDCN scheduling



Contributions

End-to-End Challenges

Challenge: BW Fluct.

Solution: Dynamic Buffer Resizing

Challenge: Demand Estimation

Solution: Endhost-based Estimation

Challenge: Workloads

Solution: App-specific Modification

***Etalon*, an RDCN Emulator**

Overview

End-to-End Challenges

Challenge:
BW Fluct.

Solution:
Dynamic Buffer Resizing

Challenge:
Demand Estimation

Solution:
Endhost-based Estimation

Challenge:
Workloads

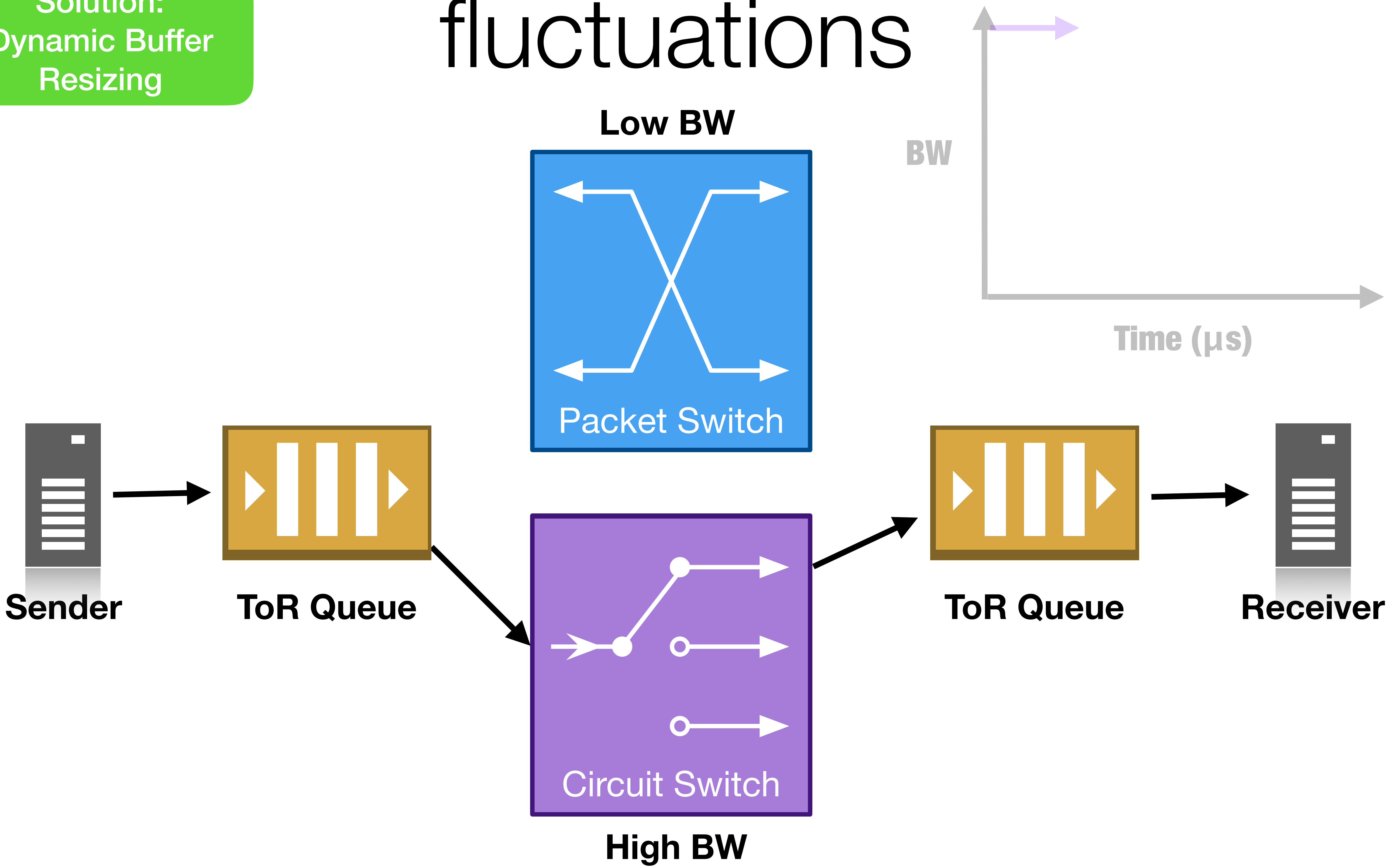
Solution:
App-specific Modification

Etalon, an RDCN Emulator

Challenge: BW Fluct.

Solution: Dynamic Buffer Resizing

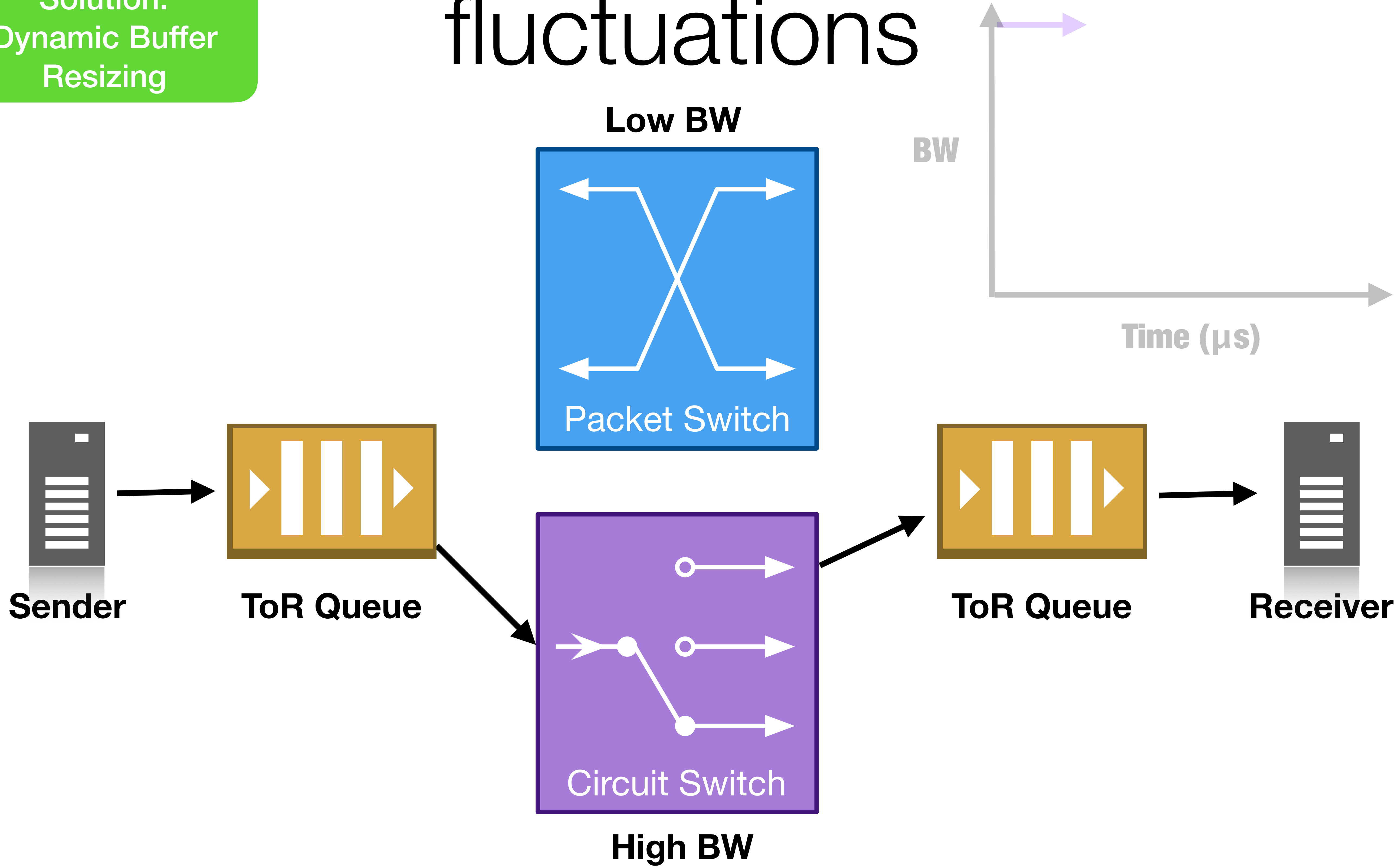
TCP and rapid bw fluctuations



Challenge: BW Fluct.

Solution: Dynamic Buffer Resizing

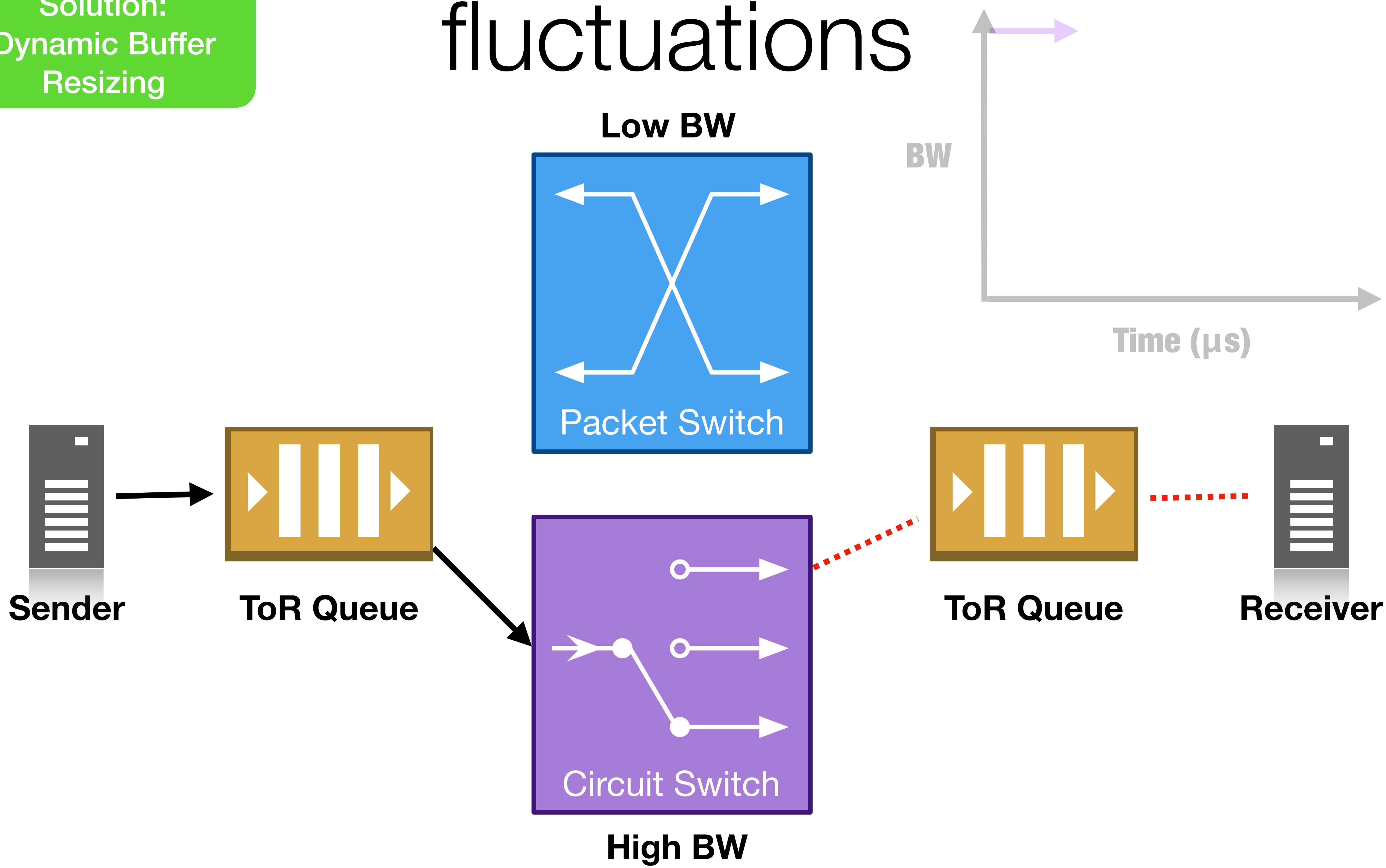
TCP and rapid bw fluctuations



Challenge:
BW Fluct.

Solution:
Dynamic Buffer
Resizing

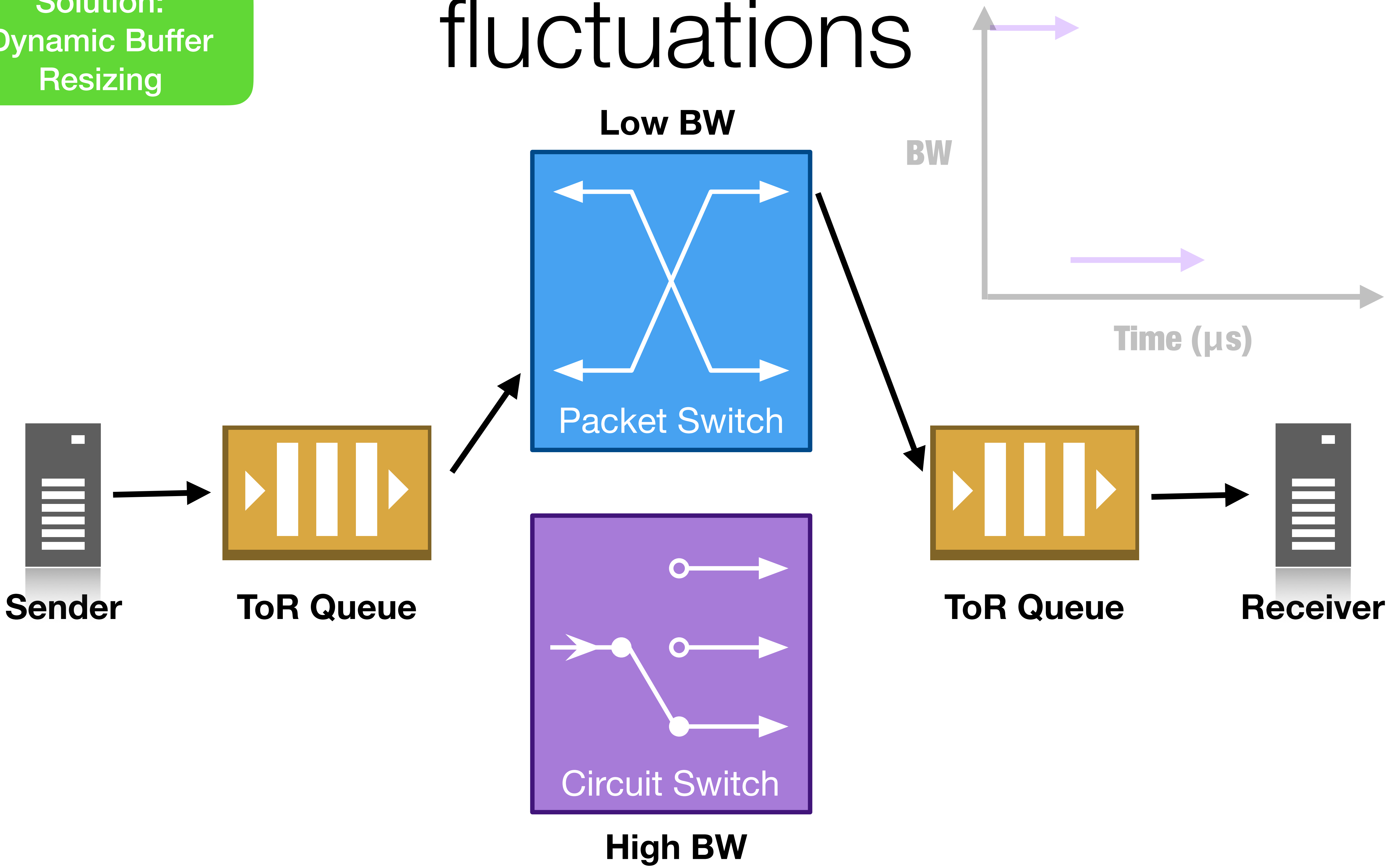
TCP and rapid bw fluctuations



Challenge: BW Fluct.

Solution: Dynamic Buffer Resizing

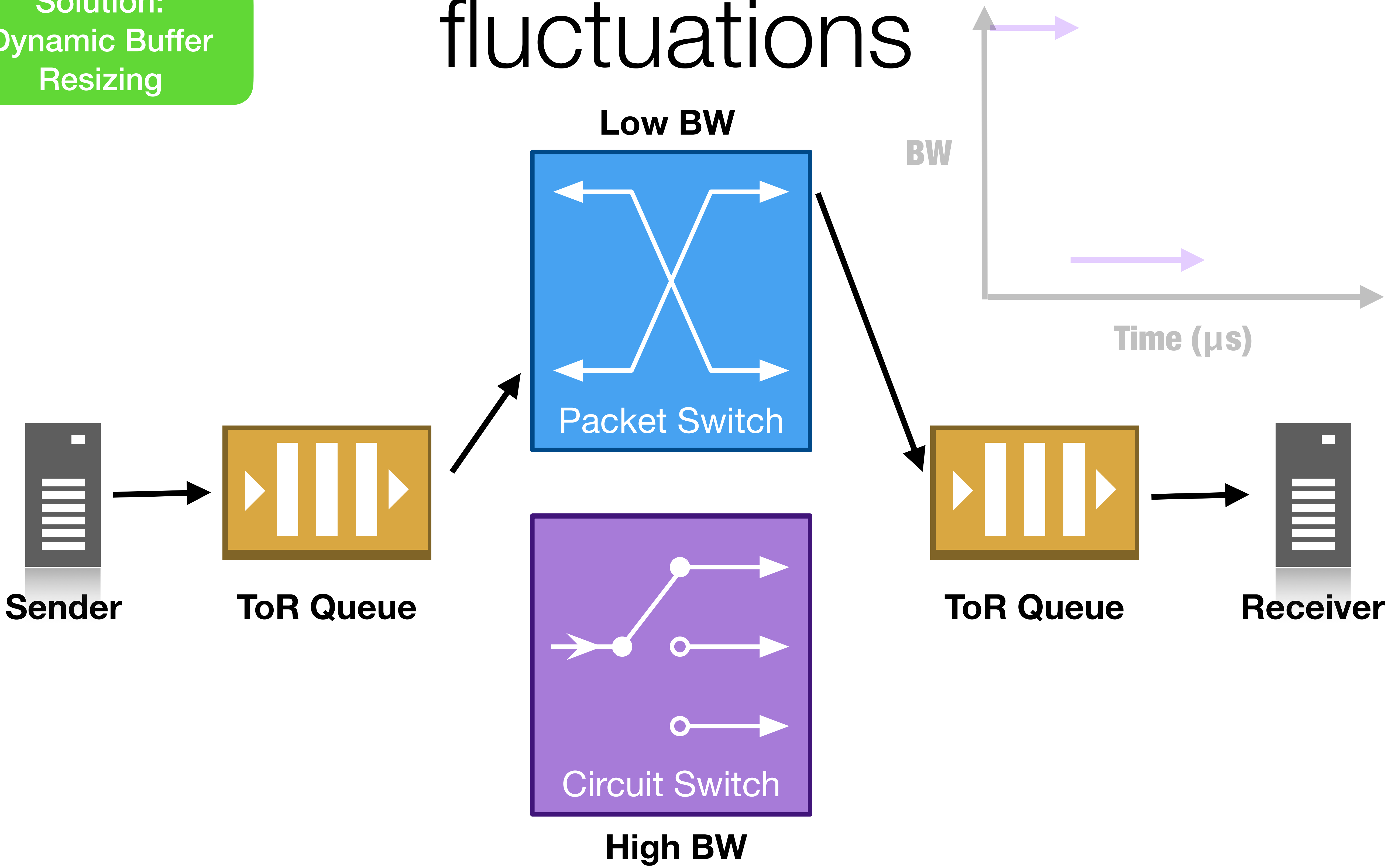
TCP and rapid bw fluctuations



Challenge:
BW Fluct.

Solution:
Dynamic Buffer
Resizing

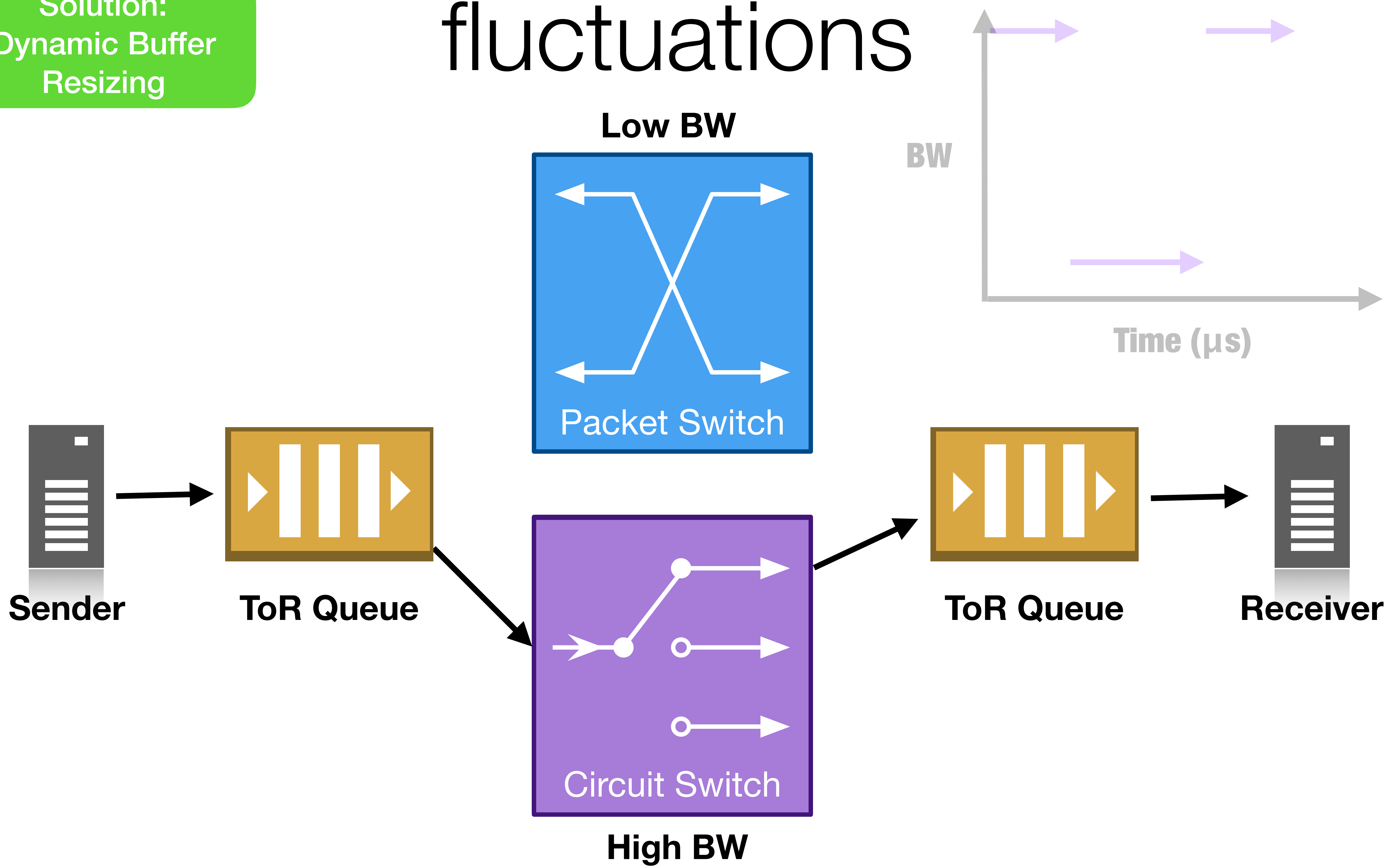
TCP and rapid bw fluctuations



Challenge: BW Fluct.

Solution: Dynamic Buffer Resizing

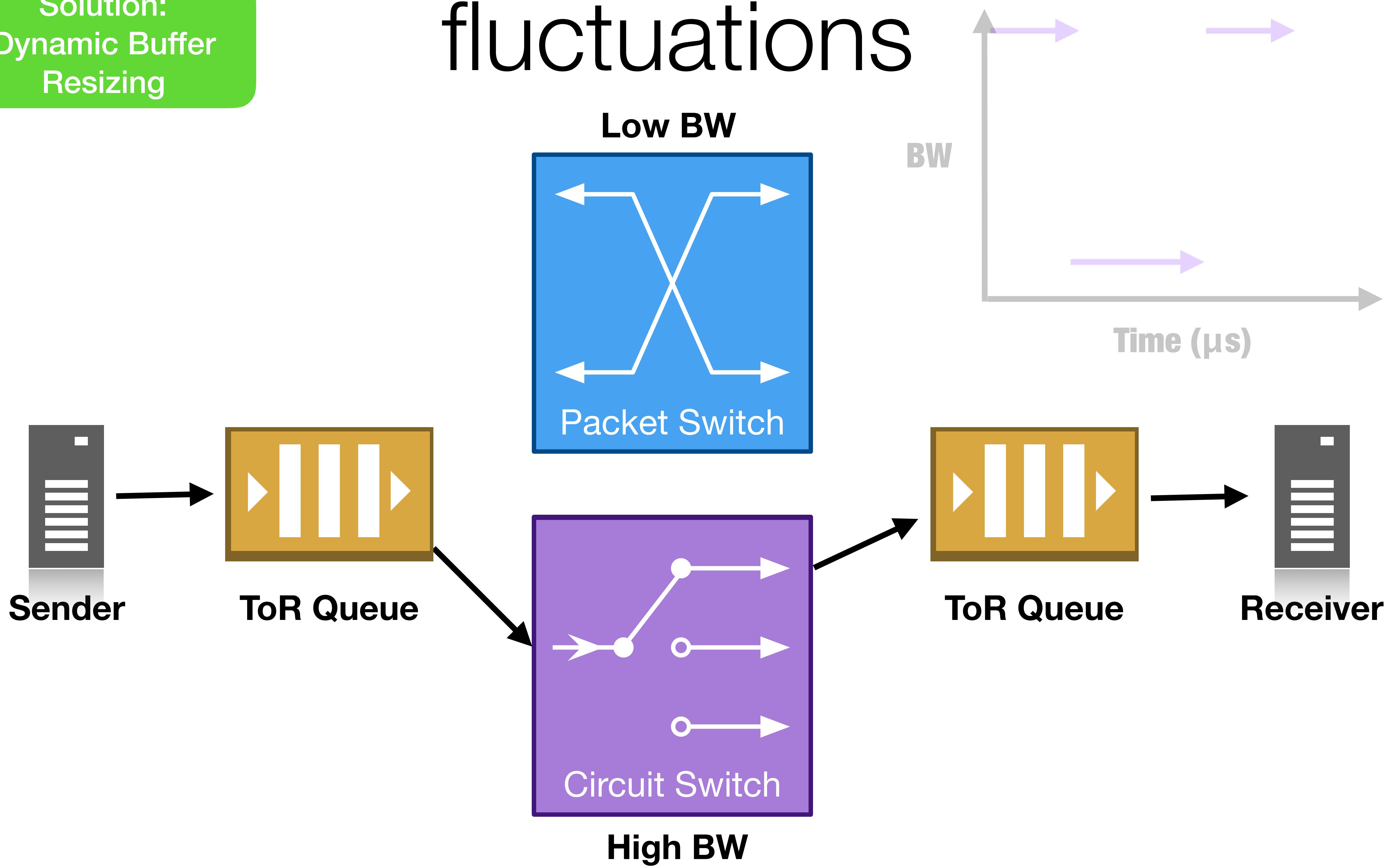
TCP and rapid bw fluctuations



TCP and rapid bw fluctuations

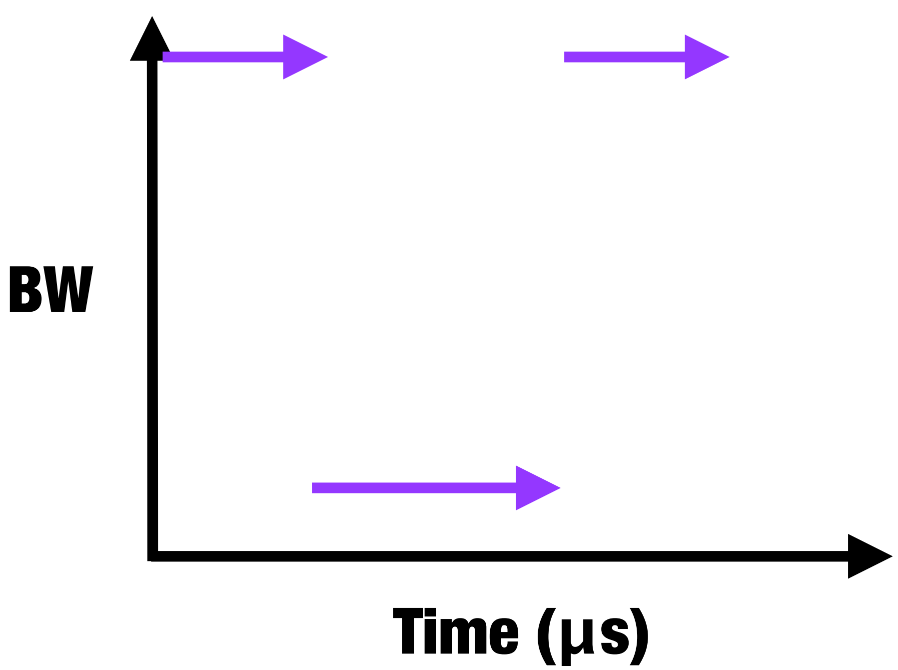
Challenge: BW Fluct.

Solution: Dynamic Buffer Resizing

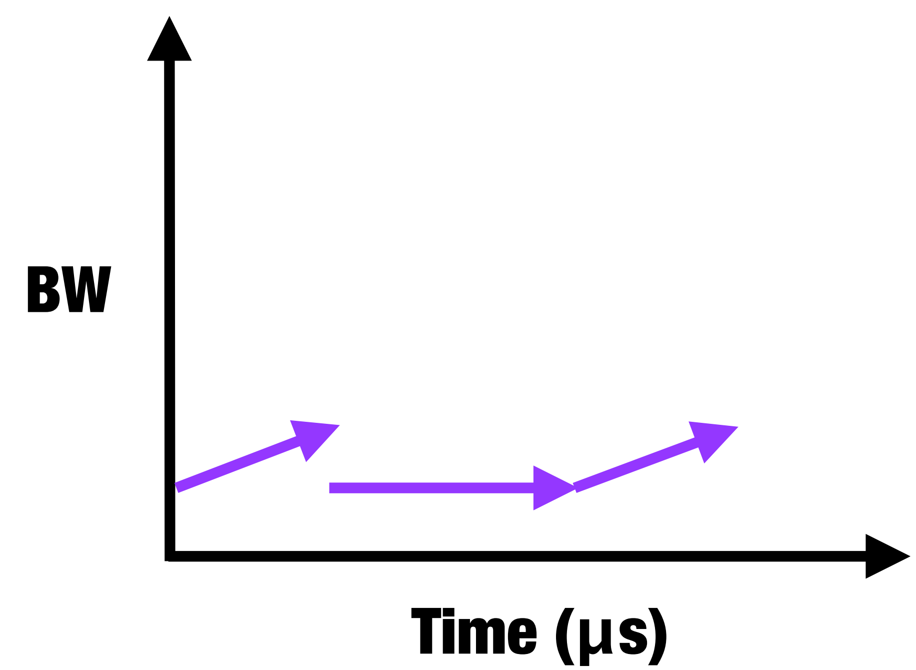


TCP and rapid bw fluctuations

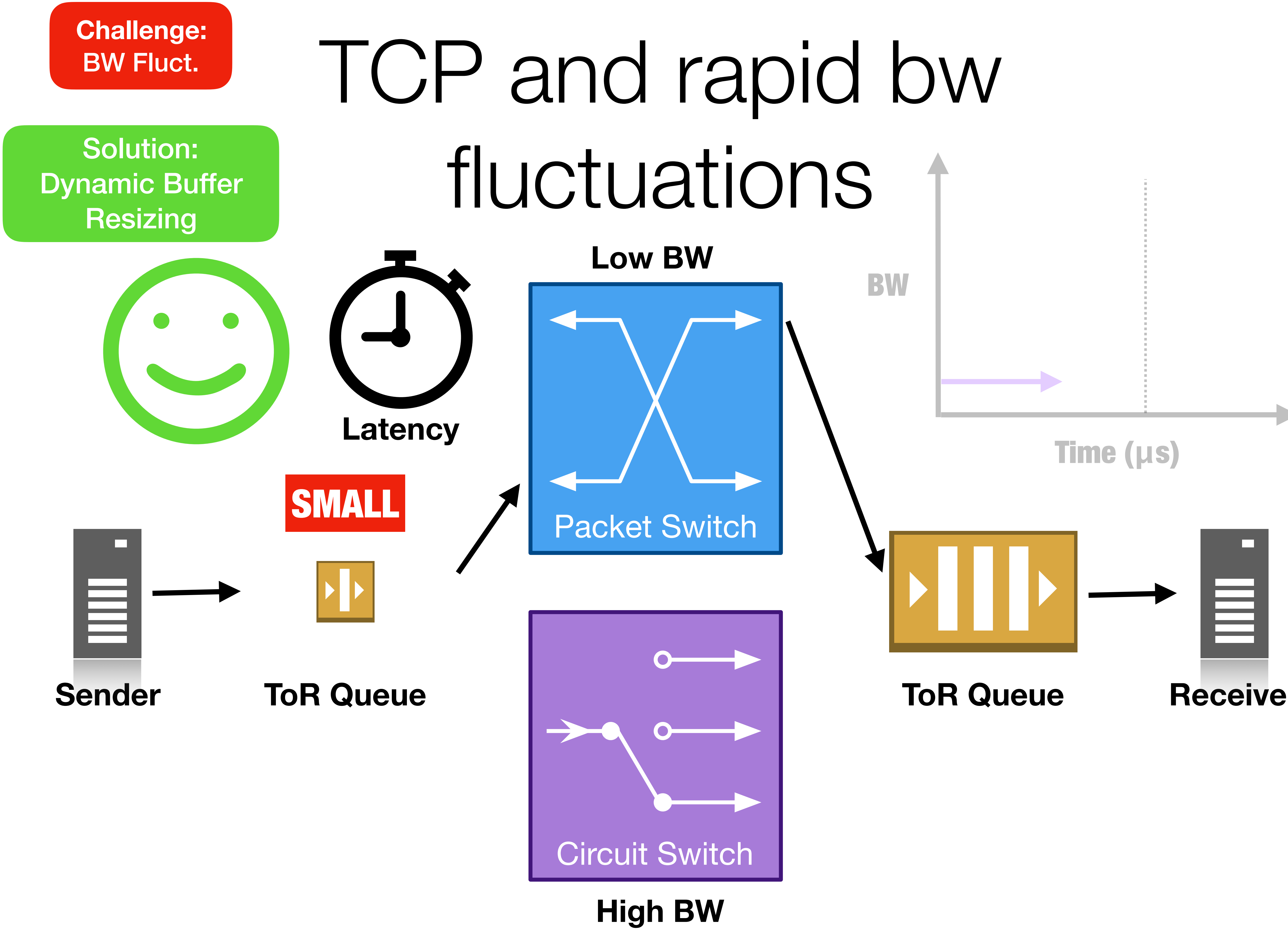
What we want



What we get



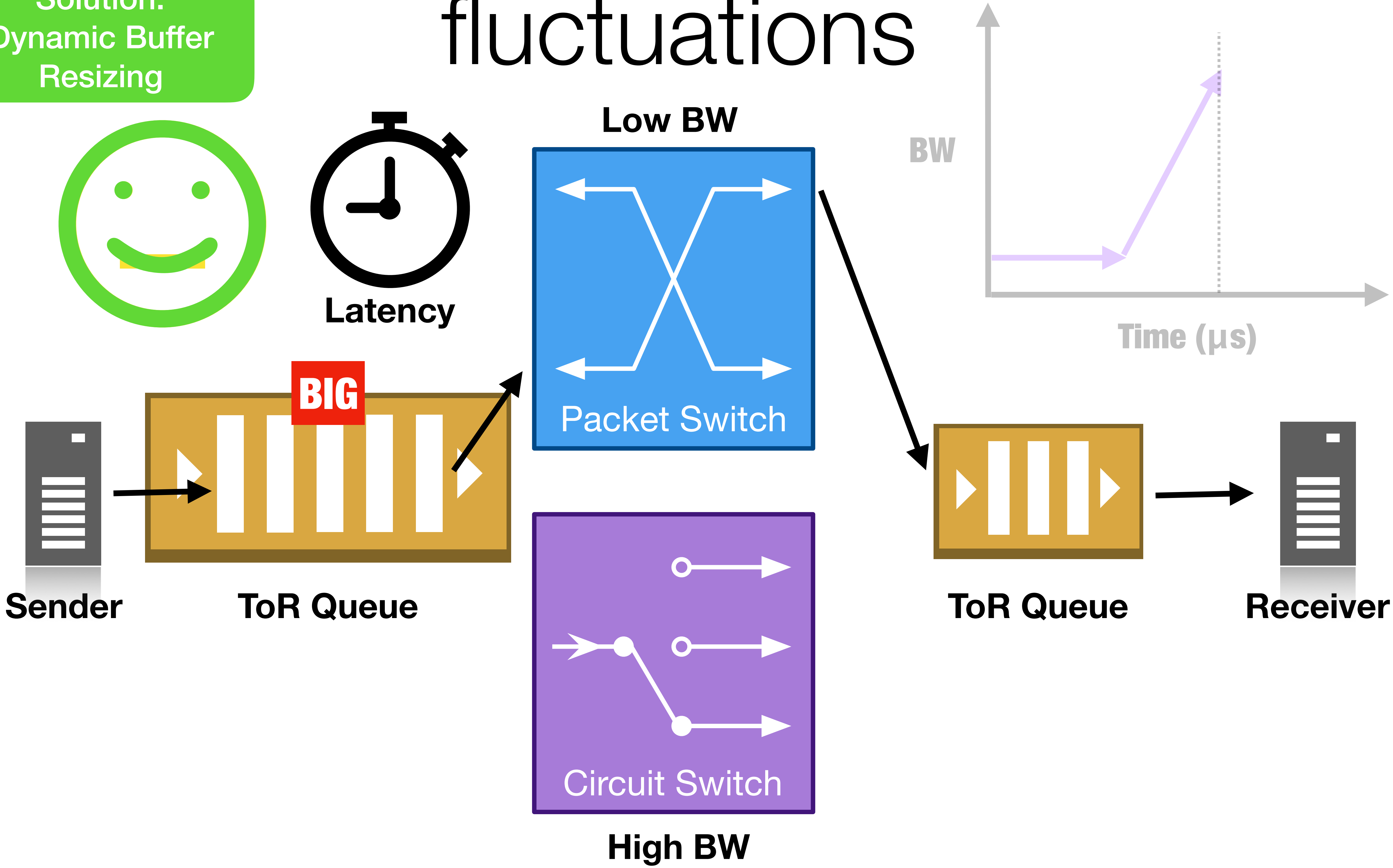
TCP and rapid bw fluctuations



Challenge:
BW Fluct.

Solution:
Dynamic Buffer
Resizing

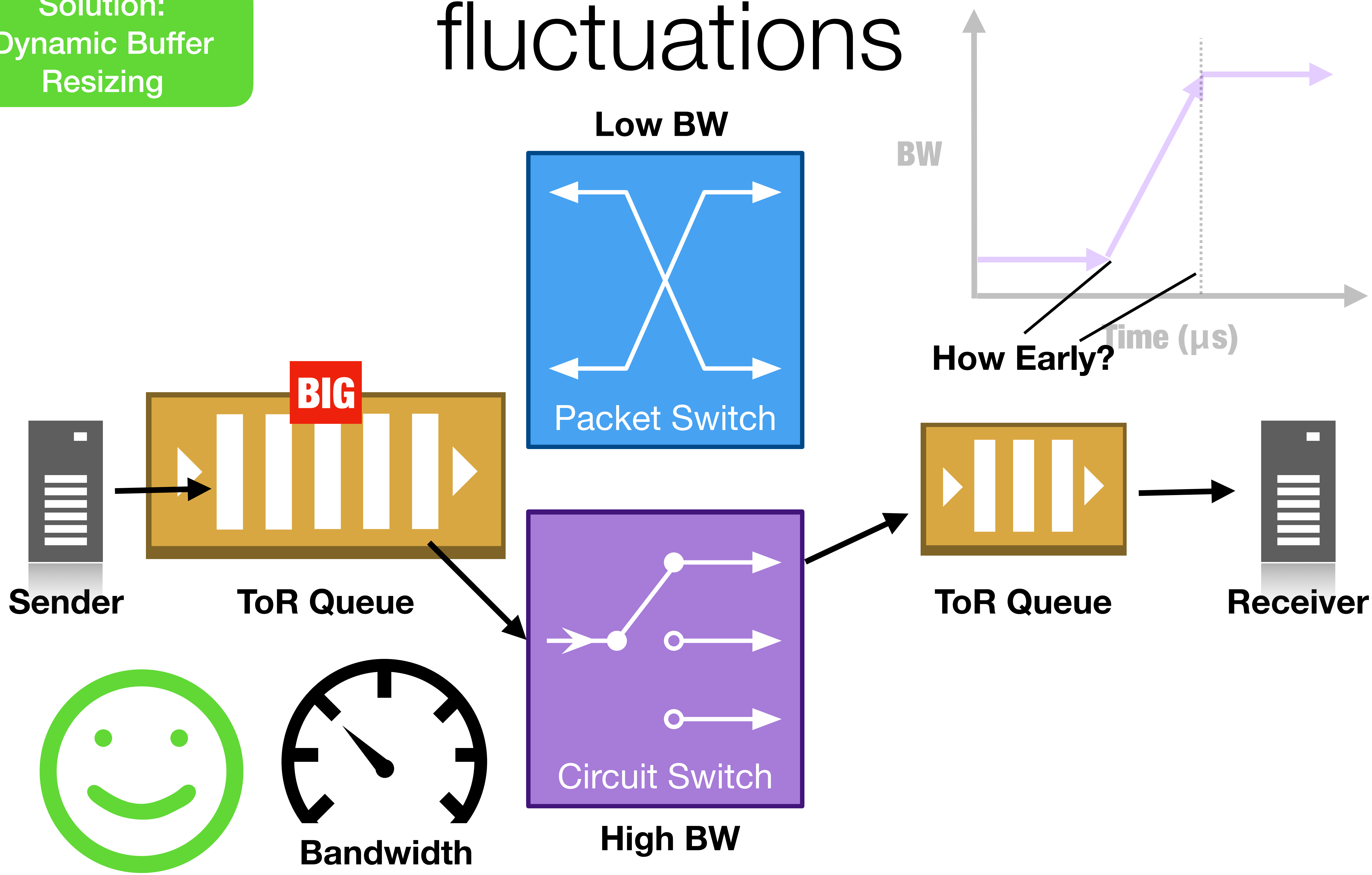
TCP and rapid bw fluctuations



Challenge: BW Fluct.

Solution: Dynamic Buffer Resizing

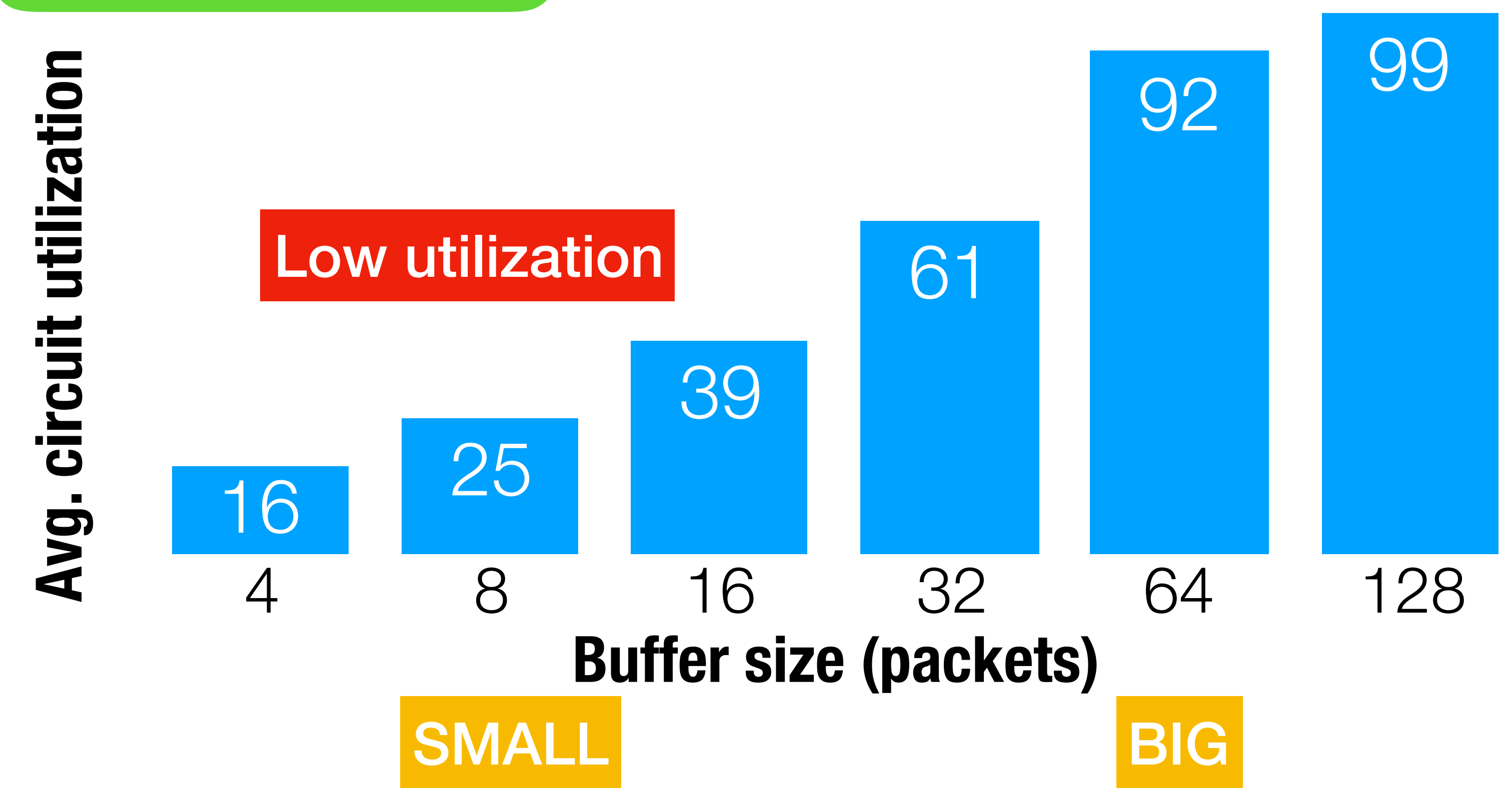
TCP and rapid bw fluctuations



Challenge:
BW Fluct.

Solution:
Dynamic Buffer
Resizing

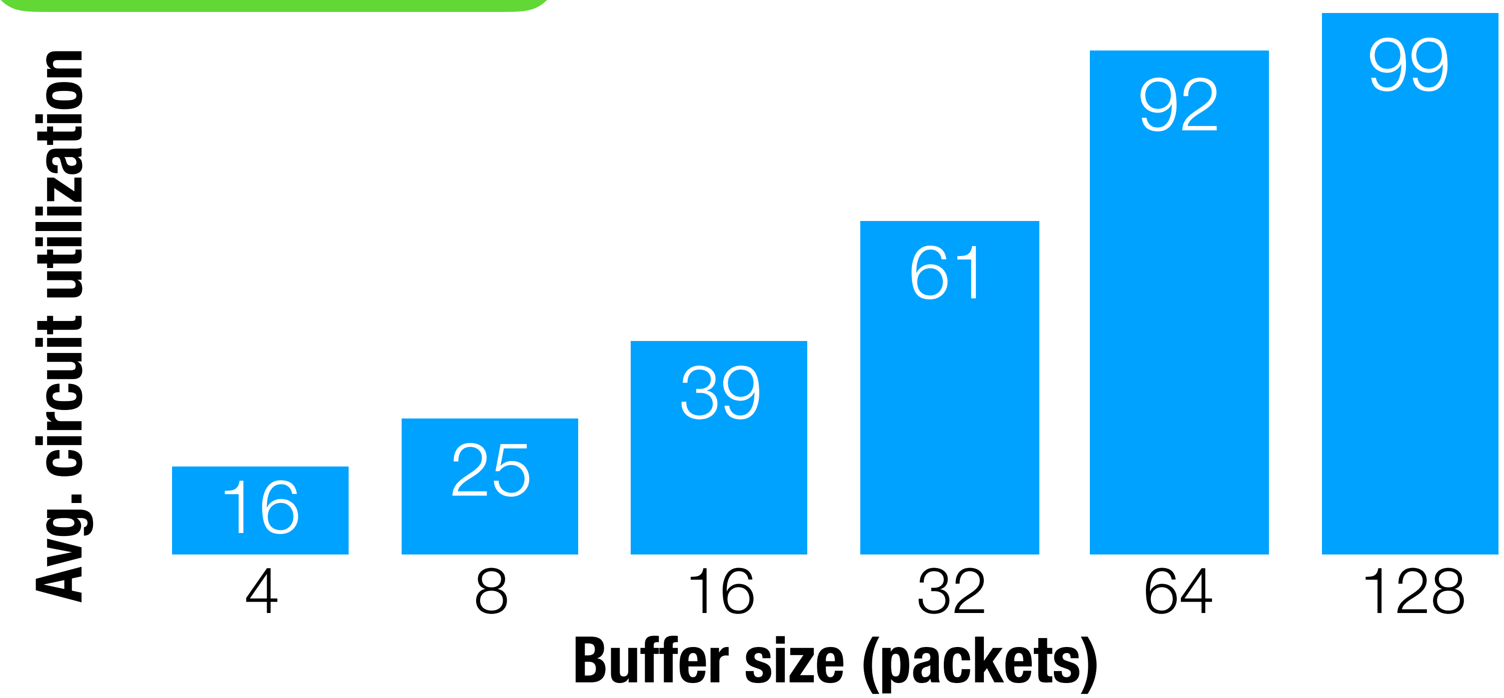
Static buffers provide good circuit util **or** latency

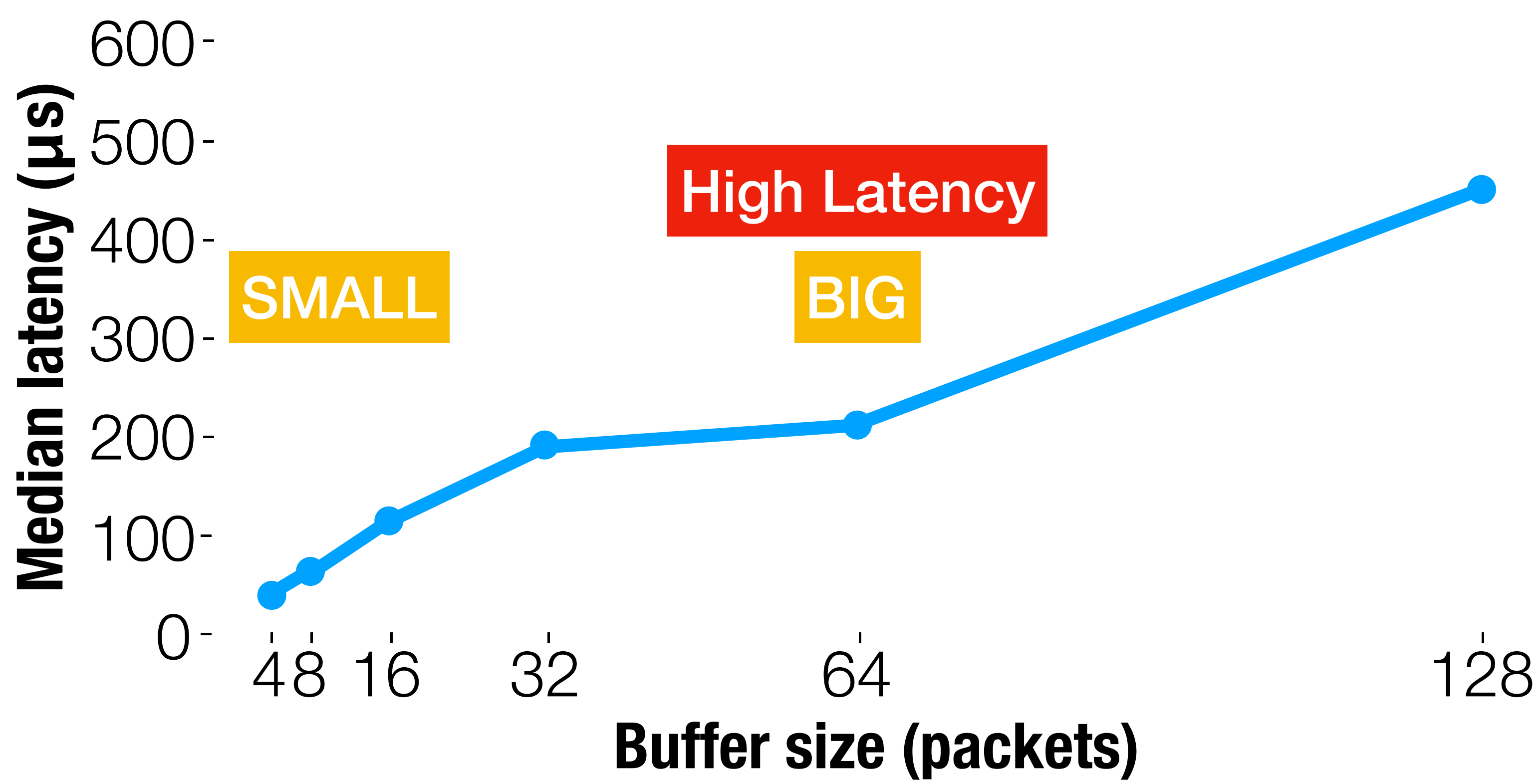


Challenge: BW Fluct.

Solution: Dynamic Buffer Resizing

Static buffers provide good circuit util **or** latency

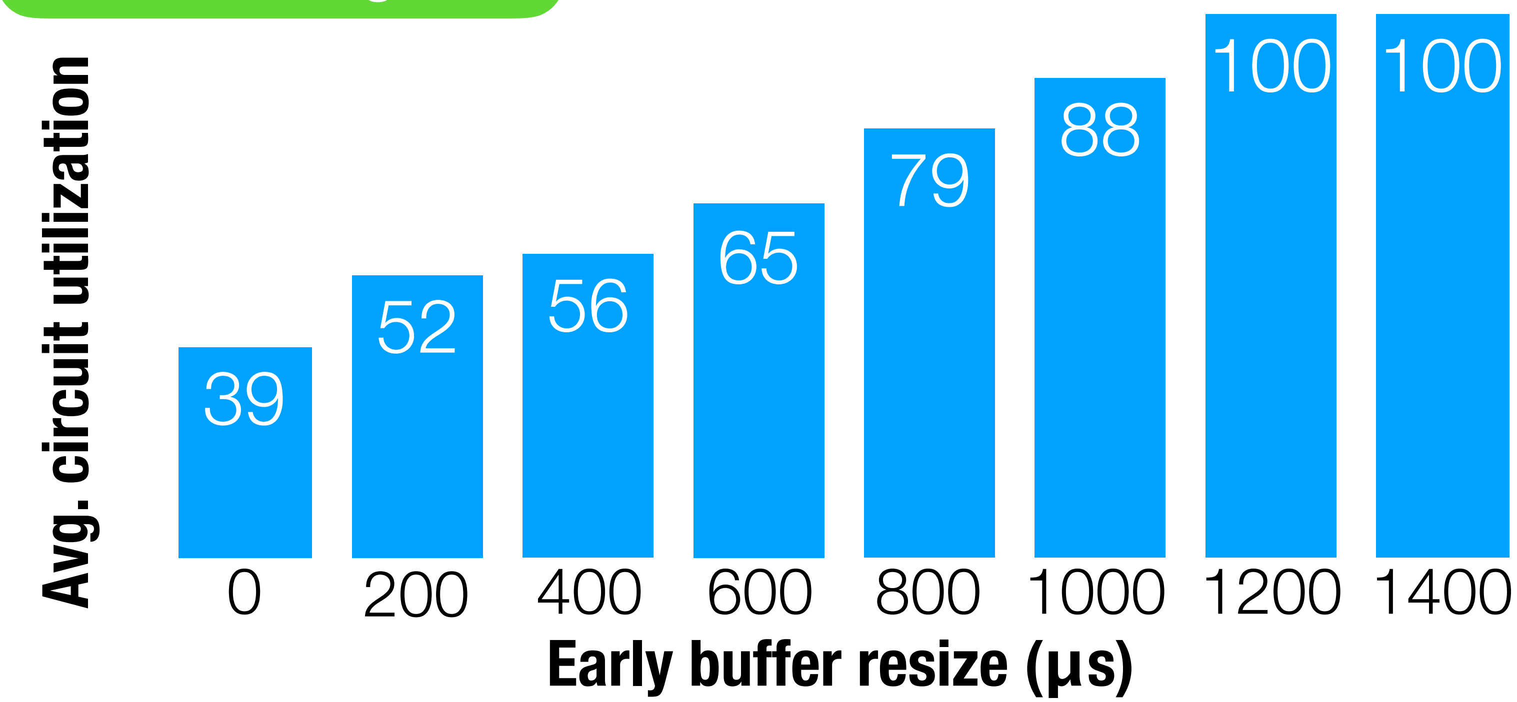




Challenge:
BW Fluct.

Solution:
Dynamic Buffer
Resizing

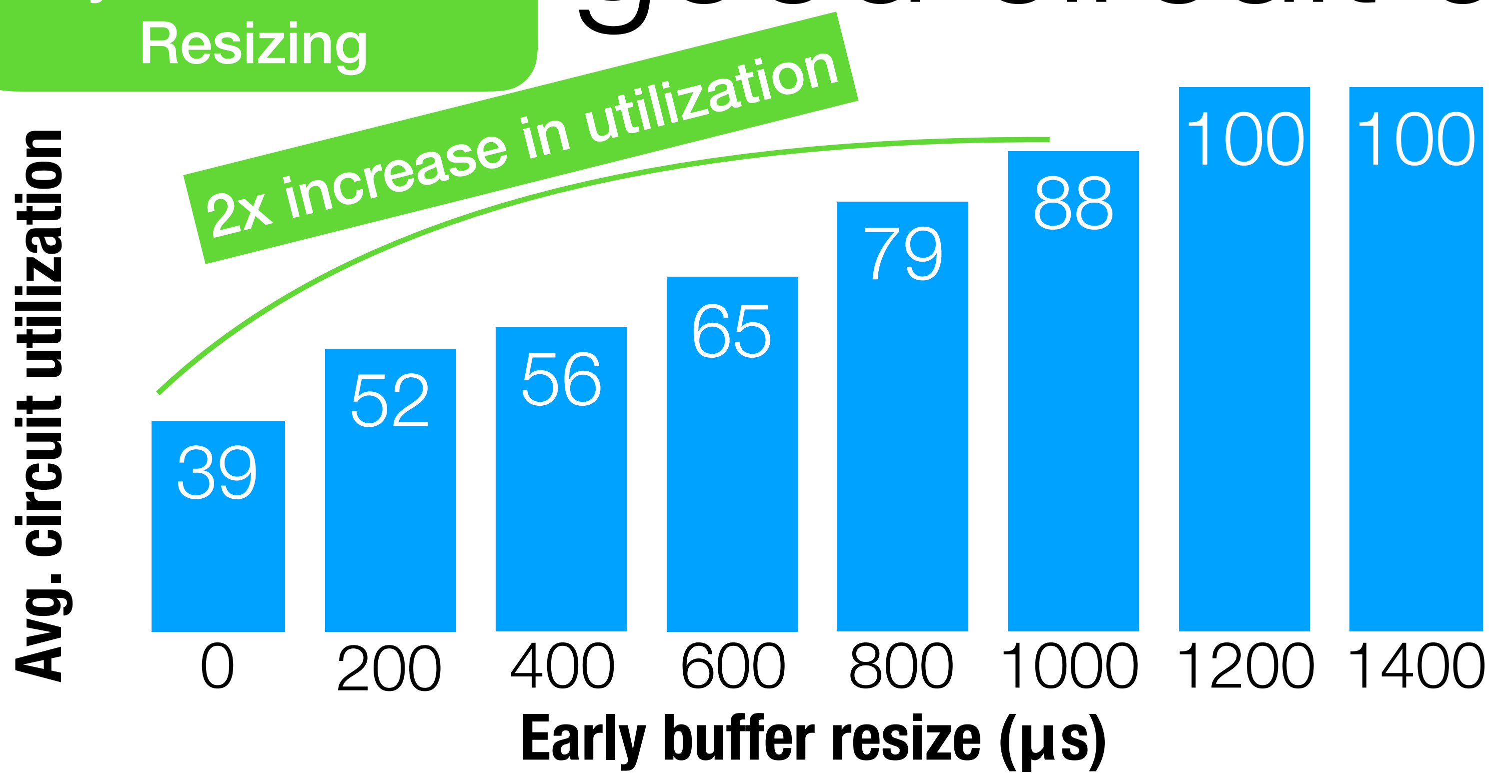
Buffer resize provides good circuit util **and** latency

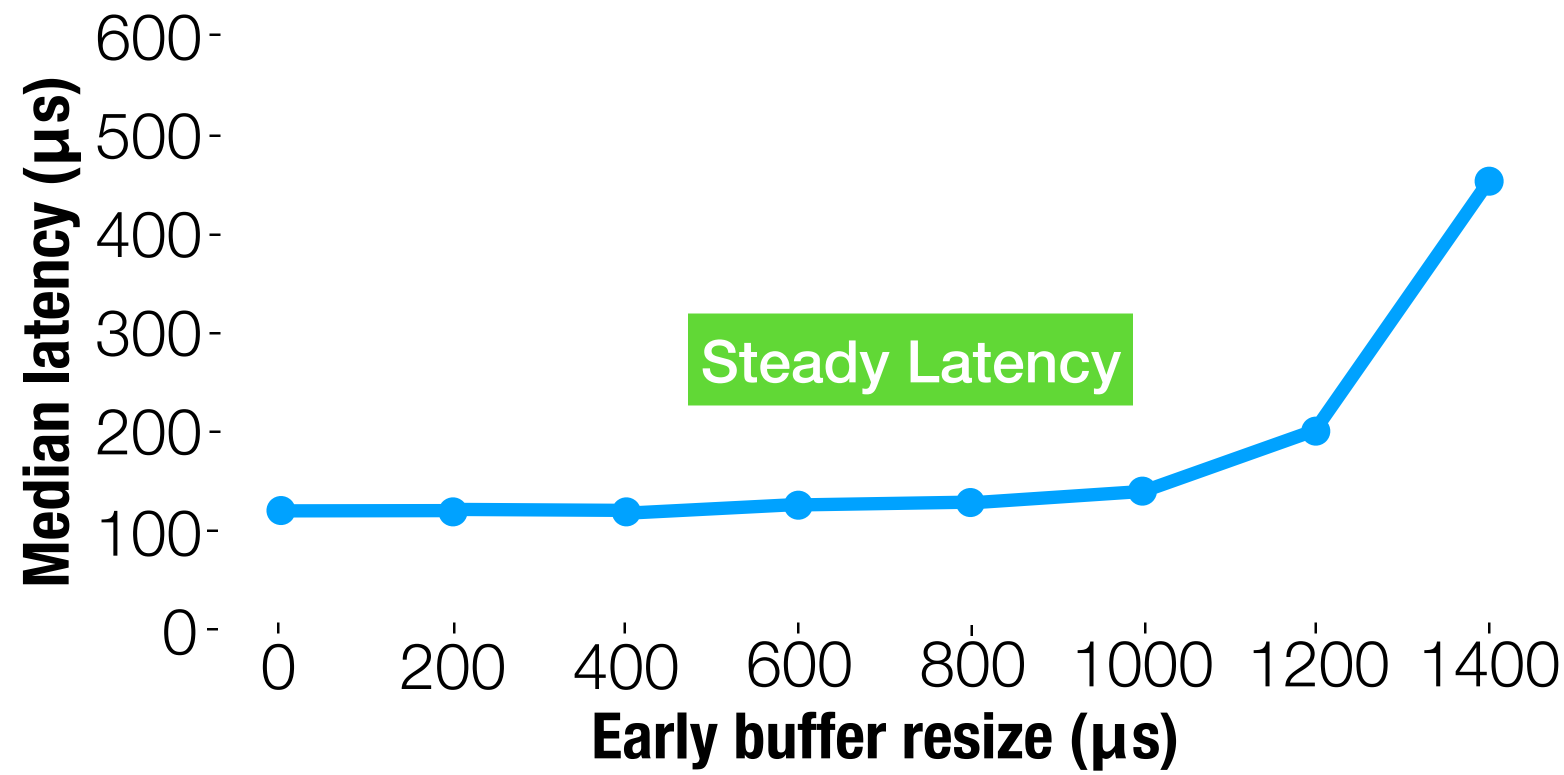


Challenge:
BW Fluct.

Buffer resize provides good circuit util **and** latency

Solution:
Dynamic Buffer Resizing





Overview

End-to-End Challenges

Challenge:
BW Fluct.

Challenge:
Demand Estimation

Challenge:
Workloads

Solution:
Dynamic Buffer Resizing

Solution:
Endhost-based Estimation

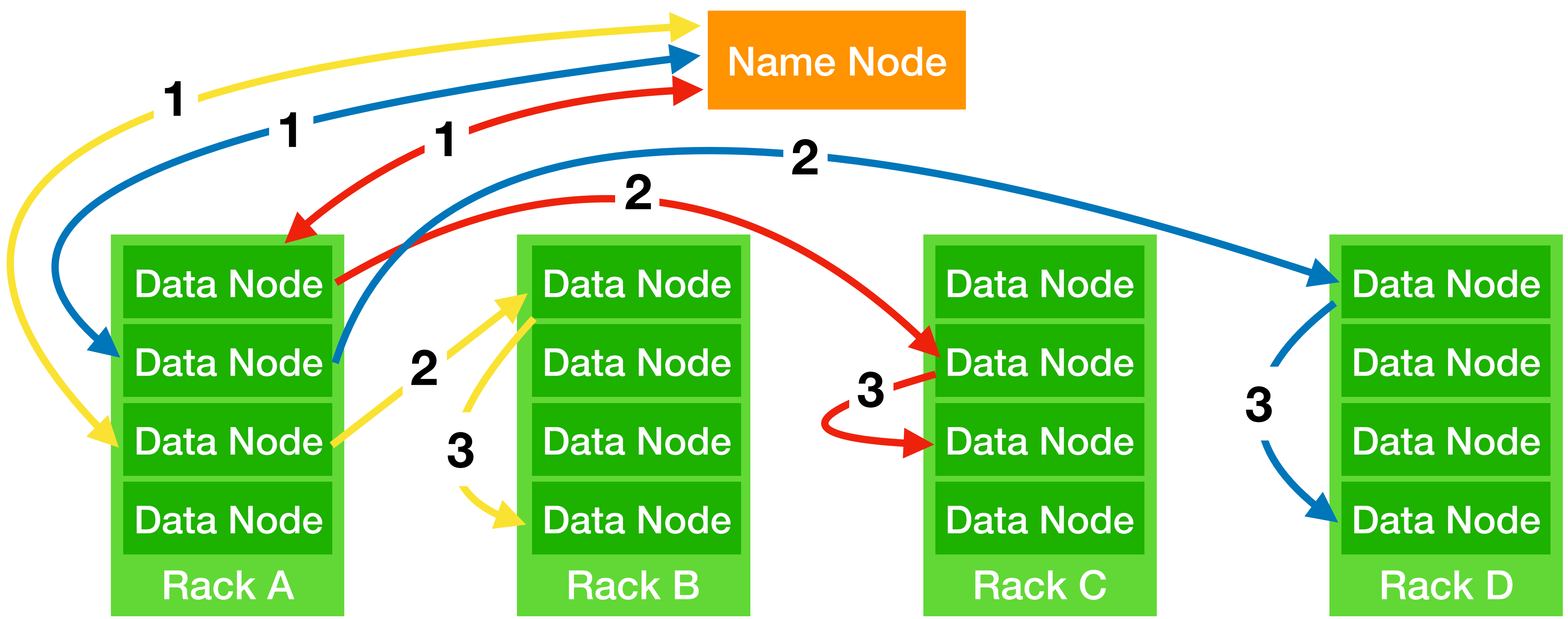
Solution:
App-specific Modification

Etalon, an RDCN Emulator

Challenge:
Workloads

Solution:
App-specific
Modification

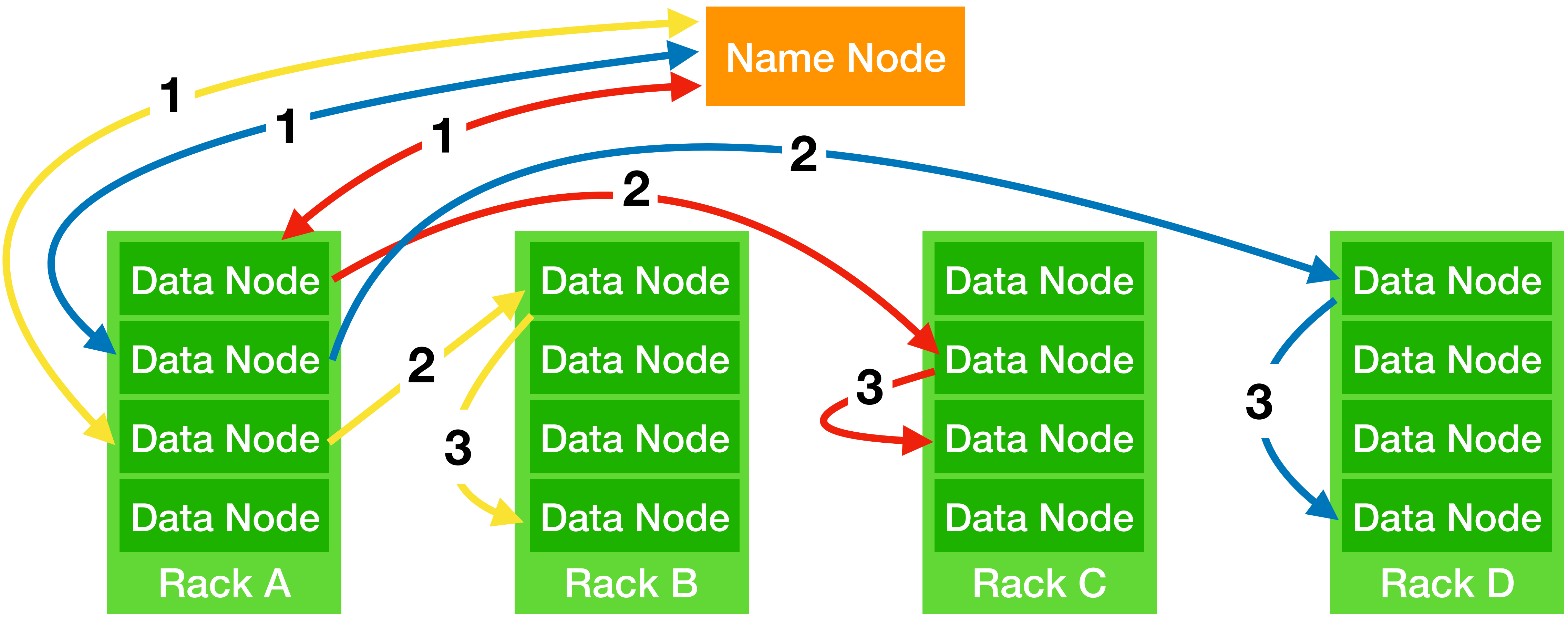
Difficult to schedule workloads

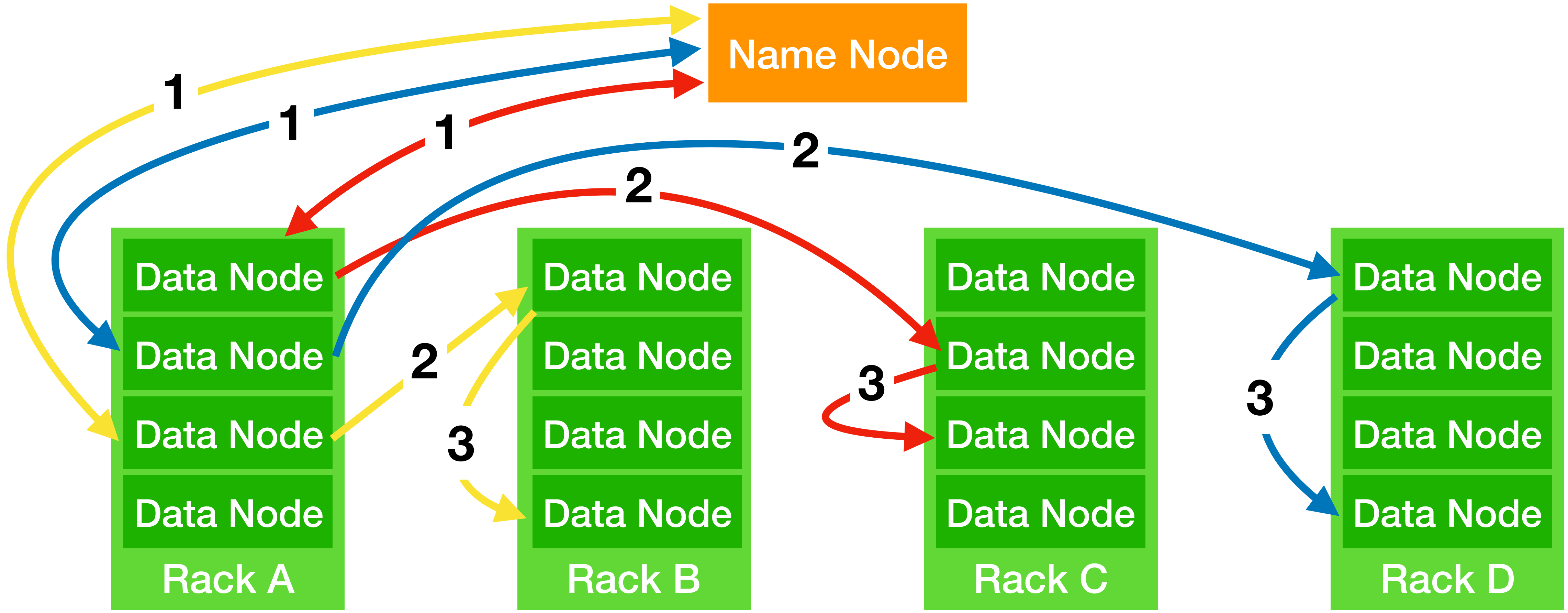


Challenge: Workloads

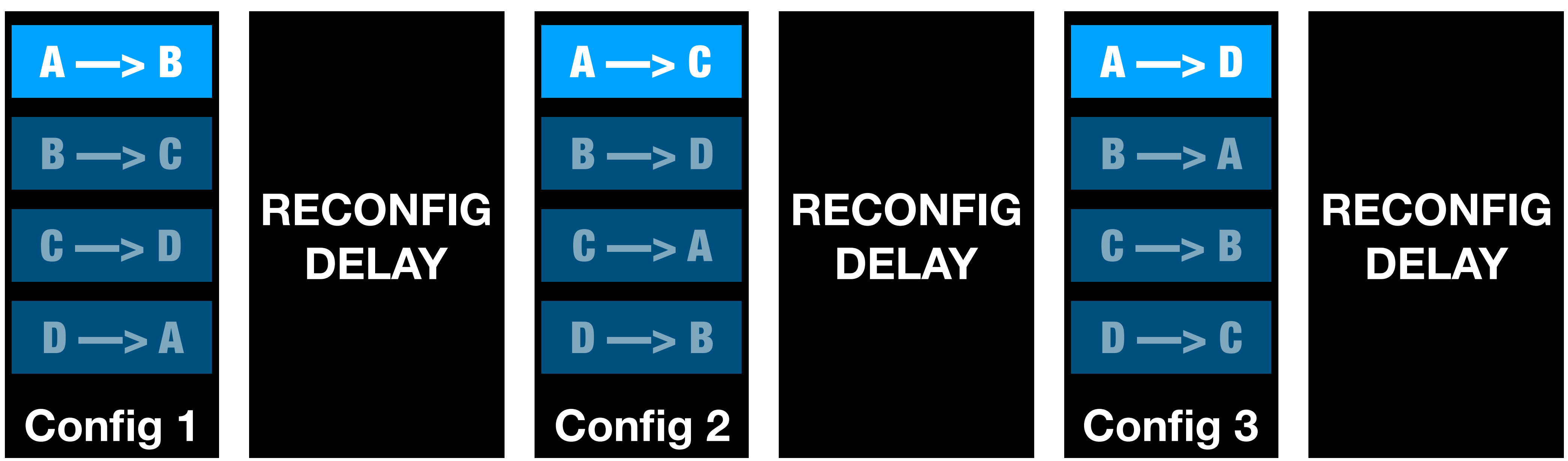
Solution: App-specific Modification

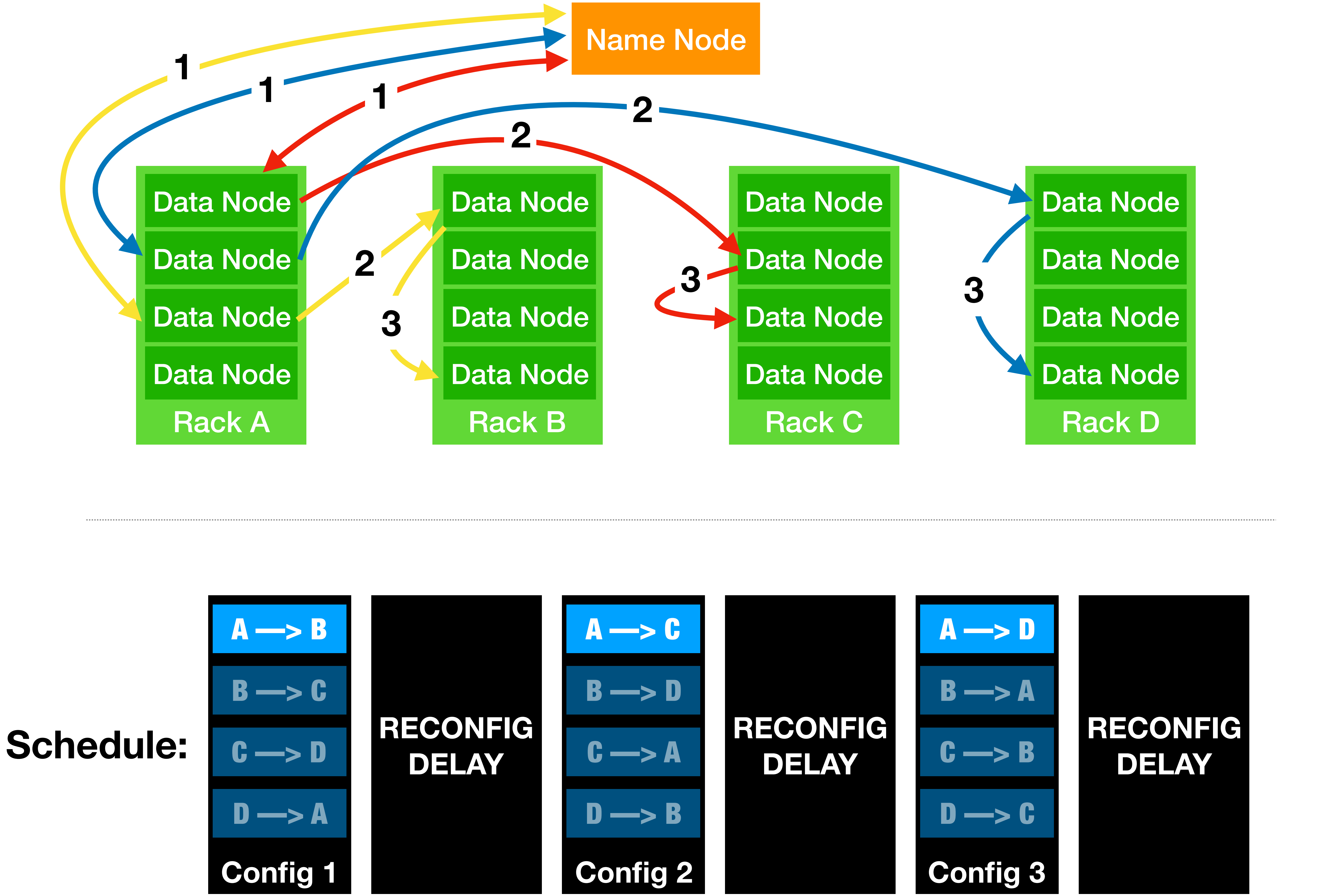
Difficult to schedule workloads




Name Node
1
1
1
2
2
2
3
3
3
Data Node
Data Node
Data Node
Data Node
Rack A
Data Node
Data Node
Data Node
Data Node
Rack B
Data Node
Data Node
Data Node
Data Node
Rack C
Data Node
Data Node
Data Node
Data Node
Rack D

Schedule:


A —> B
B —> C
C —> D
D —> A
Config 1
RECONFIG DELAY
A —> C
B —> D
C —> A
D —> B
Config 2
RECONFIG DELAY
A —> D
B —> A
C —> B
D —> C
Config 3
RECONFIG DELAY

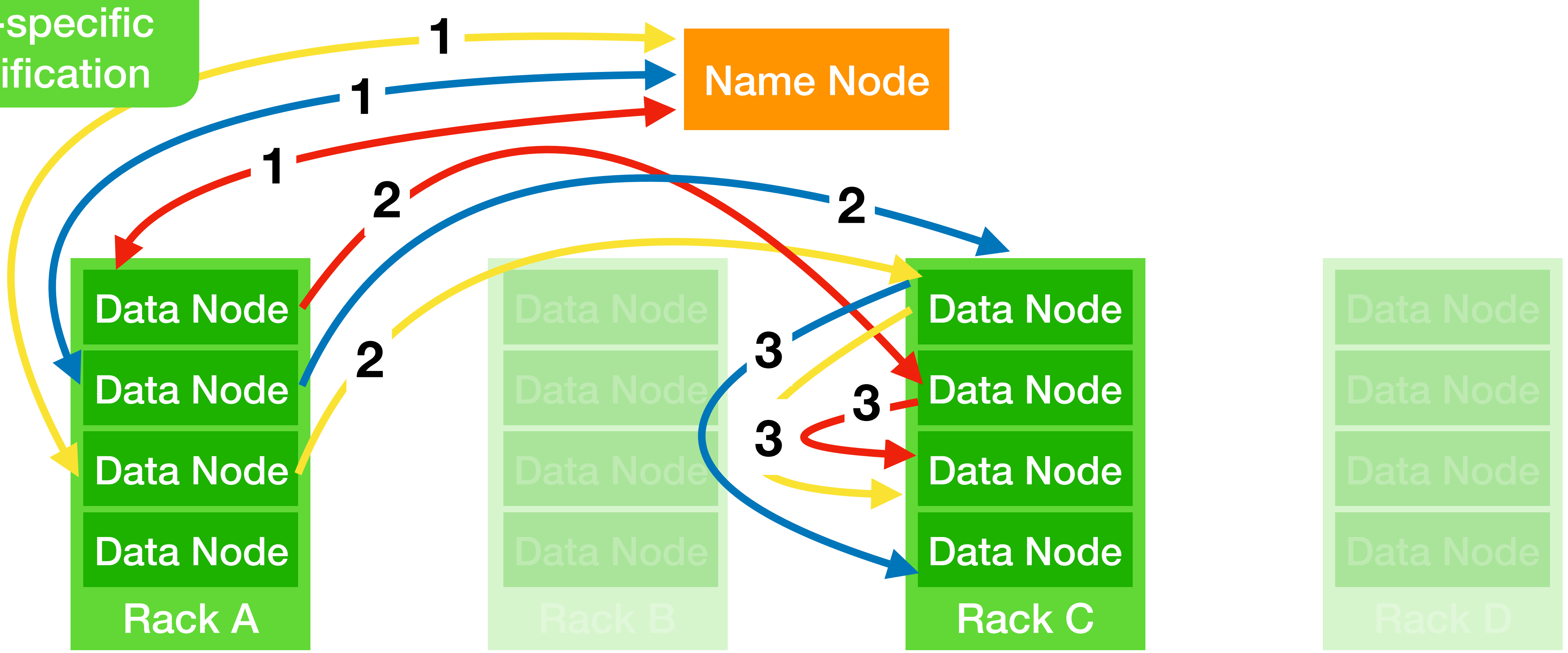


Name Node
1
1
1
2
2
2
3
3
3
Data Node
Rack A
Rack B
Rack C
Rack D
Schedule:
A —> B
B —> C
C —> D
D —> A
Config 1
RECONFIG DELAY
A —> C
B —> D
C —> A
D —> B
Config 2
RECONFIG DELAY
A —> D
B —> A
C —> B
D —> C
Config 3
RECONFIG DELAY

Challenge: Workloads

Solution: App-specific Modification

reHDFS



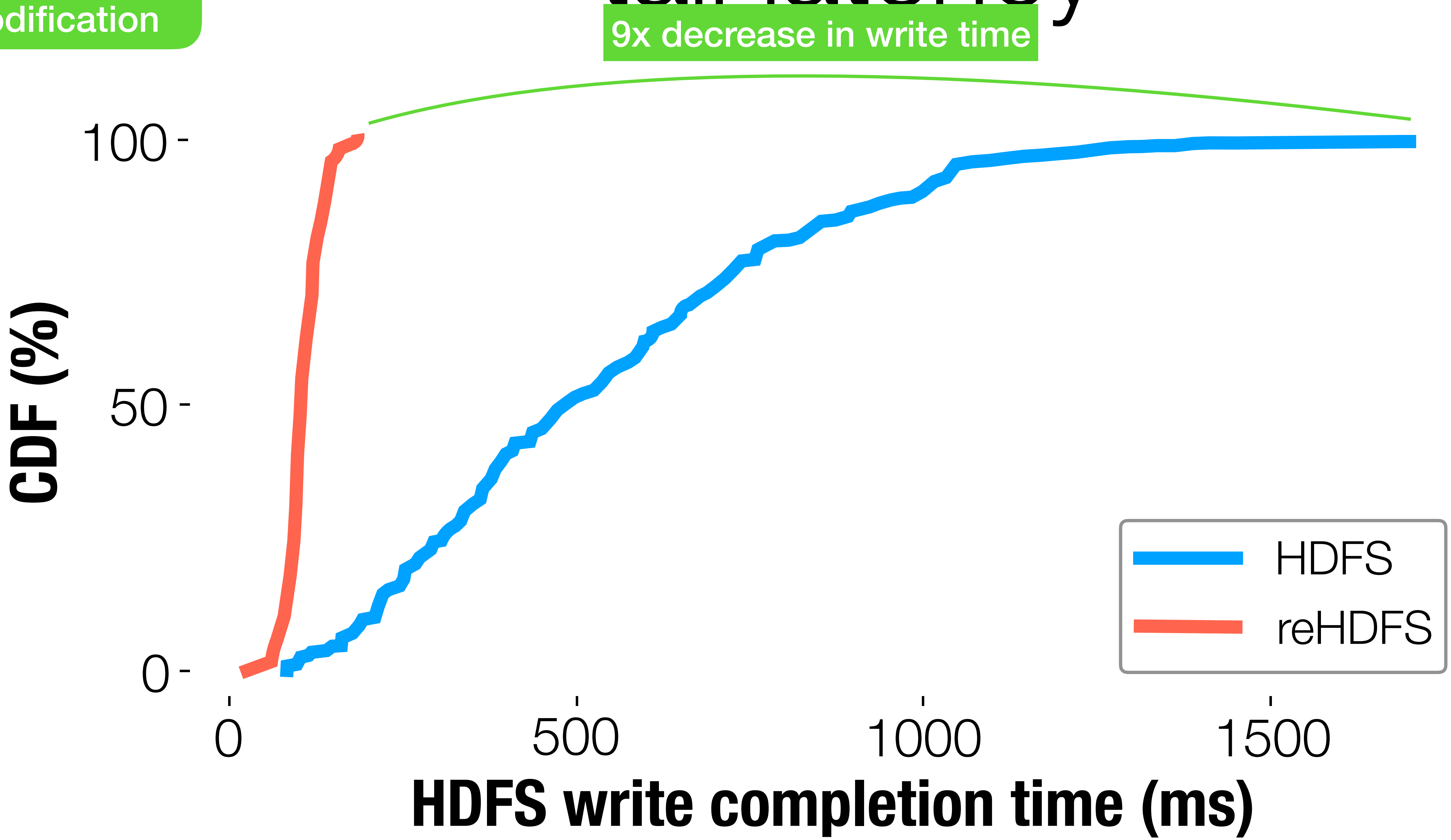
Schedule:

A —> C
B —> D
C —> A
D —> B
Config 1

Challenge:
Workloads

Solution:
App-specific
Modification

reHDFS reduces tail latency



Control Coordination

Scenario: **Layering**

Etalon

Scenario: **Admin**

VDX

Scenario: **Scalability**

VDN

App TE + ISP TE **Reaction**	Coflow **Transparency**
BGP + BGP **Priority Ranking**	Internet-scale Routing **Hierarchical Partitioning**

Control Coordination

Scenario:
Admin

VDX

Scenario:
Scalability

VDN

App TE + ISP TE

Reaction

BGP + BGP

Priority Ranking

Internet-scale Routing

Hierarchical Partitioning

Coflow

Etalon

Transparency

Scenario:
Layering

Control Coordination

Scenario:
Admin

VDX

Scenario:
Scalability

VDN

App TE + ISP TE

Reaction

BGP + BGP

Priority Ranking

Internet-scale Routing

Hierarchical Partitioning

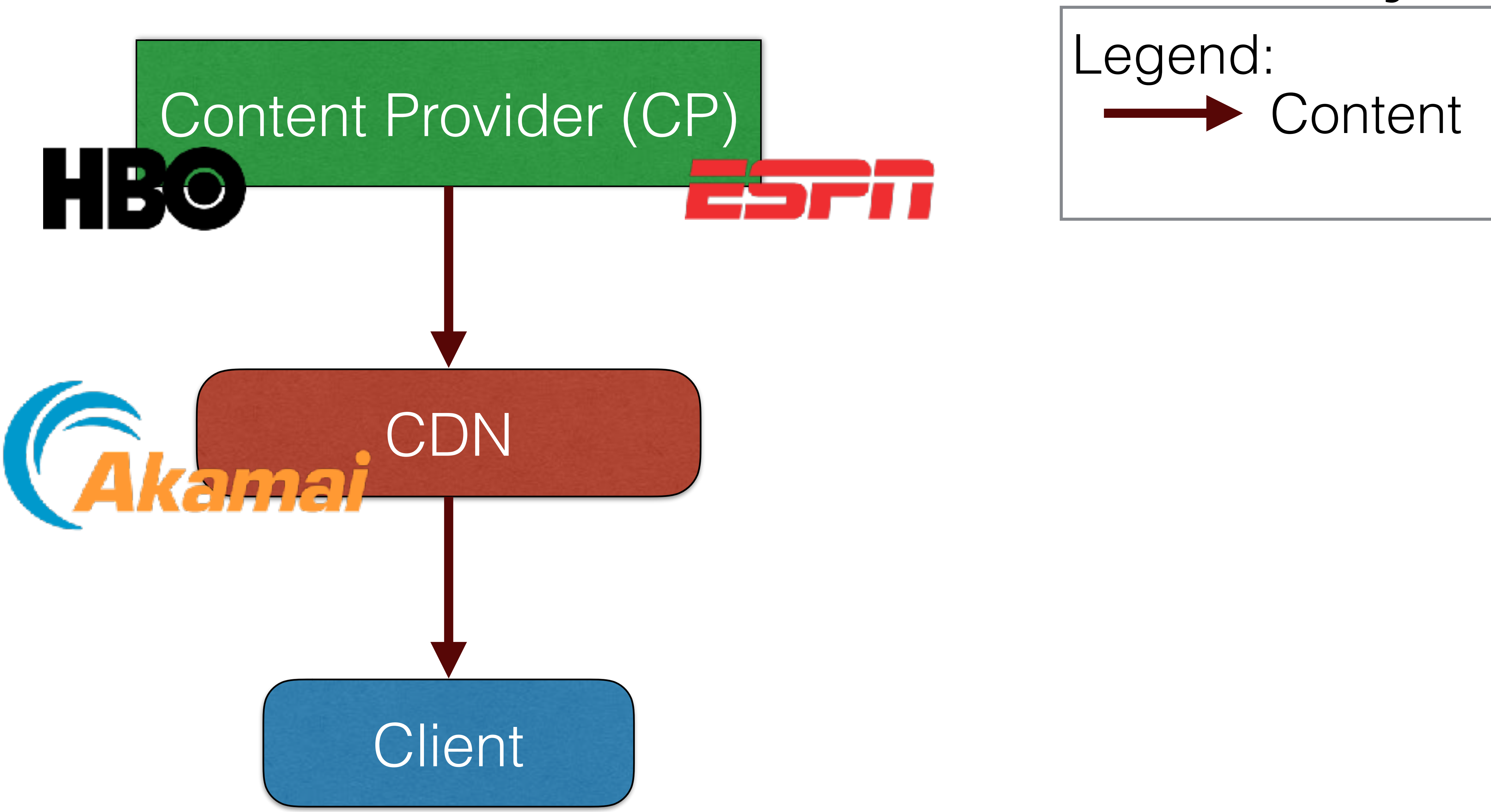
Coflow

Etalon

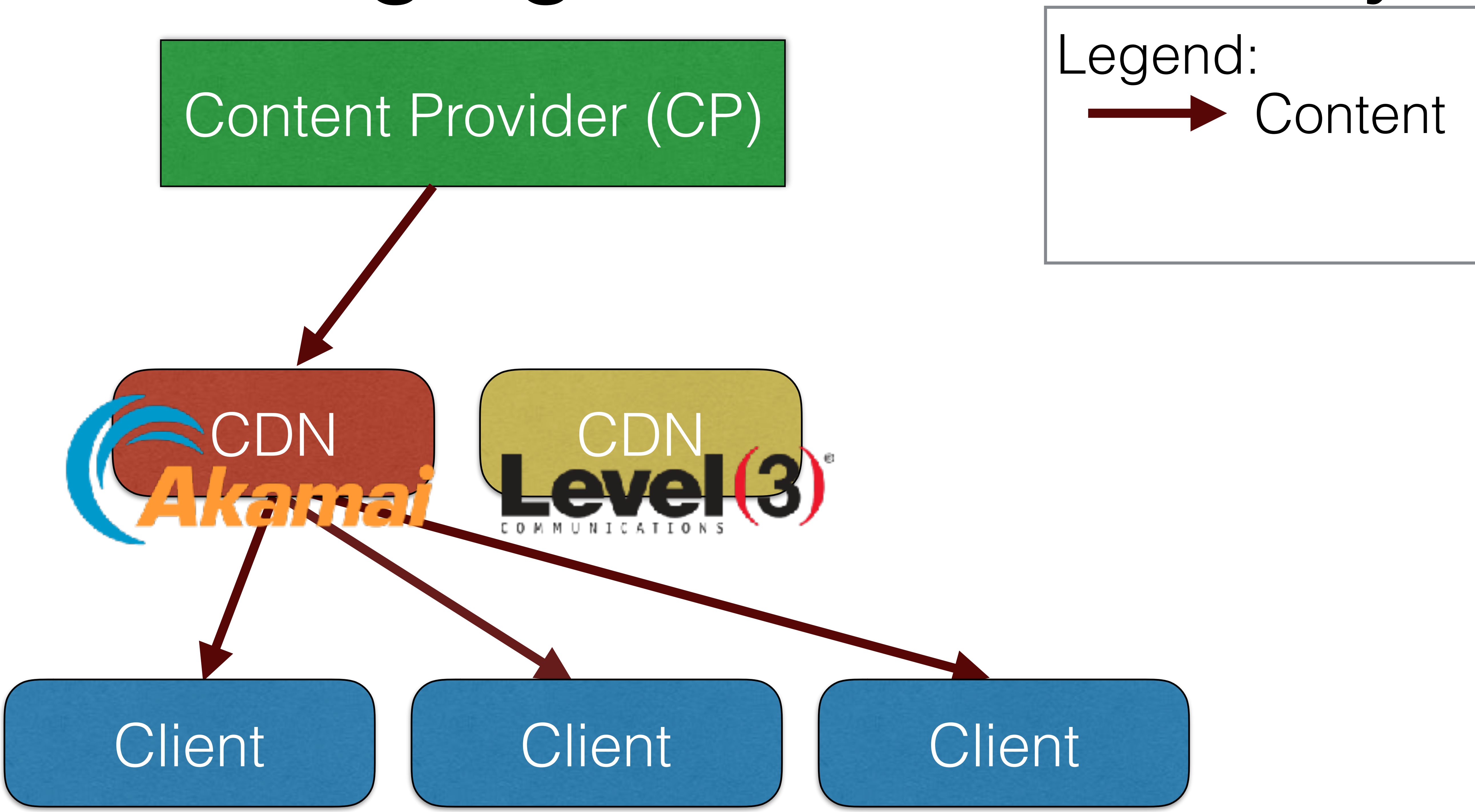
Transparency

Scenario:
Layering

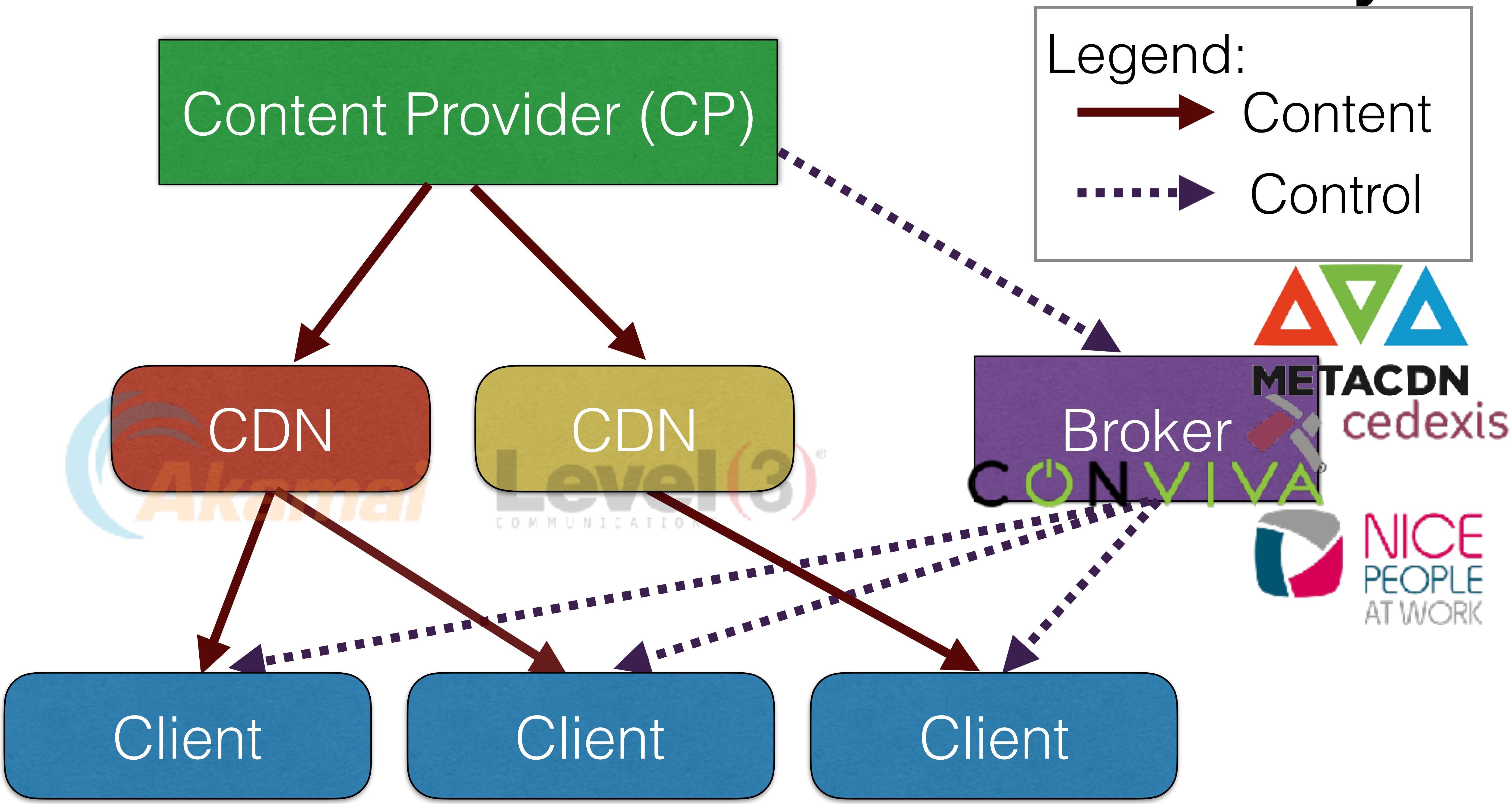
Traditional Content Delivery



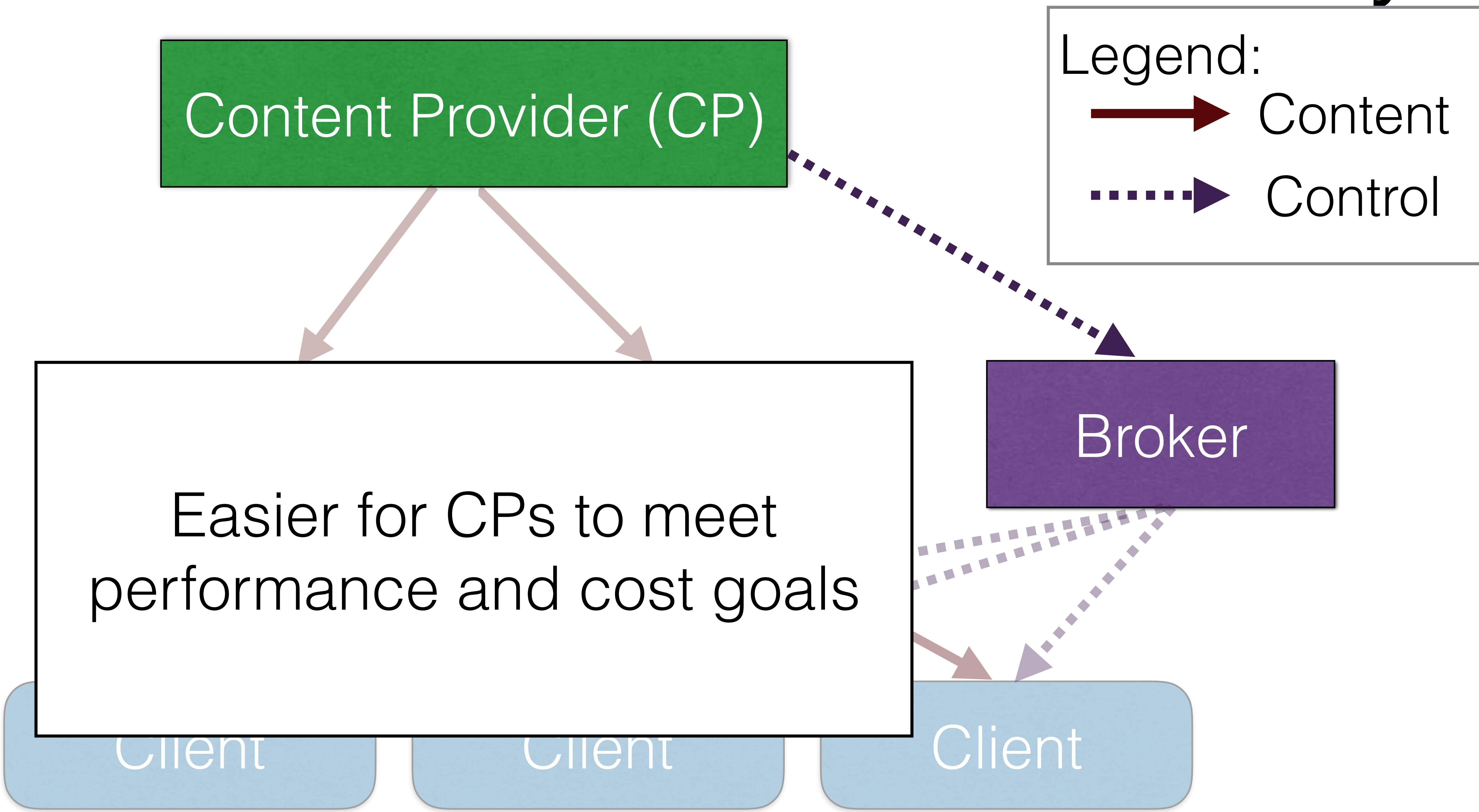
Changing Content Delivery



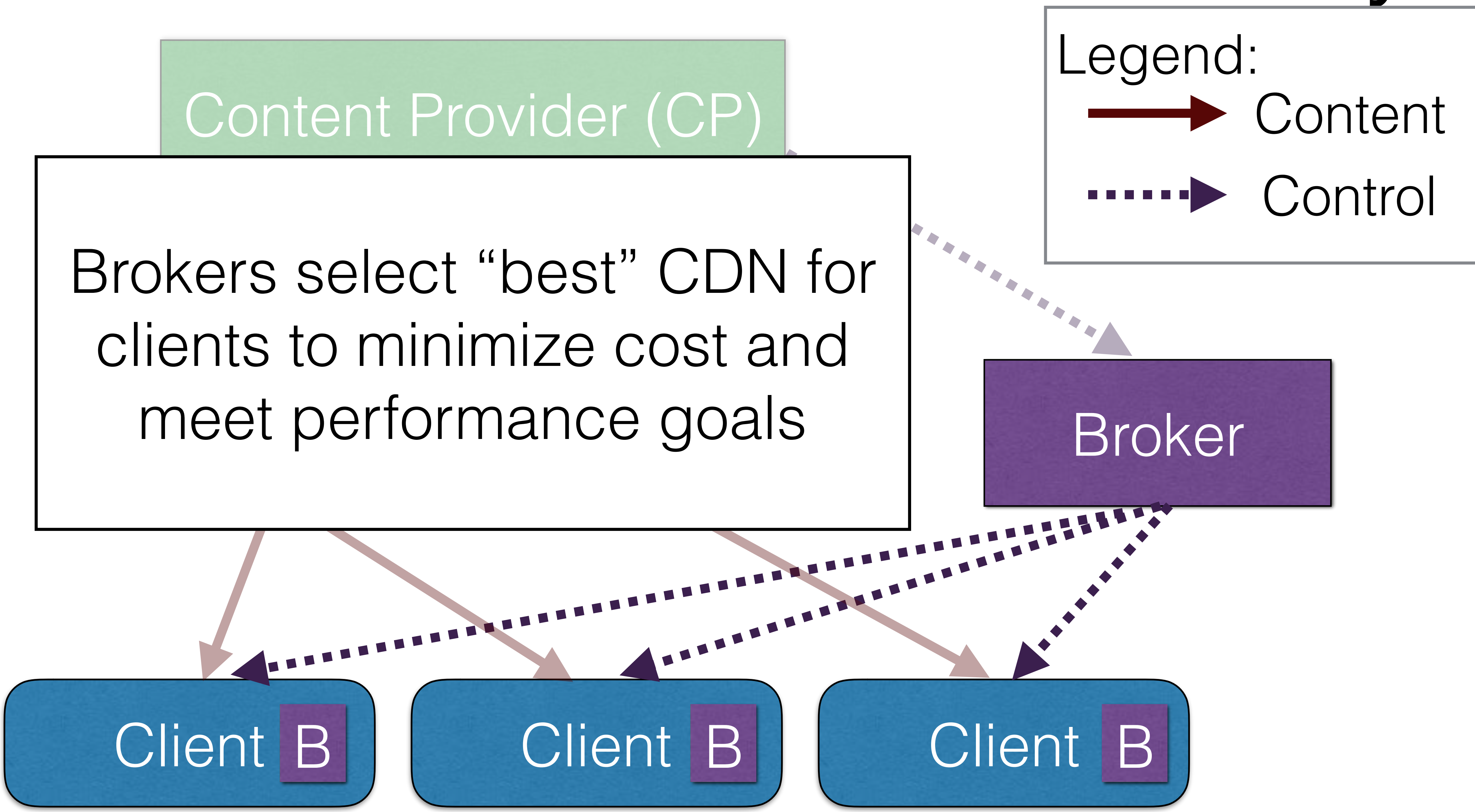
Brokered Content Delivery



Brokered Content Delivery

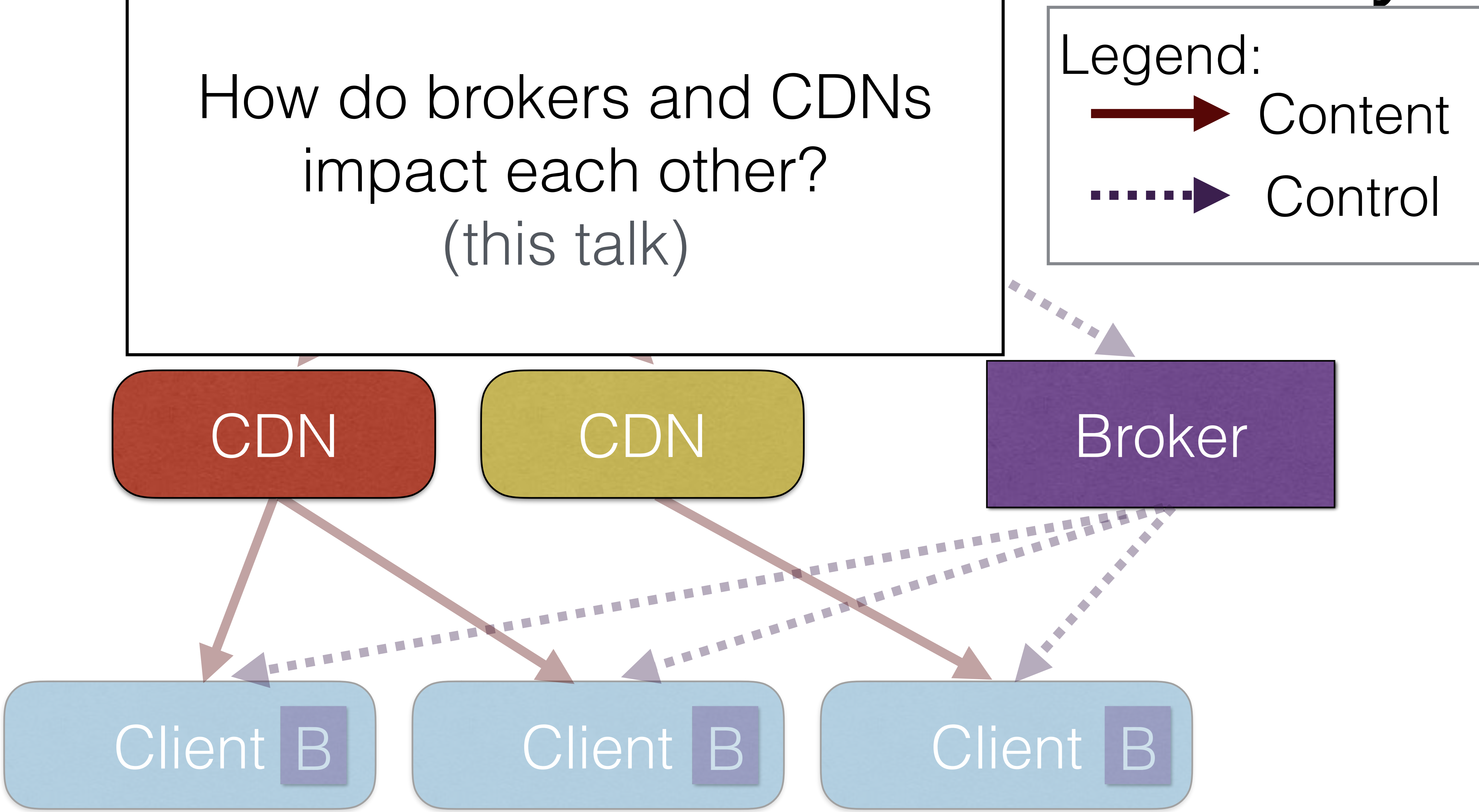


Brokered Content Delivery



Brokered Content Delivery

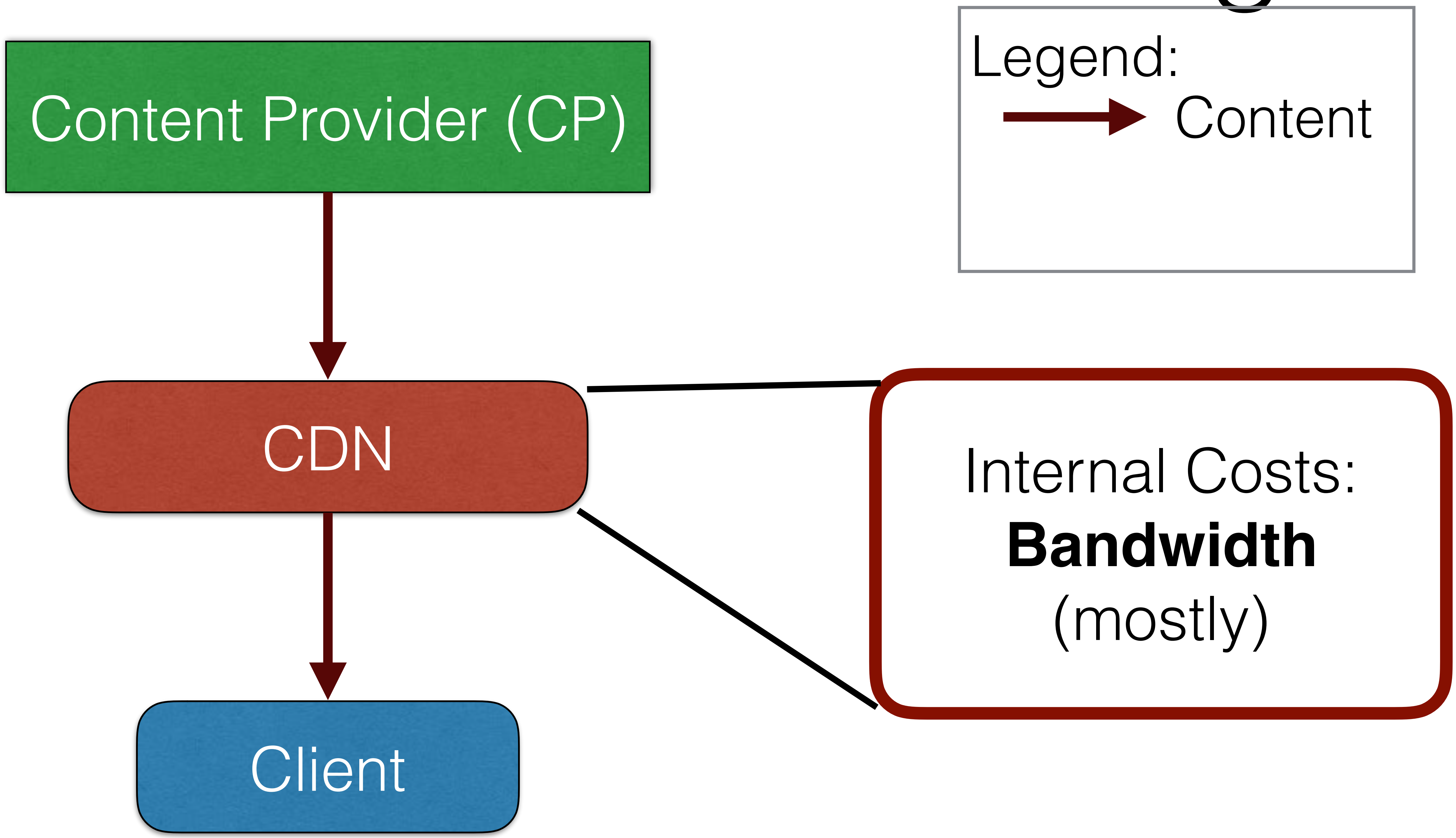
How do brokers and CDNs impact each other?
(this talk)



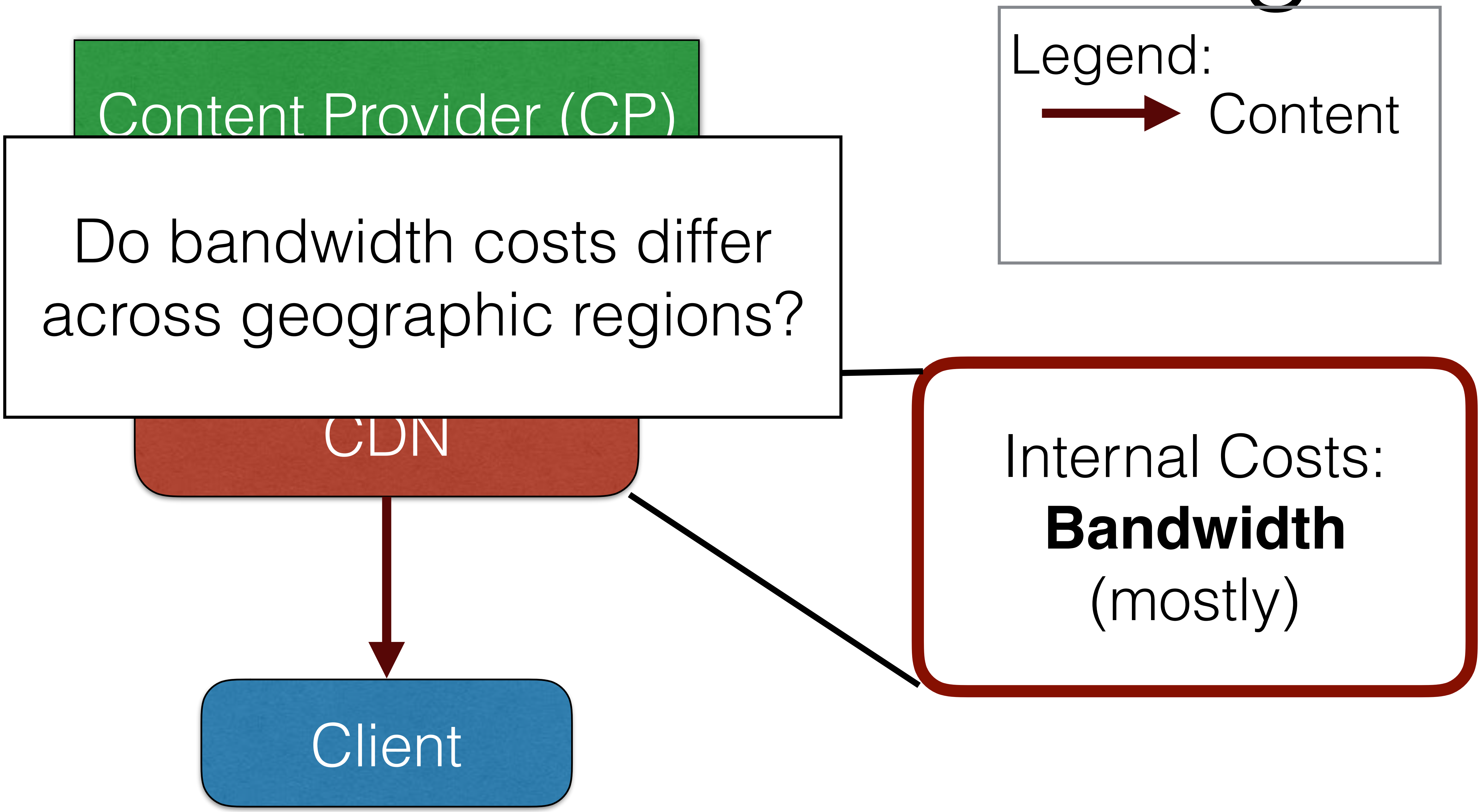
Contributions

- Identify challenges that brokers and CDNs create for each other by analyzing data from both
- Examine the design space of CDN-broker interfaces
- Evaluate the efficacy of different designs

CDN Cost and Pricing



CDN Cost and Pricing



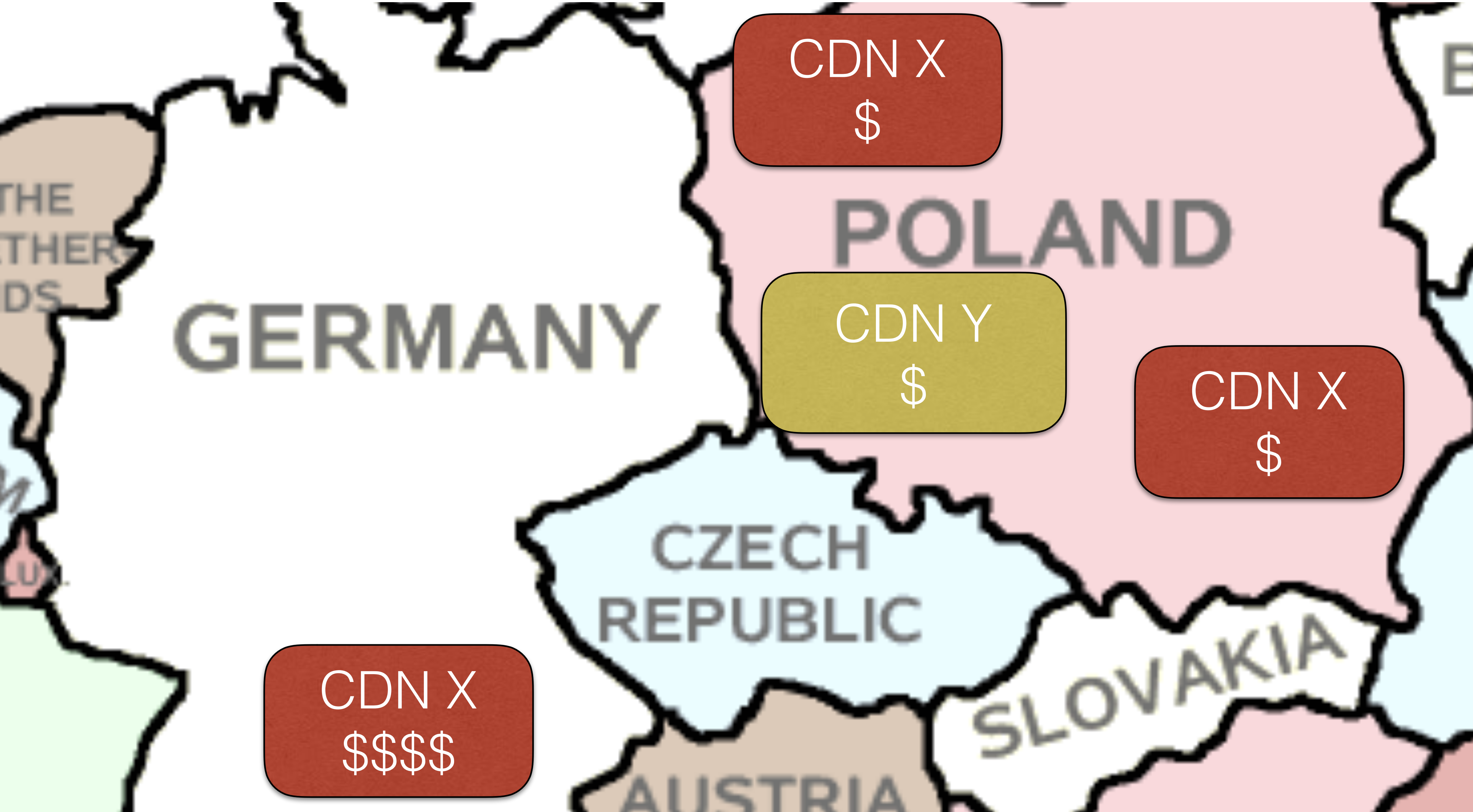
CDN Cost / Byte Delivered

30x

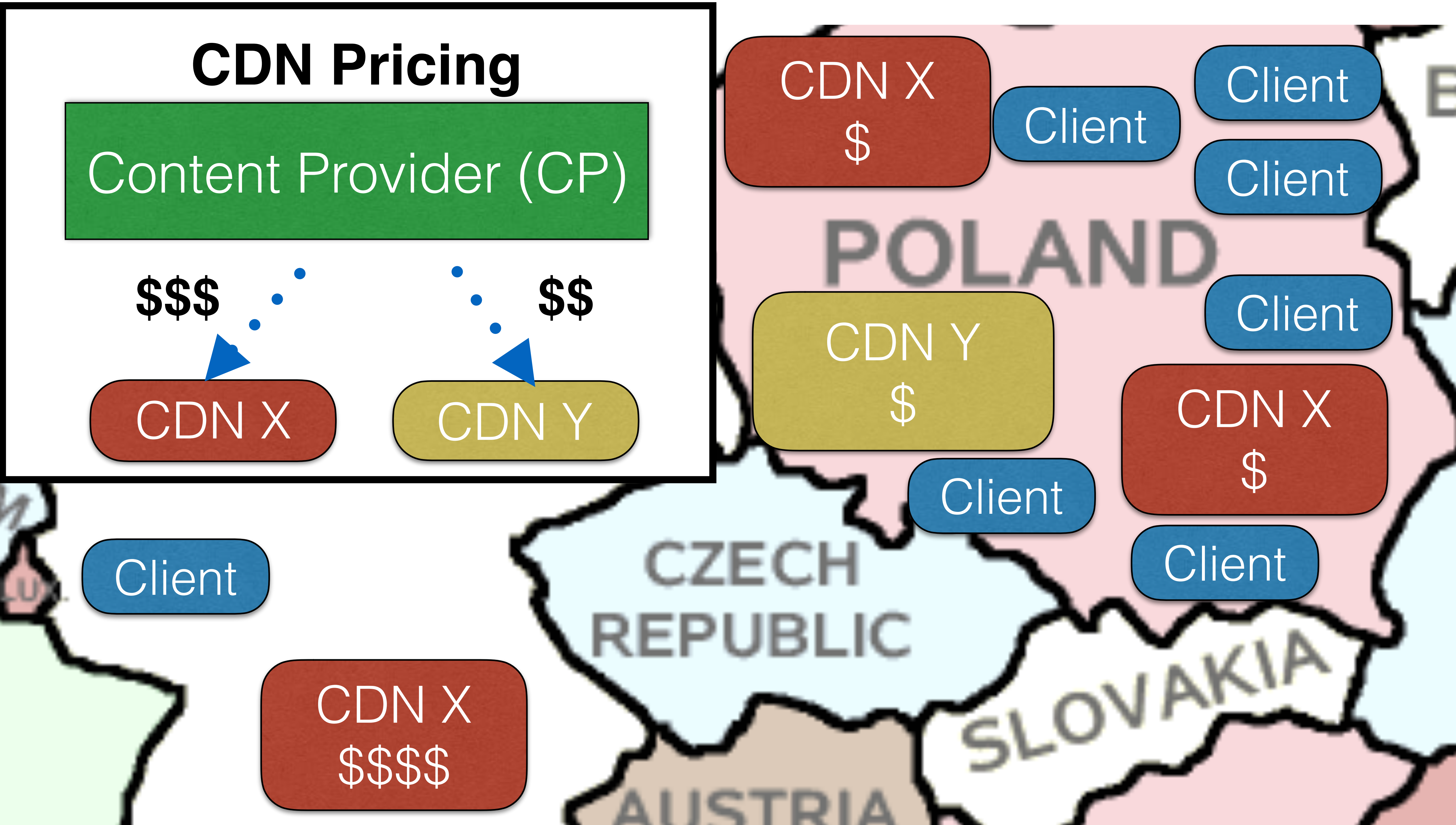
difference in cost per byte
between the most expensive
and least expensive countries

CDN Internal Cost

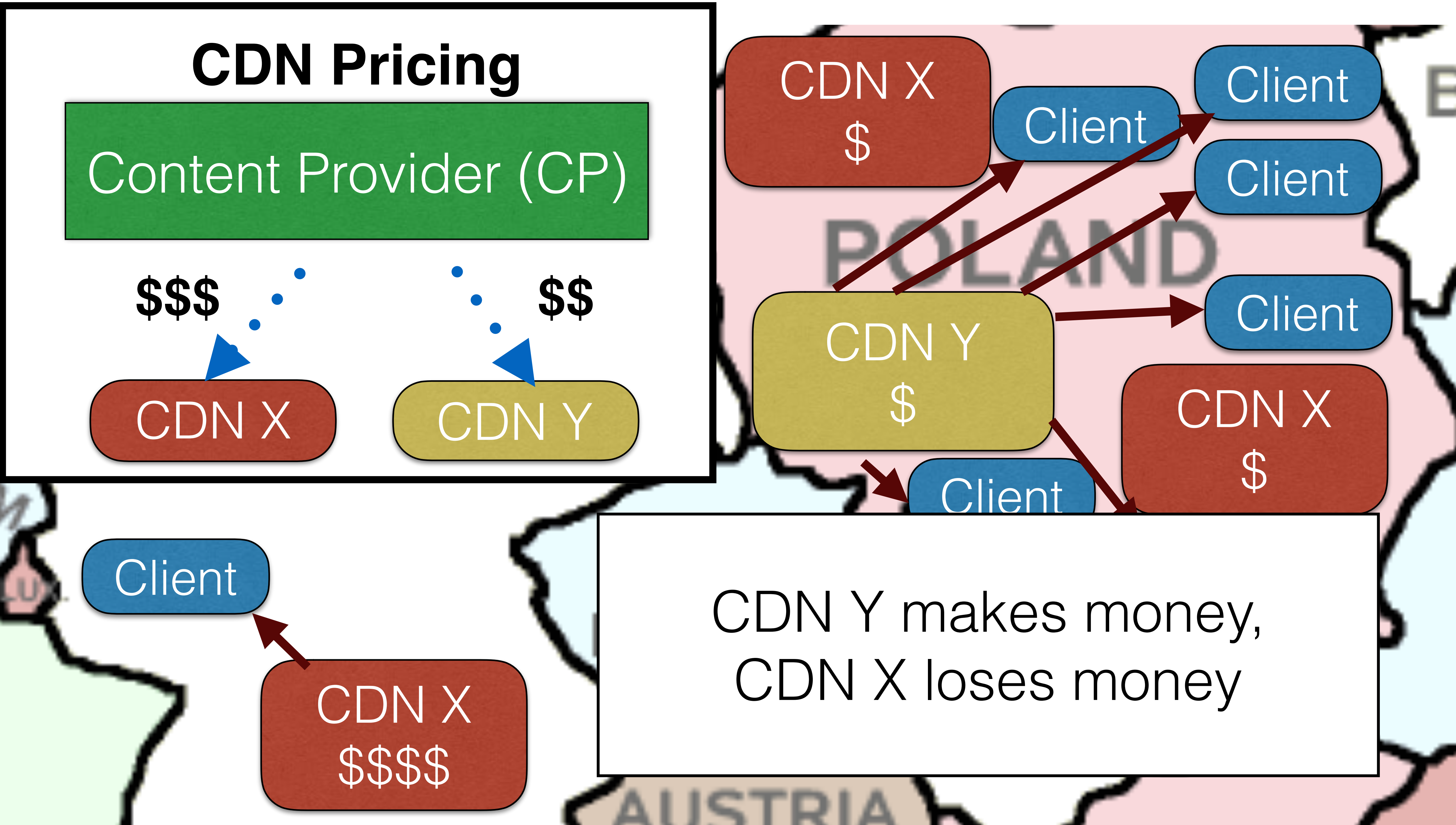
CDN Internal Cost



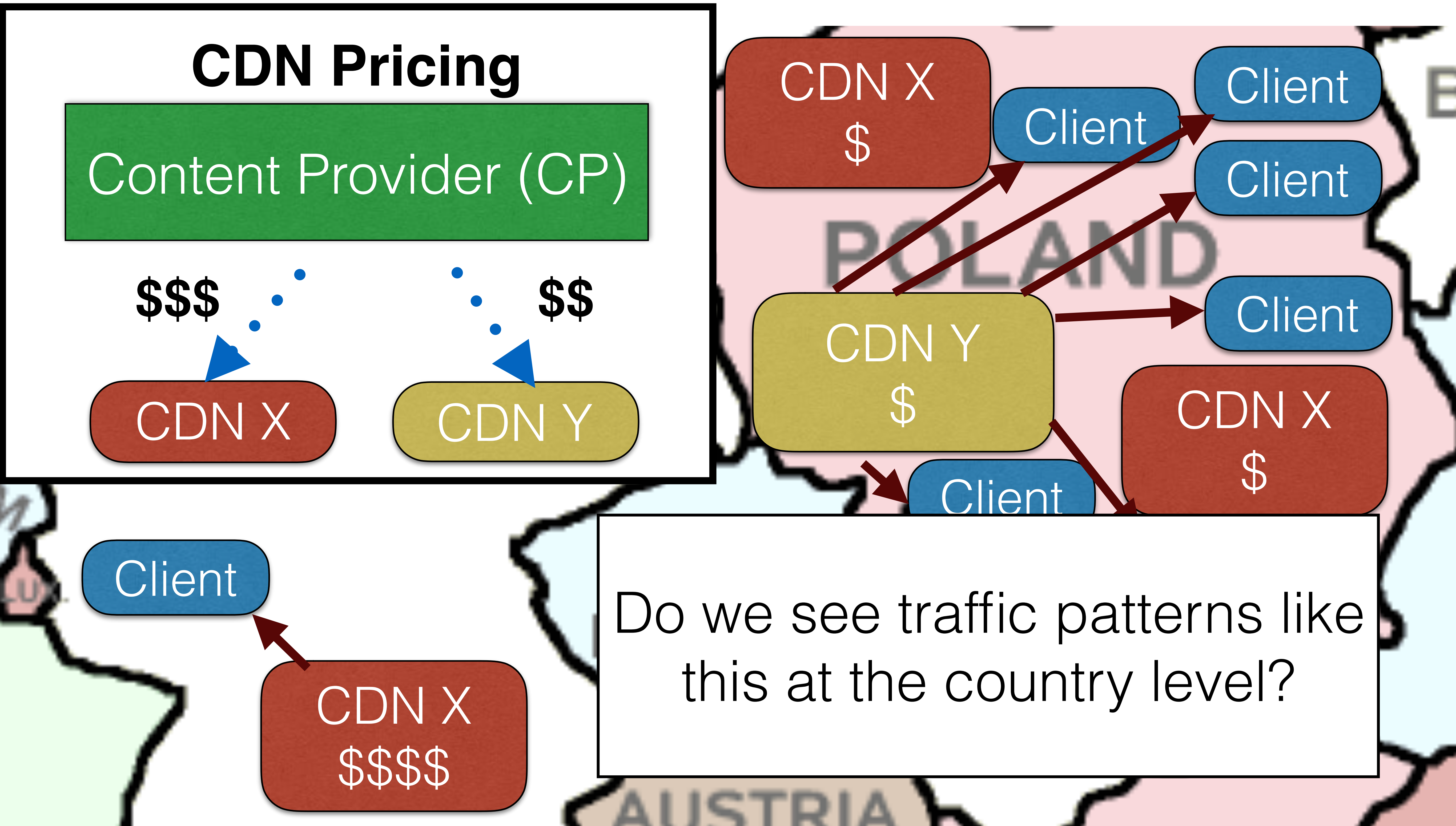
CDN External Price

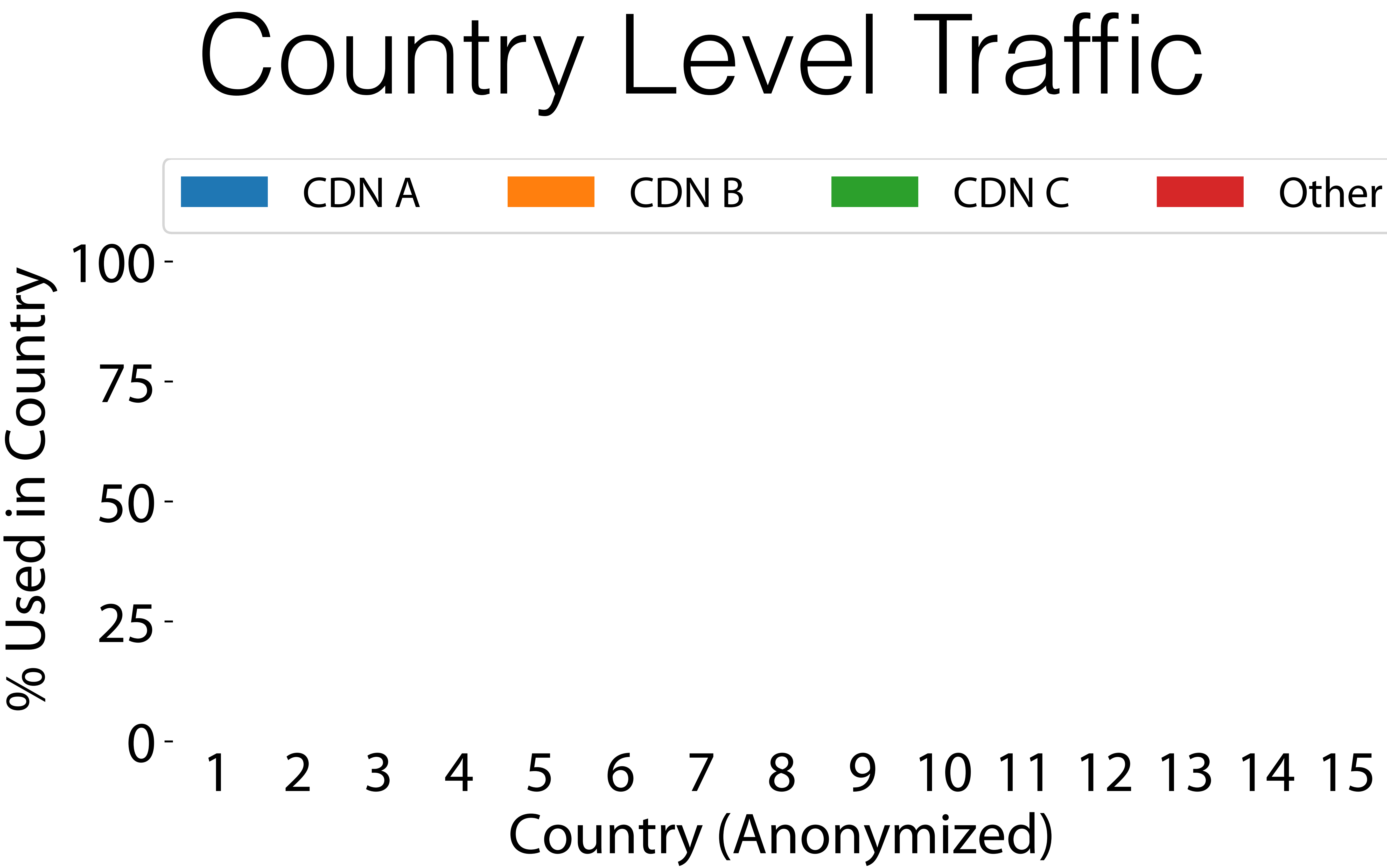


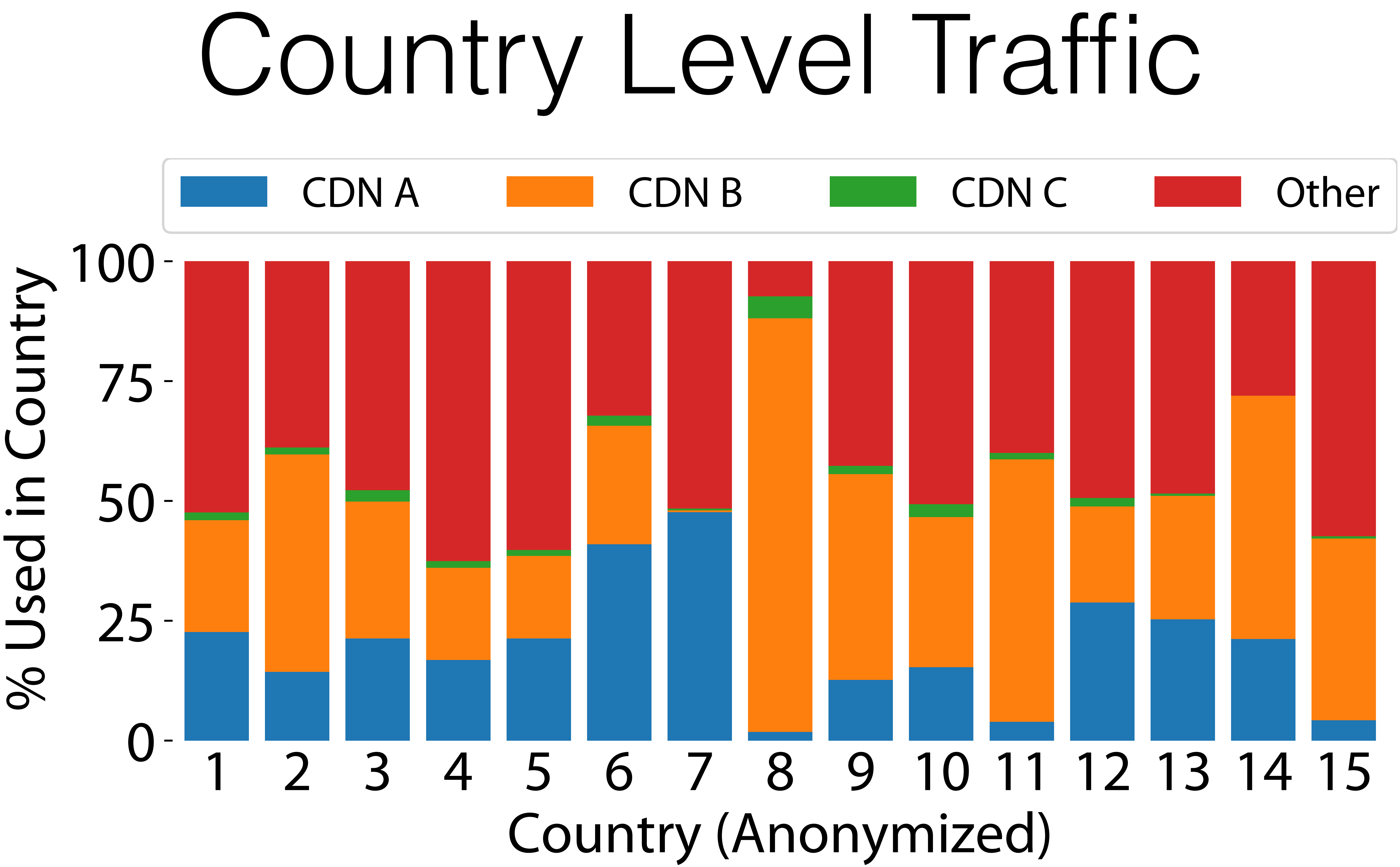
CDN External Price

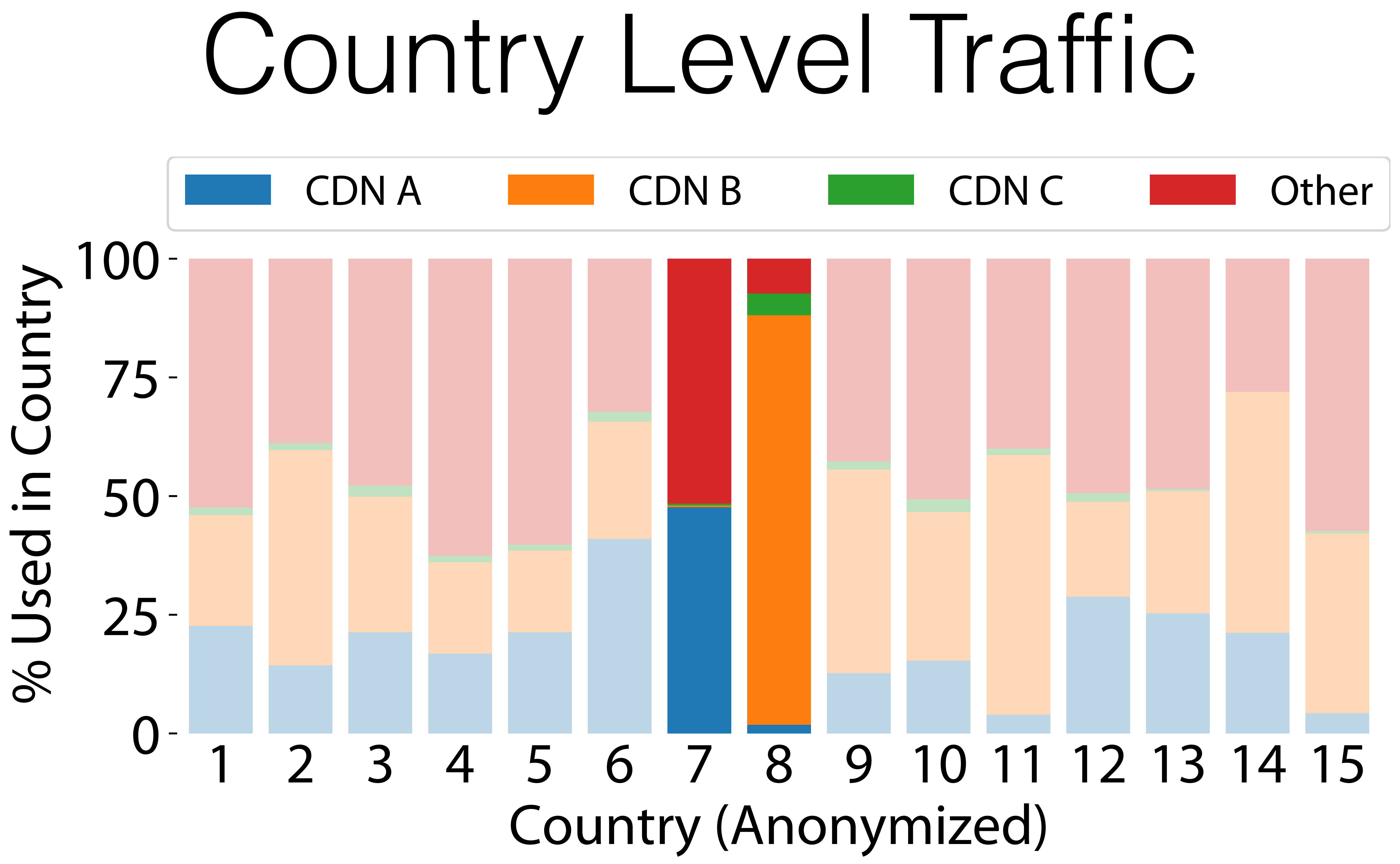


CDN External Price



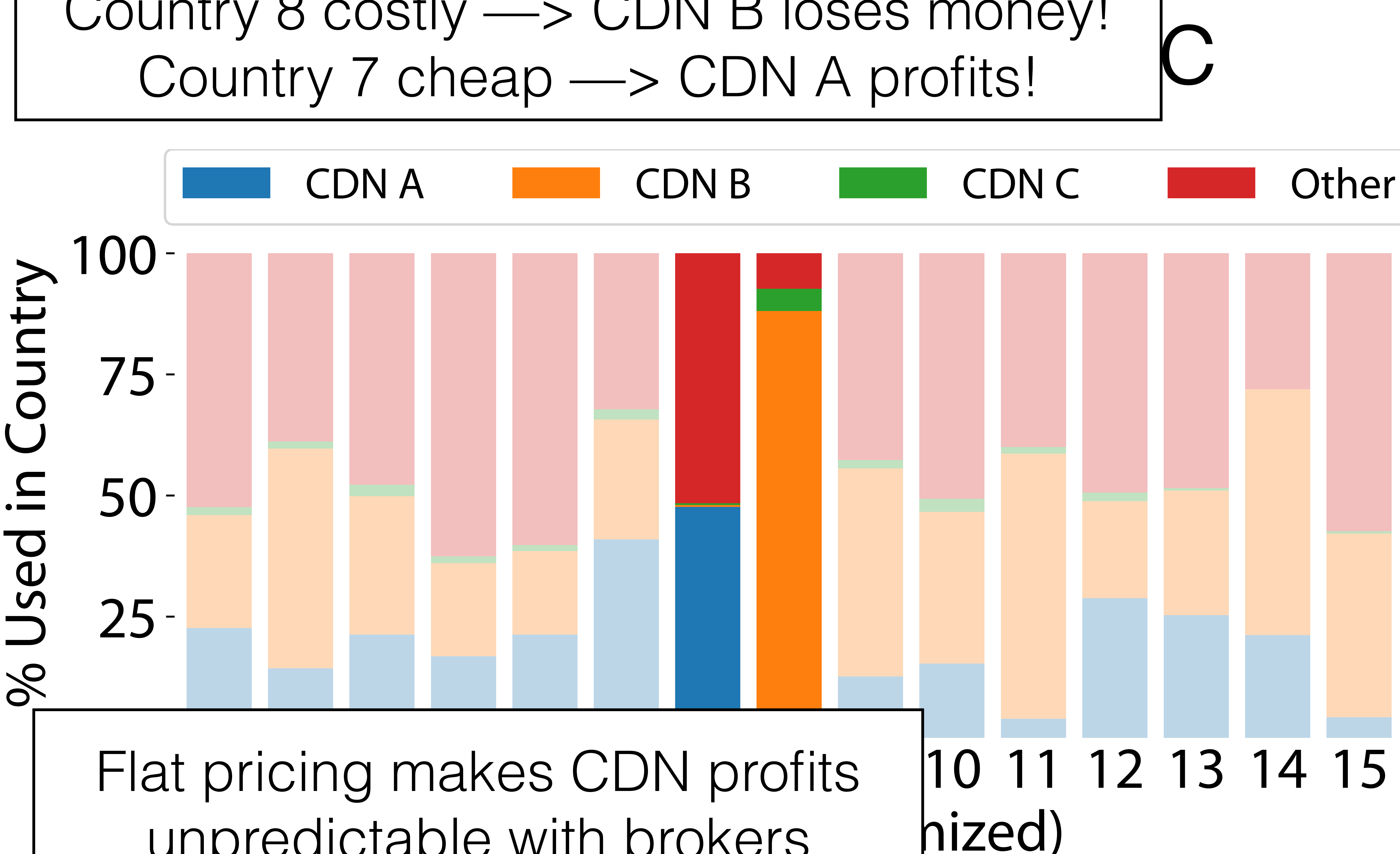

Country Level Traffic
CDN A
CDN B
CDN C
Other
100
75
50
25
0
% Used in Country
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Country (Anonymized)


Country Level Traffic
CDN A
CDN B
CDN C
Other
% Used in Country
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Country (Anonymized)



Country Level Traffic
CDN A
CDN B
CDN C
Other
% Used in Country
100
75
50
25
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Country (Anonymized)

Country 8 costly —> CDN B loses money!
Country 7 cheap —> CDN A profits!

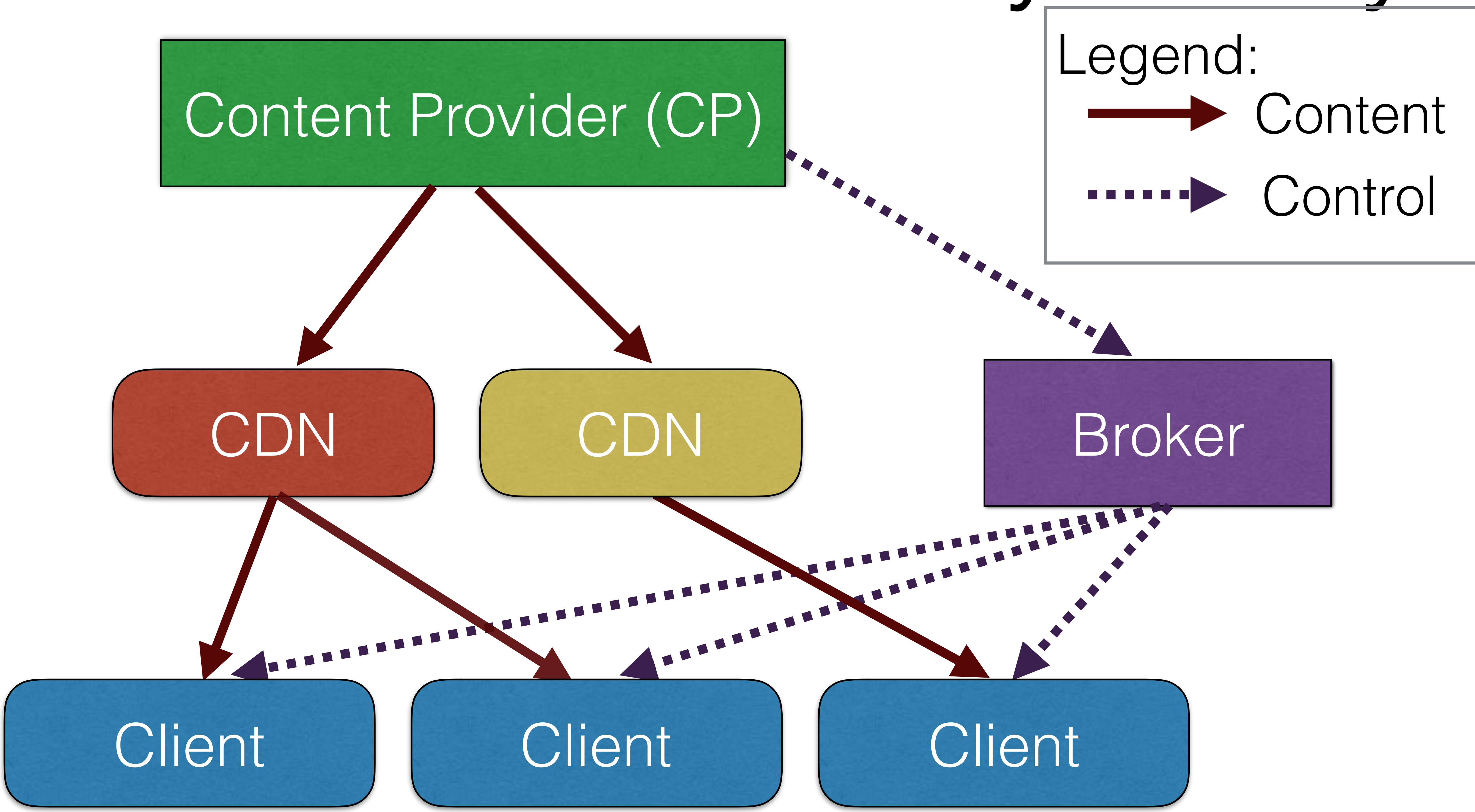


Flat pricing makes CDN profits unpredictable with brokers

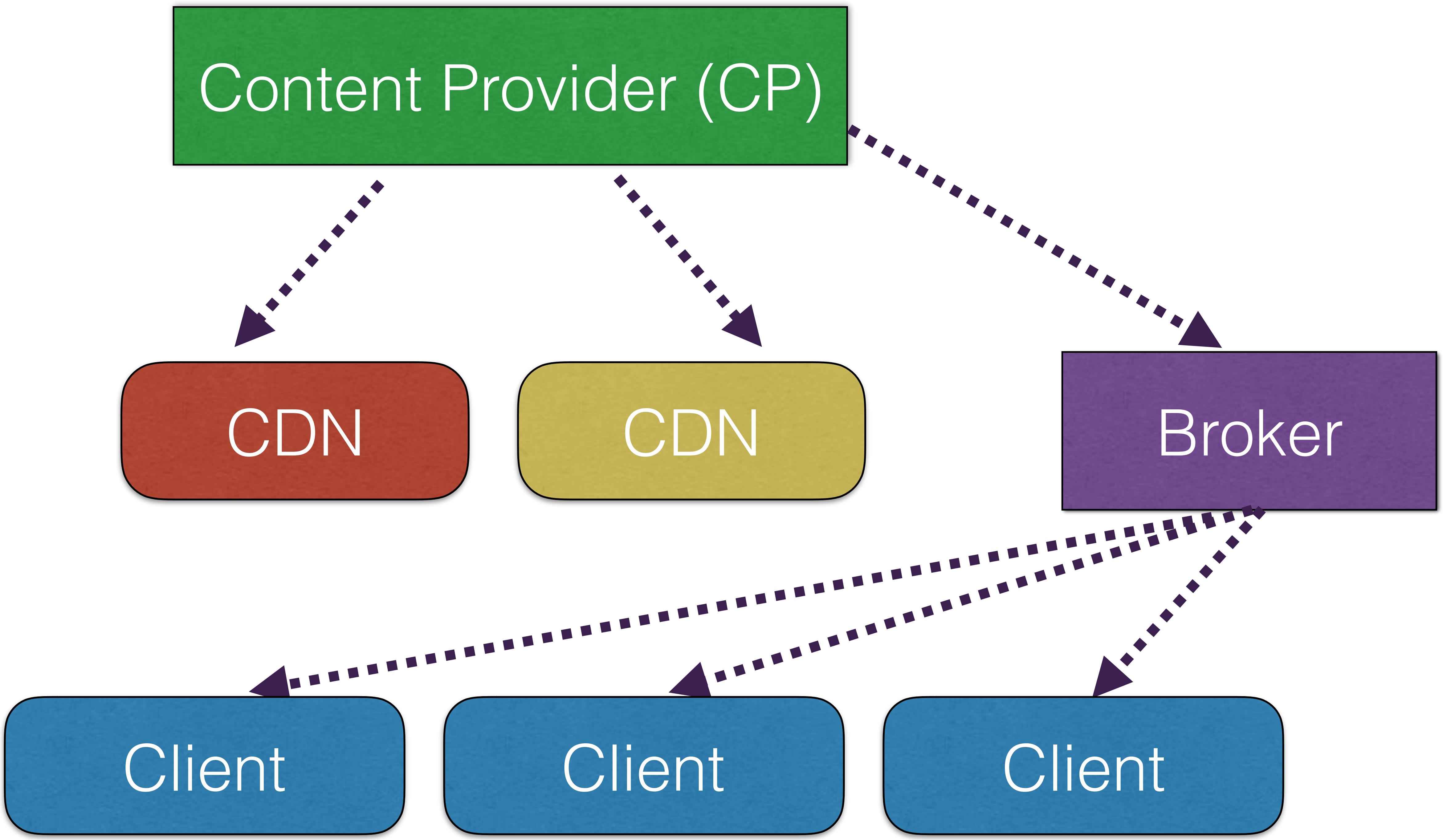
Contributions

- Identify challenges that brokers and CDNs create for each other by analyzing data from both
- Examine the design space of CDN-broker interfaces
- Evaluate the efficacy of different designs

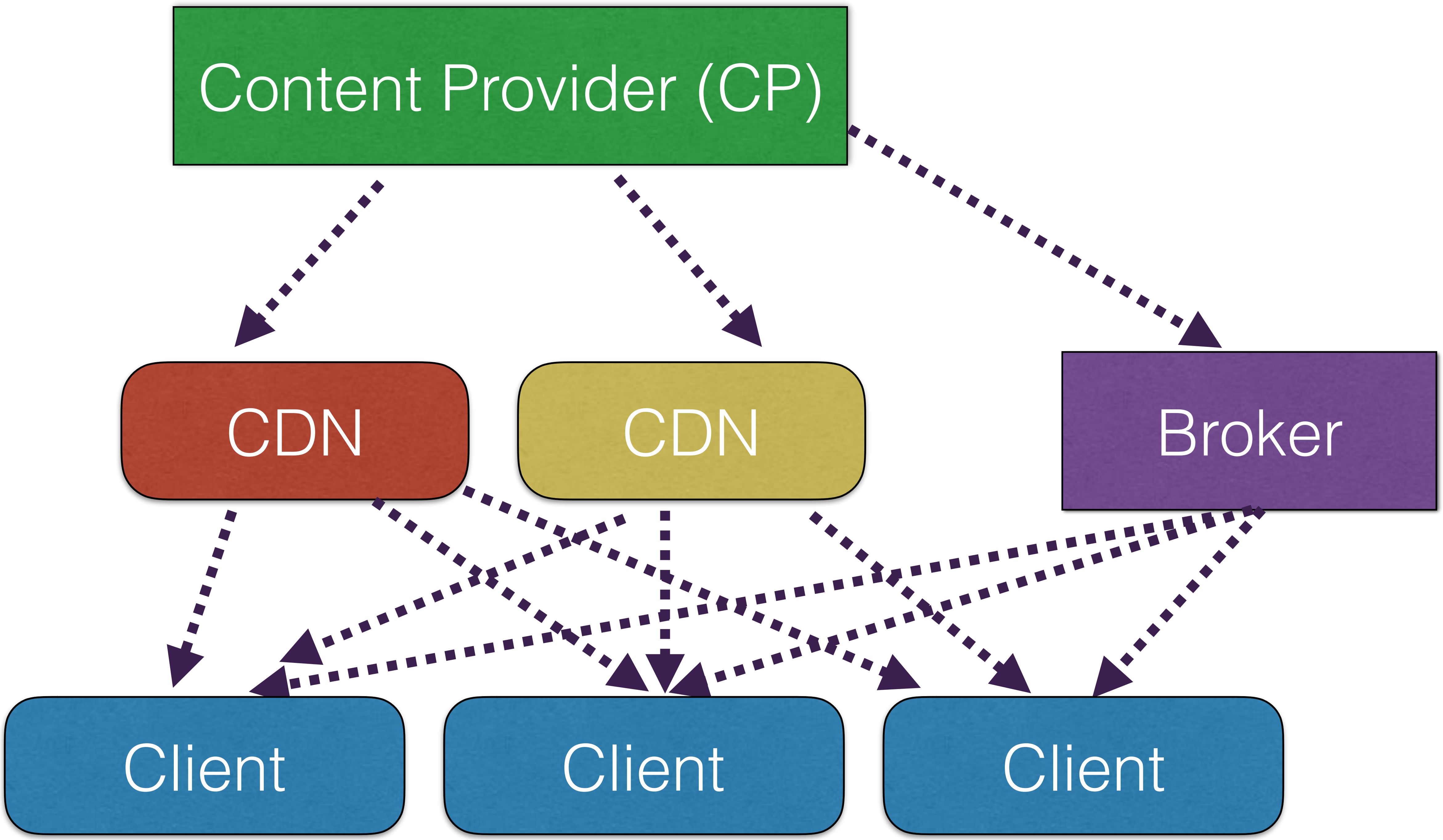
Brokered Delivery Today



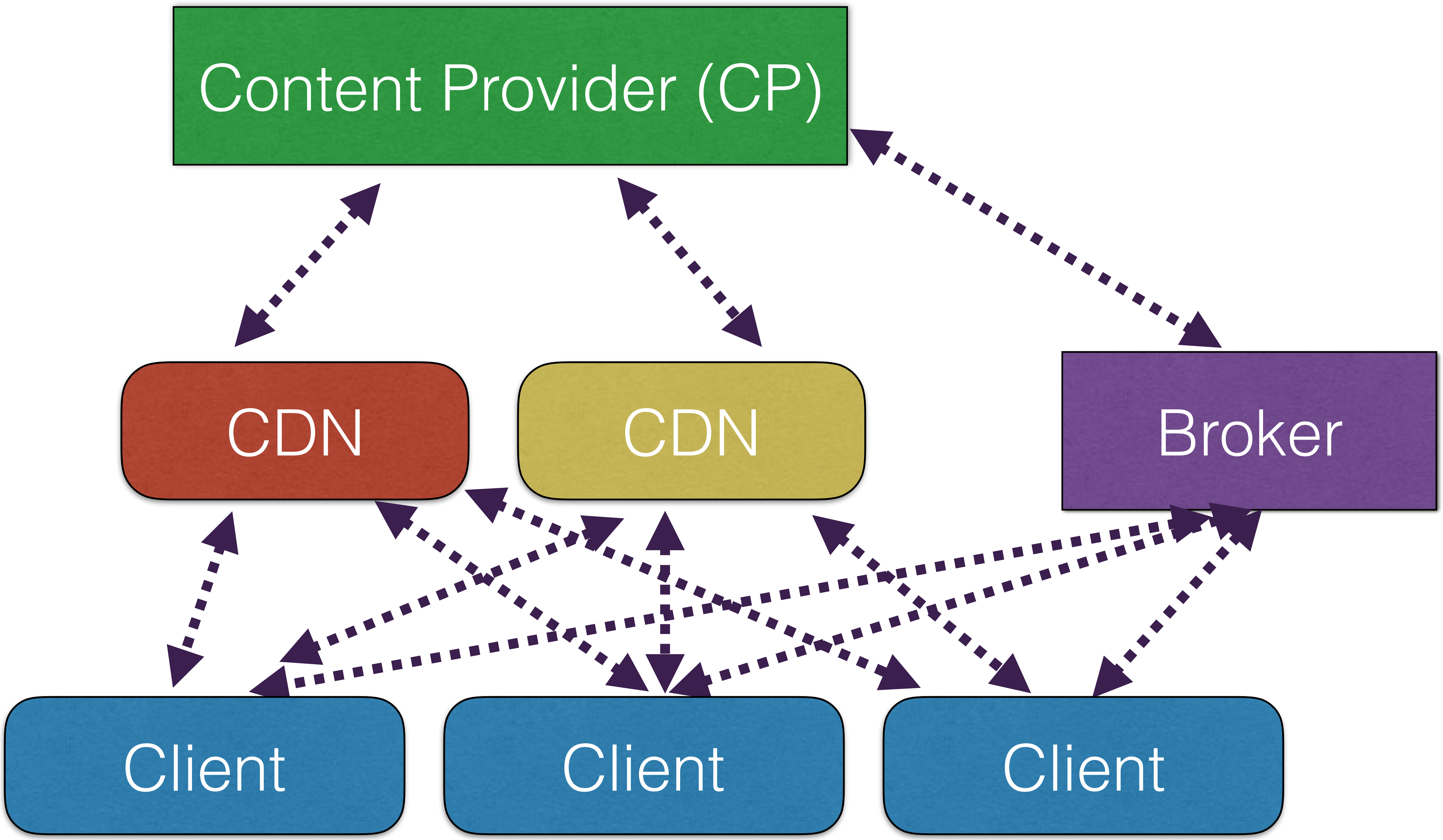
Brokered Delivery Today



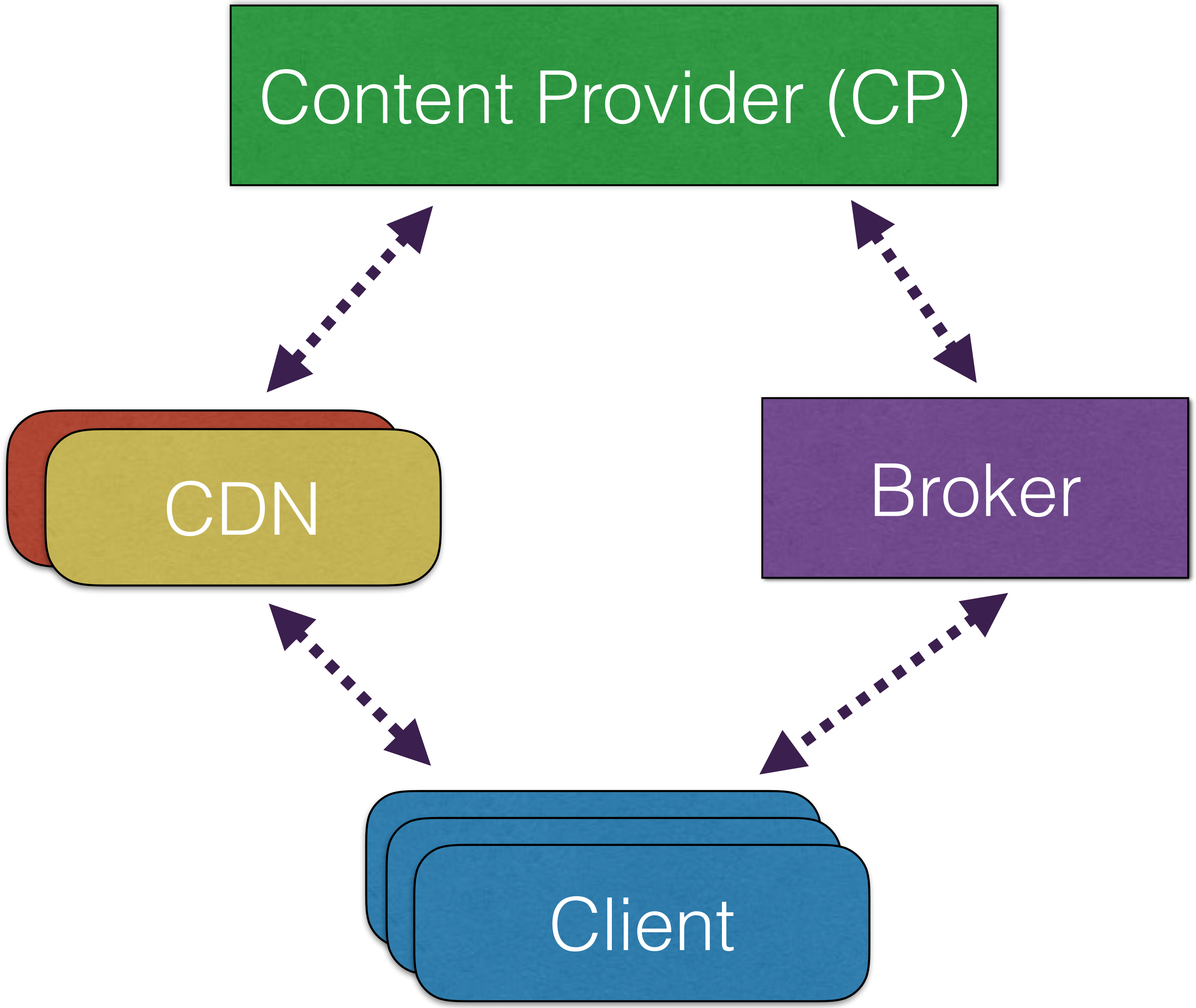
Brokered Delivery Today



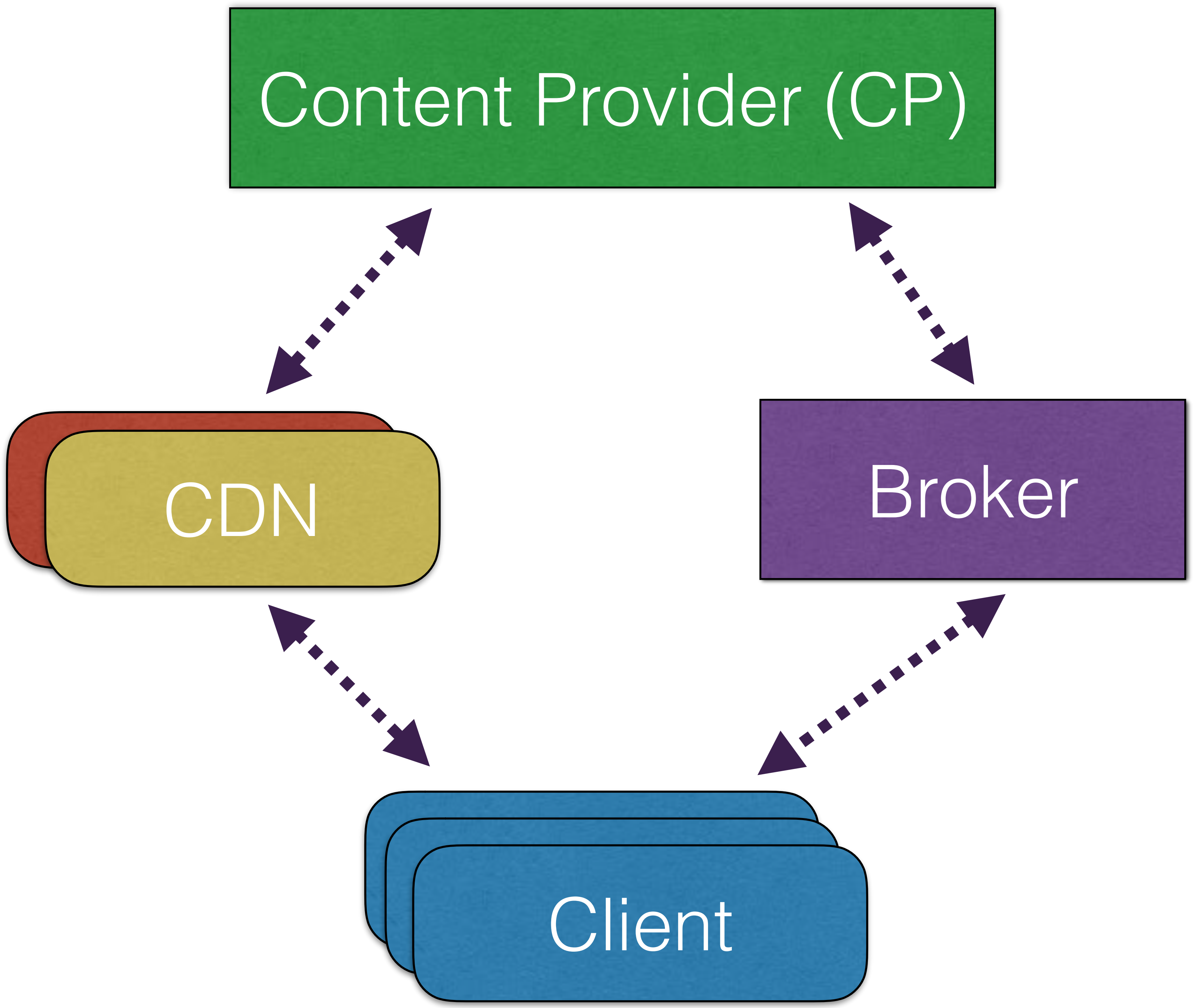
Brokered Delivery Today



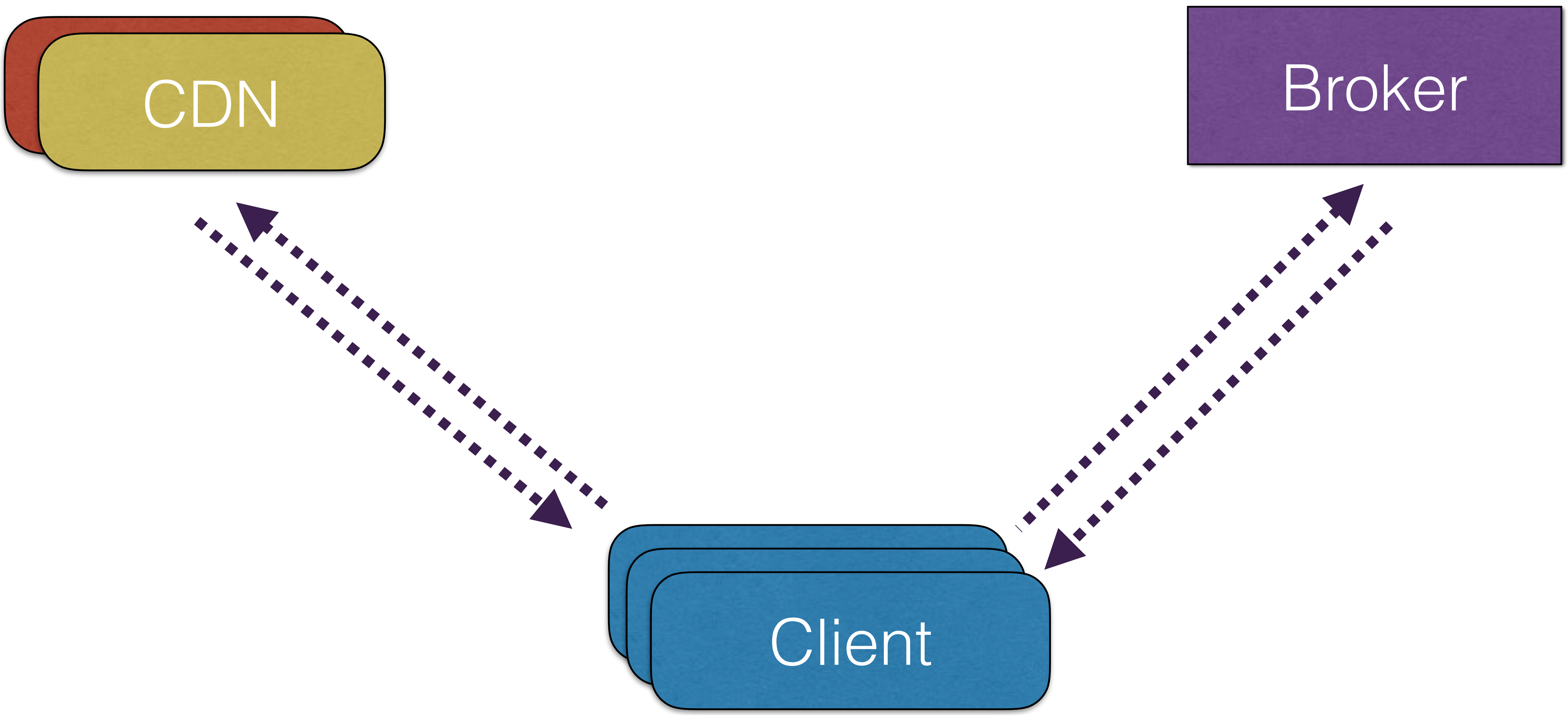
Brokered Delivery Today



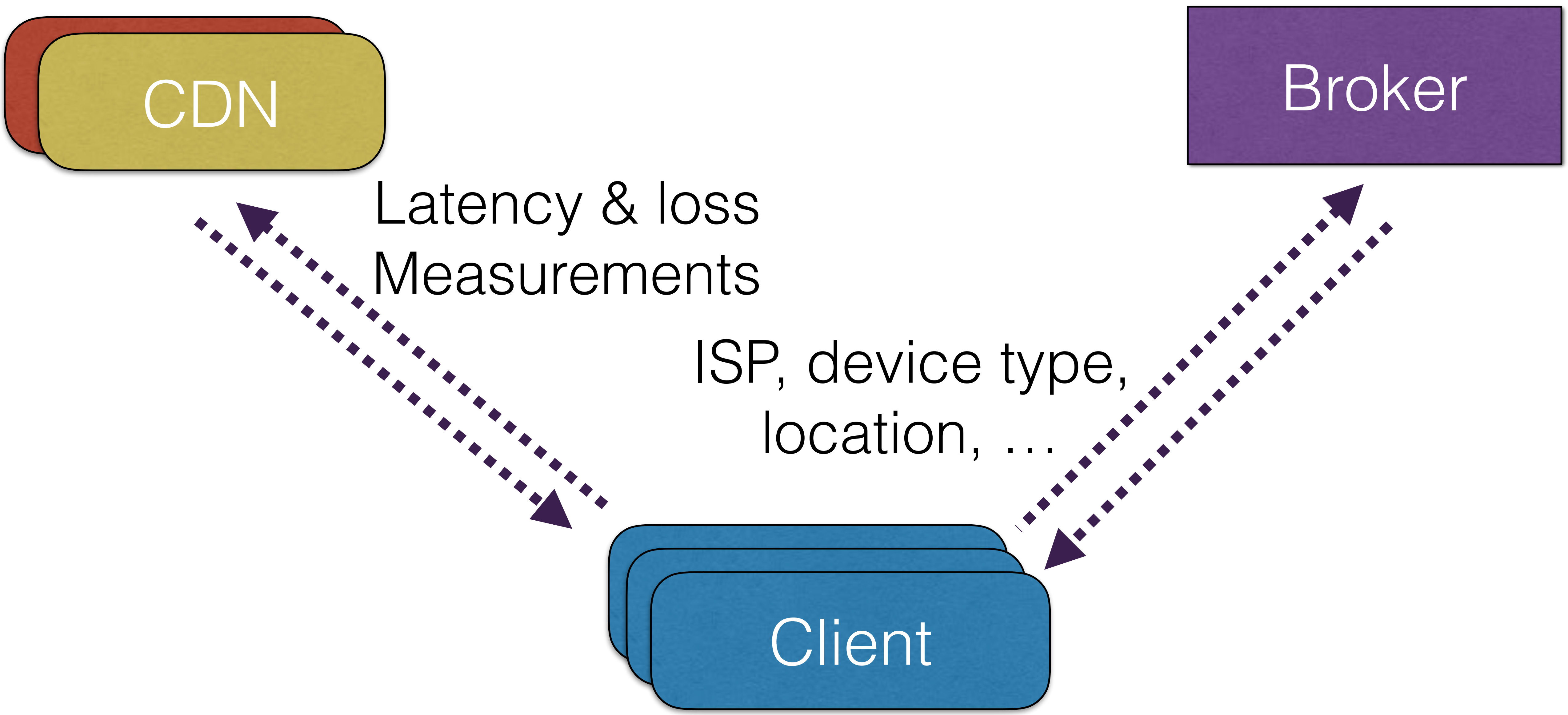
Brokered Delivery Today



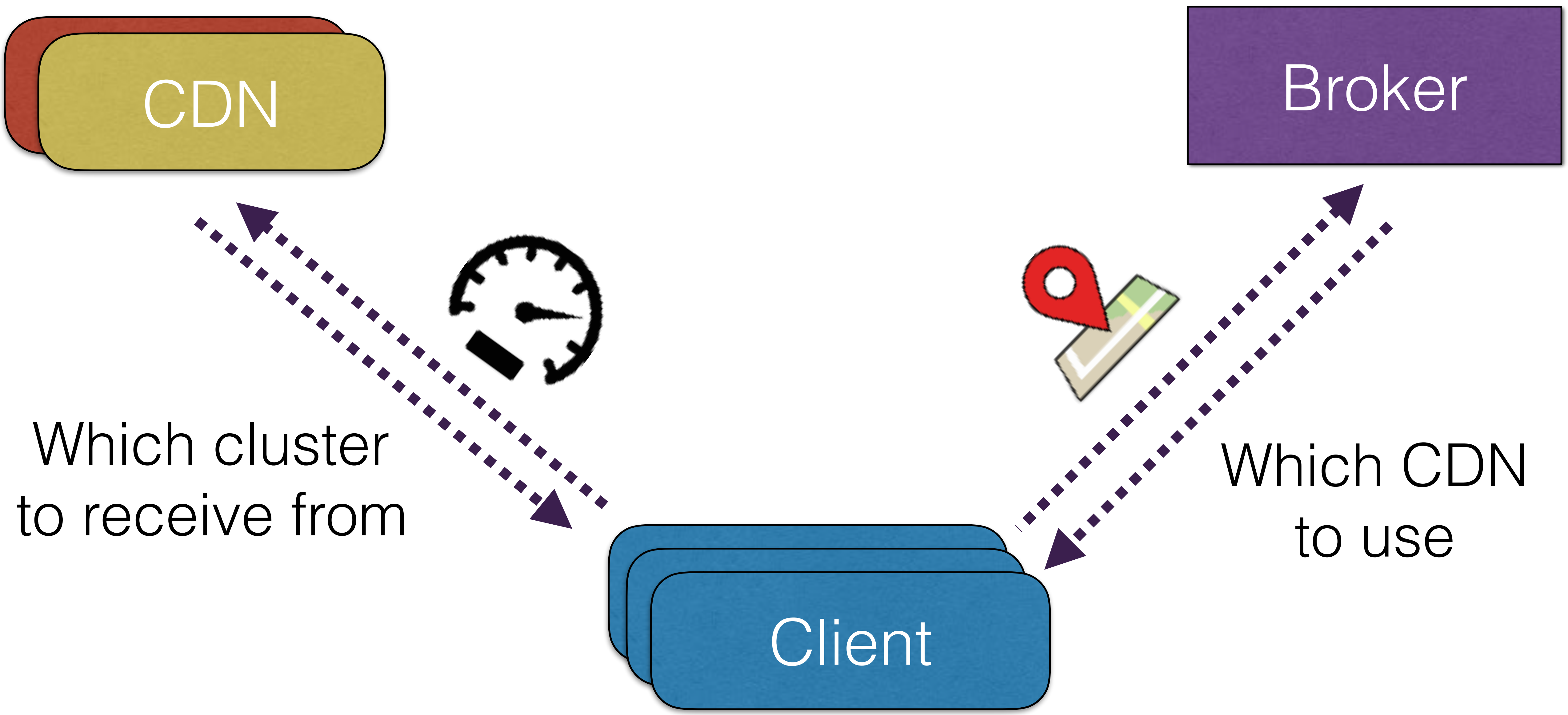
Brokered Delivery Today



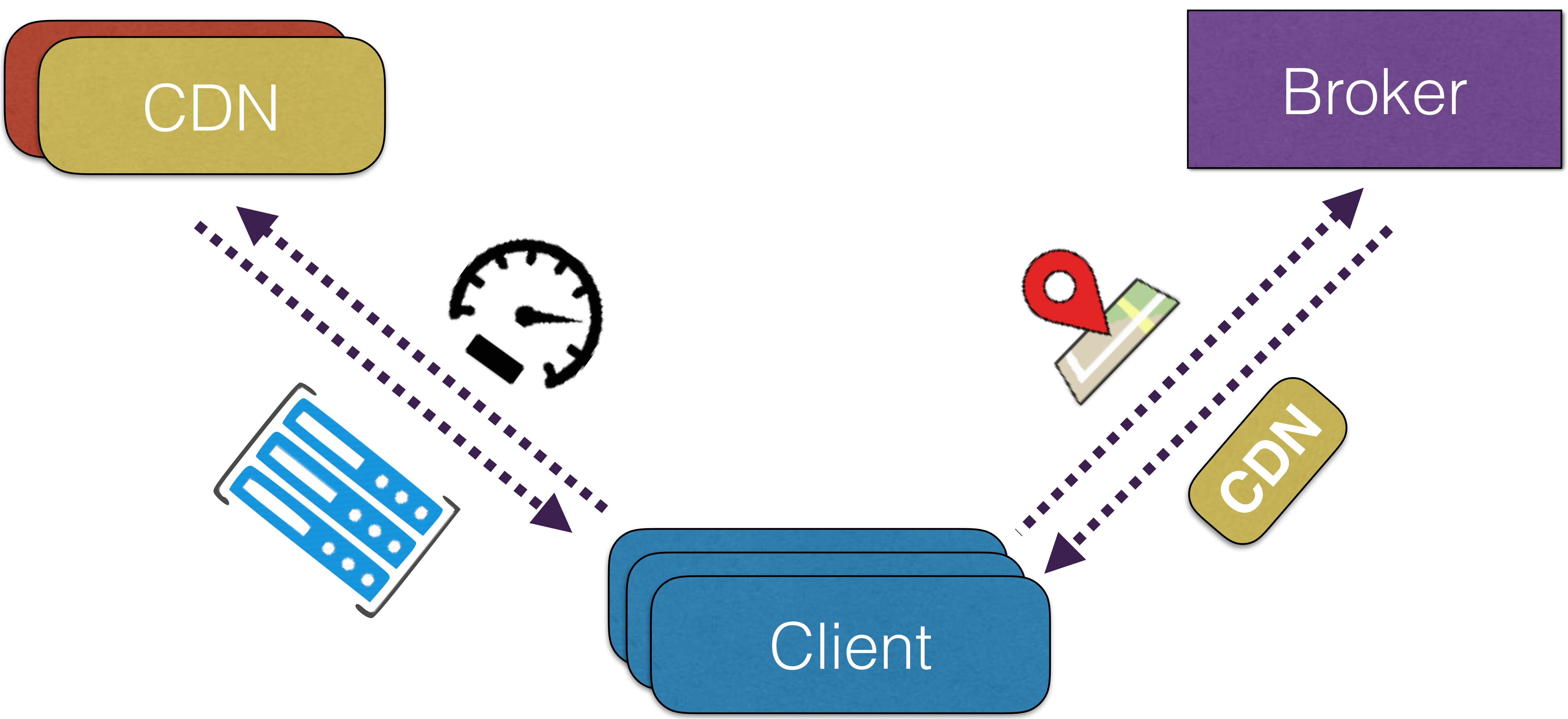
Brokered Delivery Today



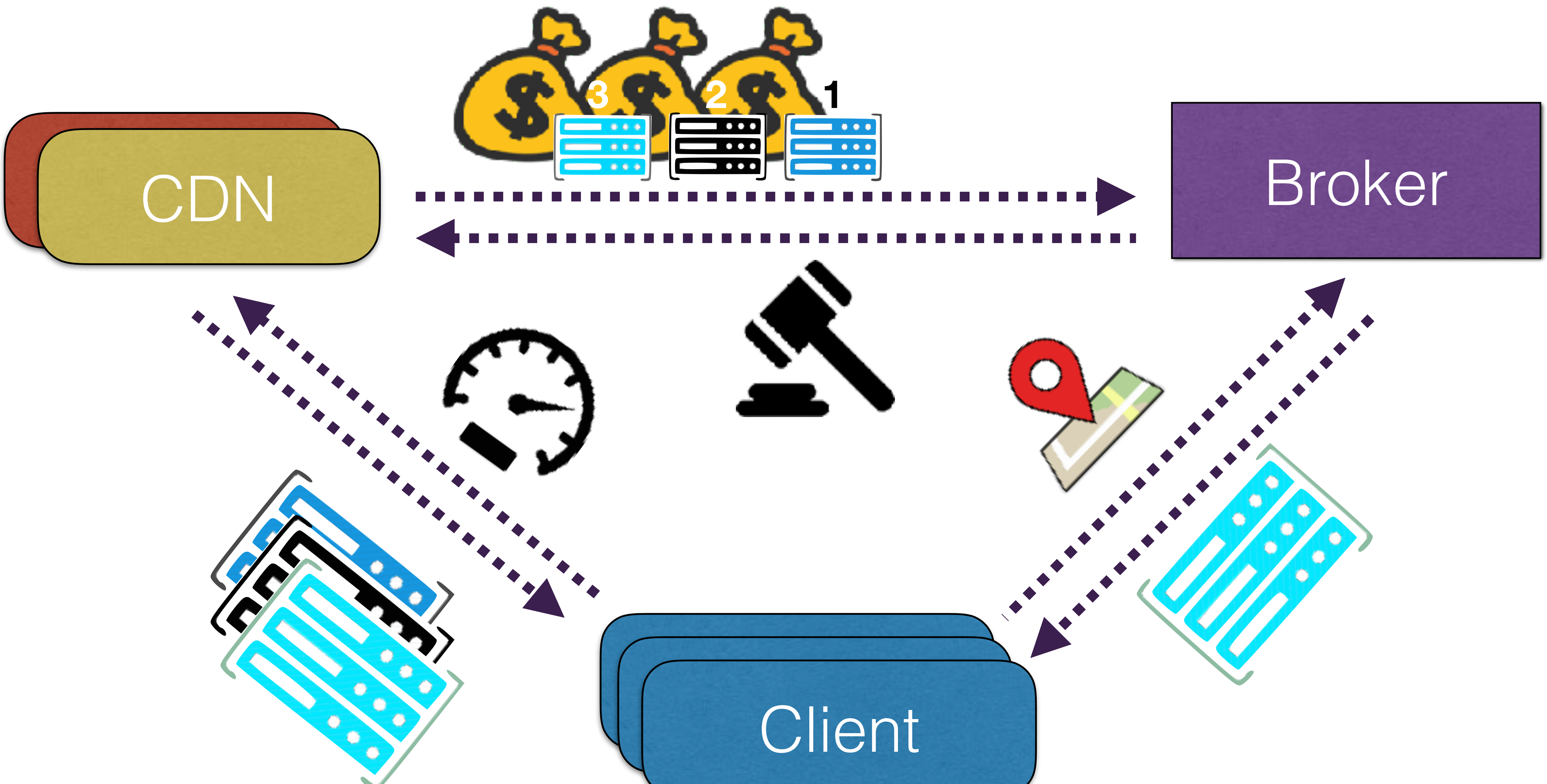
Brokered Delivery Today



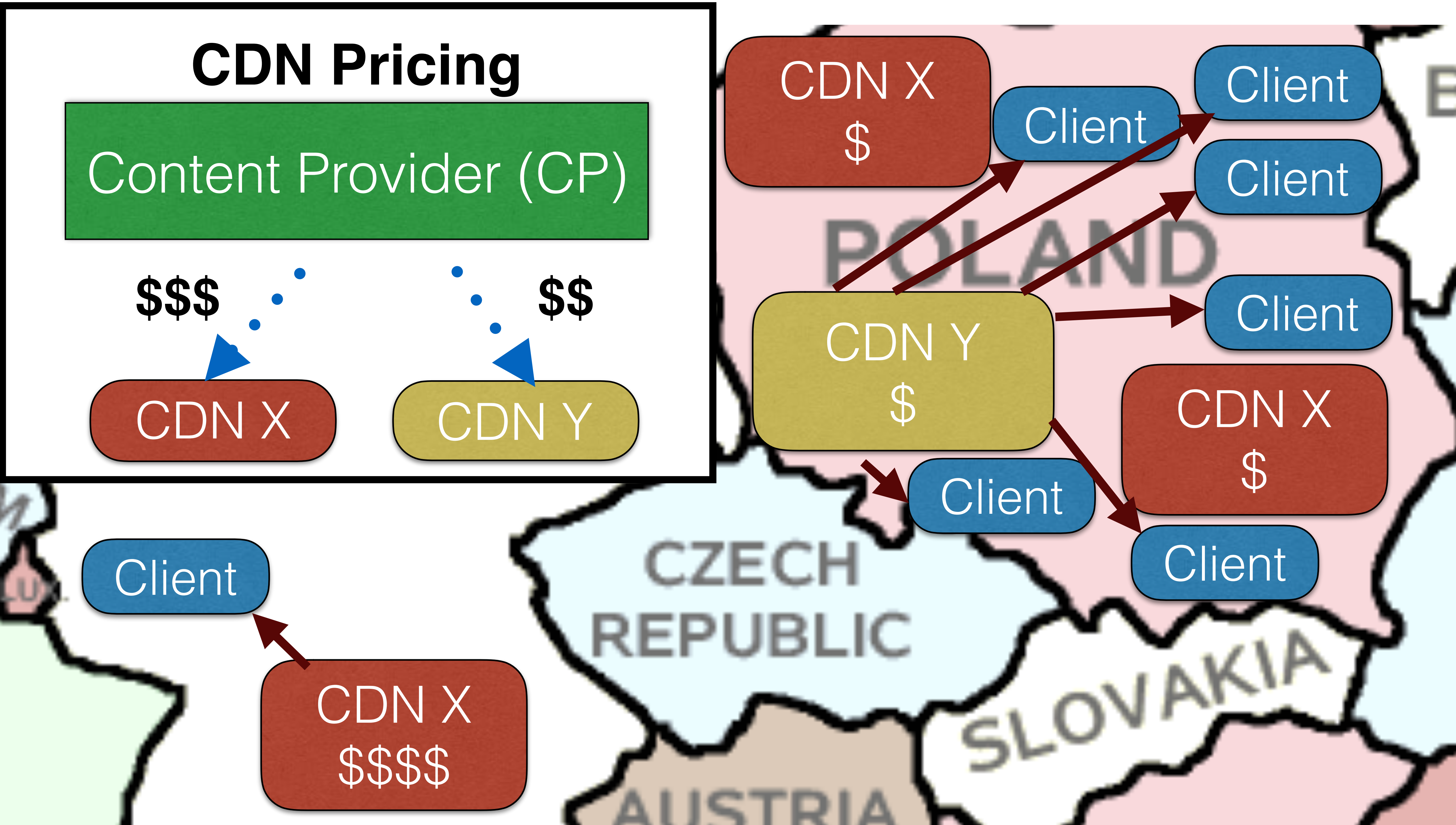
Brokered Delivery Today



VDX

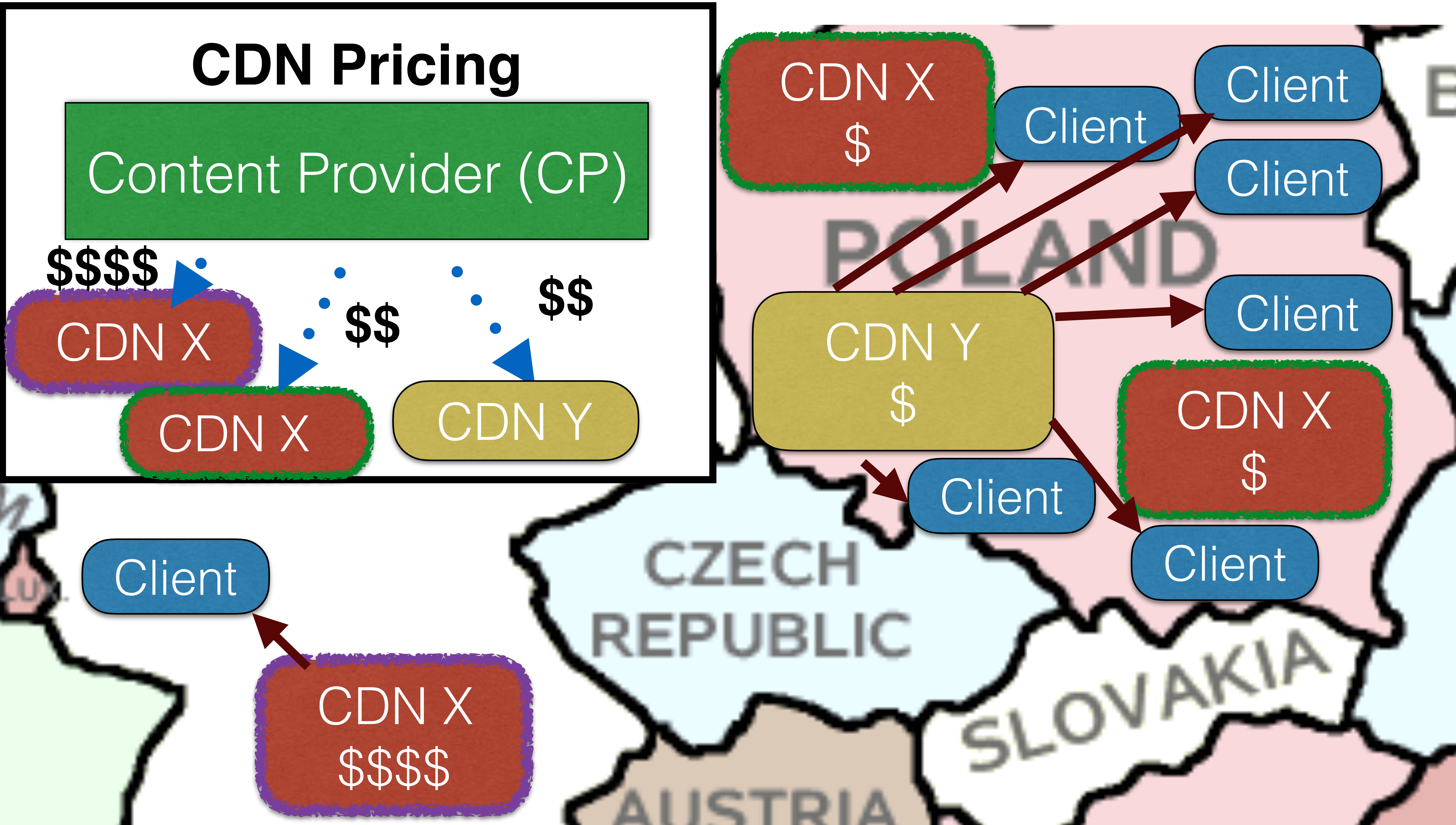


Example

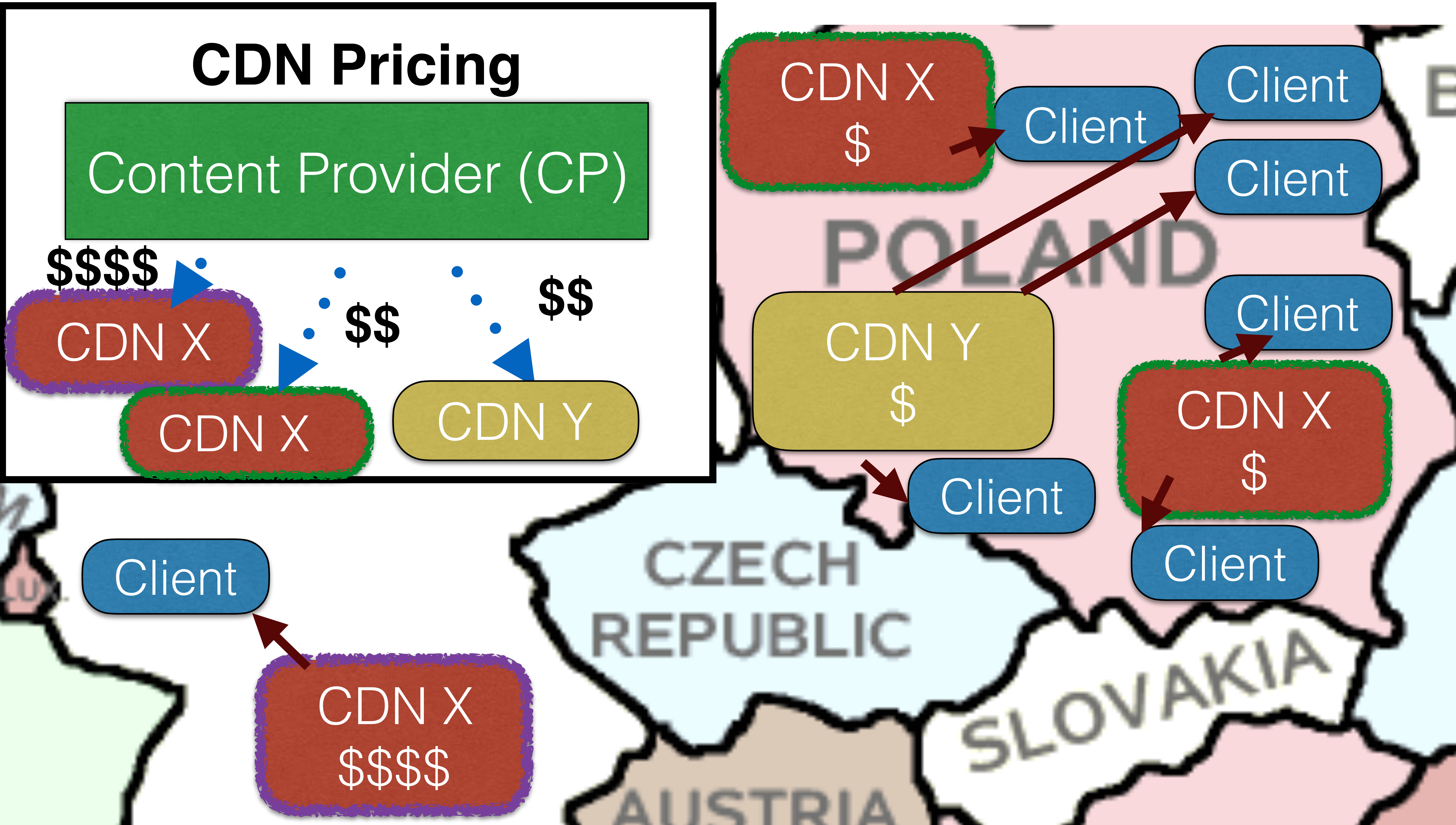


CDN Pricing
Content Provider (CP)
$$$
$$
CDN X
CDN Y
CDN X
$
Client
Client
Client
POLAND
CDN Y
$
Client
CDN X
$
Client
Client
Client
CZECH
REPUBLIC
SLOVAKIA
AUSTRIA
CDN X
$$$$

Example

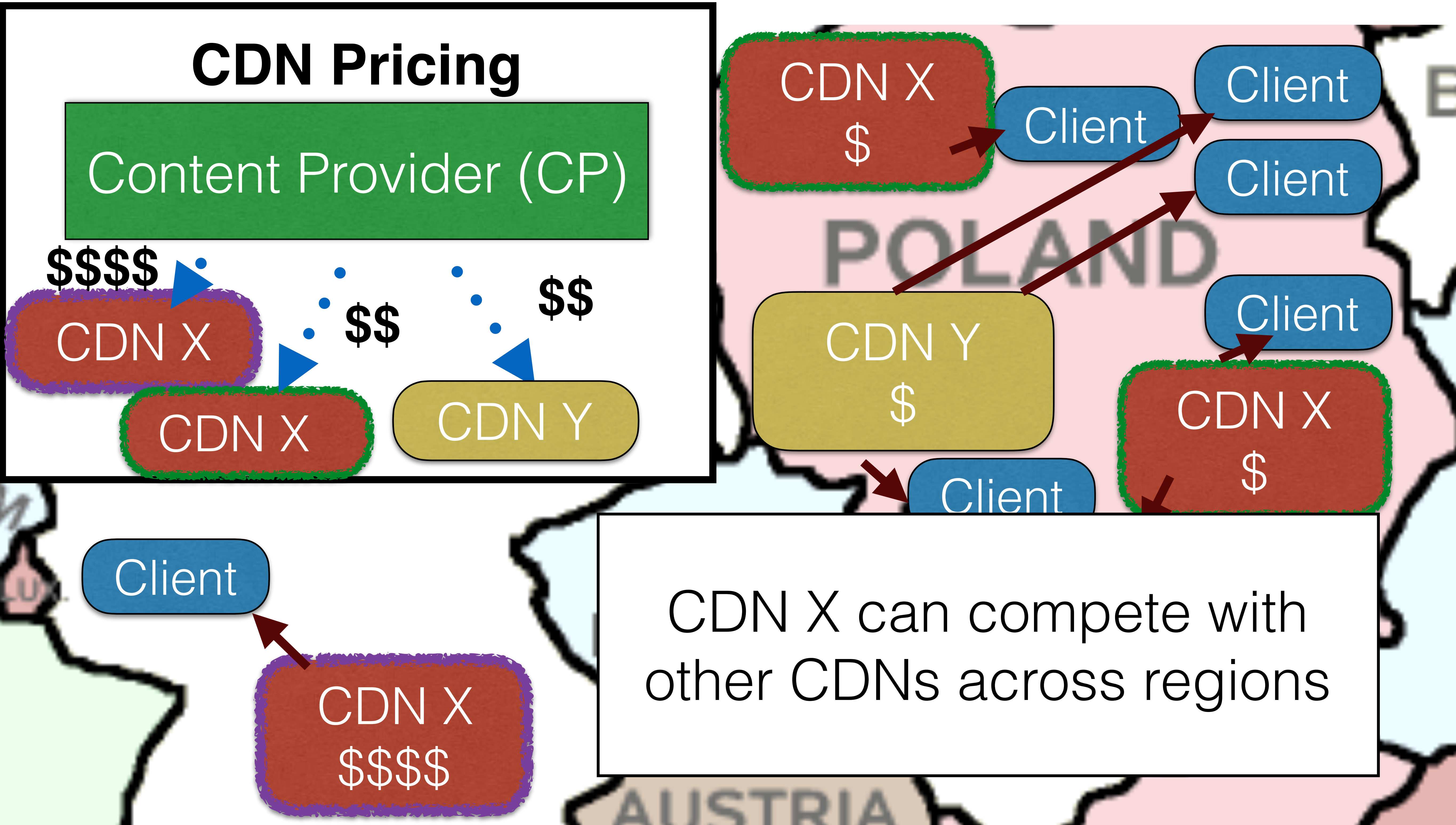

CDN Pricing
Content Provider (CP)
$$$$
$$
$$
CDN X
CDN X
CDN Y
CDN X
$
Client
Client
Client
POLAND
CDN Y
$
Client
CDN X
$
Client
Client
Client
CZECH
REPUBLIC
SLOVAKIA
AUSTRIA
CDN X
$$$$

Example



CDN Pricing
Content Provider (CP)
$$$$
CDN X
$$
$$
CDN X
CDN Y
CDN X
$
Client
Client
Client
POLAND
CDN Y
$
Client
CDN X
$
Client
Client
Client
CZECH
REPUBLIC
SLOVAKIA
AUSTRIA
CDN X
$$$$

Example

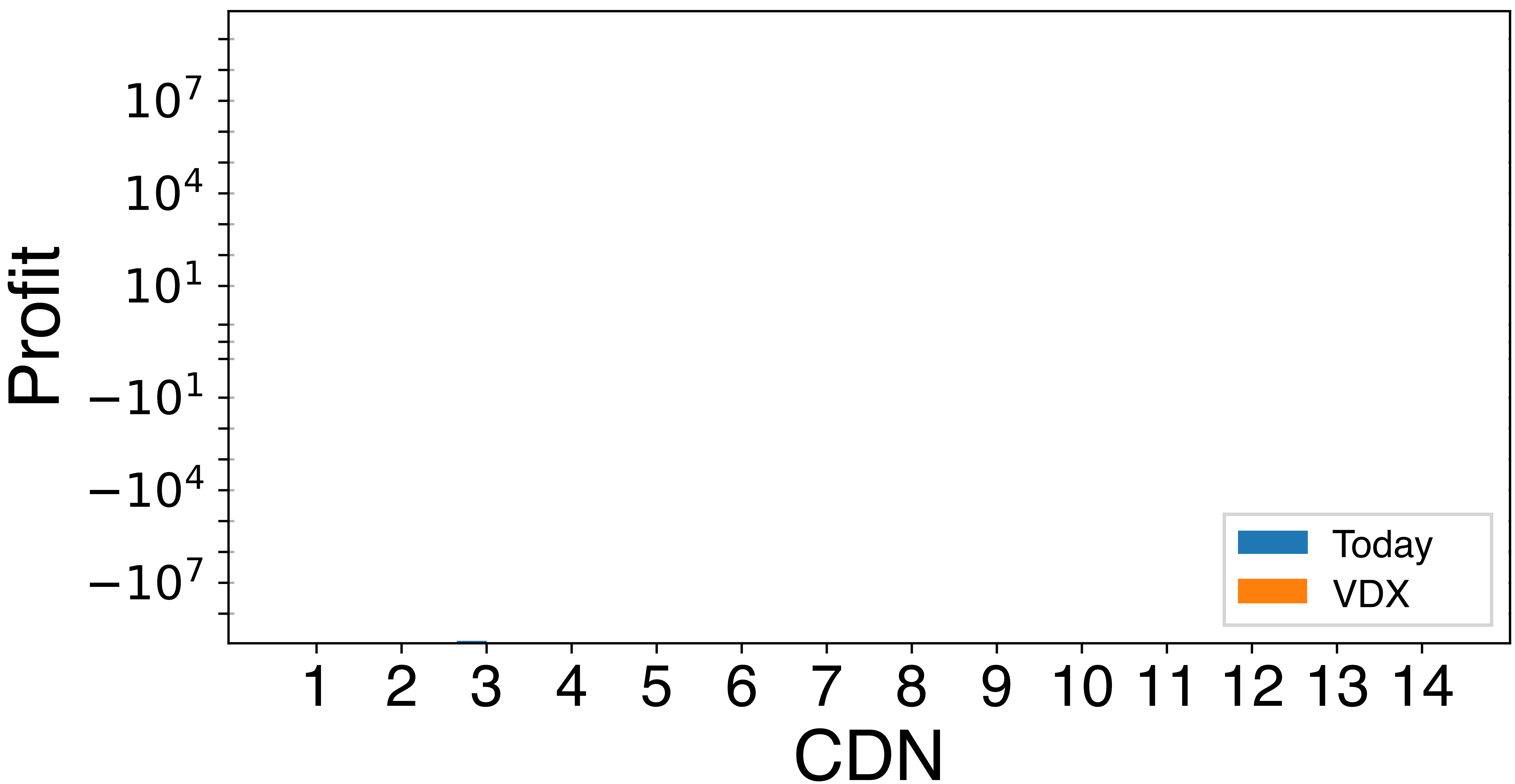


CDN X can compete with other CDNs across regions

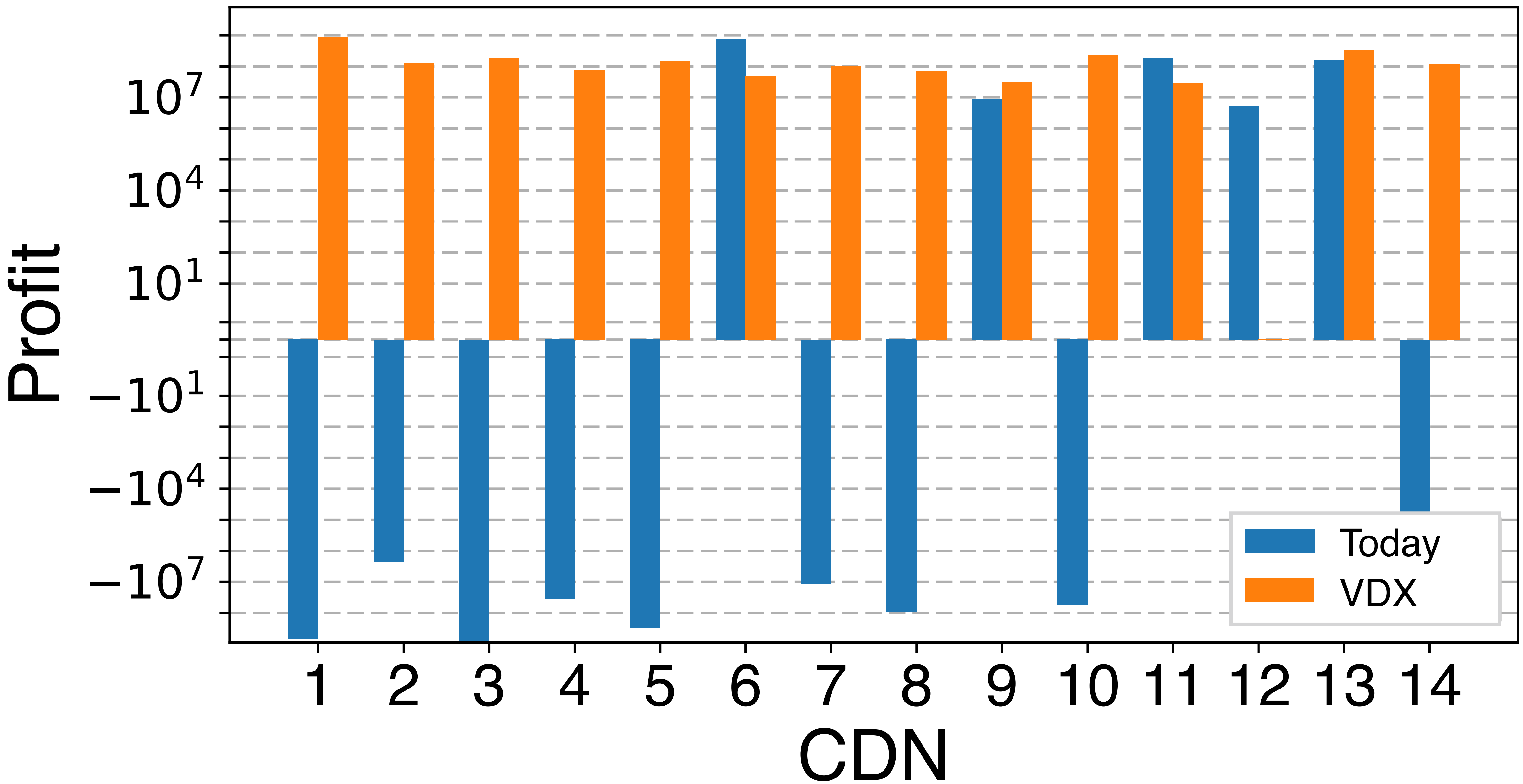
Evaluation

- Simulator using data from a broker & CDN, as well as public data from 13 other CDNs
- CDN data provides cluster locations, cluster-to-client performance, delivery costs, etc.
- Broker data provides client locations, request distributions, etc.

Per-CDN Profits



Per-CDN Profits



Evaluation Takeaways

- Today's world (Brokered) is pretty broken (performance can be better; most CDNs lose money on brokered video delivery)

- Marketplace (VDX) fixes this by exposing clusters and cost

Control Coordination

Scenario:
Admin

VDX

Scenario:
Scalability

VDN

App TE + ISP TE

Reaction

BGP + BGP

Priority Ranking

Internet-scale Routing

Hierarchical Partitioning

Coflow

Etalon

Transparency

Scenario:
Layering

Control Coordination

Scenario:
Admin

VDX

Scenario:
Scalability

VDN

App TE + ISP TE

Reaction

BGP + BGP

Priority Ranking

Internet-scale Routing

Hierarchical Partitioning

Coflow

Etalon

Transparency

Scenario:
Layering

Control Coordination

Scenario: **Scalability**

VDN

App TE + ISP TE

Reaction

Internet-scale Routing

Hierarchical Partitioning

BGP + BGP

VDX

Priority Ranking

Scenario: **Admin**

Coflow

Etalon

Transparency

Scenario: **Layering**

Some

Full

Information Sharing

Control Coordination

Scenario:
Scalability

VDN

App TE + ISP TE
Reaction

Internet-scale Routing
Hierarchical Partitioning

BGP + BGP
VDX
Priority Ranking

Scenario:
Admin

Coflow
Etalon
Transparency

Scenario:
Layering

Some
Full

Information Sharing

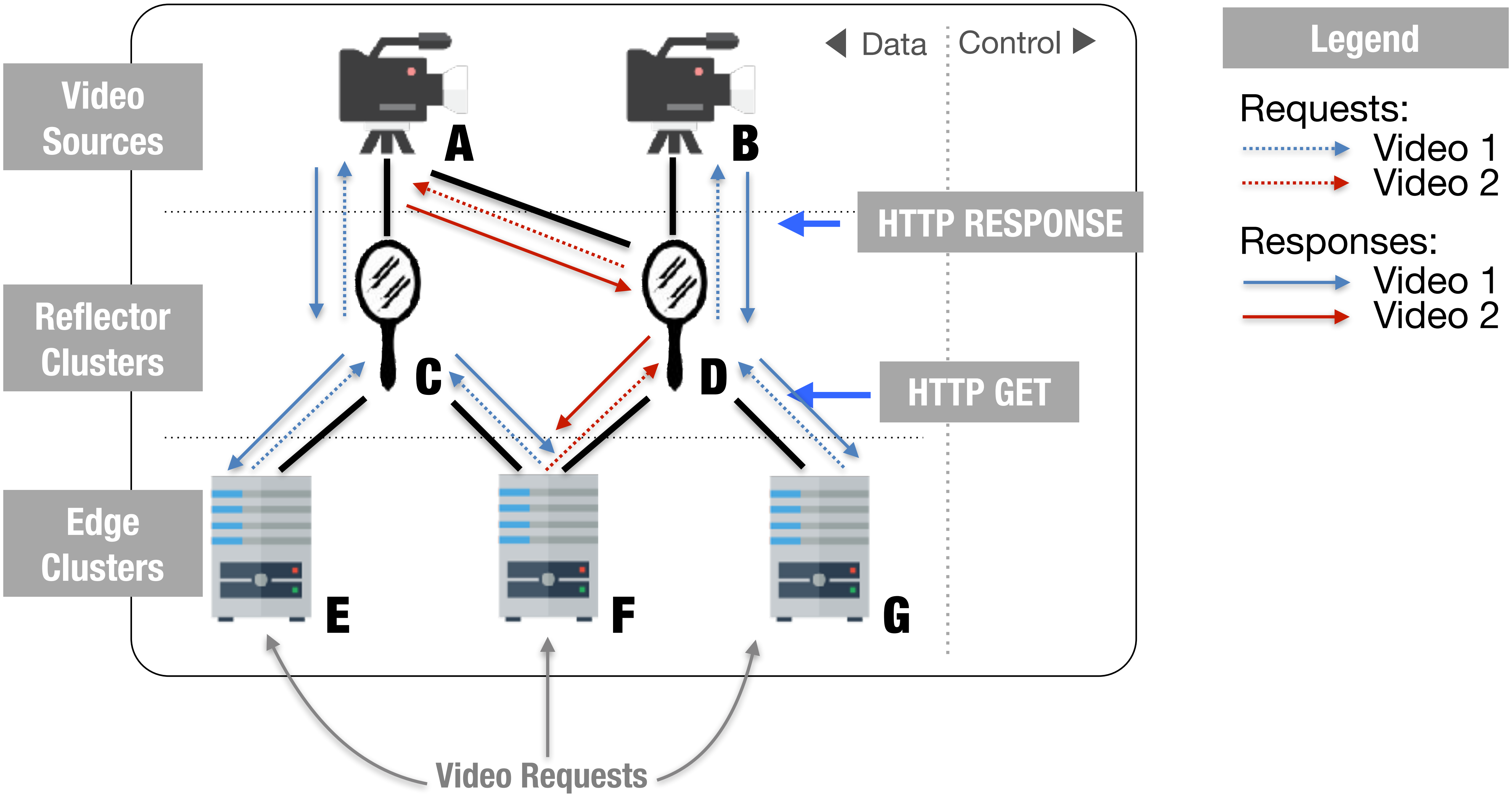
Live Video is Becoming Wildly Popular

- Commercial sports streams
- User-generated streams

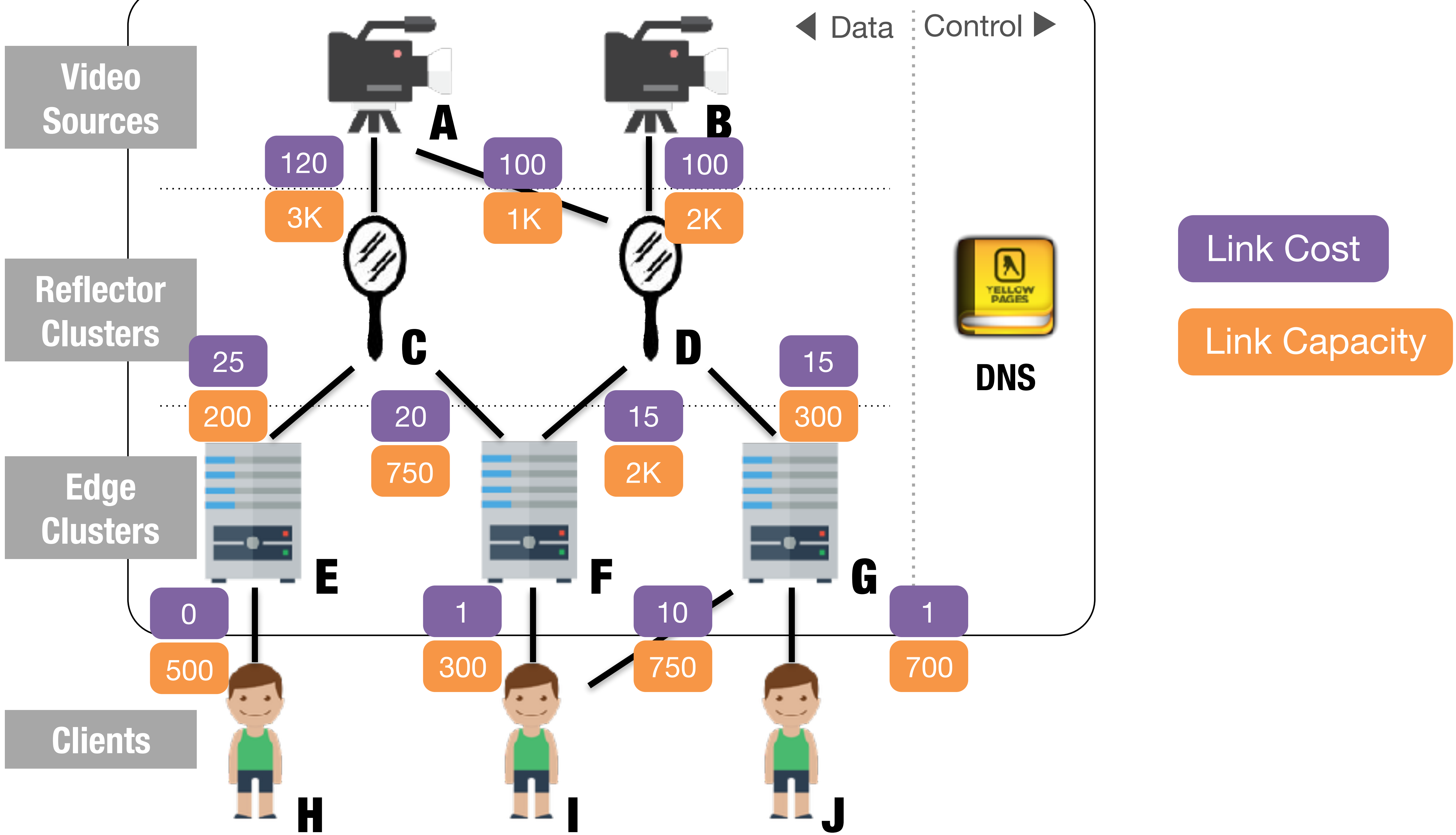
Live Video is Becoming Wildly Popular

- Commercial sports streams
  - **Single** World Cup stream = **40%** global Internet traffic
- User-generated streams (e.g., Twitch)
  - Users watch **150b min of live video per month**
  - Amazon buys Twitch for ~**$1Billion**

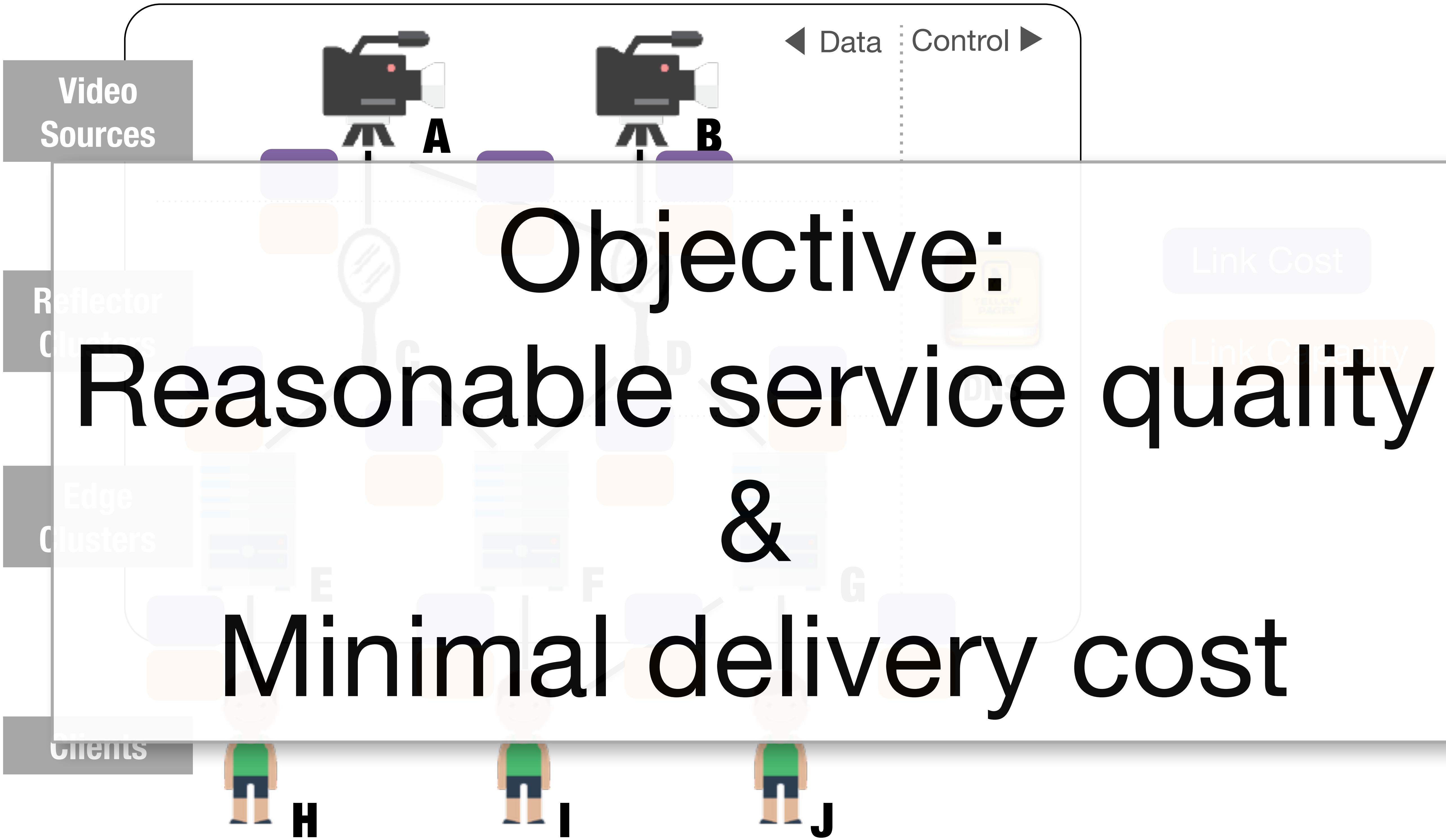
CDN Live Video Delivery Background



CDN Live Video Delivery Background



CDN Live Video Delivery Background



Objective:

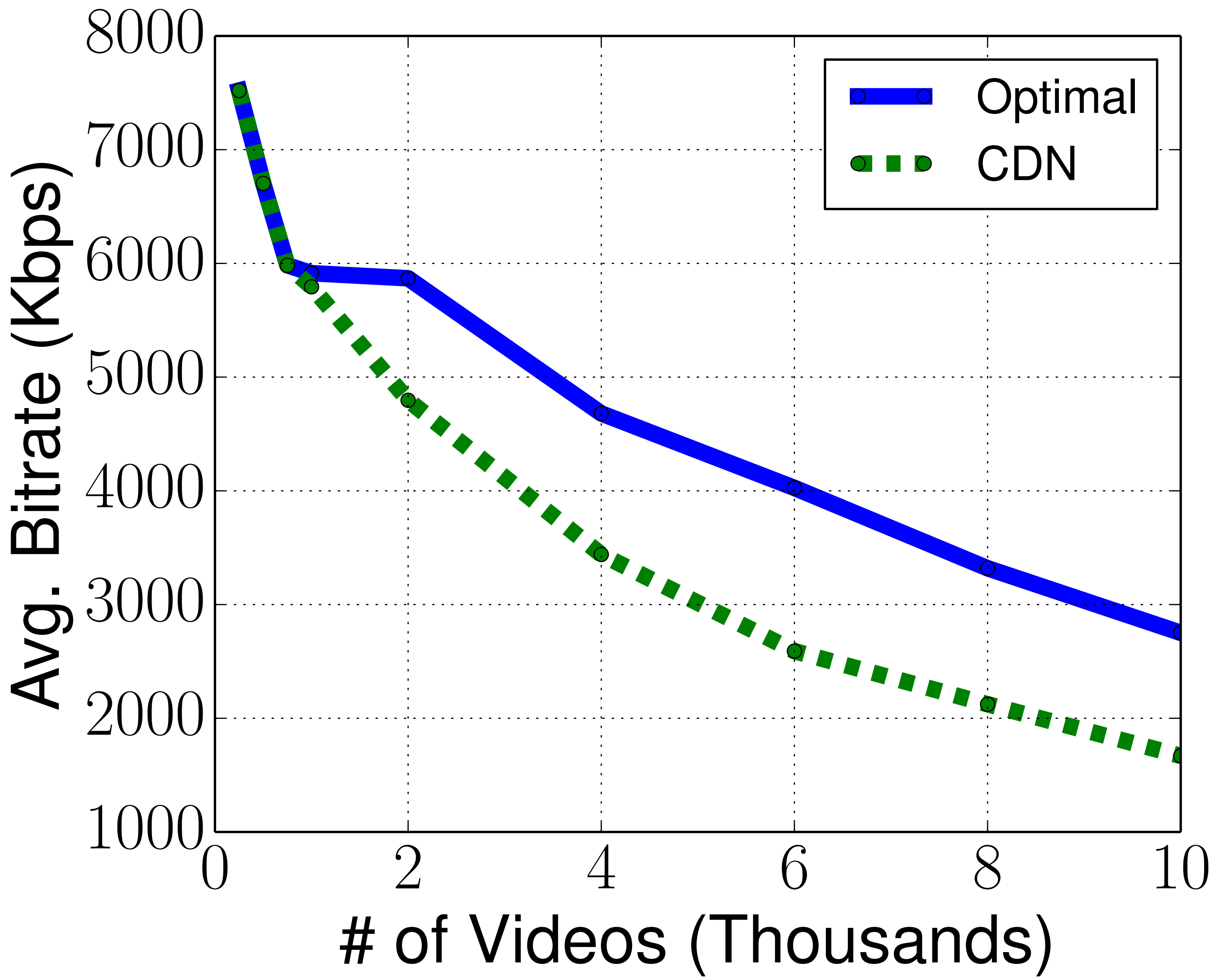
Reasonable service quality

&

Minimal delivery cost

Problems with CDNs Today

Service Quality



Simulation using Conviva traces, modeling user-generated content

Delivery Cost

(per request)

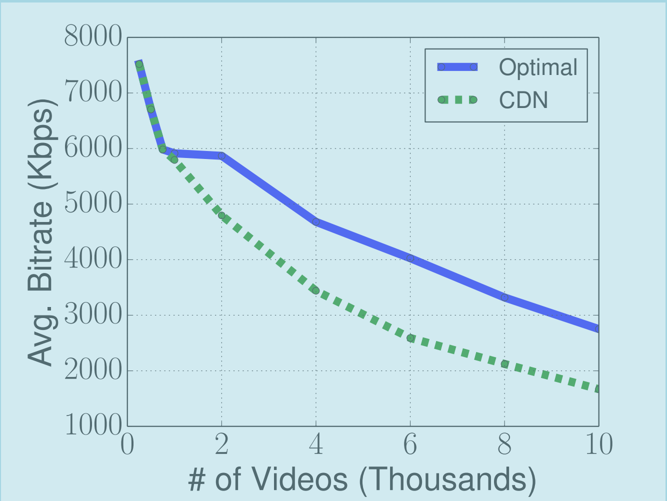
CDN
2.0x

OPTIMAL
1.0x

Simulation using Conviva traces, modeling large sports events

Problems with CDNs Today

Service Quality



Delivery Cost

CDN 2.0x

OPTIMAL 1.0x

QUANTITATIVE

Not Fine-Grained

Videos aggregated into large groups

Slow DNS Updates

Can't push updates

DNS entries get cached

QUALITATIVE

Solution?

Service Quality

Not Fine-Grained

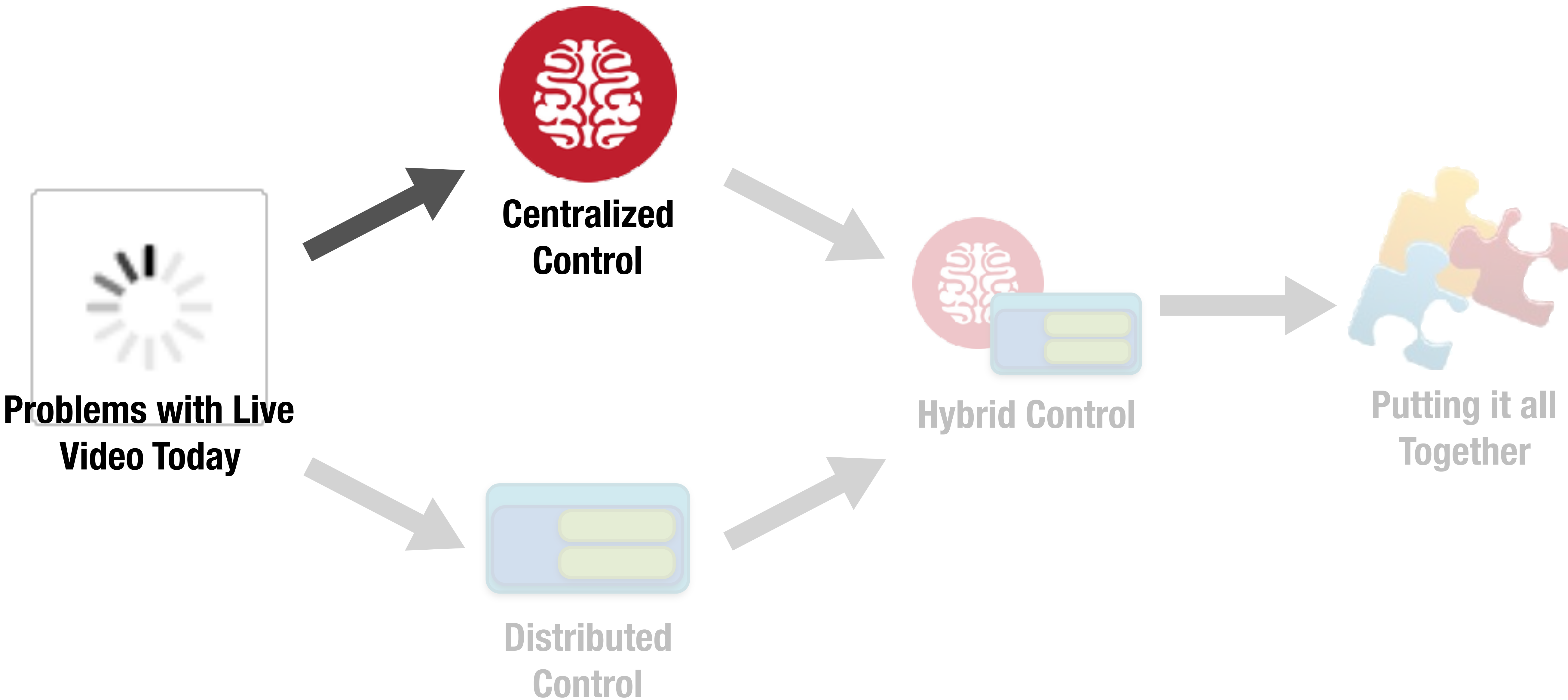
Centralization!

[Liu, Xi et. al. A Case for a Coordinated Video Control Plane. SIGCOMM 2012]

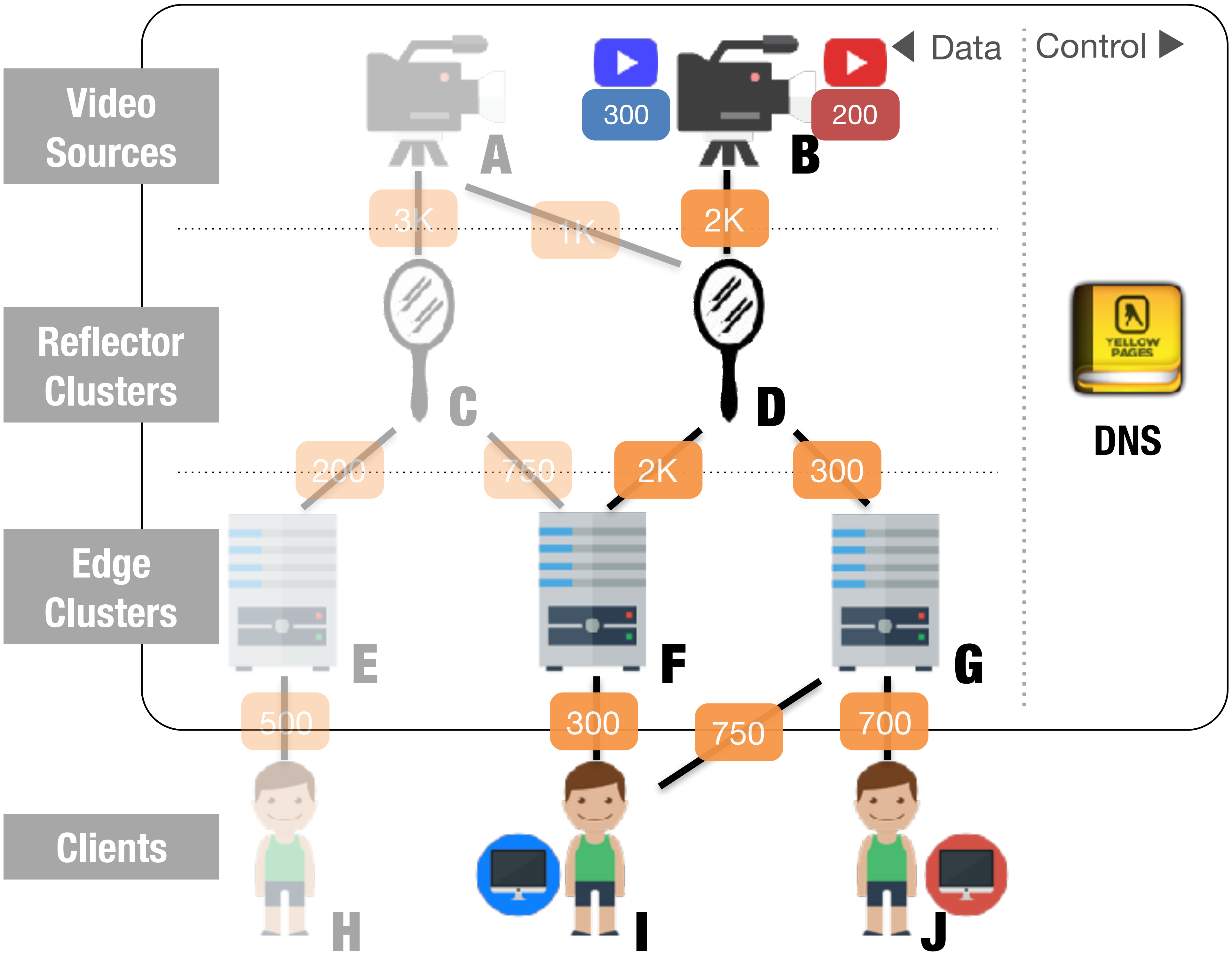
QUANTITATIVE

QUALITATIVE

Outline

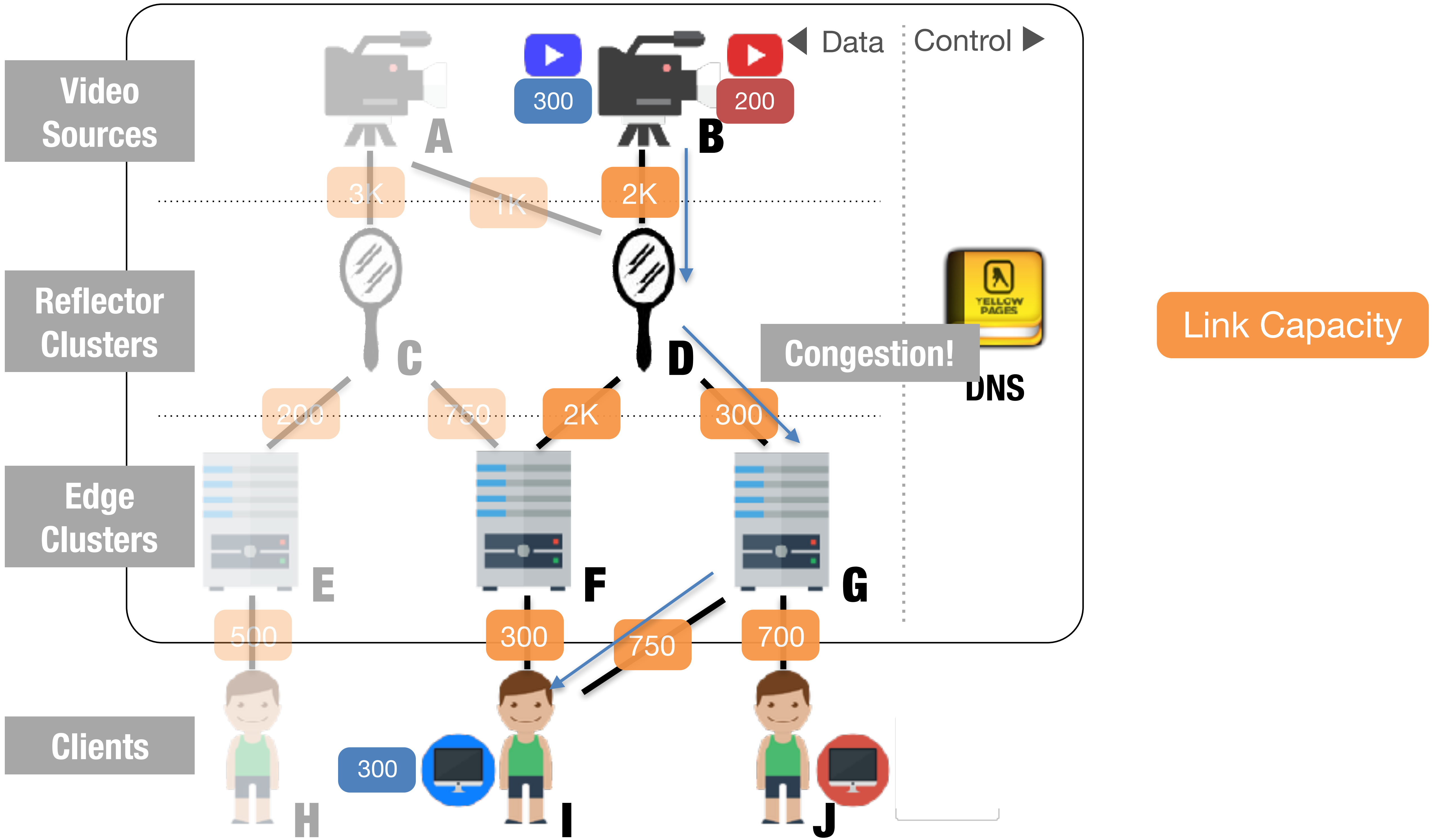

Problems with Live Video Today
Centralized Control
Distributed Control
Hybrid Control
Putting it all Together

Motivating Centralized Optimization

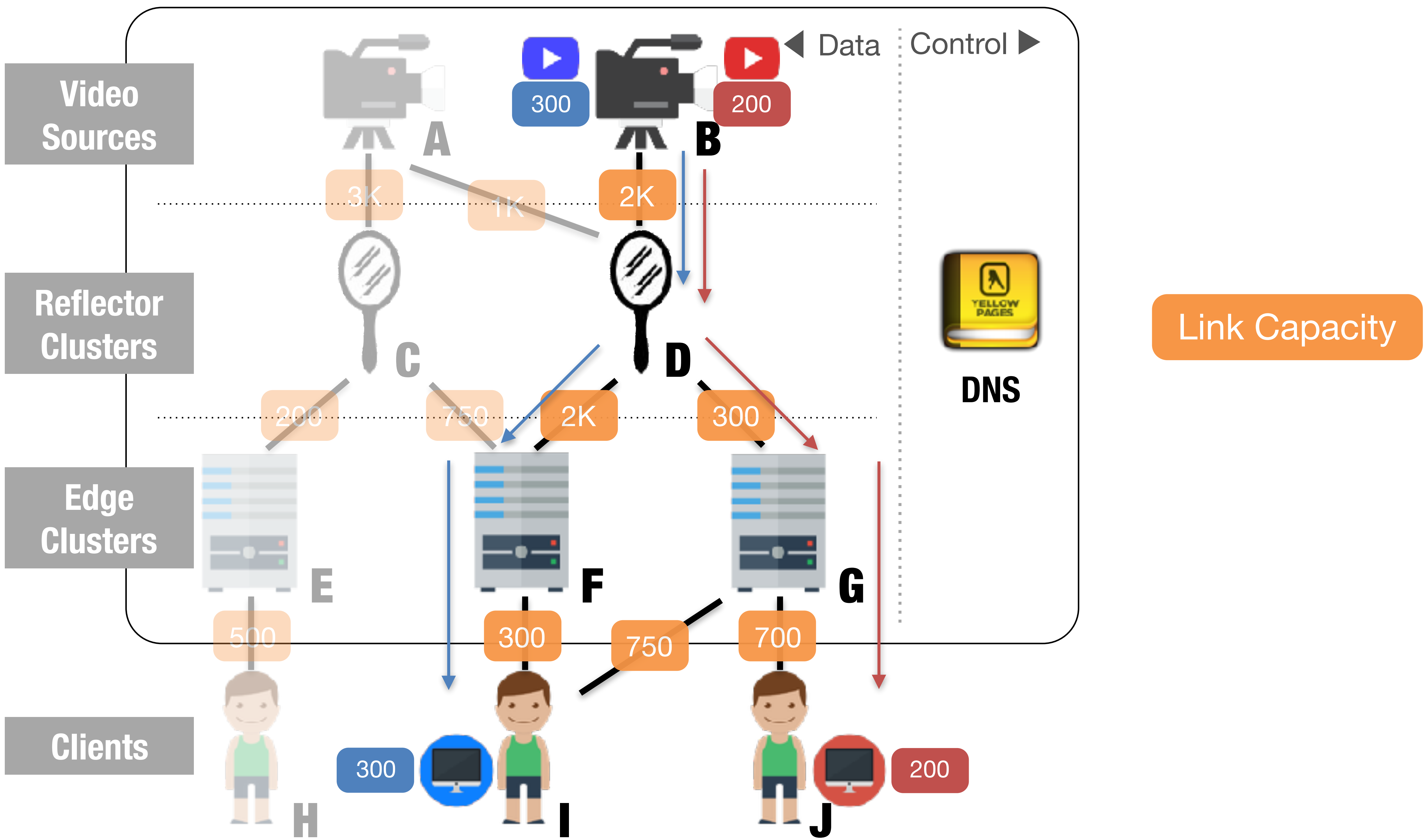


Link Capacity

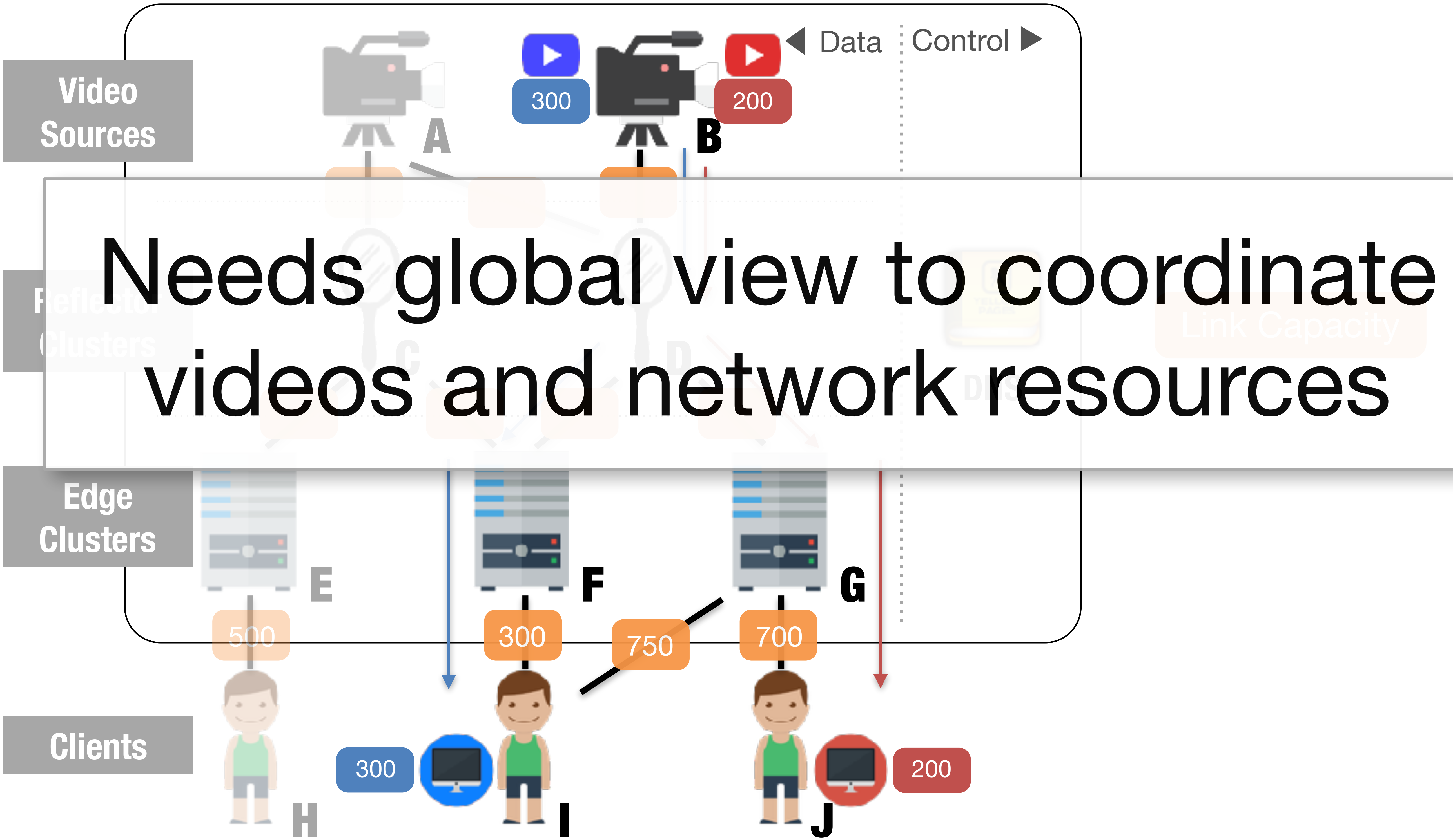
Motivating Centralized Optimization



Motivating Centralized Optimization

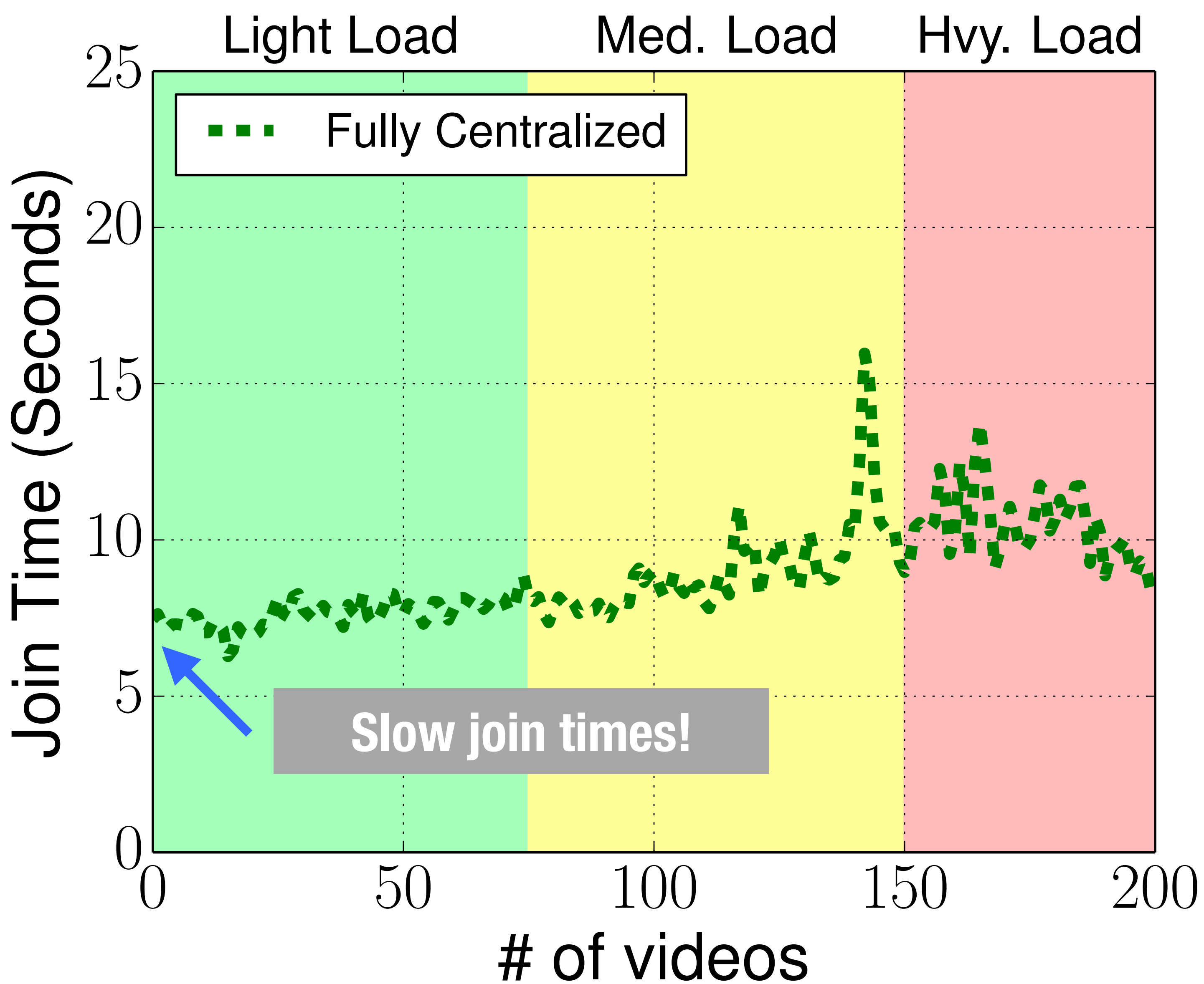


Motivating Centralized Optimization



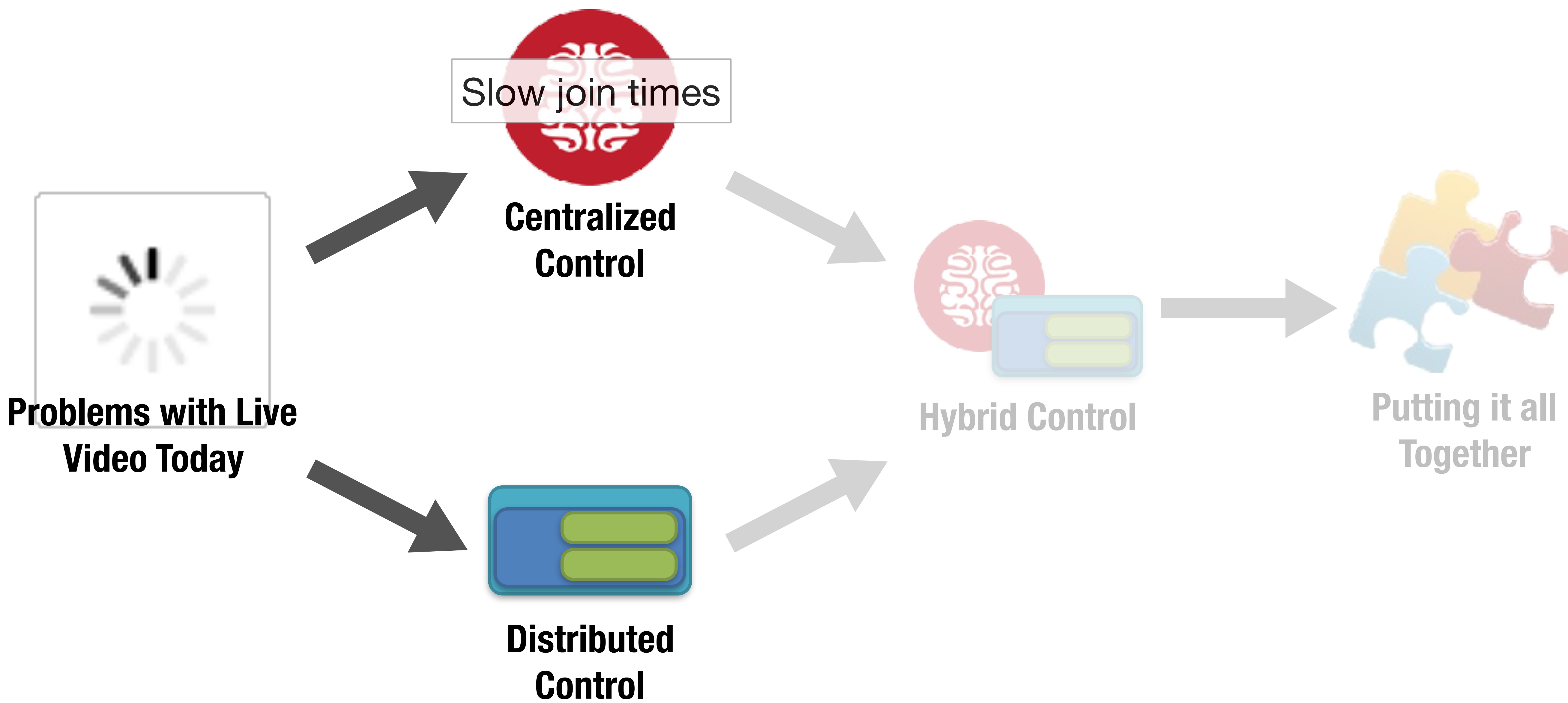
Needs global view to coordinate videos and network resources

Unfortunately… No Free Lunch

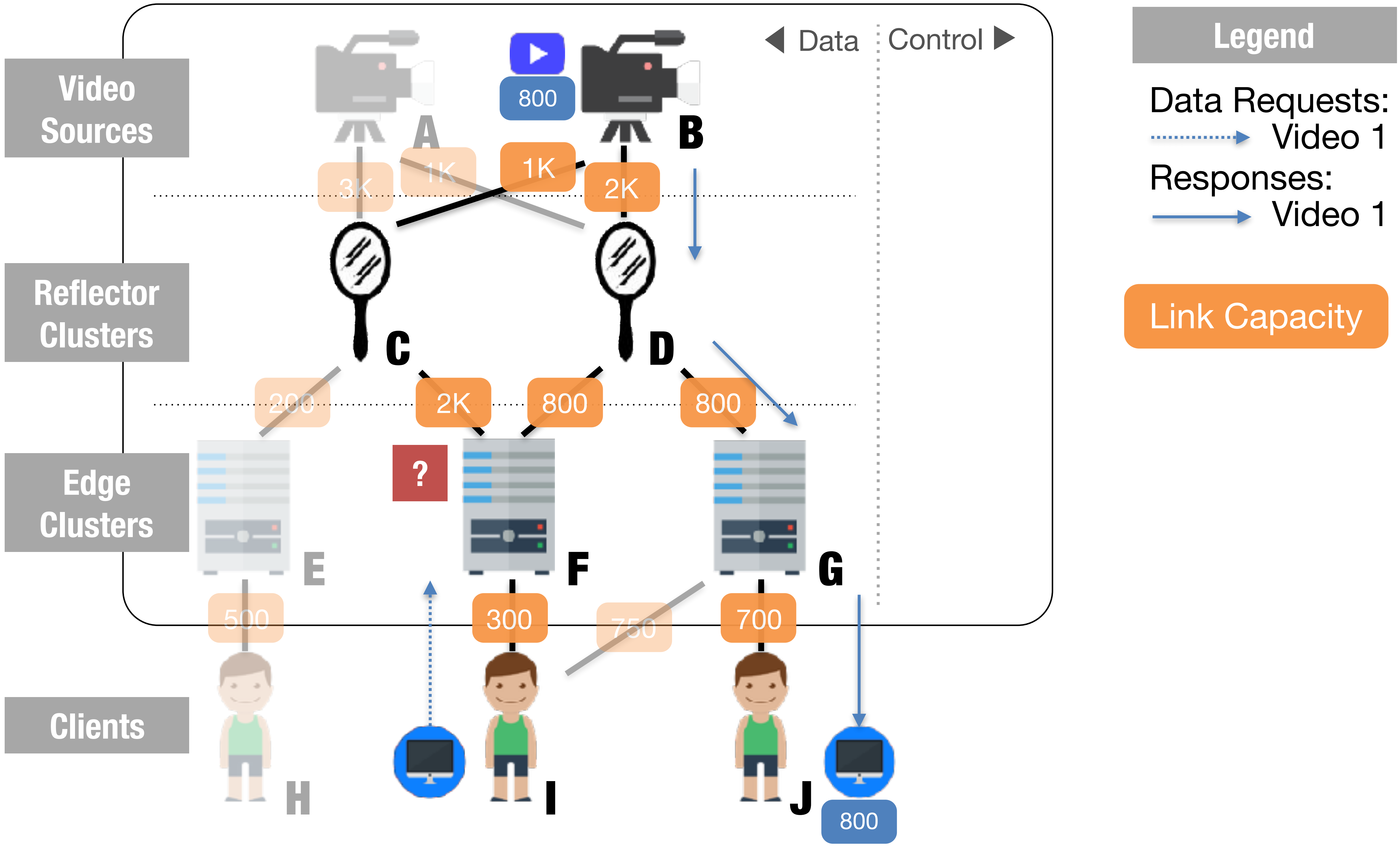


Experiments on EC2 nodes with a
centralized controller at CMU across the Internet

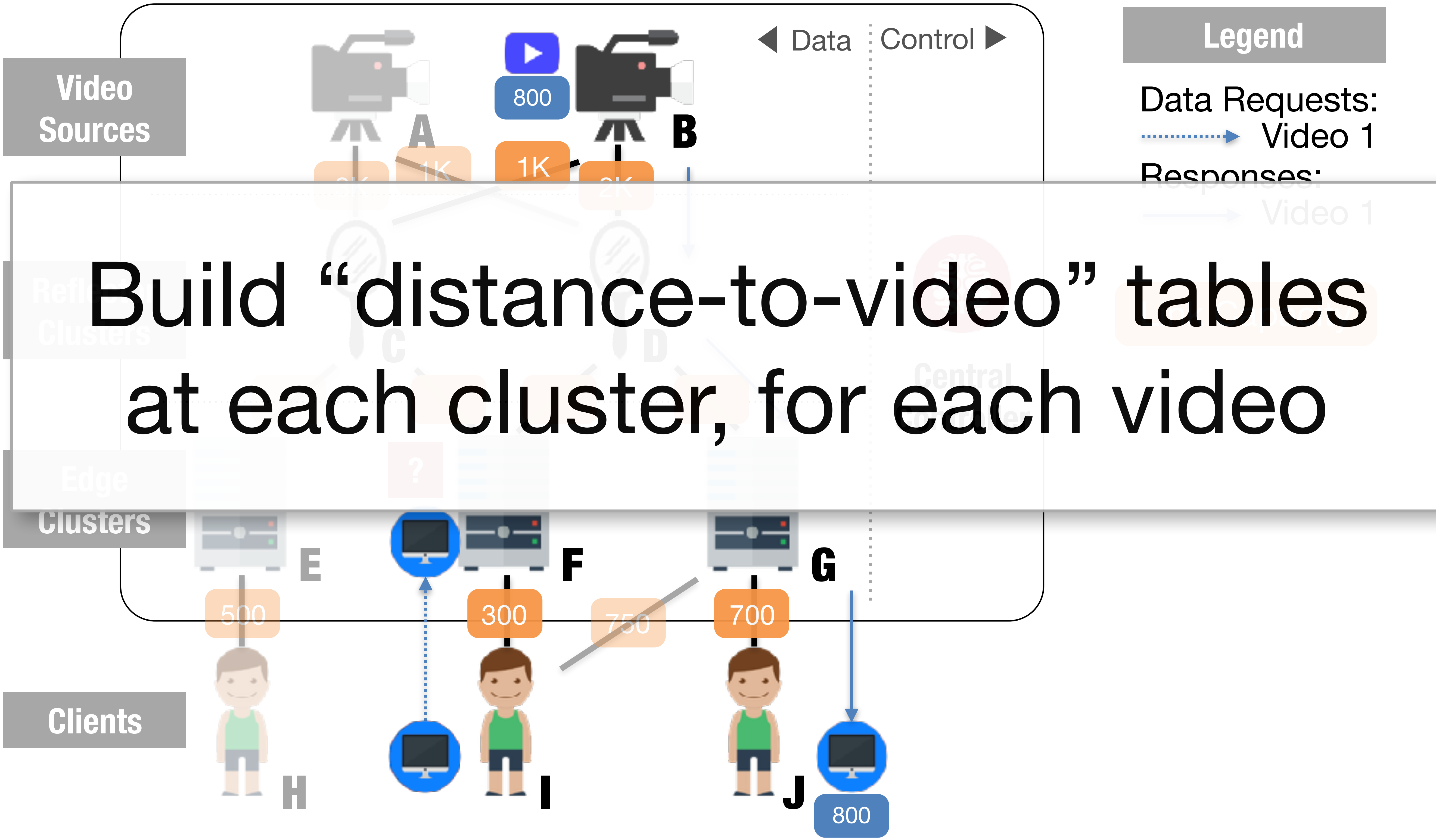
Outline


Slow join times
Centralized Control
Problems with Live Video Today
Distributed Control
Hybrid Control
Putting it all Together

Alternate Approach: Distributed

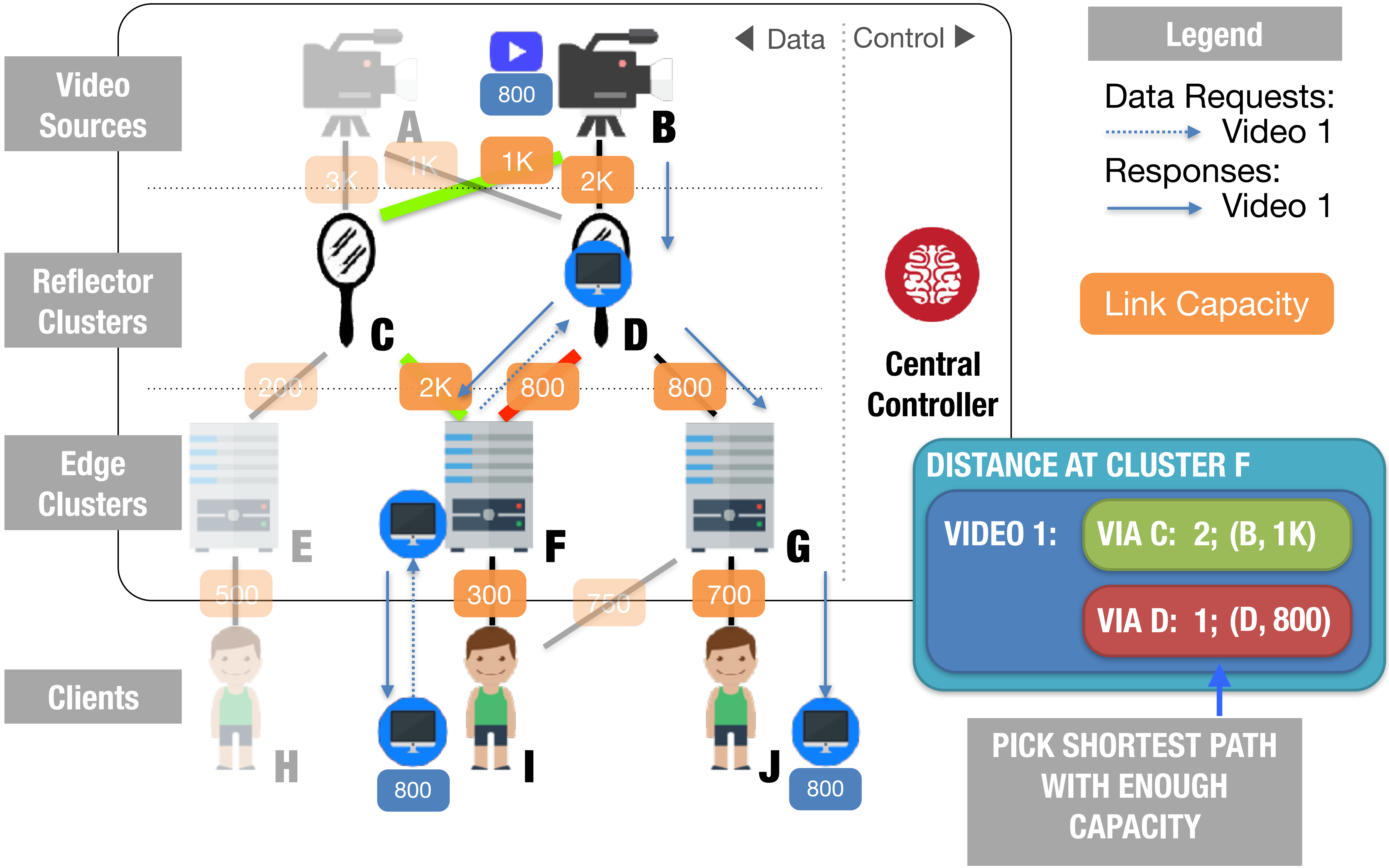


Alternate Approach: Distributed

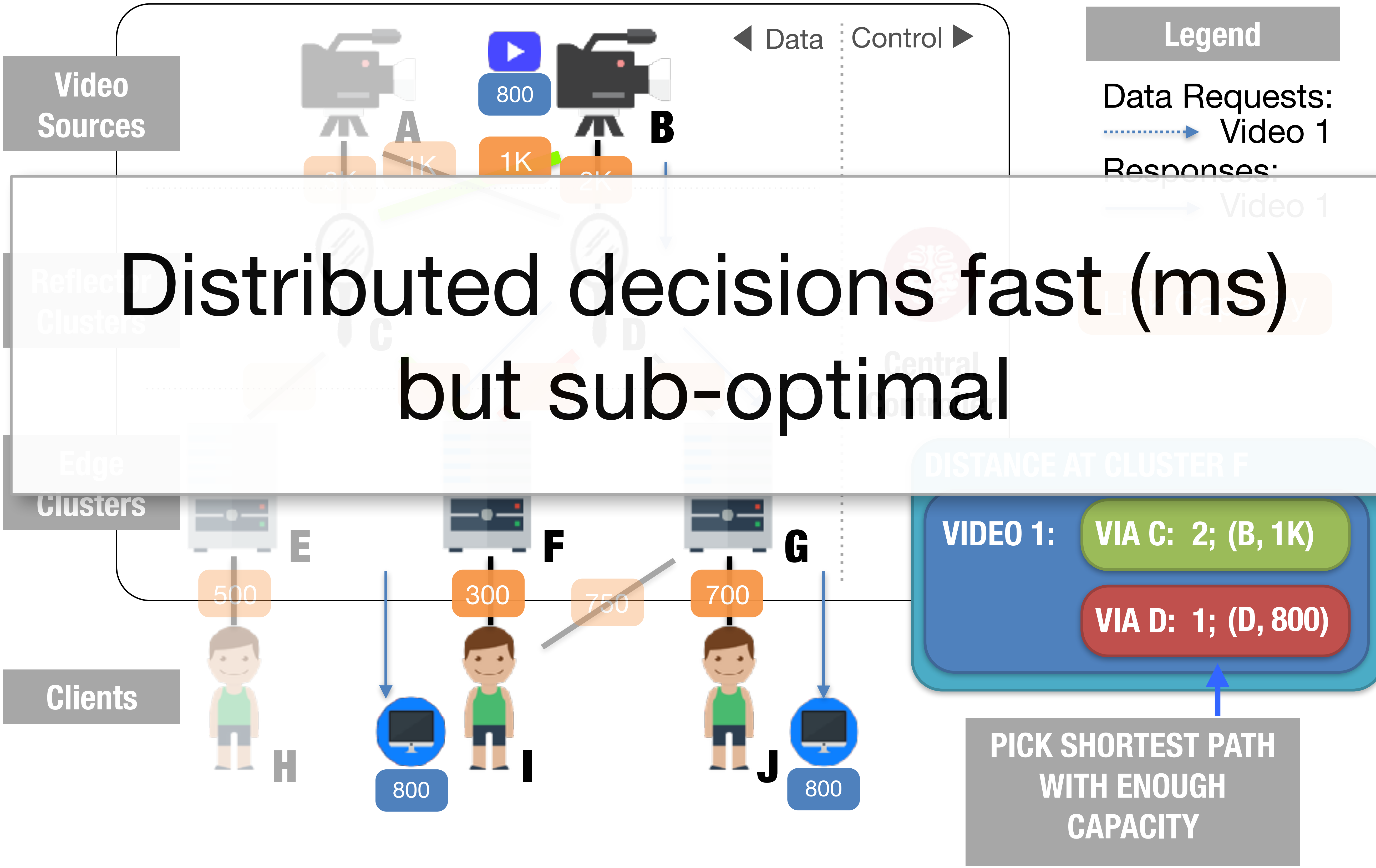


Build "distance-to-video" tables at each cluster, for each video

Alternate Approach: Distributed

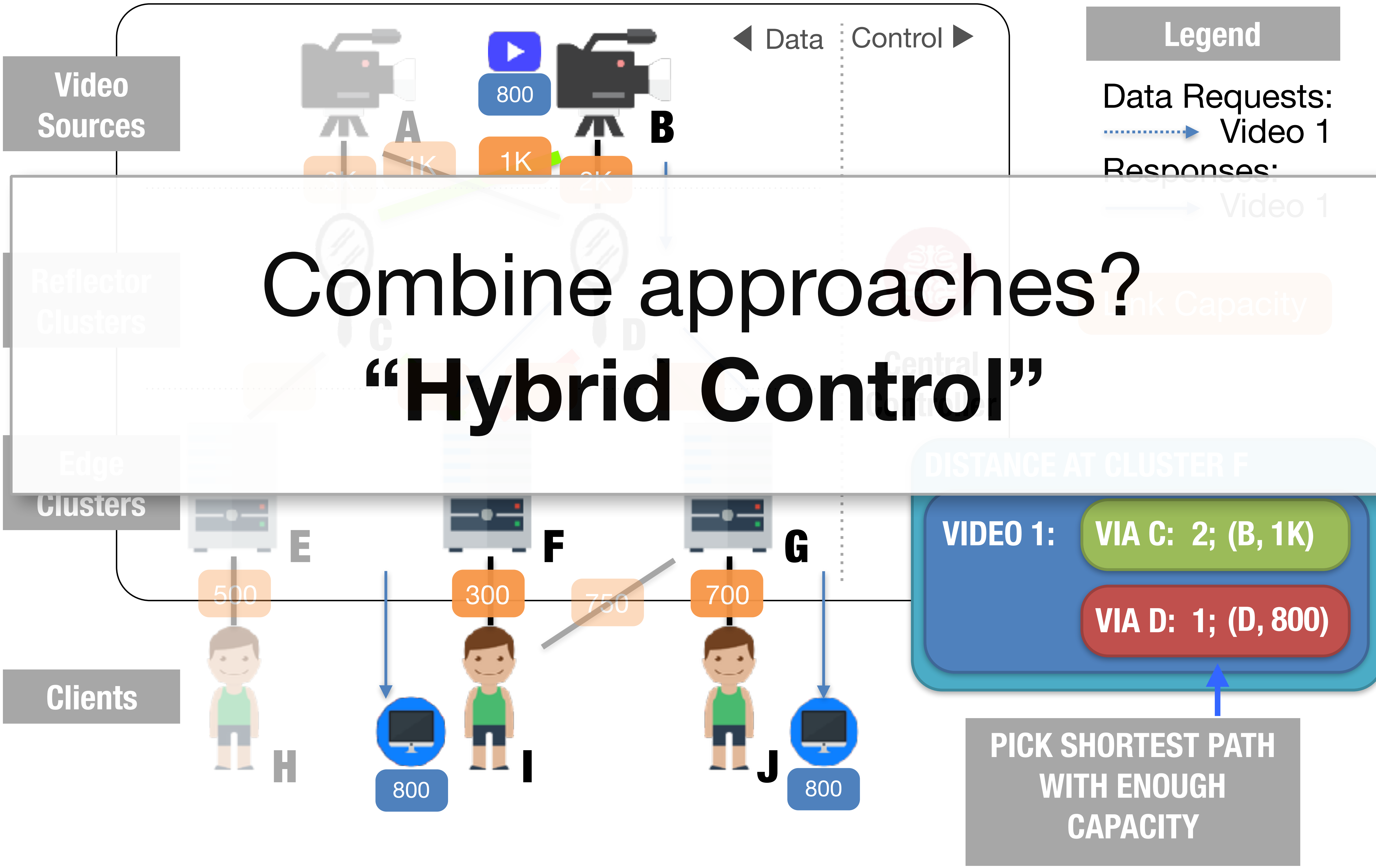


Alternate Approach: Distributed



Distributed decisions fast (ms) but sub-optimal

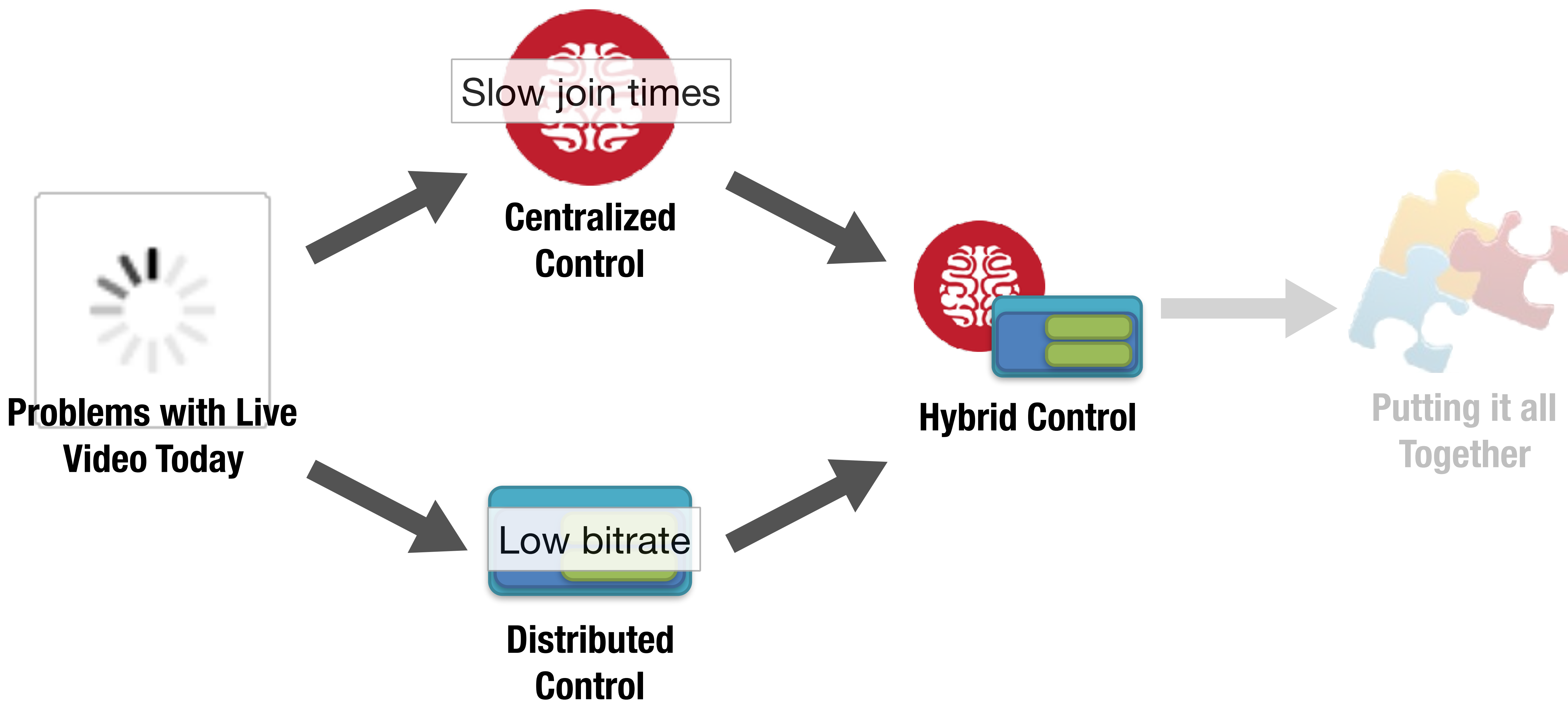
Alternate Approach: Distributed



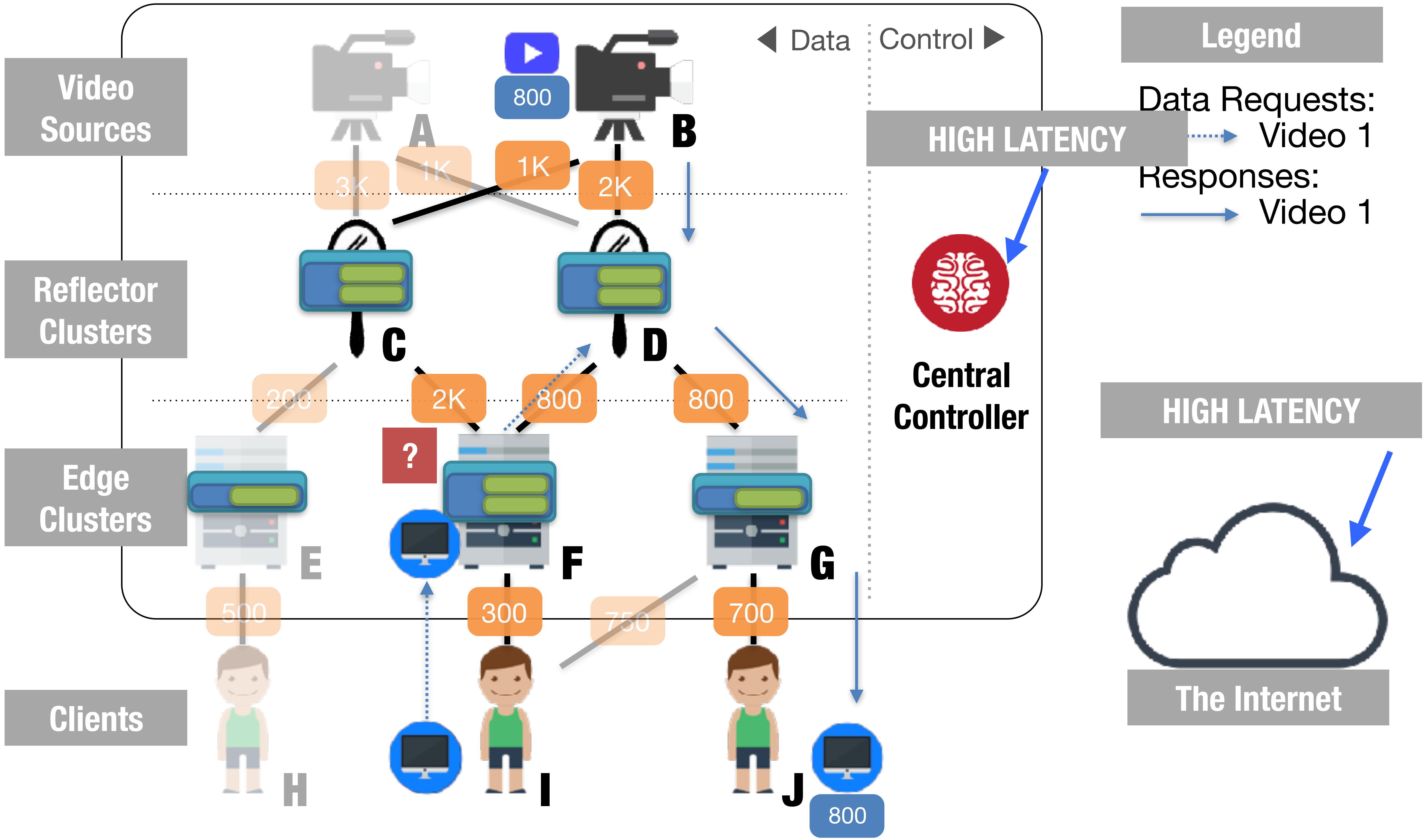
Combine approaches?

“Hybrid Control”

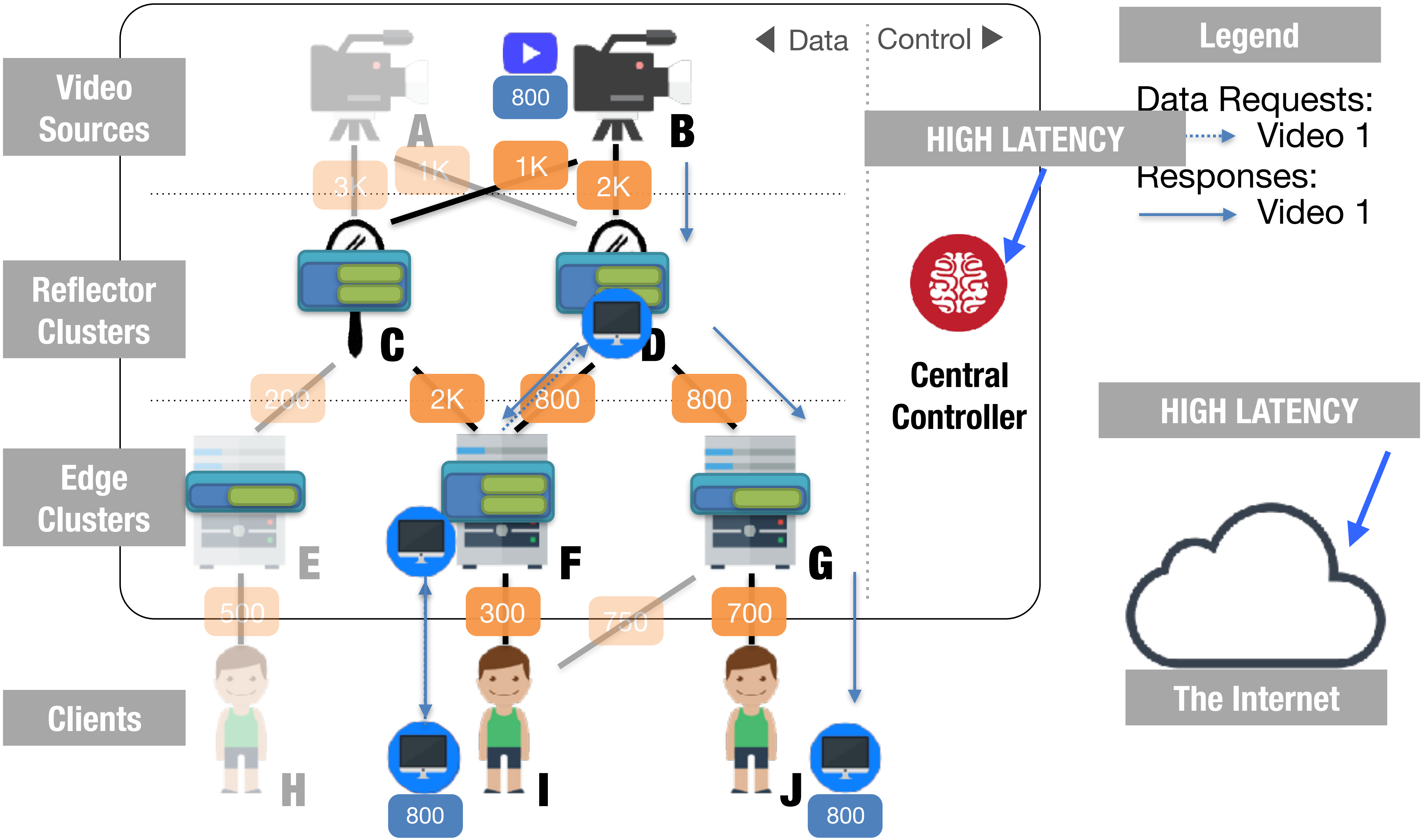
Outline


Slow join times
Centralized Control
Problems with Live Video Today
Hybrid Control
Putting it all Together
Low bitrate
Distributed Control

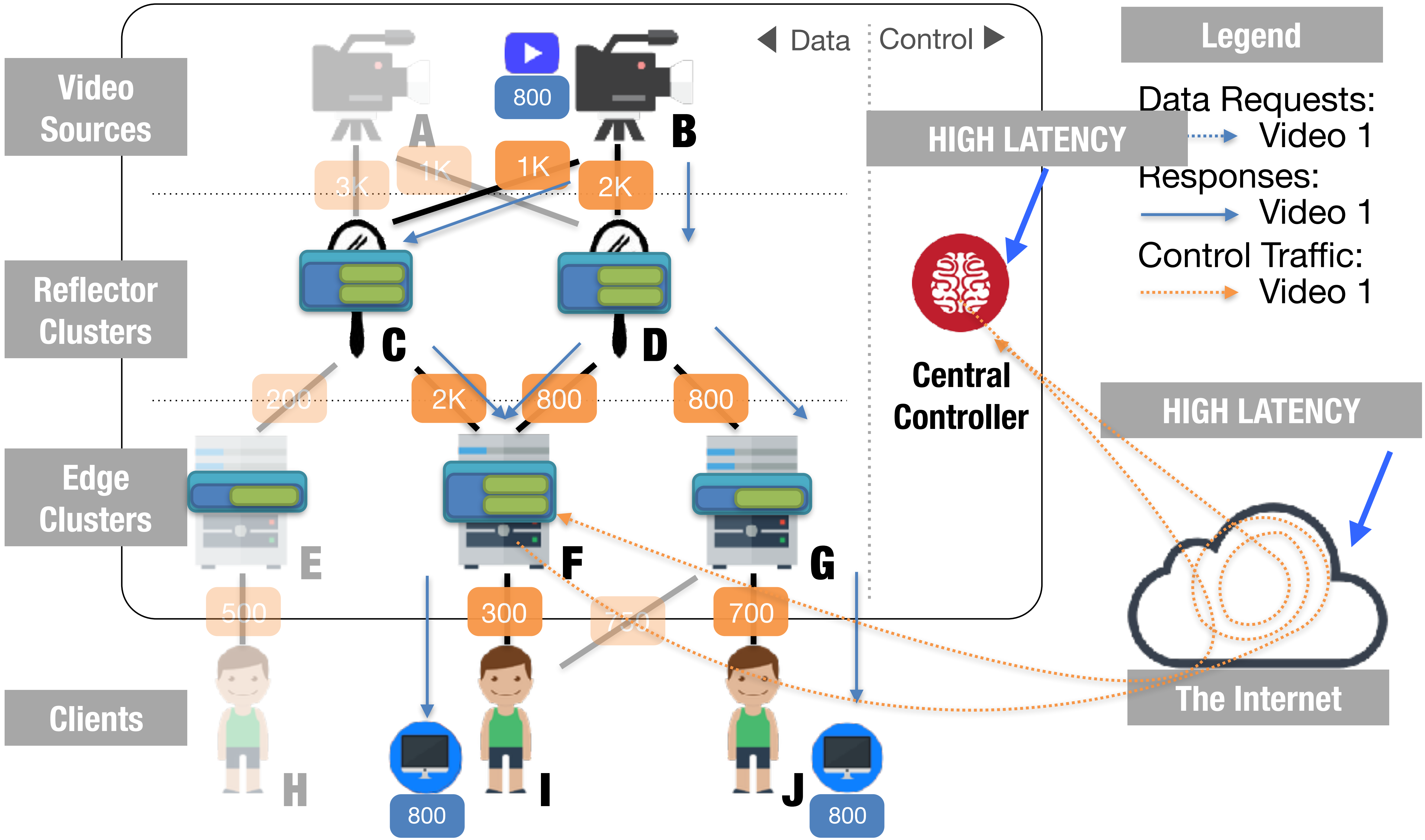
Combining Approaches: **VDN**



Combining Approaches: **VDN**



Combining Approaches: **VDN**



Challenges of Hybrid Control

- Forwarding loops
  - Always forward requests upwards

TRIVIAL

- State transitions
  - Versioning and “shadow FIBS”

PRIOR WORK

- Avoid bad control loop interactions

CHALLENGING

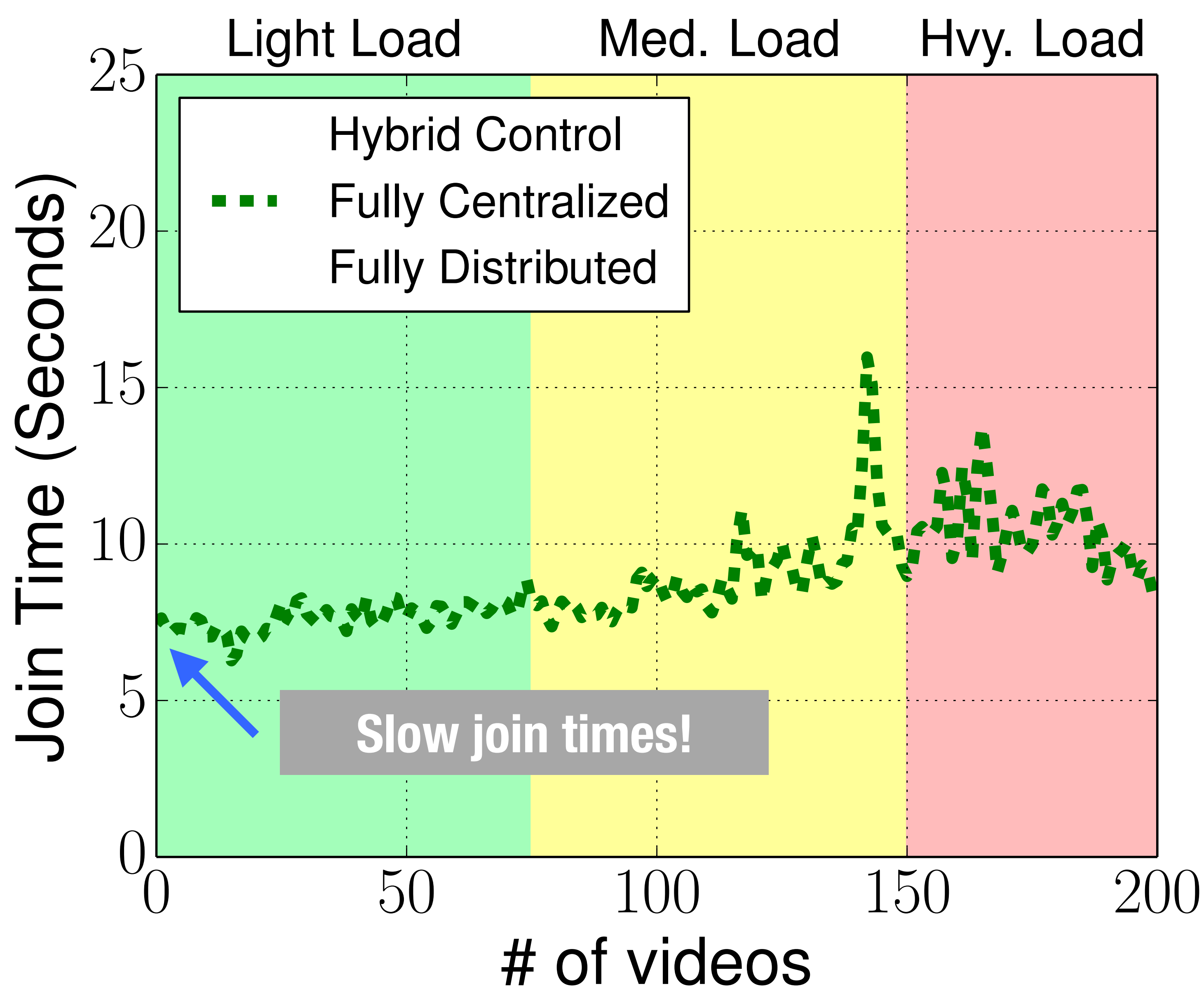
Challenges of Hybrid Control

- Avoid bad control loop interactions

CHALLENGING

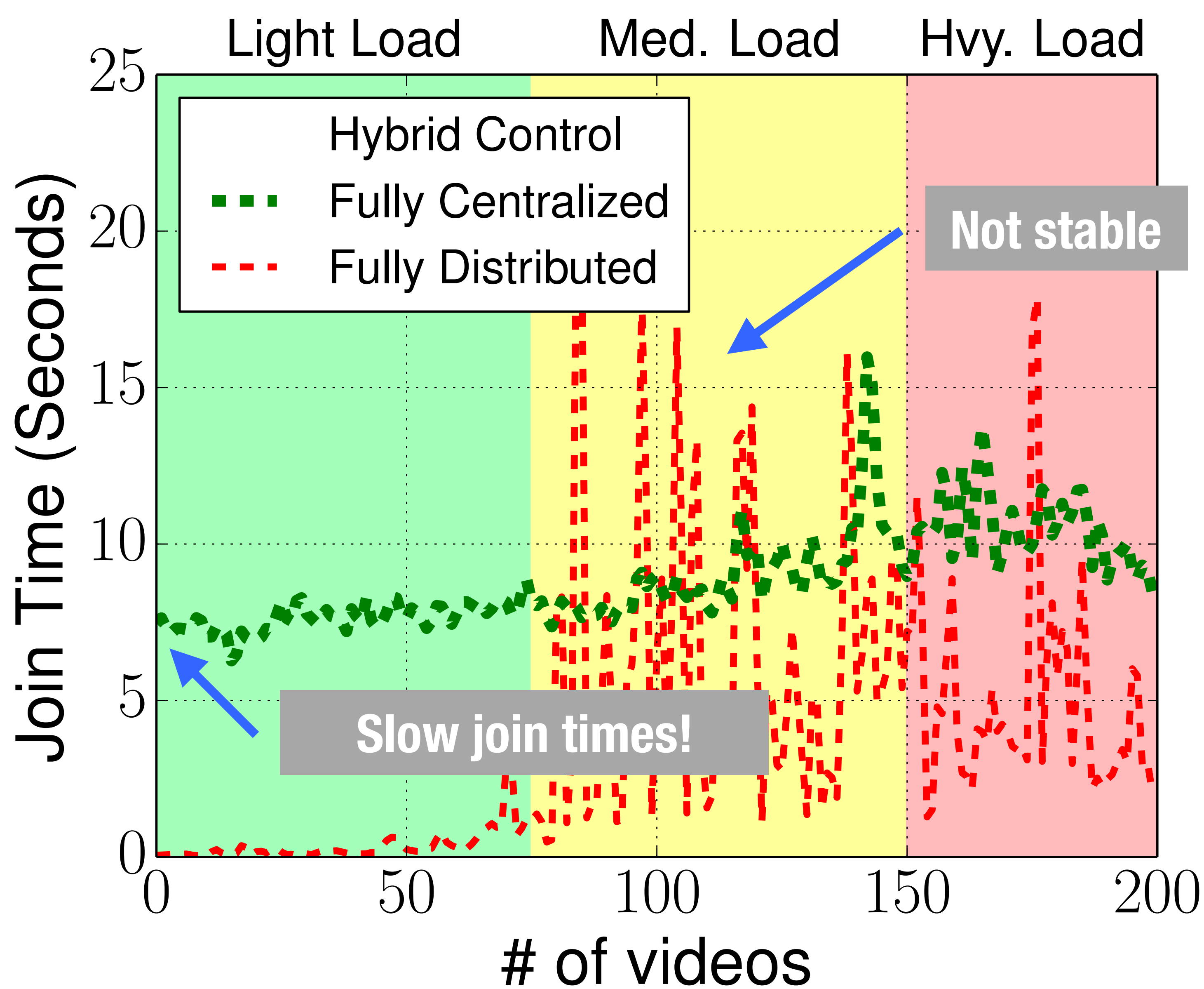
1. Centralized decision has priority
2. Distributed uses slack in network

Hybrid Control and Responsiveness



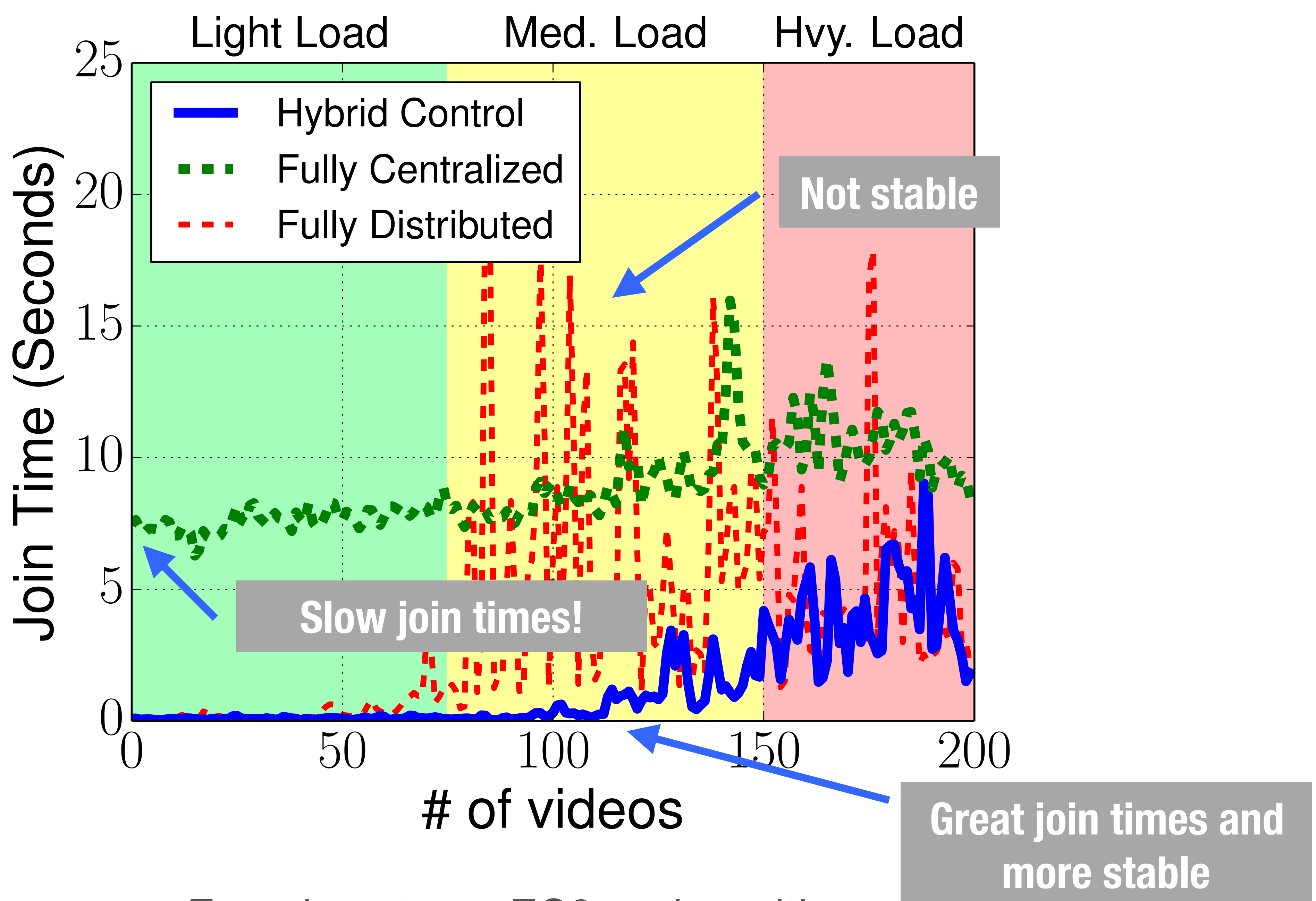
Experiments on EC2 nodes with a
centralized controller at CMU across the Internet

Hybrid Control and Responsiveness



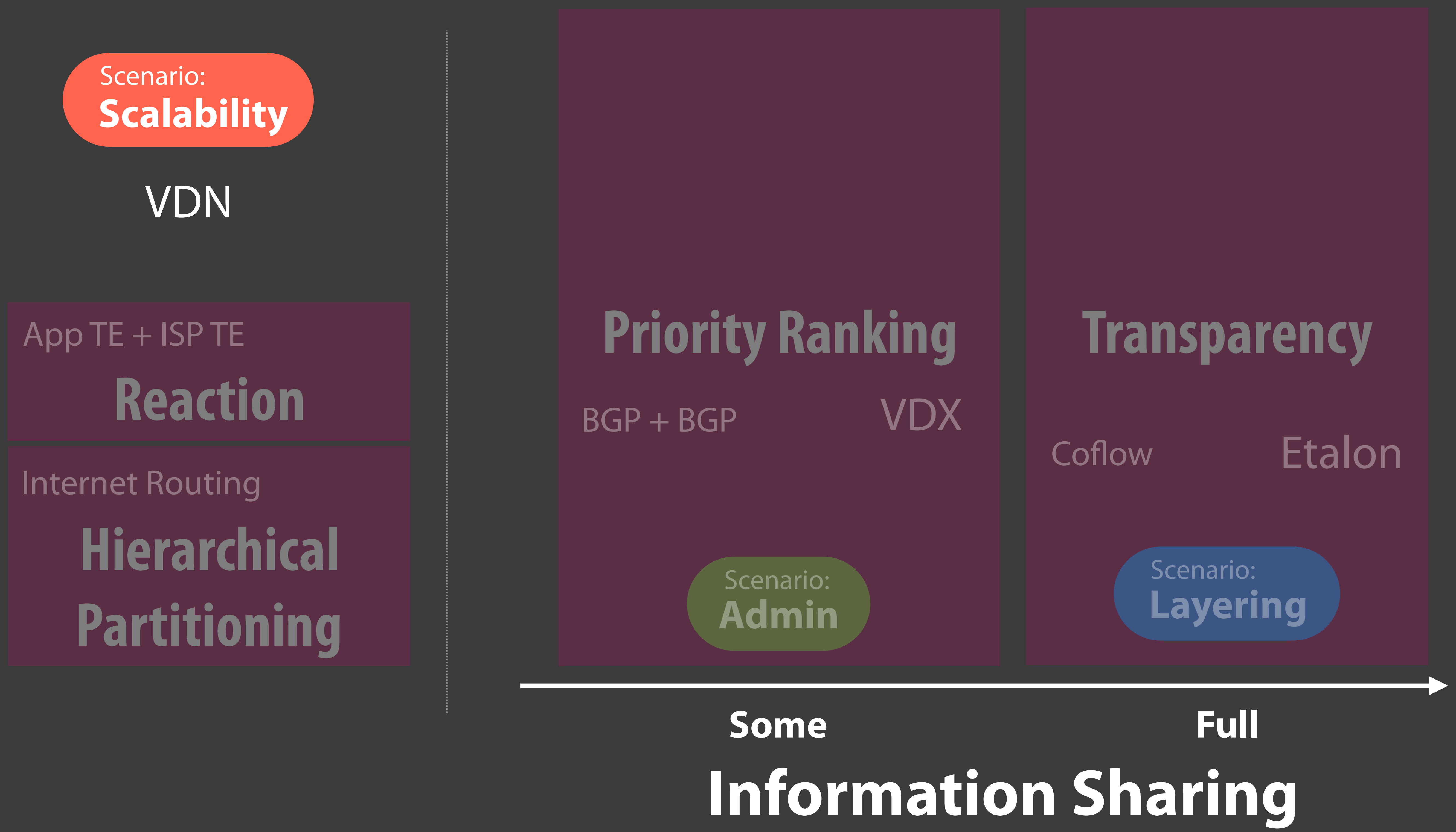
Experiments on EC2 nodes with a
centralized controller at CMU across the Internet

Hybrid Control and Responsiveness

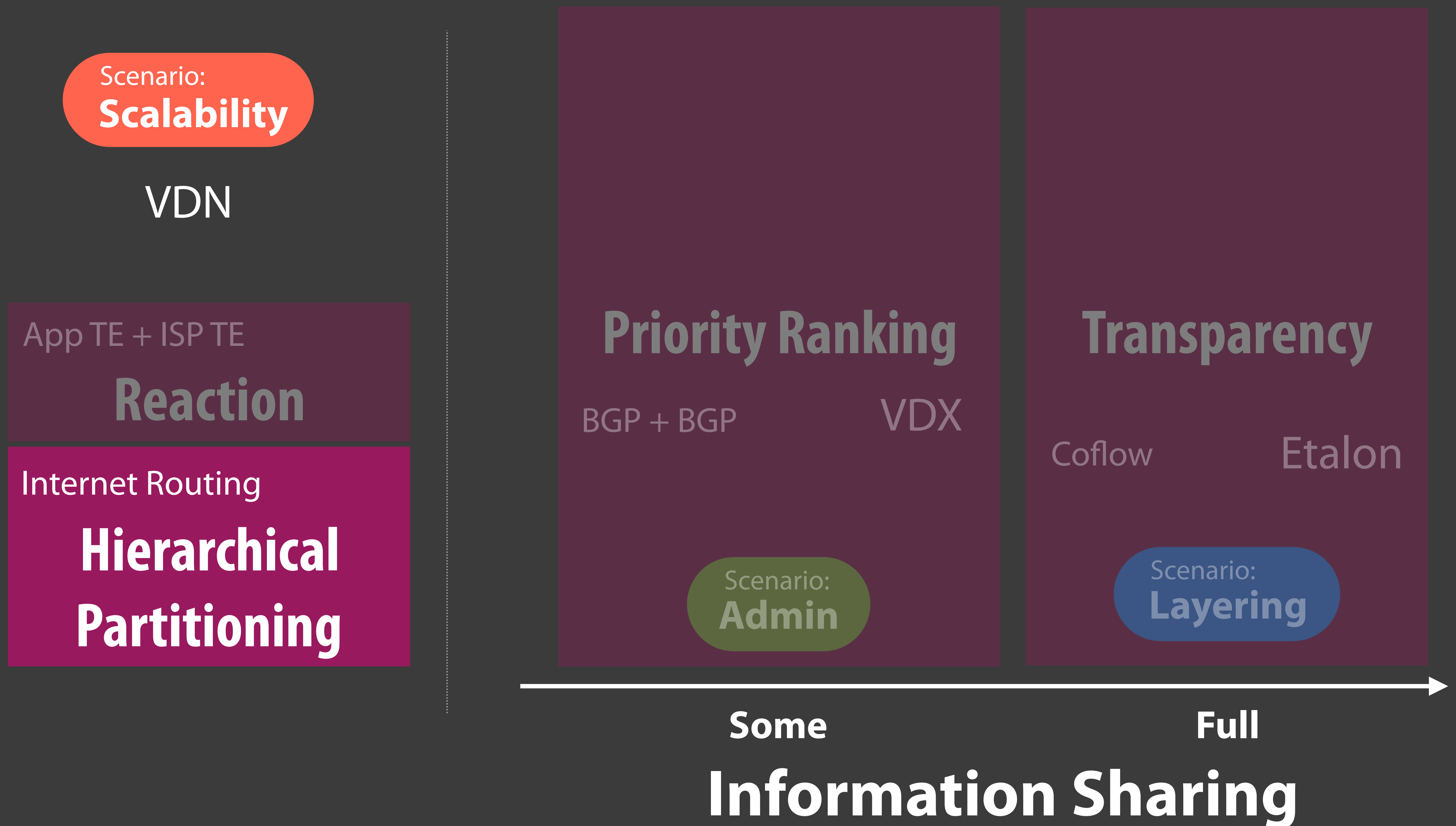


Experiments on EC2 nodes with a centralized controller at CMU across the Internet

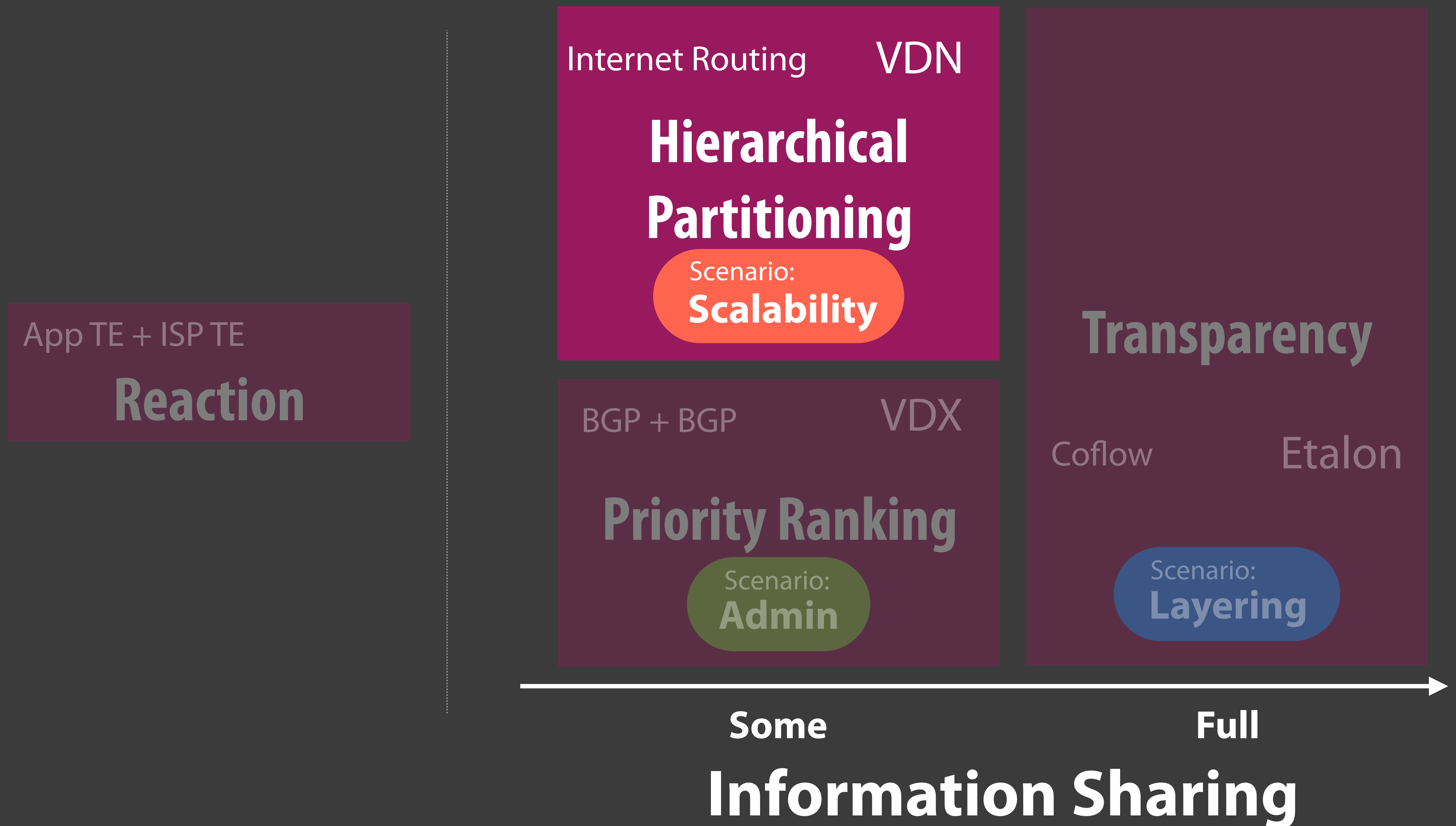
Control Coordination



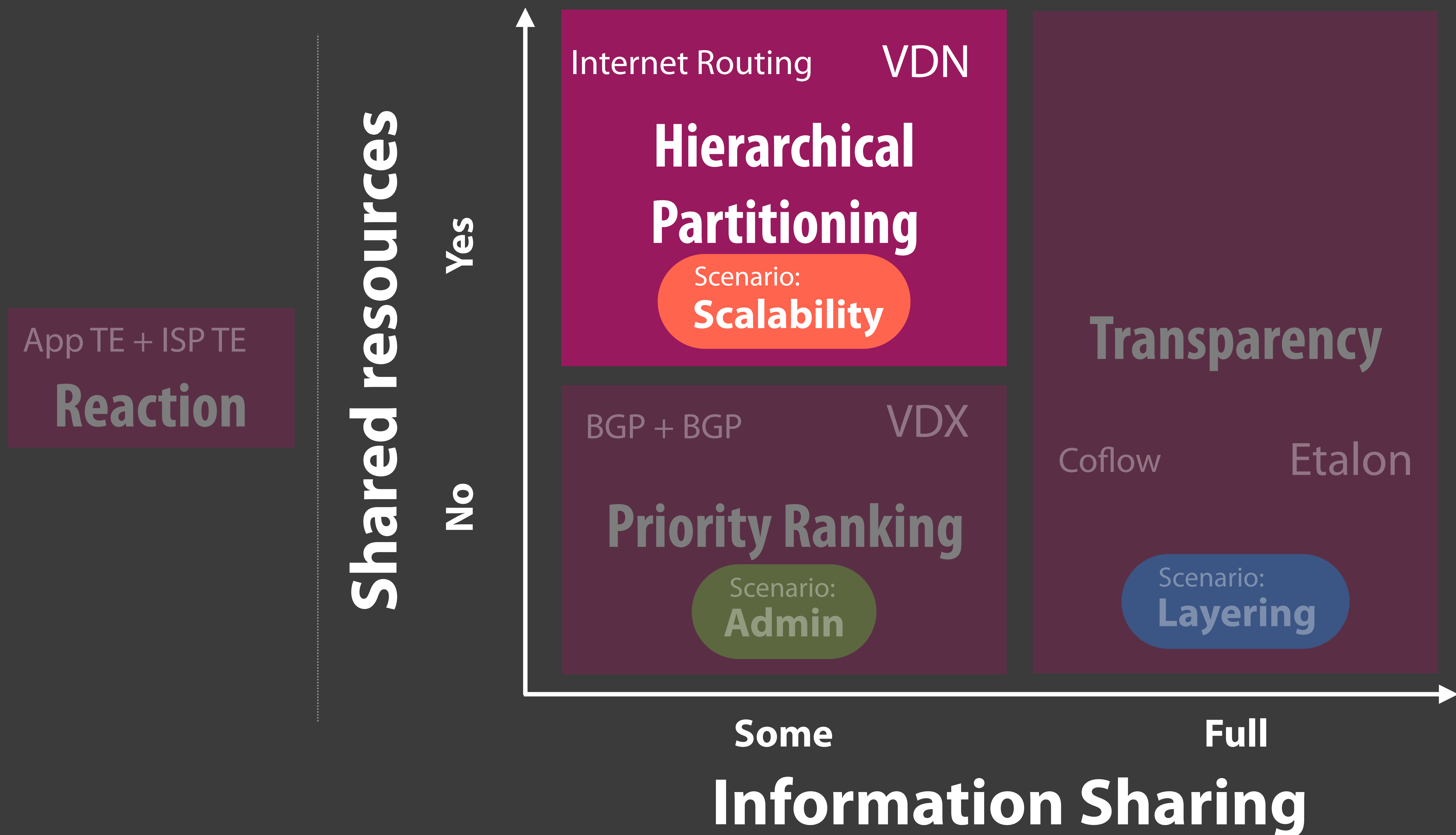
Control Coordination



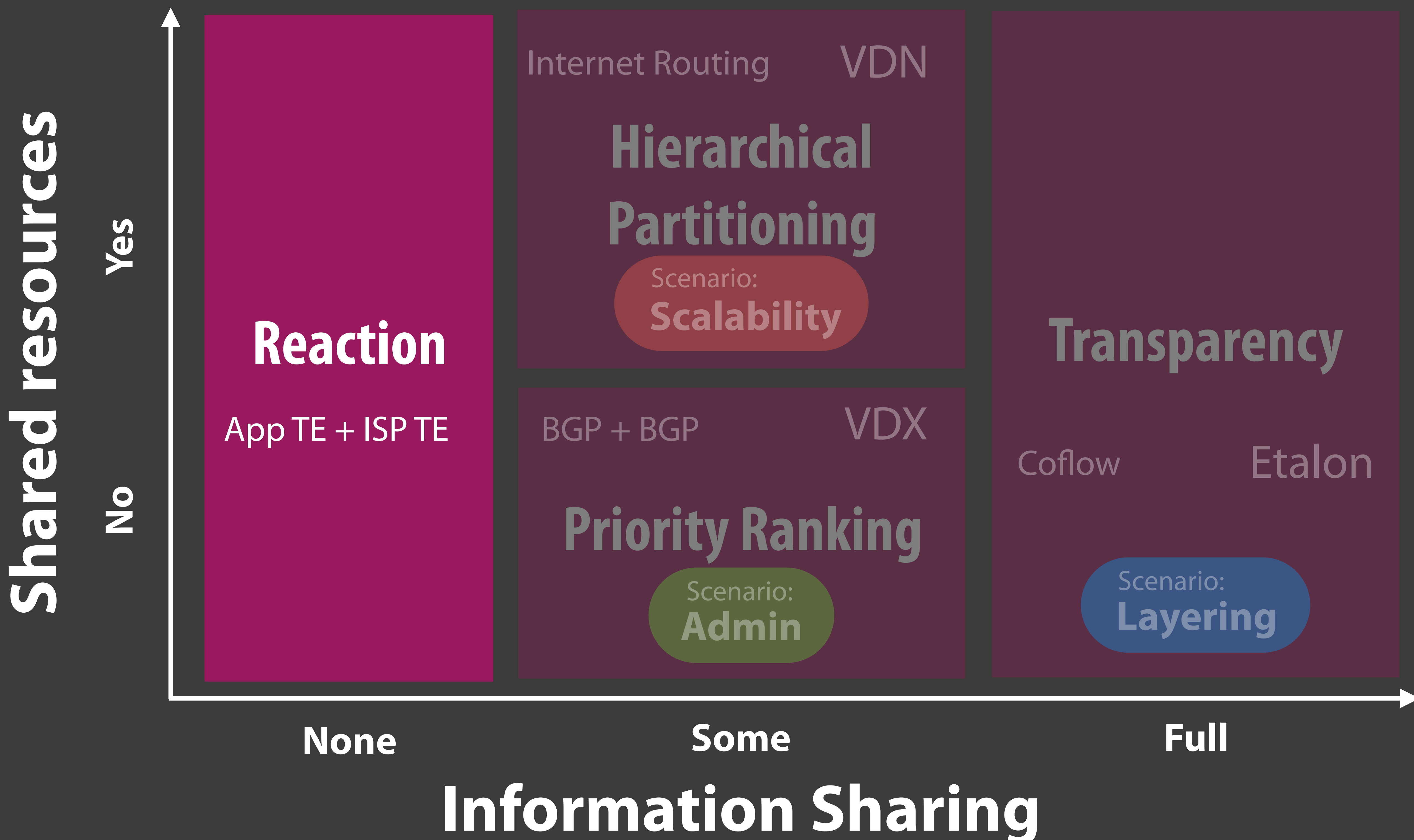
Control Coordination



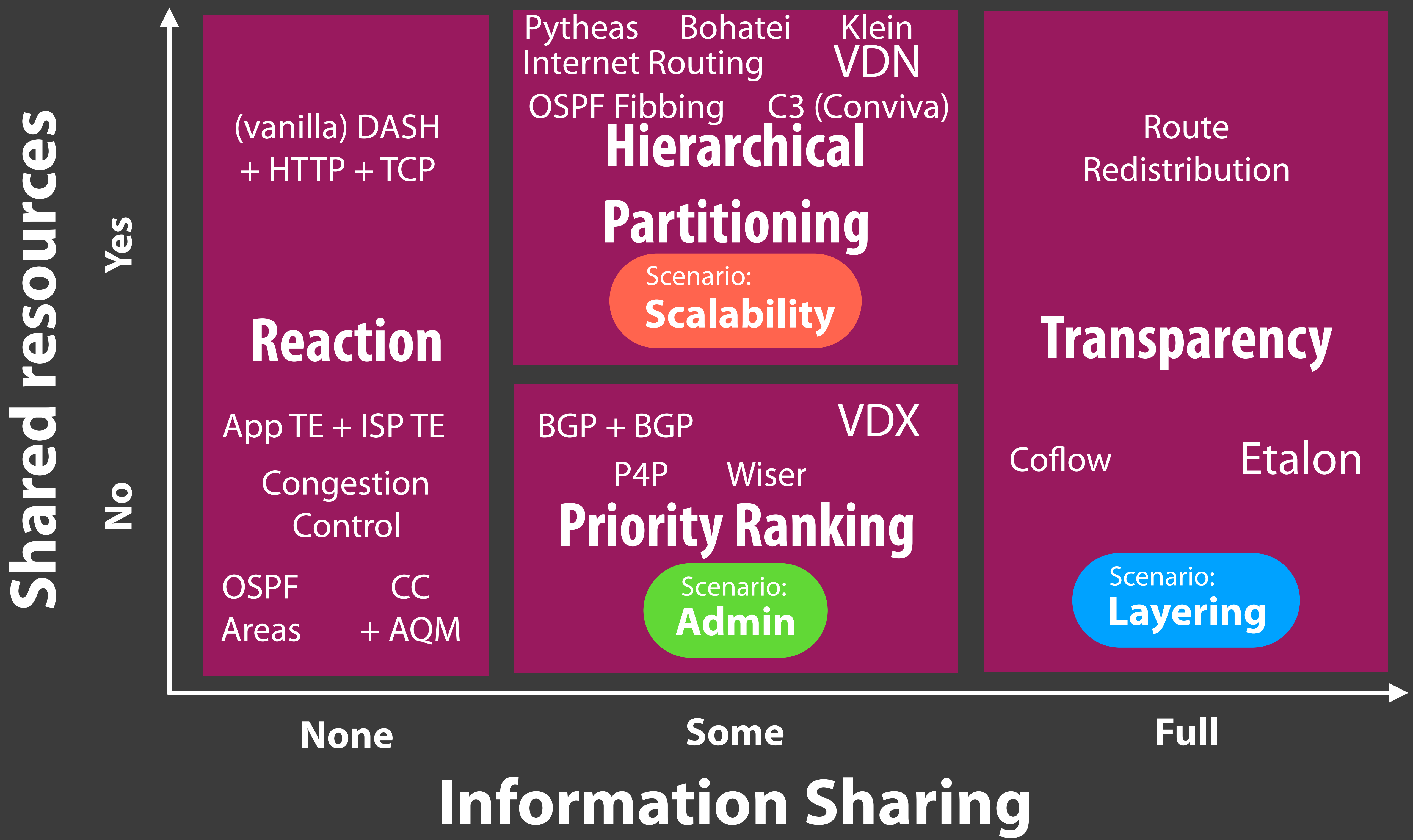
Control Coordination



Control Coordination



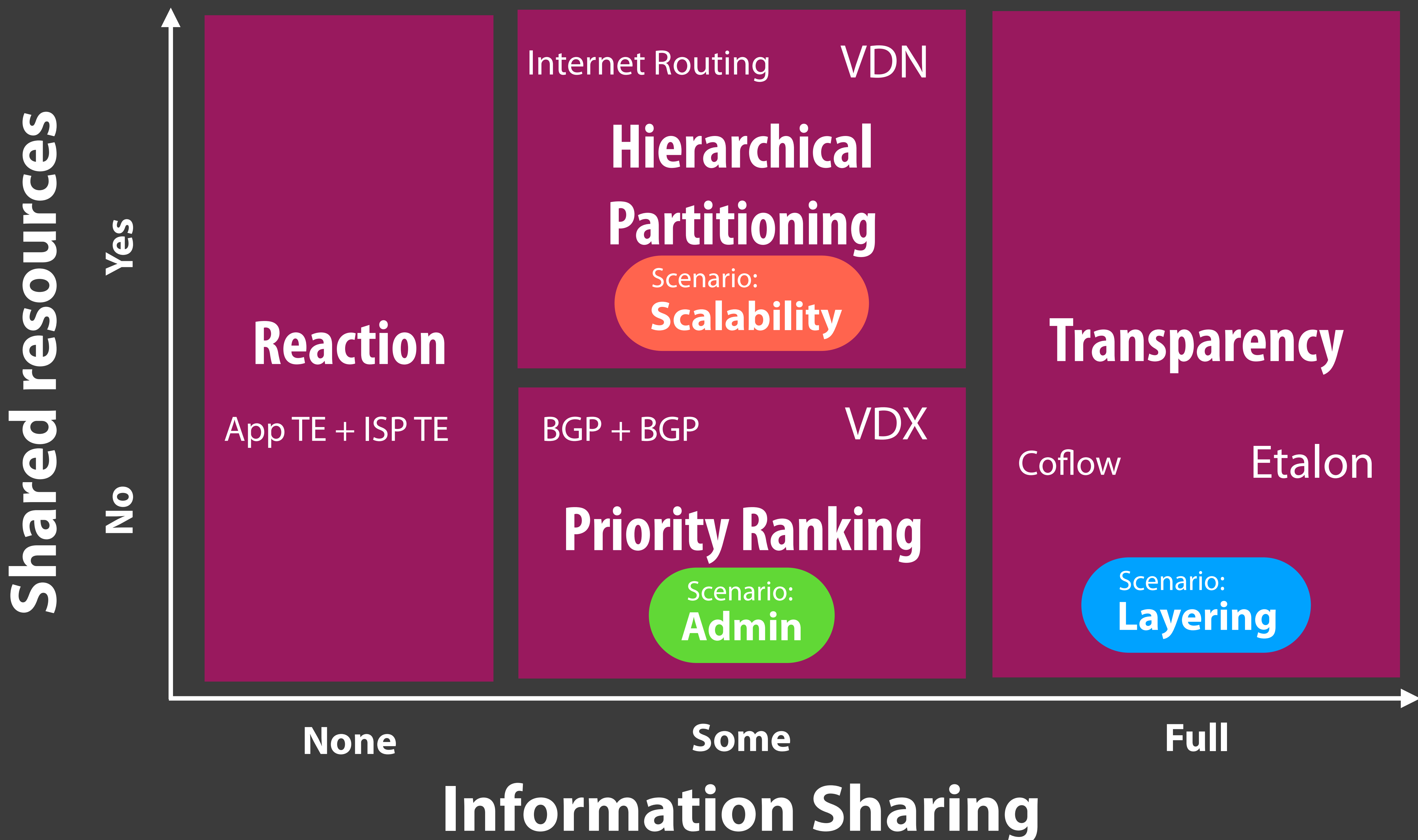
Control Coordination



Future Work

- Control theory / verification approach
- Validating VDN
- Extending VDX to multi-broker
- Principled approach to reconfigurable datacenters
- Network / endhost co-design
  - e.g., network-aware applications

Control Coordination



Eliminating Adverse Control Plane Interactions in Independent Network Systems

Matthew K. Mukerjee

Computer Science PhD Thesis Defense

May 1st, 2018

Carnegie Mellon University