

TCP ex Machina: Computer-Generated Congestion Control

Keith Winstein

MIT Computer Science and Artificial Intelligence Laboratory

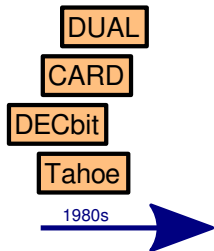
<http://mit.edu/remy>

March 3, 2014

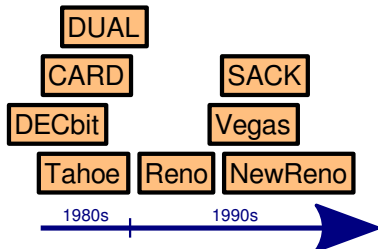


Joint work with Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

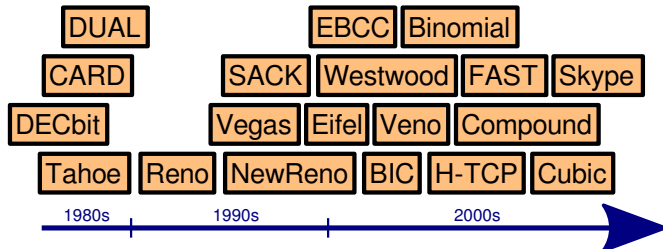
The march of congestion-control protocols



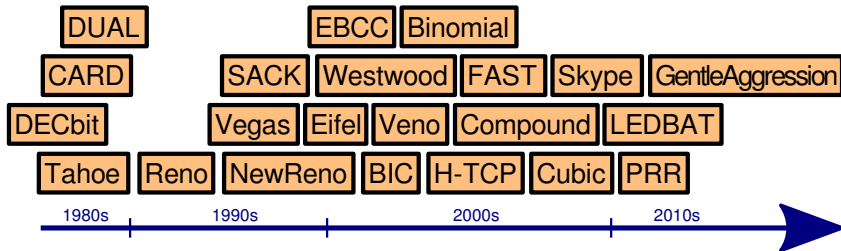
The march of congestion-control protocols



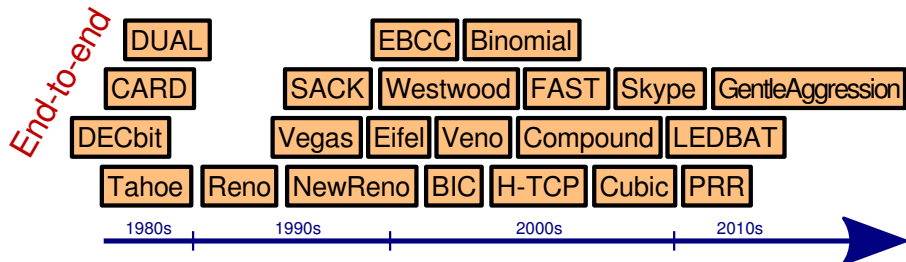
The march of congestion-control protocols



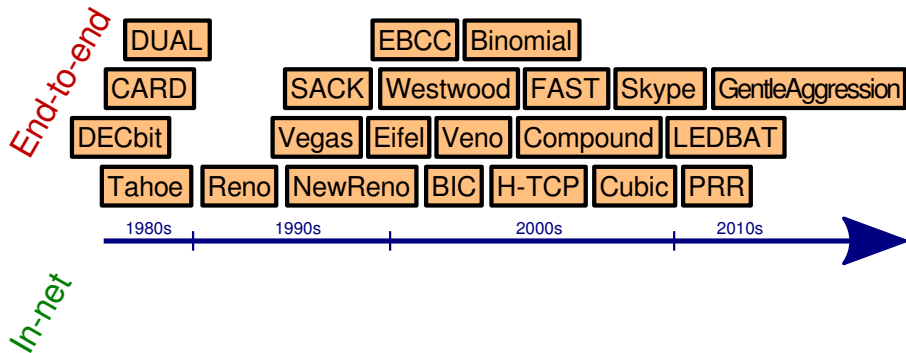
The march of congestion-control protocols



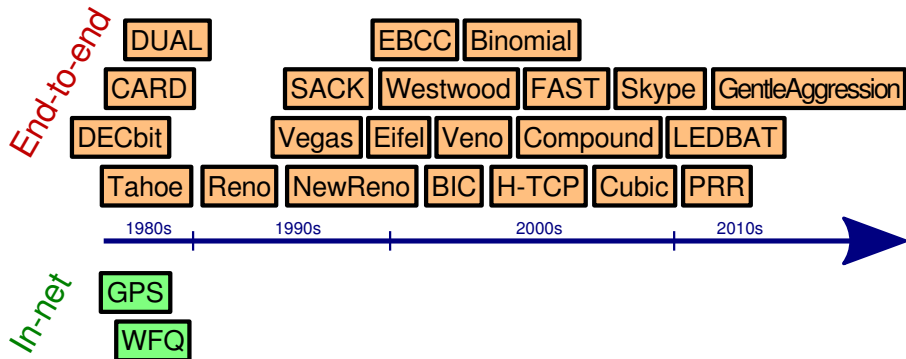
The march of congestion-control protocols



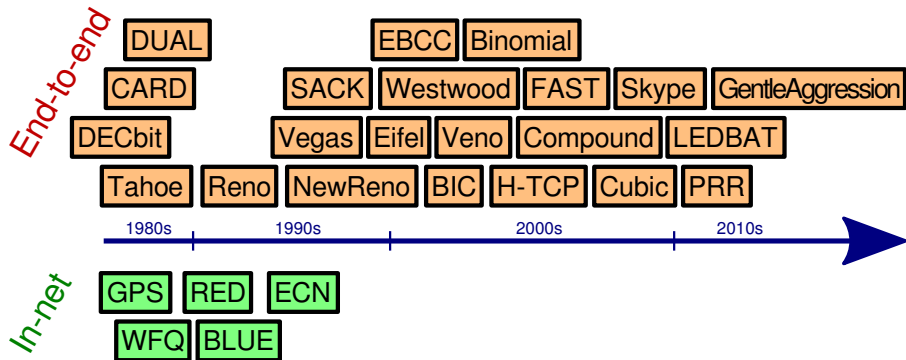
The march of congestion-control protocols



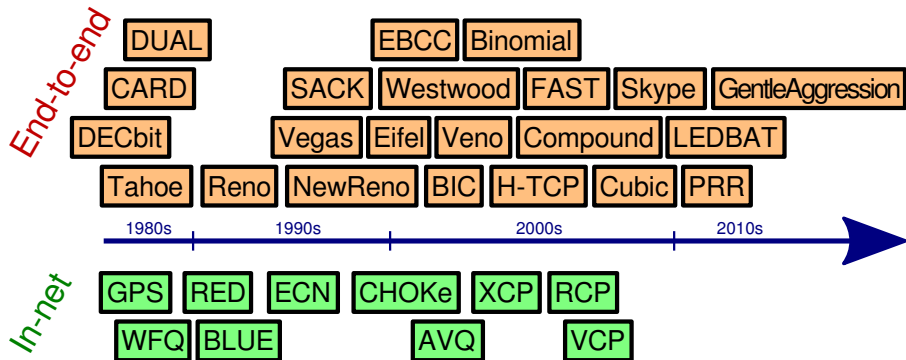
The march of congestion-control protocols



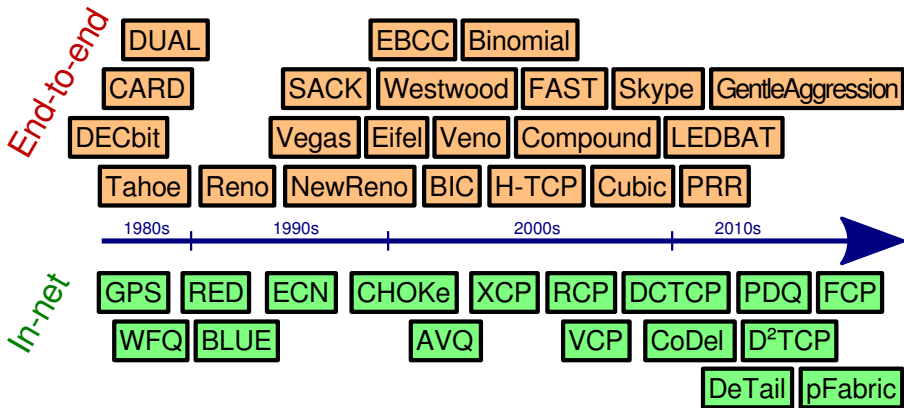
The march of congestion-control protocols



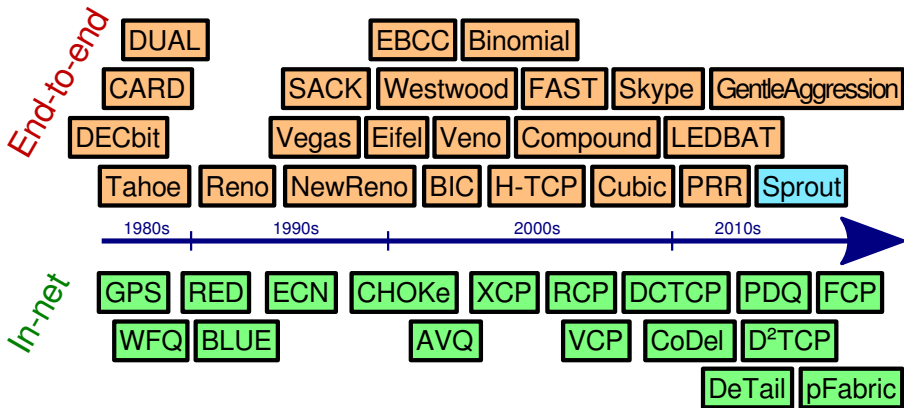
The march of congestion-control protocols



The march of congestion-control protocols



The march of congestion-control protocols



Rational choice of scheme is challenging

Cubic vs. **Compound**

- ▶ Different missions?
- ▶ Different assumptions about network?
- ▶ One scheme just plain better?

Networks constrained by a fuzzy idea of TCP's assumptions

- ▶ Mask stochastic loss
- ▶ Bufferbloat
- ▶ Mask out-of-order delivery
- ▶ No parallel/multipath routing

Advice for Internet Subnetwork Designers
(RFC 3819) is 21,000 words!

Apps hack around TCP

- ▶ Open lots of flows
- ▶ Goose slow start
- ▶ Add pacing
- ▶ Give up and do it yourself

Apps hack around TCP

- ▶ Open lots of flows
- ▶ Goose slow start
- ▶ Add pacing
- ▶ Give up and do it yourself



Apps hack around TCP

- ▶ Open lots of flows
- ▶ Goose slow start
- ▶ Add pacing
- ▶ Give up and do it yourself



Google MICROSOFT.

Apps hack around TCP

- ▶ Open lots of flows
- ▶ Goose slow start
- ▶ Add pacing
- ▶ Give up and do it yourself



Google MICROSOFT

You Tube

Apps hack around TCP

- ▶ Open lots of flows
- ▶ Goose slow start
- ▶ Add pacing
- ▶ Give up and do it yourself



Google MICROSOFT

YouTube

Chrome (QUIC)

BitTorrent (μ TP)

Mosh (SSP)

IBM Aspera (fasp)

Idea: computer-generated protocols

Transport layer should adapt to **whatever**:

- ▶ network does
- ▶ application wants

Idea: computer-generated protocols

Transport layer should adapt to **whatever**:

- ▶ network does (**model**)
- ▶ application wants (**mission**)

What we built

Remy: a program that generates congestion-control schemes offline

Input:

- ▶ Assumptions about network and workload (**model**)
- ▶ Application's objective (**mission**)

Output: CC algorithm for a TCP sender (RemyCC)

Time: hours to days

The basic question of congestion control

At this moment, do I:

- ▶ send a packet
- ▶ not send a packet?

Missions of congestion control

Maximize

- ▶ $\sum_i \log [\text{throughput}_i]$ (proportionally fair throughput)

Missions of congestion control

Maximize

- ▶ $\sum_i \log [\text{throughput}_i]$ (proportionally fair throughput)

- ▶ $\sum_i \log \left[\frac{\text{throughput}_i}{\text{delay}_i} \right]$ (proportionally fair throughput/delay)

Missions of congestion control

Maximize

▶ $\sum_i \log [\text{throughput}_i]$ (proportionally fair throughput)

▶ $\sum_i \log \left[\frac{\text{throughput}_i}{(\text{delay}_i)^\delta} \right]$ (proportionally fair throughput/delay)

Missions of congestion control

Maximize

- ▶ $\sum_i \log [\text{throughput}_i]$ (proportionally fair throughput)

- ▶ $\sum_i \log \left[\frac{\text{throughput}_i}{(\text{delay}_i)^\delta} \right]$ (proportionally fair throughput/delay)

- ▶ $\min_i \text{throughput}_i$ (max-min throughput)

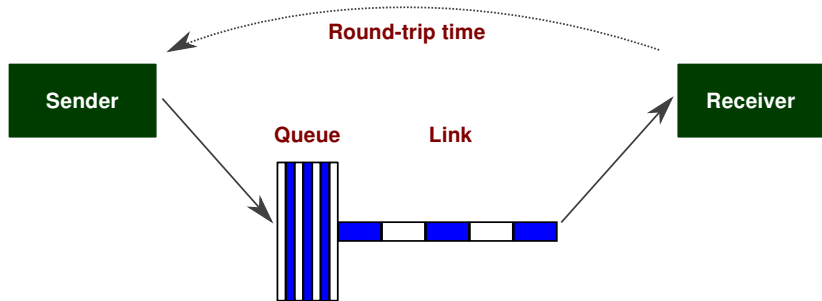
Minimize

- ▶ average flow completion time
- ▶ page load time
- ▶ tail completion time

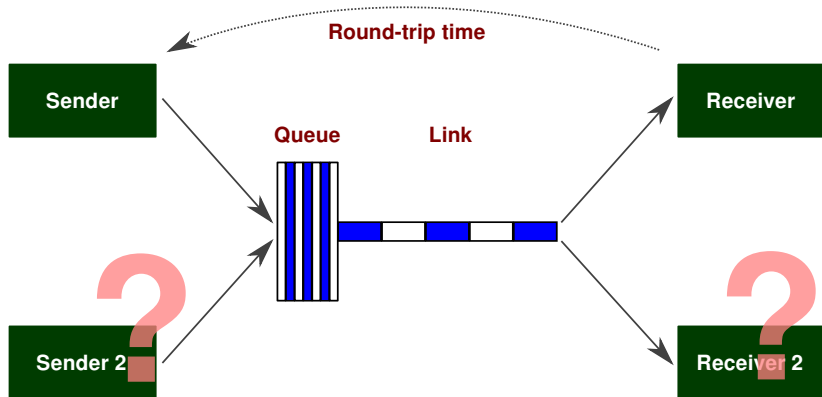
Encoding the designer's prior assumptions

- ▶ **Model** of network uncertainty
 - ▶ Link speed distribution
 - ▶ Delay distribution
 - ▶ Topology distribution
- ▶ **Model** of workload
 - ▶ Web browsing
 - ▶ MapReduce
 - ▶ videoconferencing
 - ▶ streaming video (YouTube/Netflix)

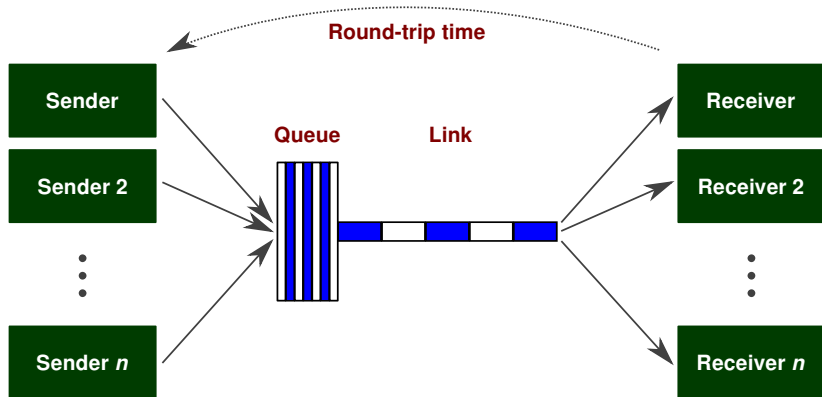
Dumbbell network



Dumbbell network



Dumbbell network



Superrational congestion control

At this moment, * do I:

- ▶ send a packet
- ▶ not send a packet?

Superrational congestion control

At this moment,* do I:

- ▶ send a packet
- ▶ not send a packet?

* Assuming every node is running the same algorithm.

Internet congestion control as a Dec-POMDP

I: independent endpoint computers

Internet congestion control as a Dec-POMDP

I: independent endpoint computers

S: packets in net + whether each computer has data to send now

Internet congestion control as a Dec-POMDP

- I : independent endpoint computers
- S : packets in net + whether each computer has data to send now
- A_i : { send a packet, don't send a packet }

Internet congestion control as a Dec-POMDP

- I*: independent endpoint computers
- S*: packets in net + whether each computer has data to send now
- A_i*: { send a packet, don't send a packet }
- T*: simple networks are deterministic, with parameters drawn from a distribution. Flows arrive per a random process.

Internet congestion control as a Dec-POMDP

- I*: independent endpoint computers
- S*: packets in net + whether each computer has data to send now
- A_i*: { send a packet, don't send a packet }
- T*: simple networks are deterministic, with parameters drawn from a distribution. Flows arrive per a random process.
- R*: overall objective awarded at end of long run

Internet congestion control as a Dec-POMDP

- I : independent endpoint computers
- S : packets in net + whether each computer has data to send now
- A_i : { send a packet, don't send a packet }
- T : simple networks are deterministic, with parameters drawn from a distribution. Flows arrive per a random process.
- R : overall objective awarded at end of long run
- Ω : acknowledgements from corresponding receiver

Internet congestion control as a Dec-POMDP

- I : independent endpoint computers
- S : packets in net + whether each computer has data to send now
- A_i : { send a packet, don't send a packet }
- T : simple networks are deterministic, with parameters drawn from a distribution. Flows arrive per a random process.
- R : overall objective awarded at end of long run
- Ω : acknowledgements from corresponding receiver
- O : deterministic

Internet congestion control as a Dec-POMDP

- I : independent endpoint computers
- S : packets in net + whether each computer has data to send now
- A_i : { send a packet, don't send a packet }
- T : simple networks are deterministic, with parameters drawn from a distribution. Flows arrive per a random process.
- R : overall objective awarded at end of long run
- Ω : acknowledgements from corresponding receiver
- O : deterministic

Complexity in general: $\mathcal{O}(2^{2^n})$ (NEXP-hard)

Remy: tractable search for best policy

- ▶ Best decision given all history: not tractable
- ▶ Instead, remember only **a summary of history**

A RemyCC tracks four congestion signals

r_ewma_α : **short-term** moving average of interval between acks
“How fast are packets arriving (now)?”

r_ewma_β : **long-term** moving average of same
“How fast are packets arriving (smoothed)?”

s_ewma : moving average of interval between acked timestamps
“How fast was I sending?”

rtt_ratio : ratio of last RTT to smallest RTT so far
“How long is the queue?”

Why these congestion signals?

- ▶ Removing any of the four hurts
 - ▶ r_ewma_α hurts the most
- ▶ More signals increase search time
- ▶ Other signals might help on other networks

A RemyCC maps each state to an action

$$\text{REMYCC}(r_ewma_{\alpha\beta}, s_ewma, rtt_ratio) \rightarrow \langle m, b, \tau \rangle$$

m Multiple to congestion window

b Increment to congestion window

τ Minimum interval between two outgoing packets

Runtime for a RemyCC

On ack:

- ▶ $\langle m, b, \tau \rangle \leftarrow \text{REMYCC}(r_ewma_{\alpha\beta}, s_ewma, rtt_ratio)$
- ▶ $cwnd \leftarrow m \cdot cwnd + b$

Send packet if:

- ▶ $cwnd > \text{FlightSize}$, and
- ▶ last packet sent $> \tau$ ago

Remy's job

Find piecewise-continuous `REMYCC()` that optimizes expected value of objective function

Remy example: 2D state space

On ack:

$$\langle m, b, \tau \rangle \leftarrow \text{REMYCC}(s_ewma, r_ewma_{\alpha}, r_ewma_{\beta}, rtt_ratio)$$

Remy example: 2D state space

On ack:

$$\langle m, b, \tau \rangle \leftarrow \text{REMYCC}(s_ewma, r_ewma_\alpha, \text{[redacted]})$$

Remy example: model

Quantity	Distribution	Units
Link speed	Uniform(10, 20)	Mbps
RTT	Uniform(100, 200)	ms
n	Uniform(1, 16)	
“On” process	$\text{exp}[\mu = 5]$	seconds
“Off” process	same	

Remy example: **mission**

$$\sum_i \log \left[\frac{\text{throughput}_i}{\text{delay}_i} \right]$$

One action for all states. Find the best value.

r_{ewma}

$\langle ?, ?, ? \rangle$

s_{ewma}

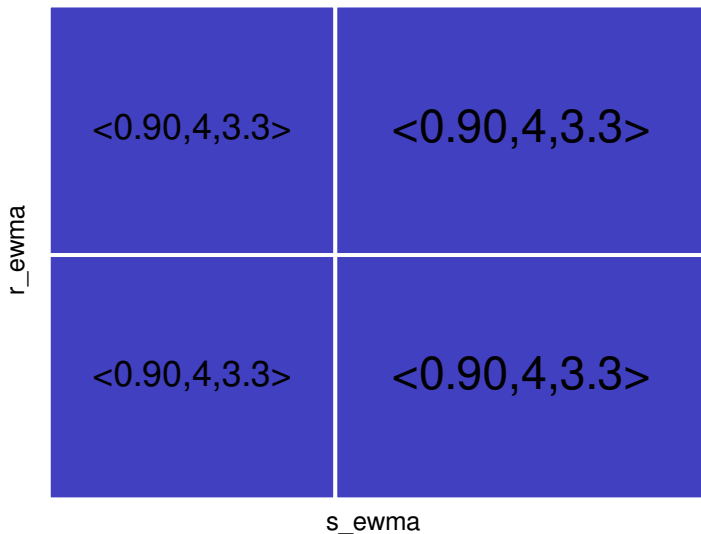
The best (single) action. Now split it on median.

r_ewma

$\langle 0.90, 4, 3.3 \rangle$

s_ewma

Simulate



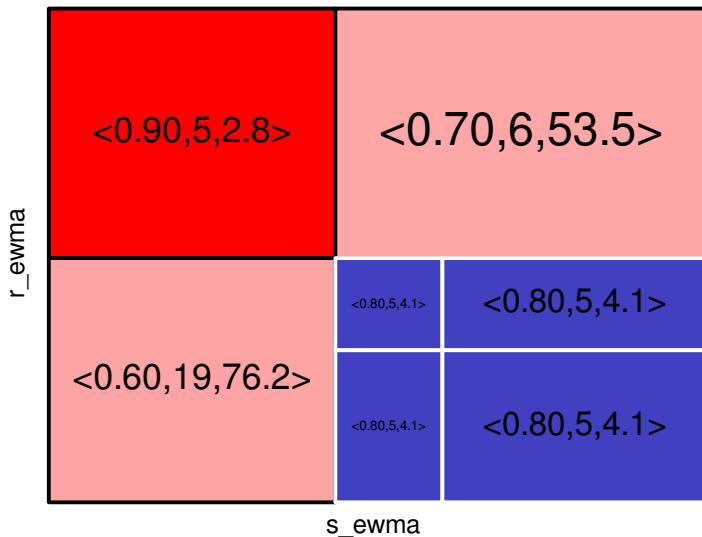
Optimize each of the new actions

r_ewma	<0.90,4,3.3>	<0.90,4,3.3>
	<0.90,4,3.3>	<0.90,4,3.3>
		s_ewma

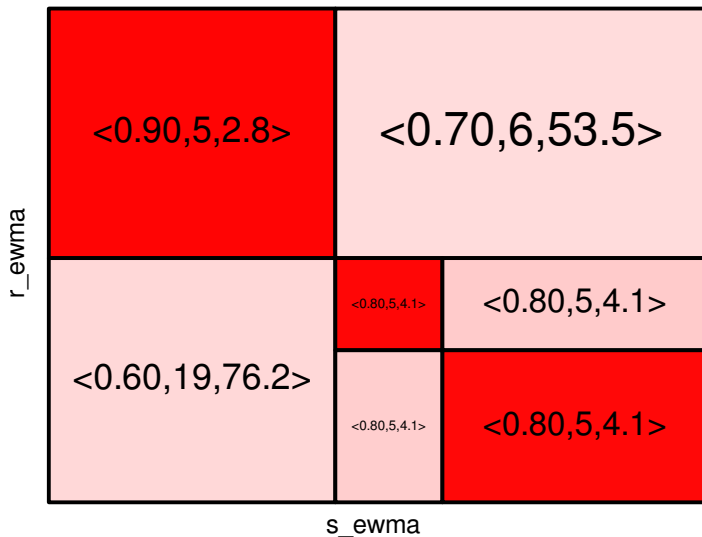
Now split the most-used rule

r_ewma	<0.90,5,2.8>	<0.70,6,53.5>
	<0.60,19,76.2>	<0.80,5,4.1>
	s_ewma	

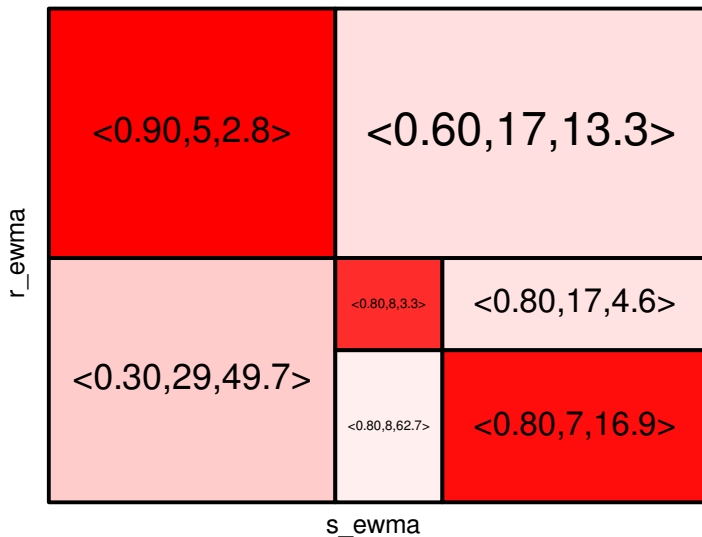
Simulate



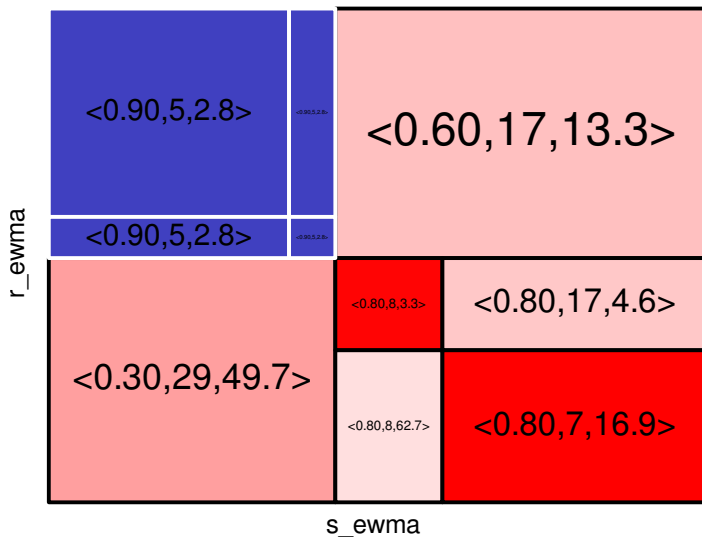
Optimize



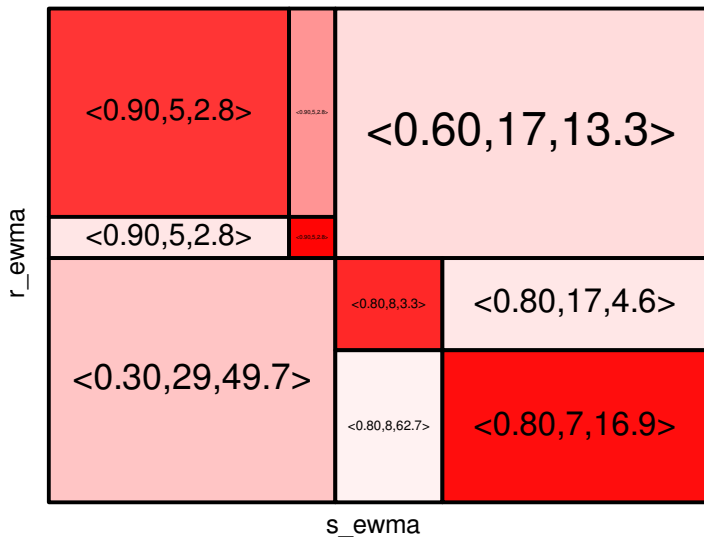
Split



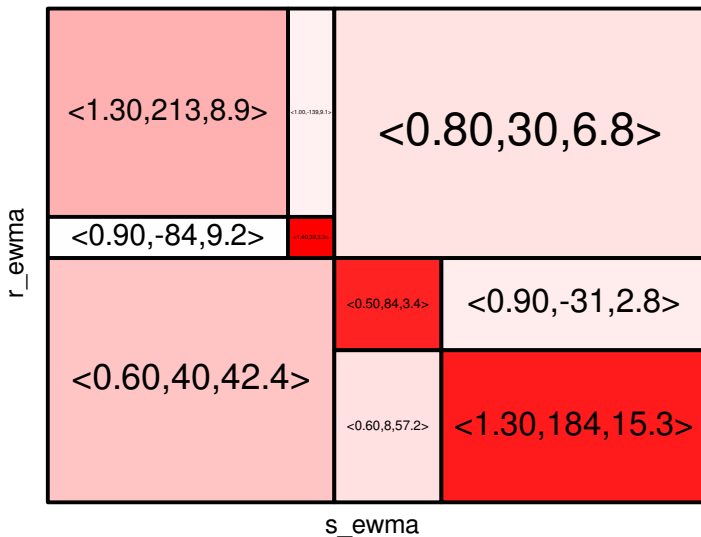
Simulate



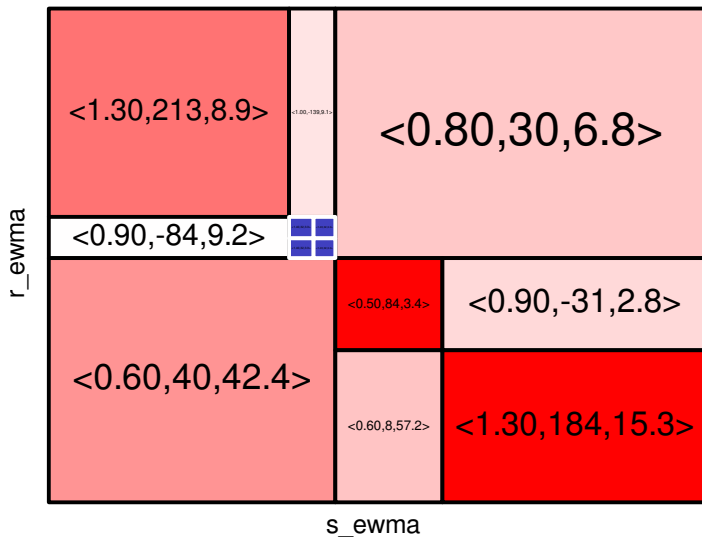
Optimize



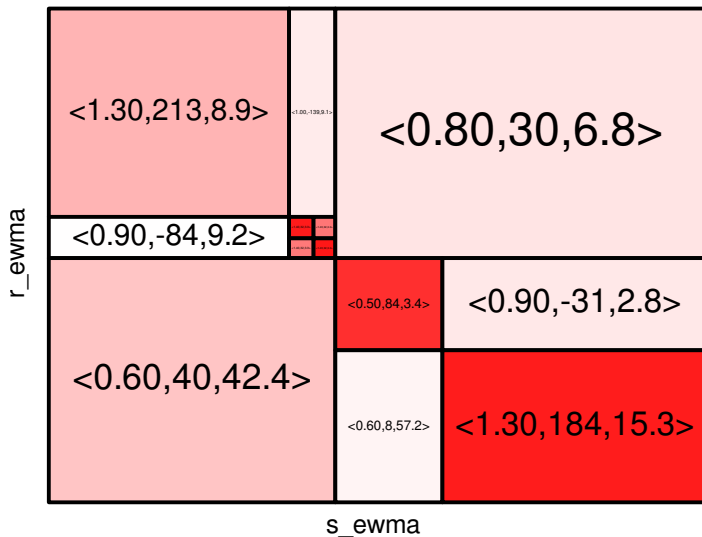
Split



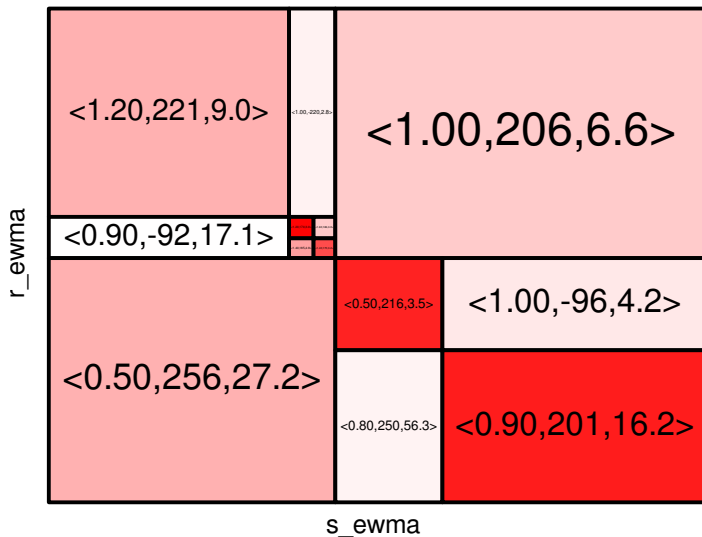
Simulate



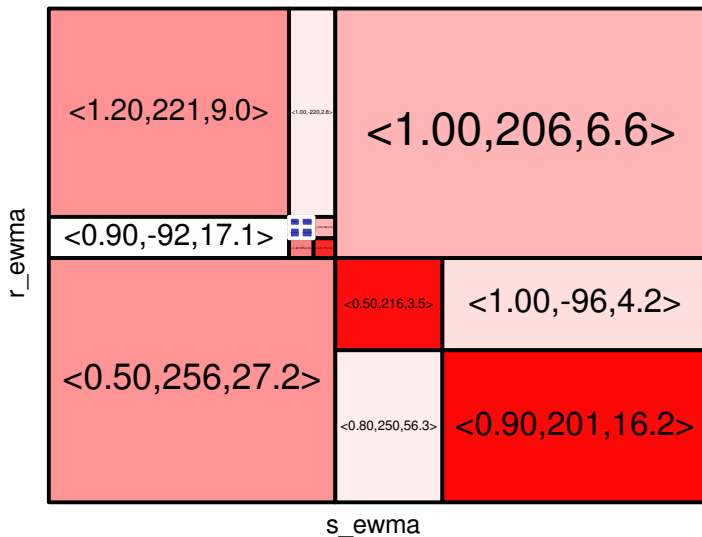
Optimize



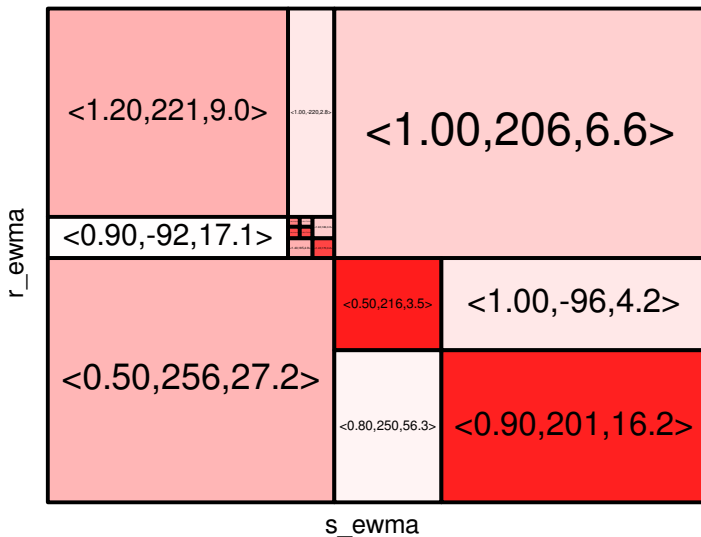
Split



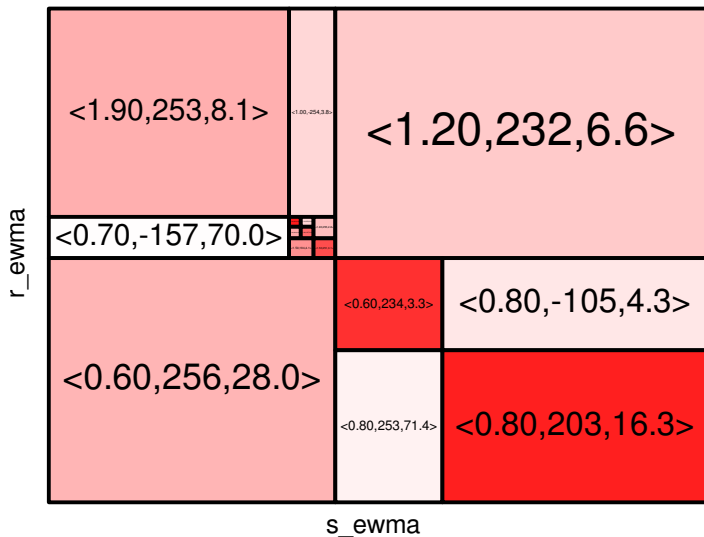
Simulate



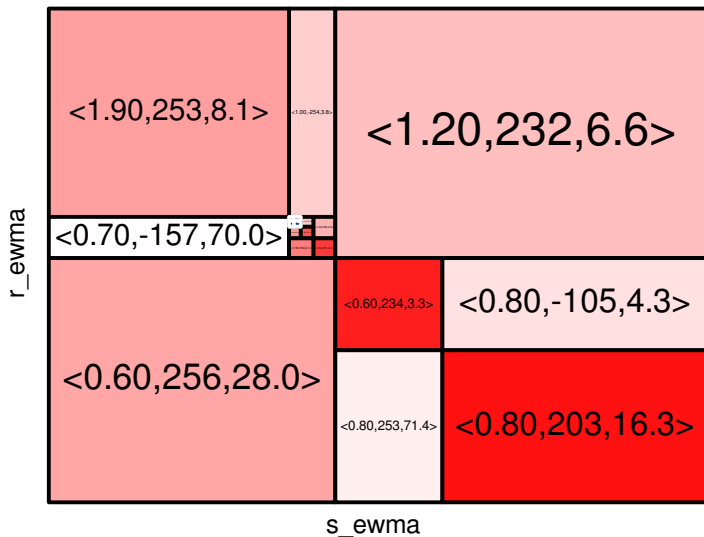
Optimize



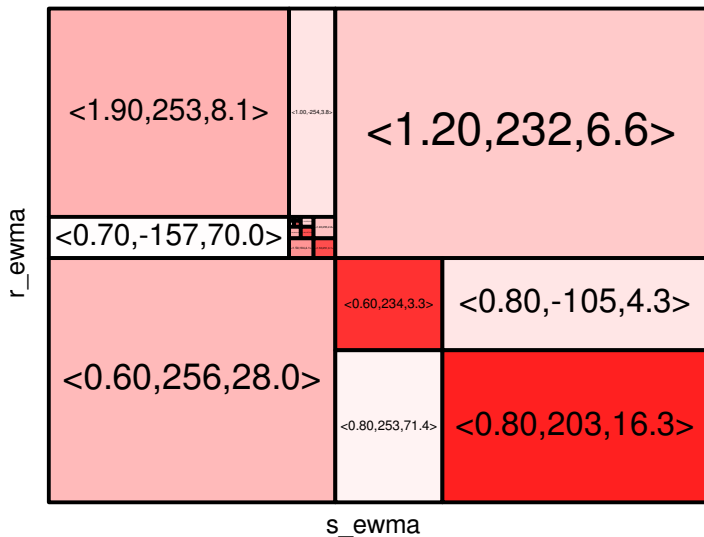
Split



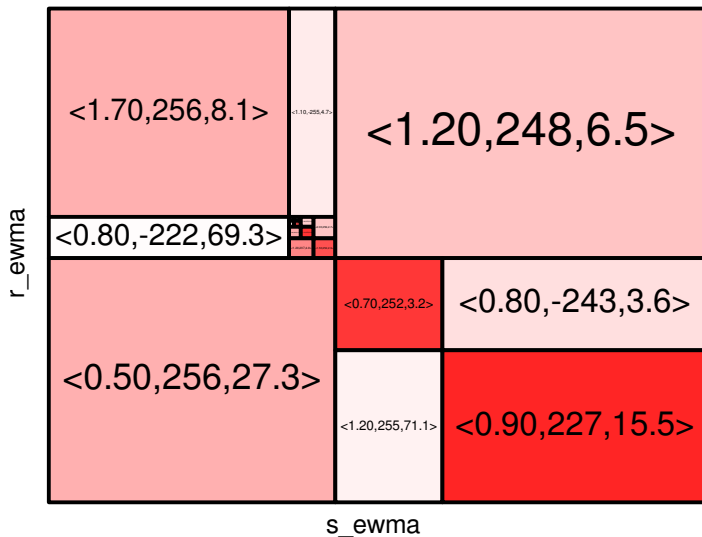
Simulate



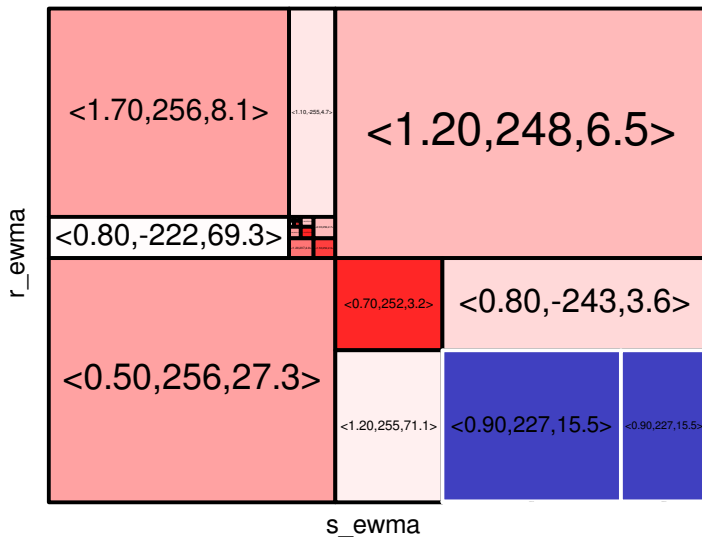
Optimize



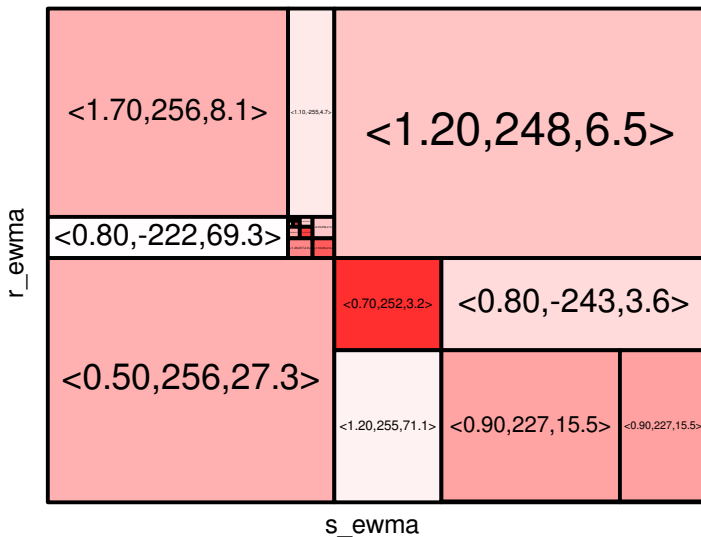
Split



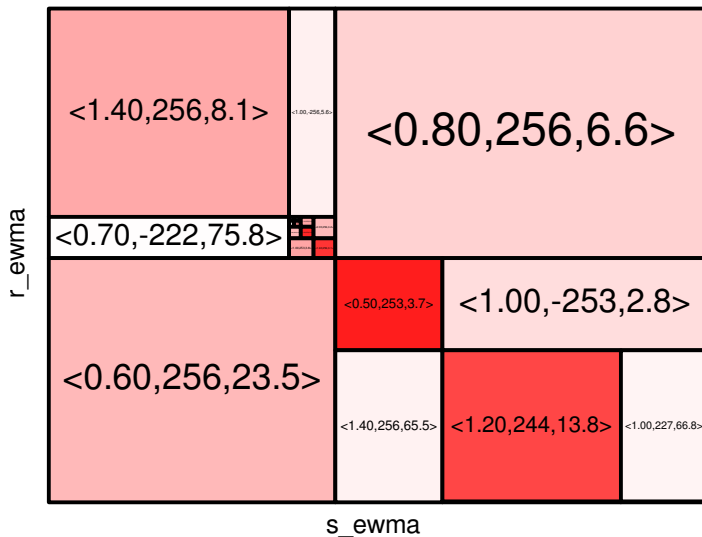
Simulate



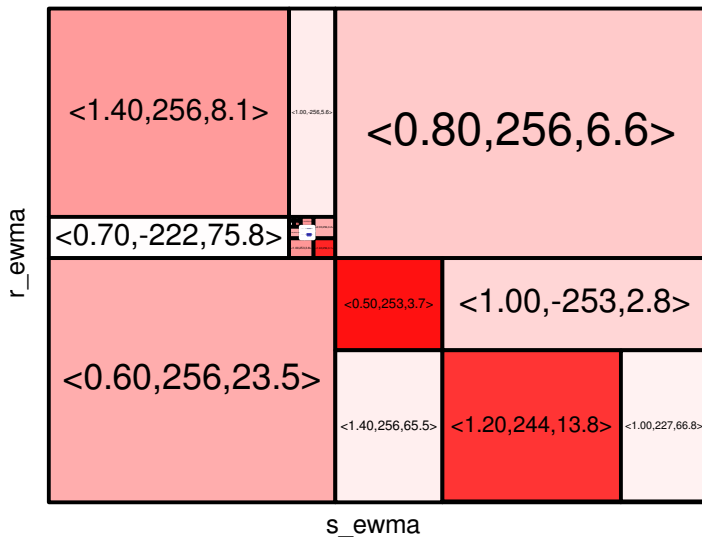
Optimize



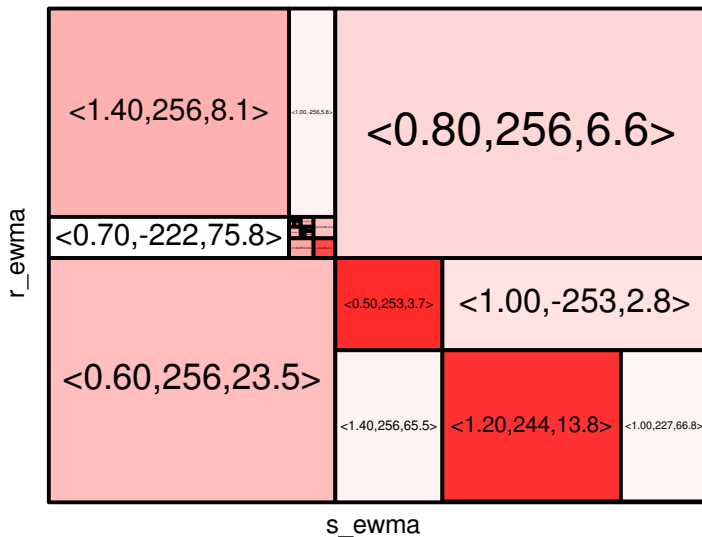
Split



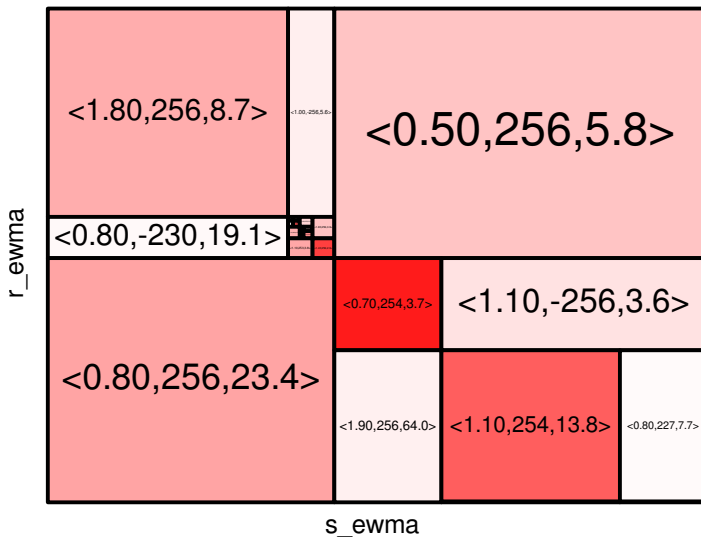
Simulate



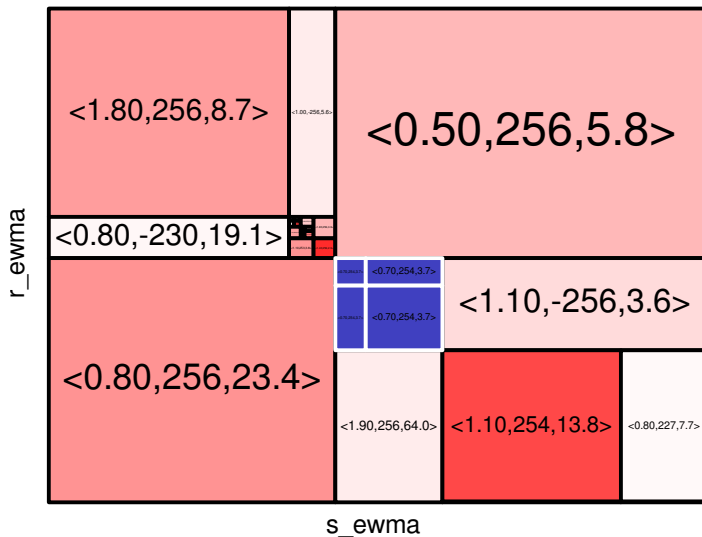
Optimize



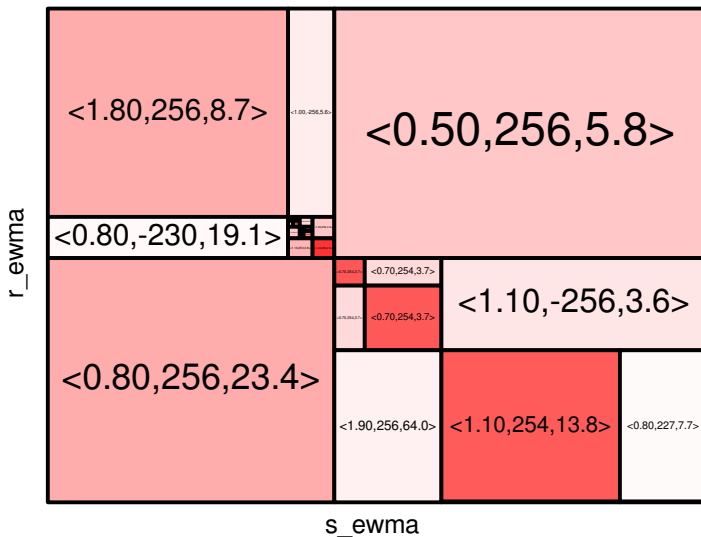
Split



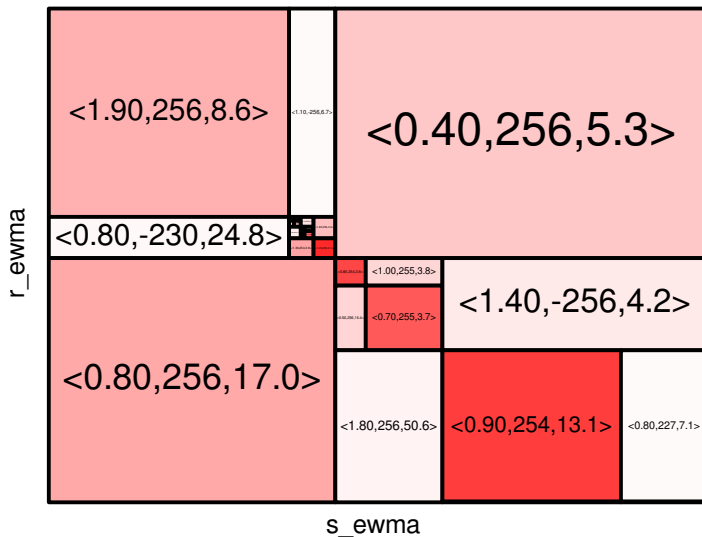
Simulate



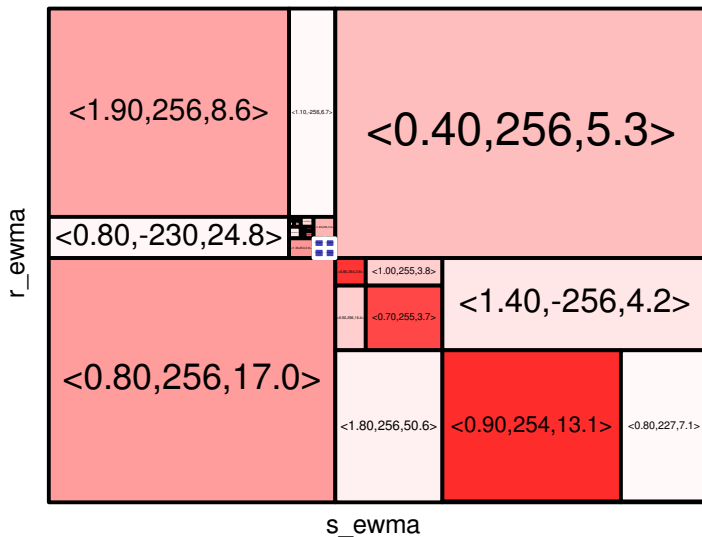
Optimize



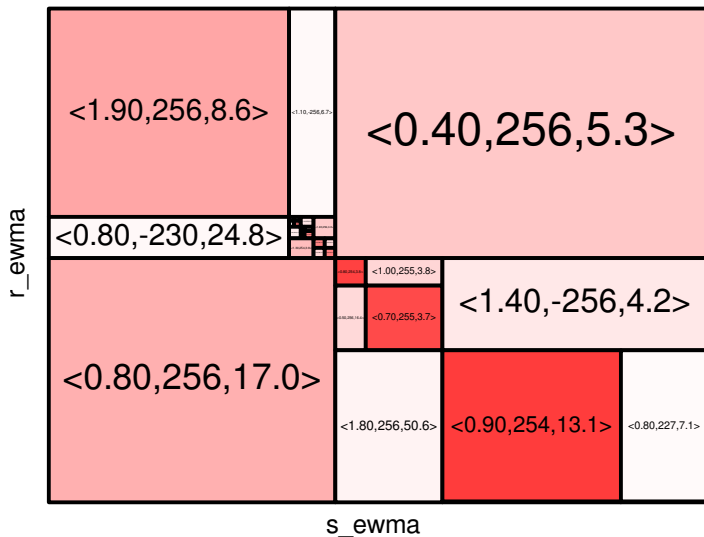
Split



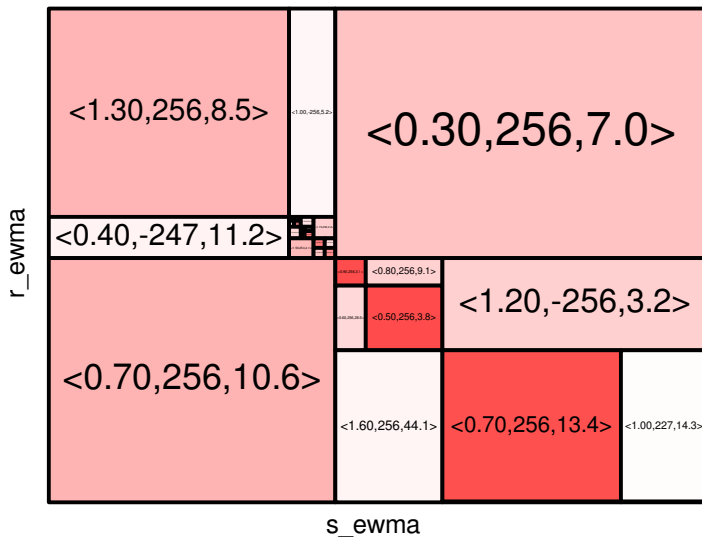
Simulate



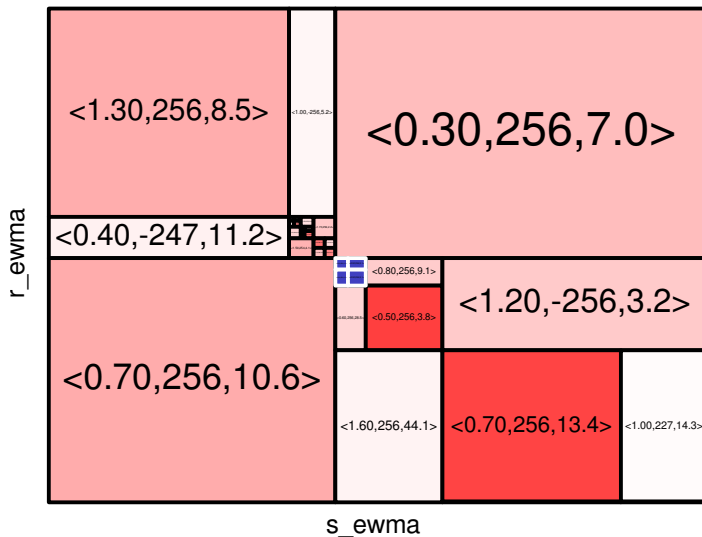
Optimize



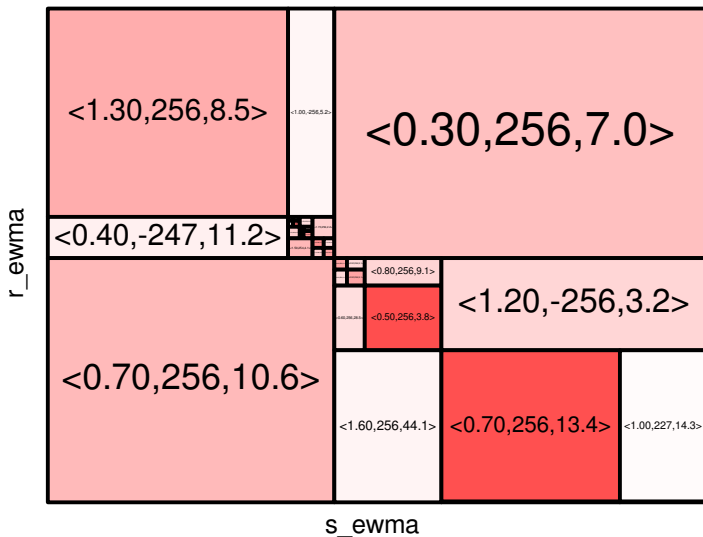
Split



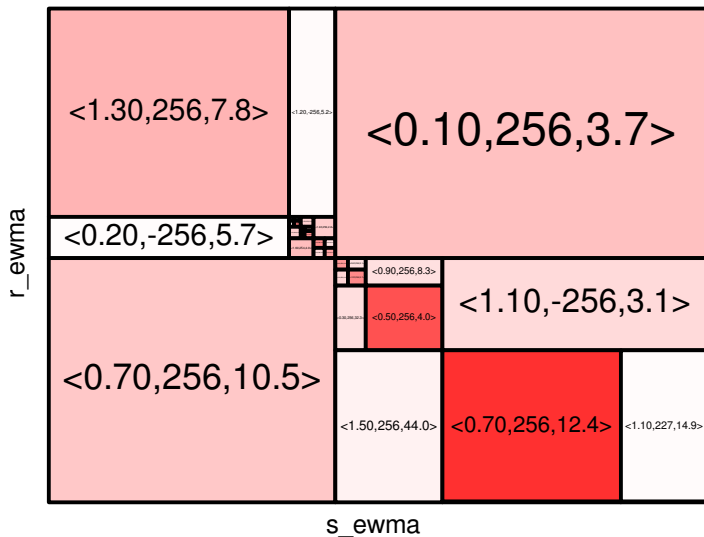
Simulate



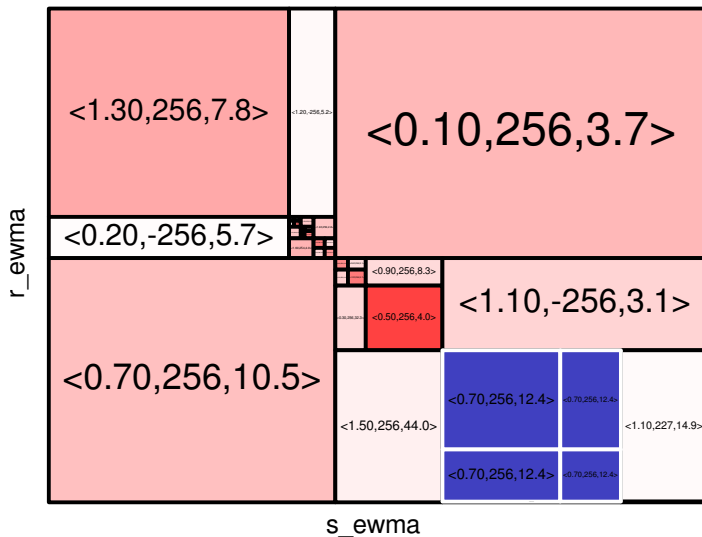
Optimize



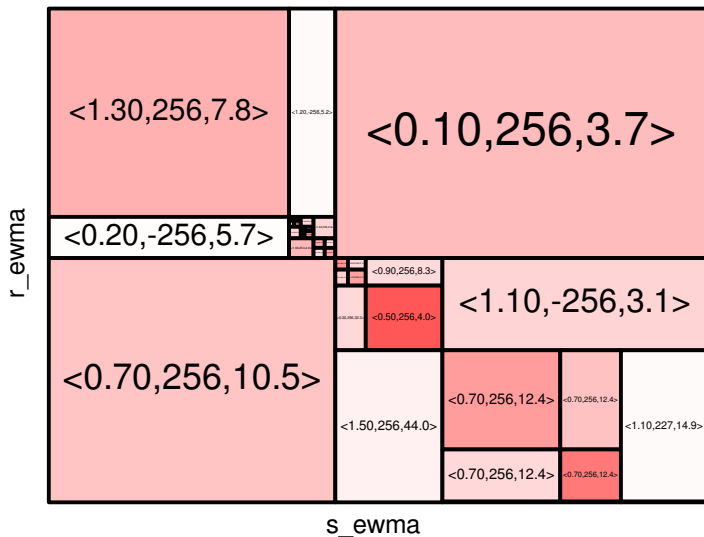
Split



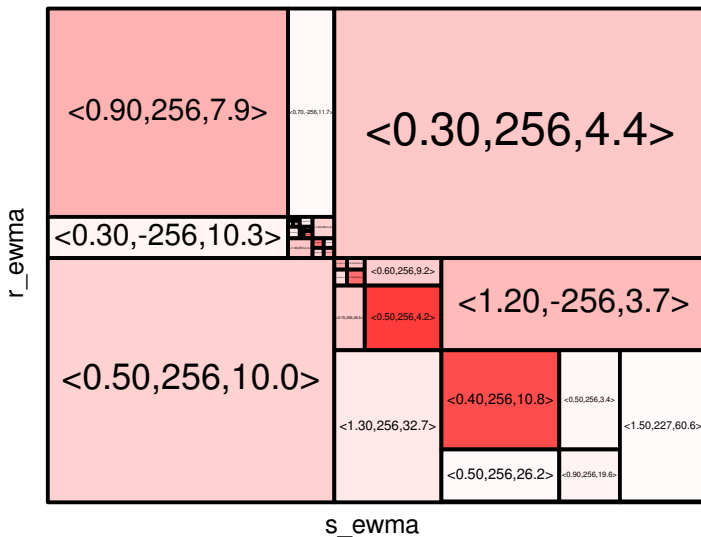
Simulate



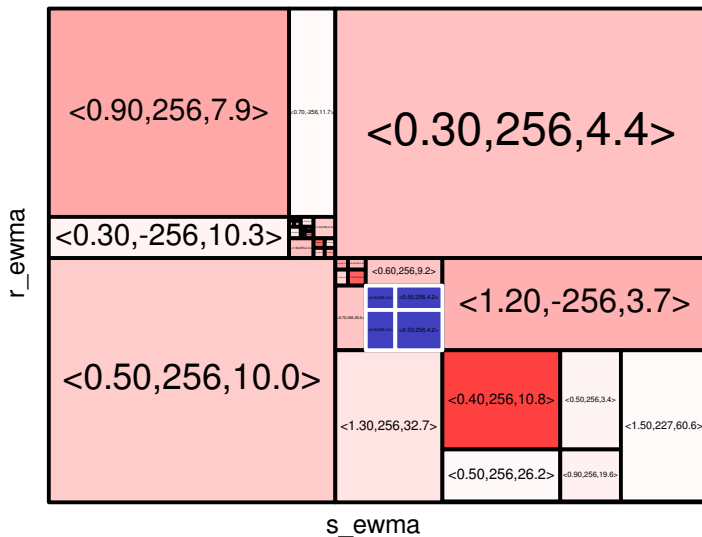
Optimize



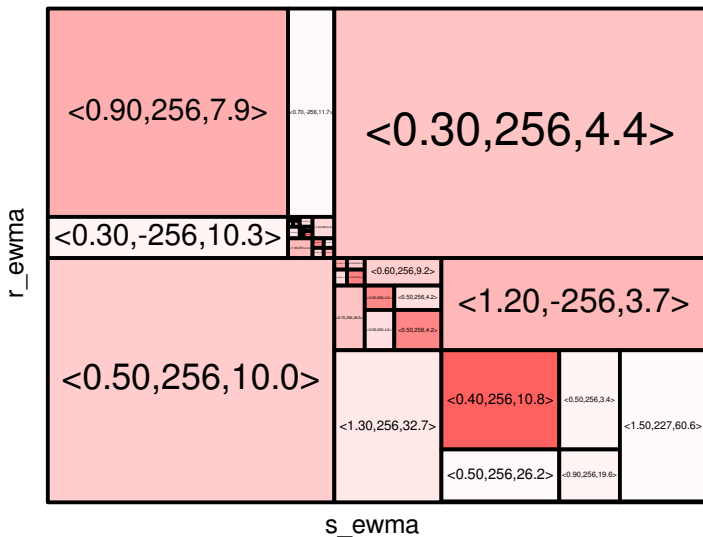
Split



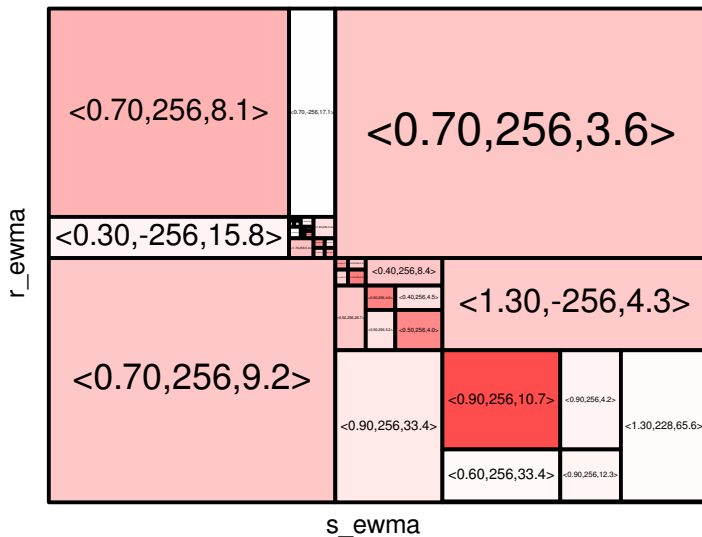
Simulate



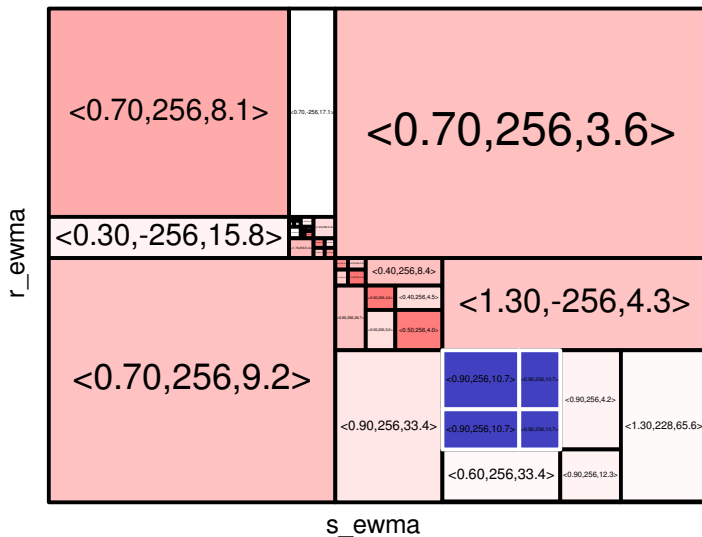
Optimize



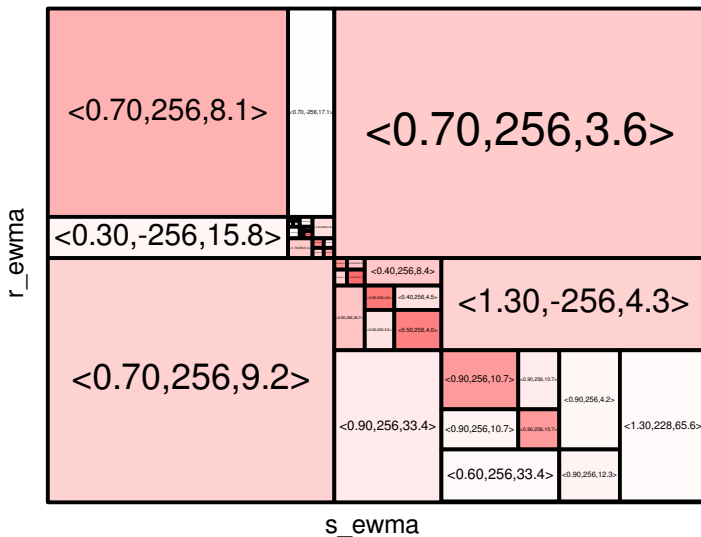
Split



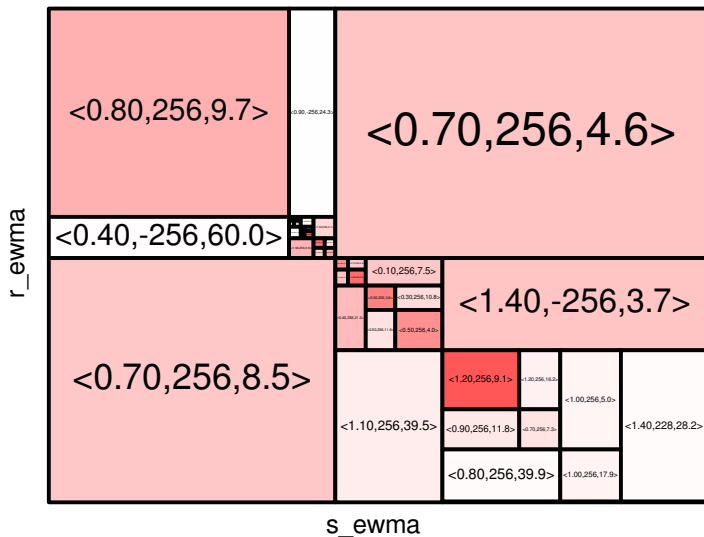
Simulate



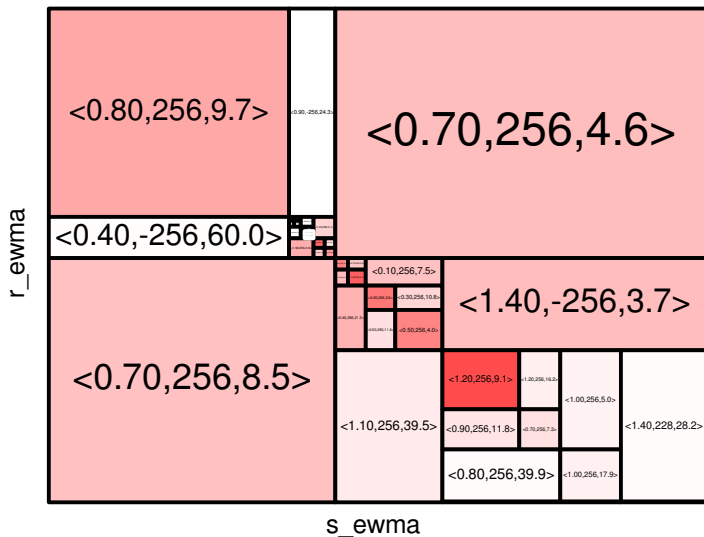
Optimize



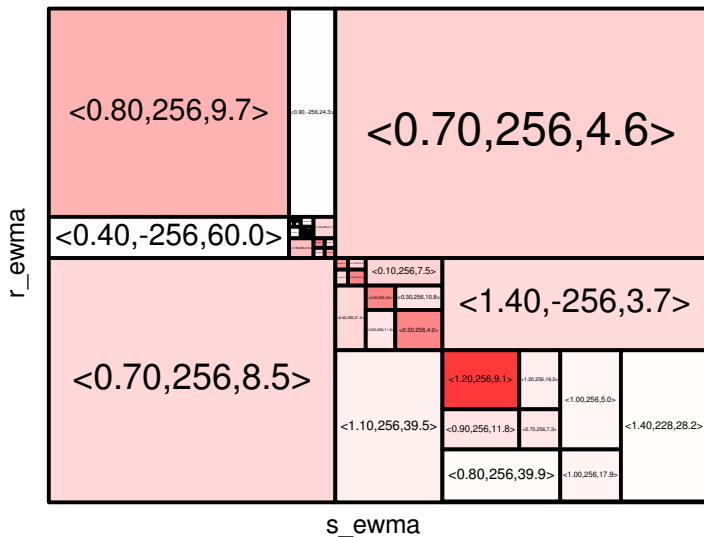
Split



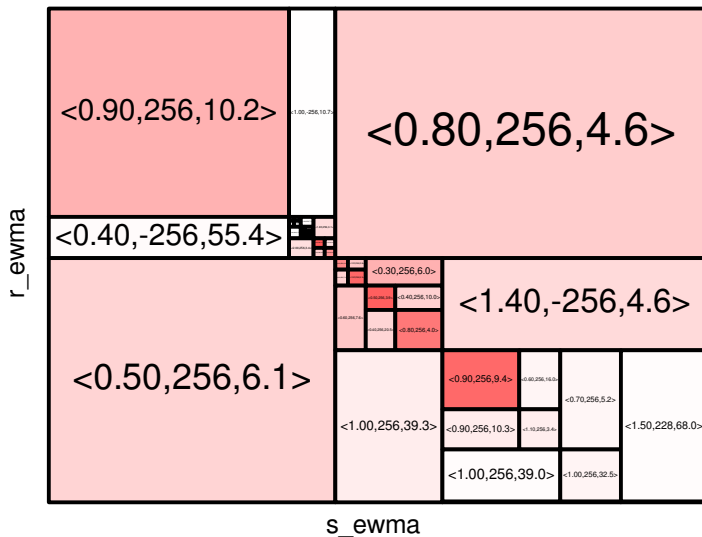
Simulate



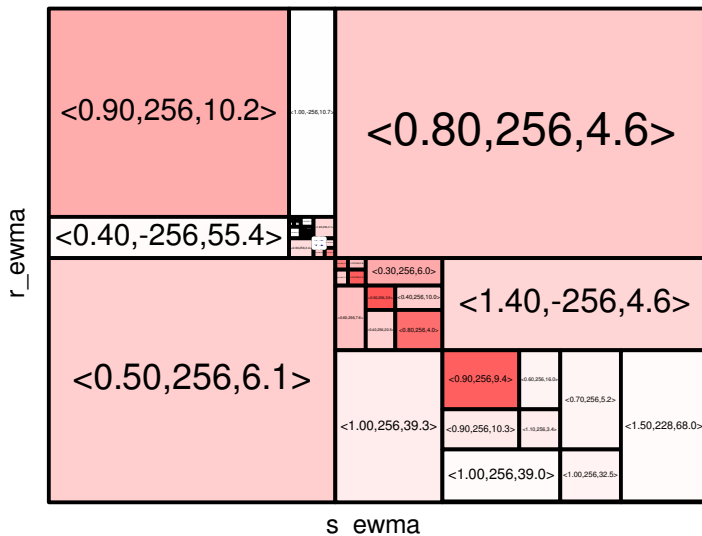
Optimize



Split



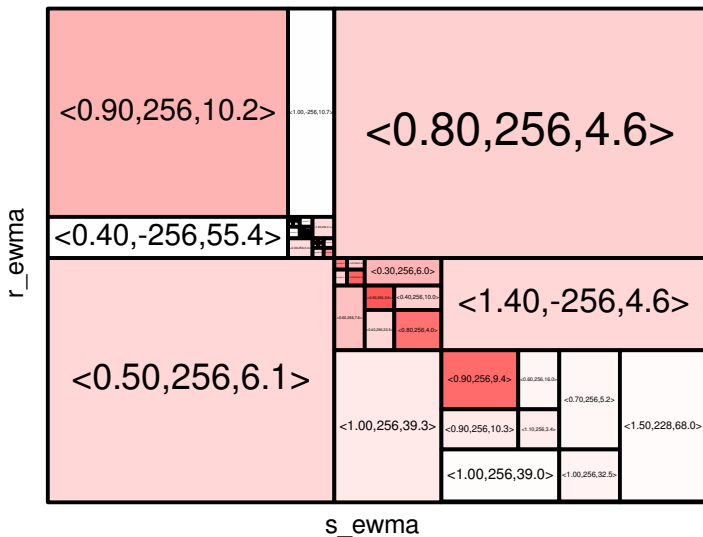
Simulate



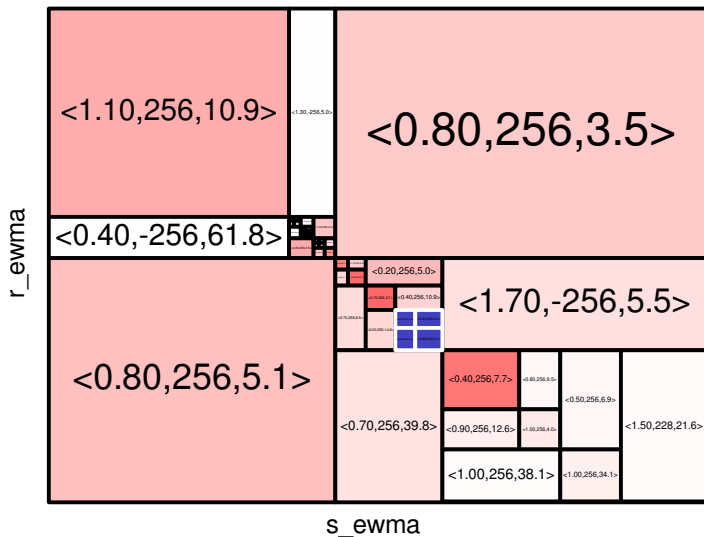
Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

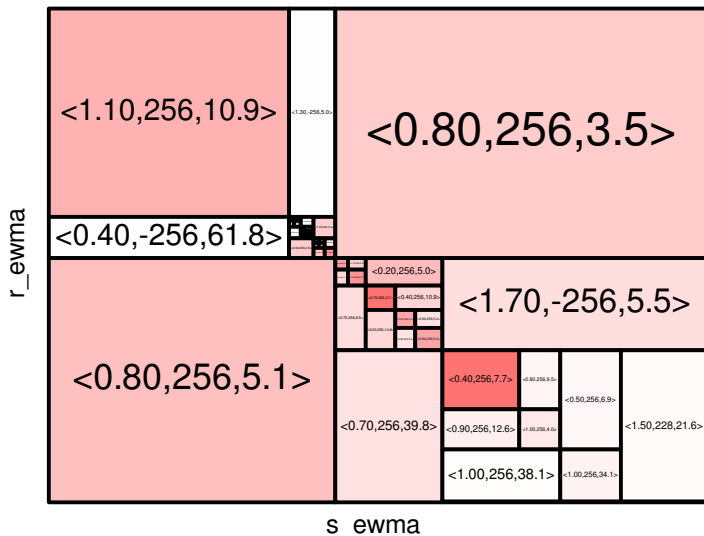
Optimize



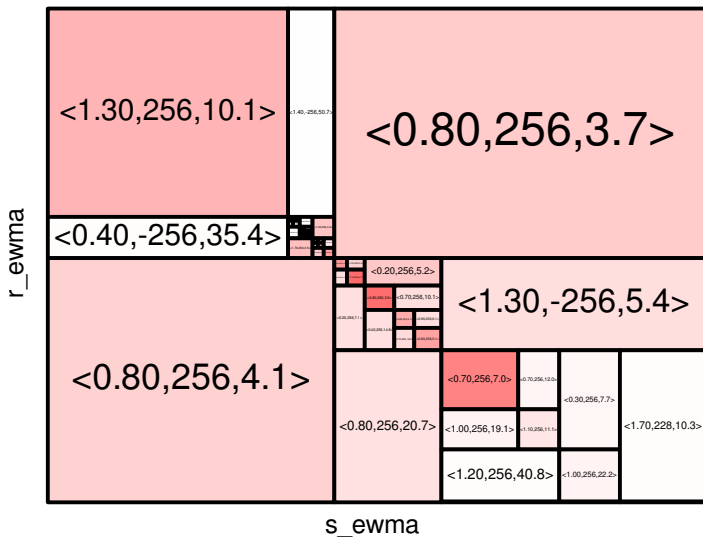
Simulate



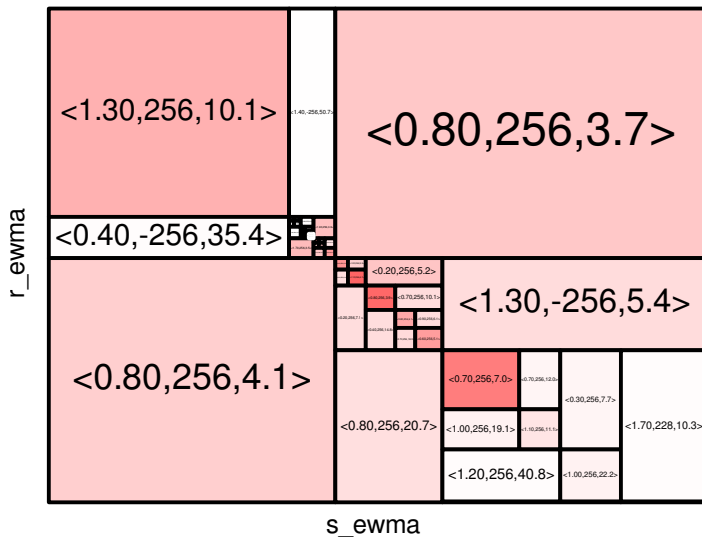
Optimize



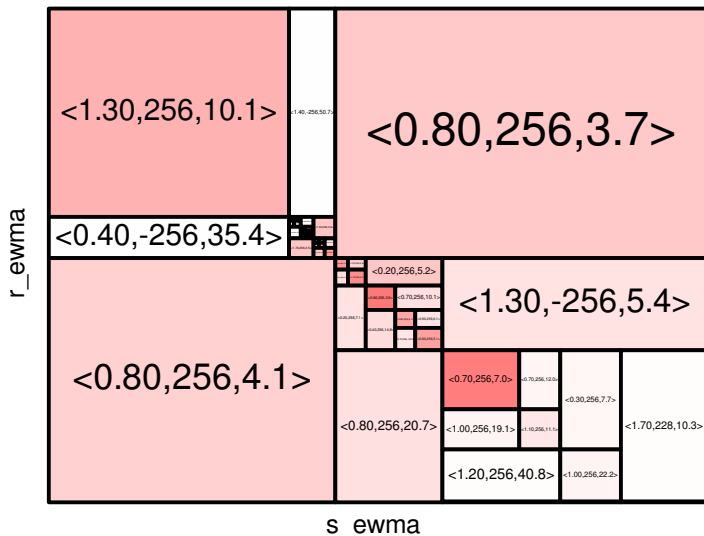
Split



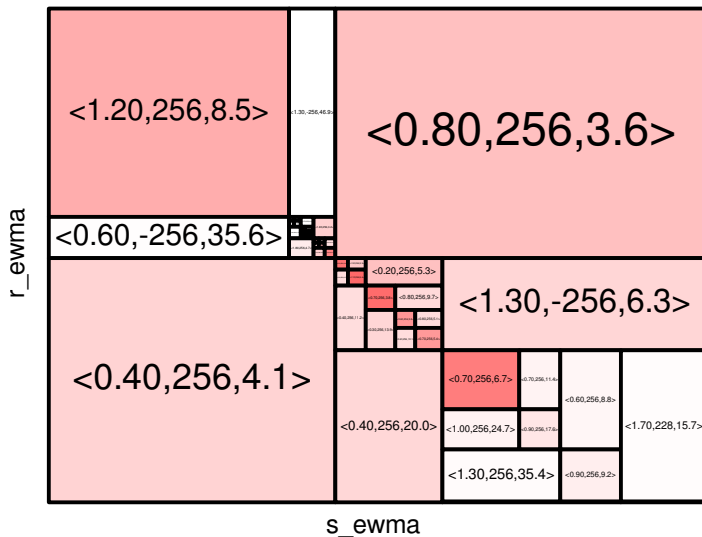
Simulate



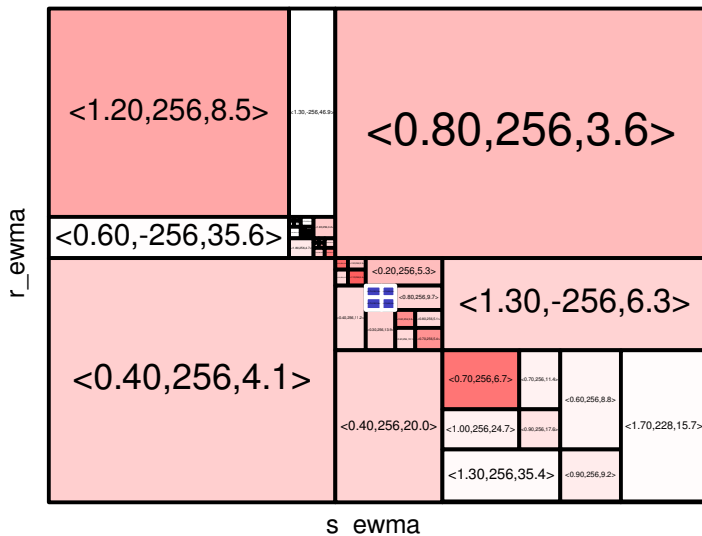
Optimize



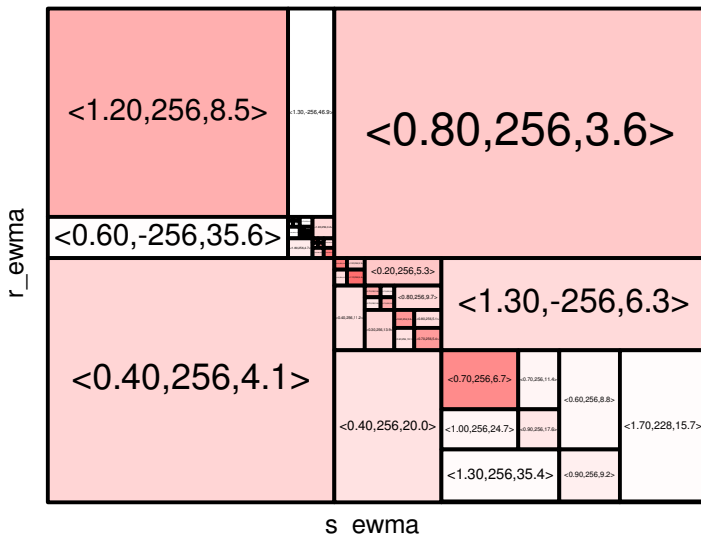
Split



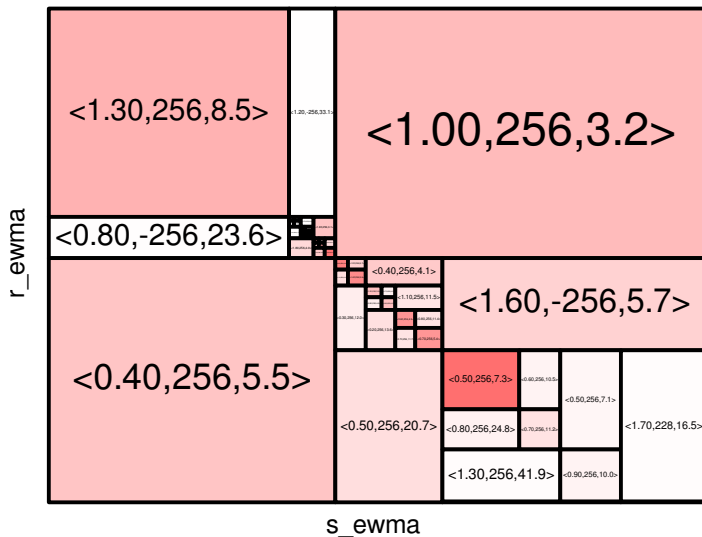
Simulate



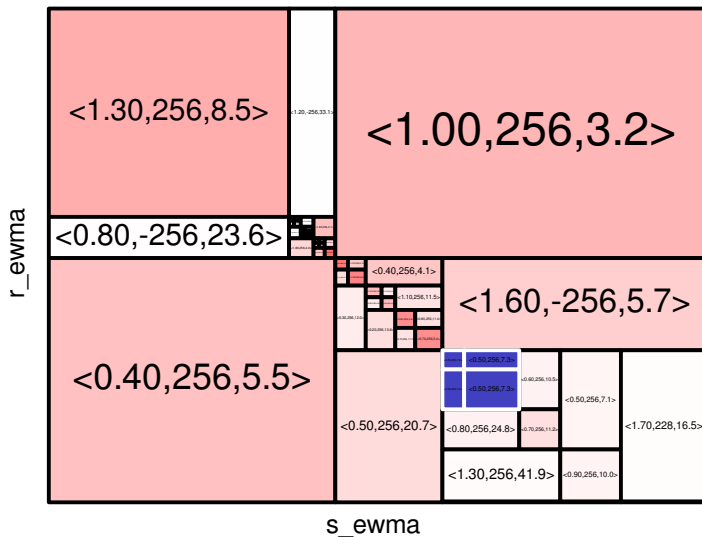
Optimize



Split



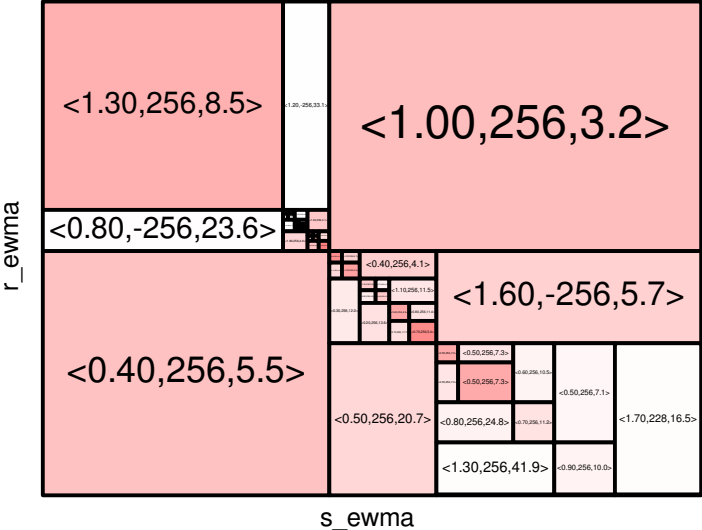
Simulate



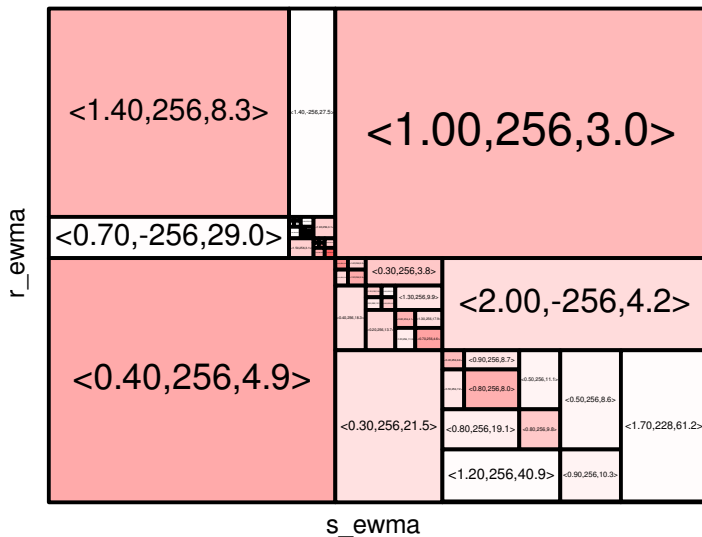
Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

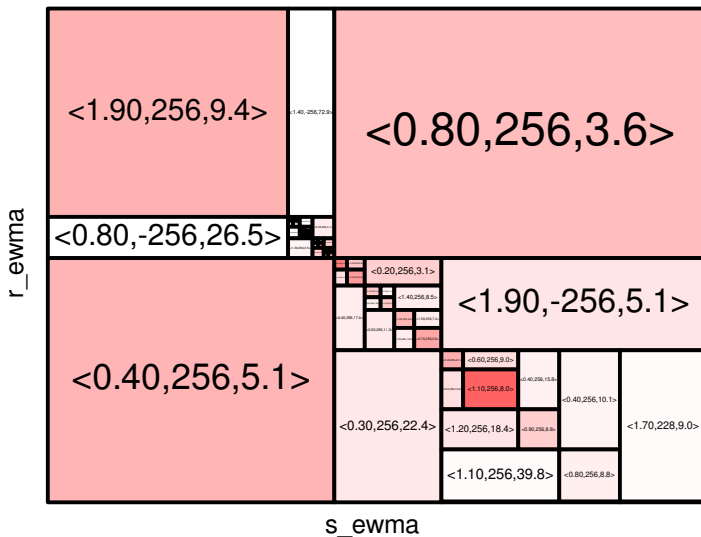
Optimize



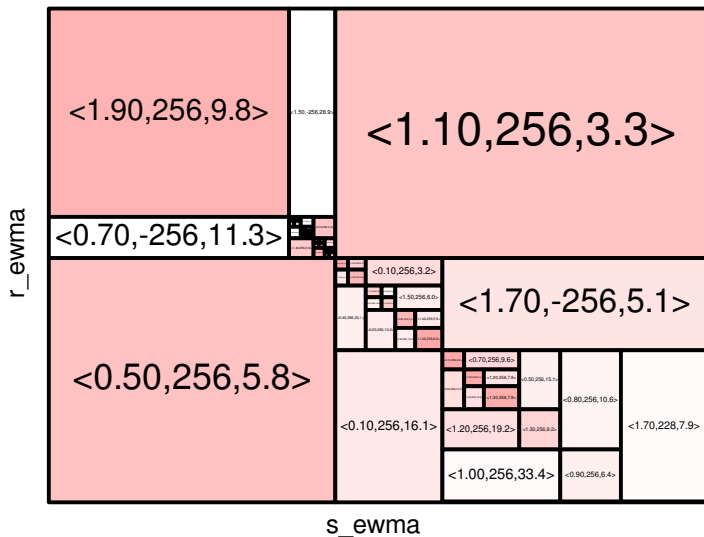
Split



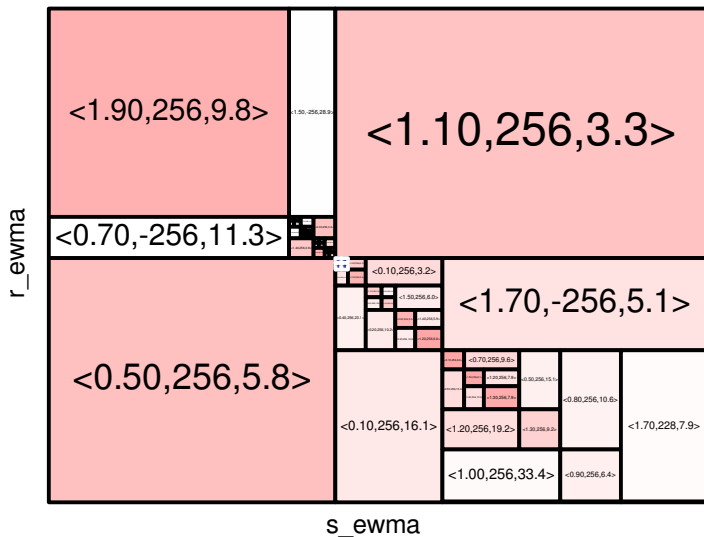
Split



Split



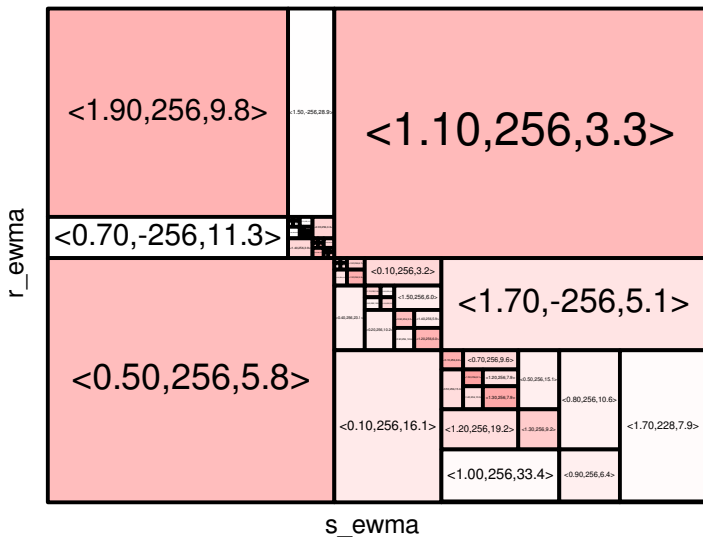
Simulate



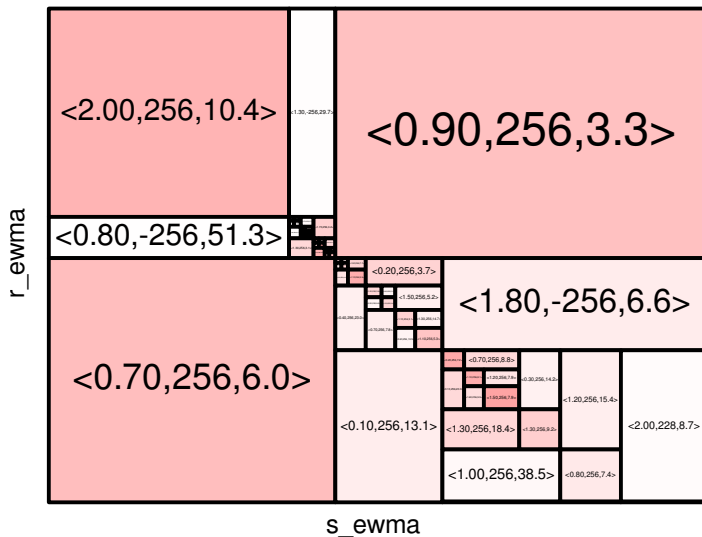
Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

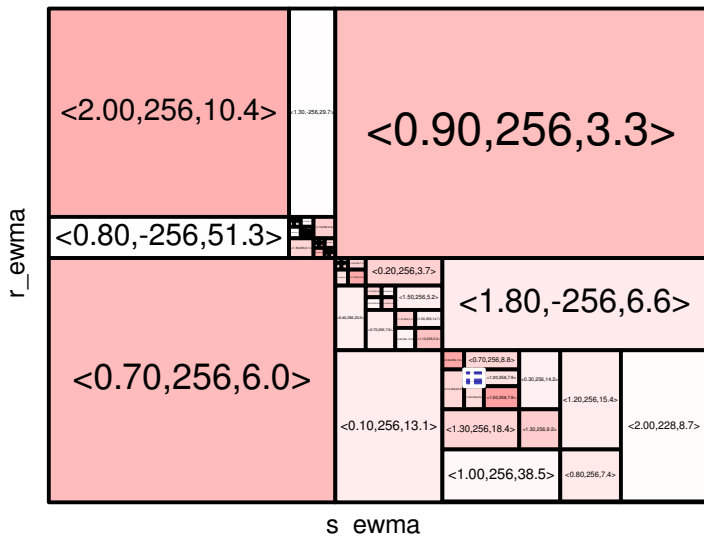
Optimize



Split



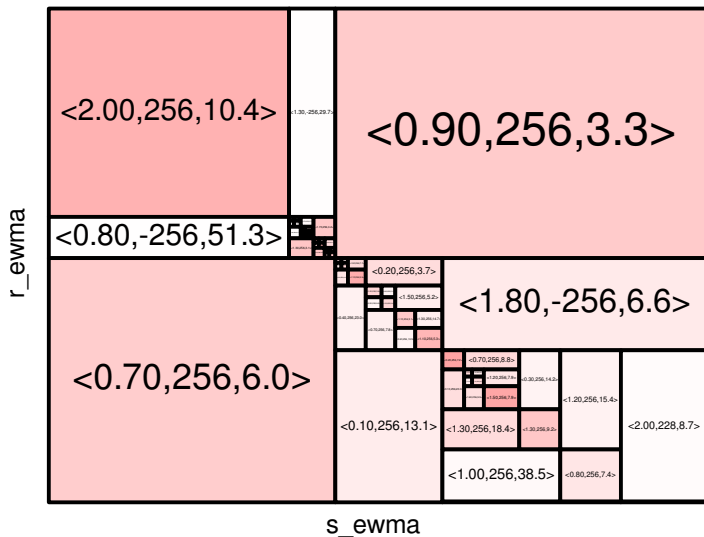
Simulate



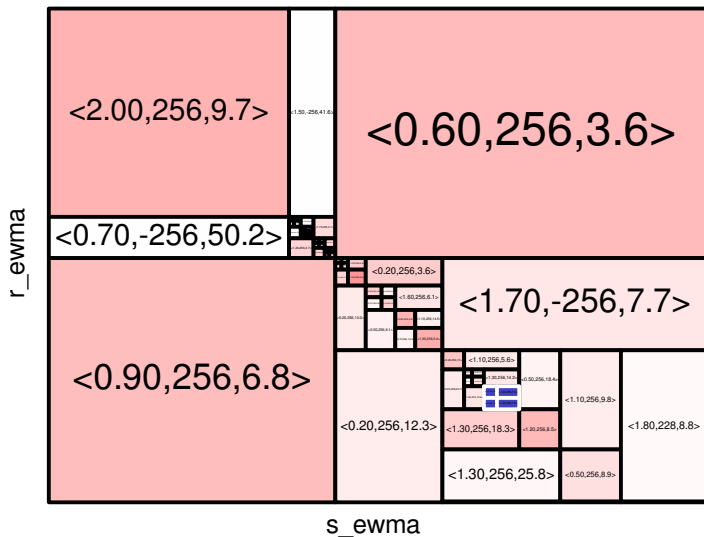
Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

Optimize



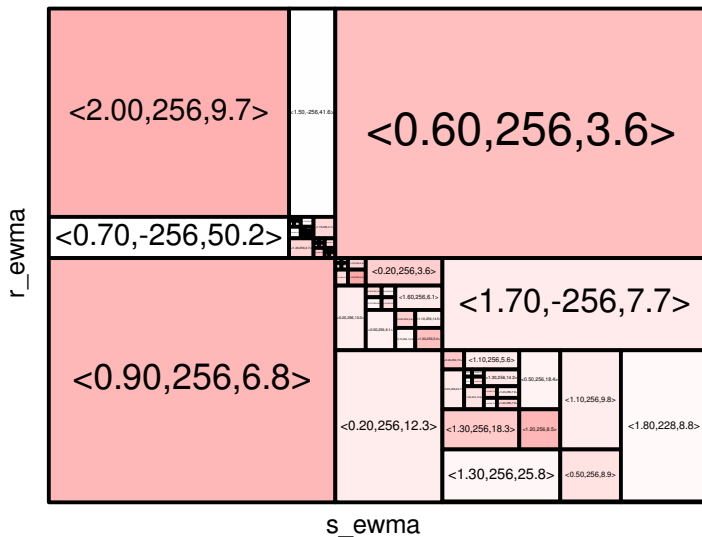
Simulate

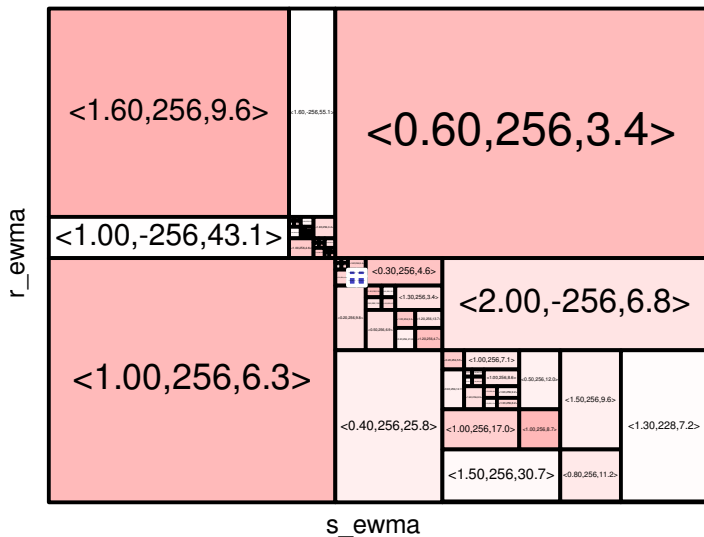


Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

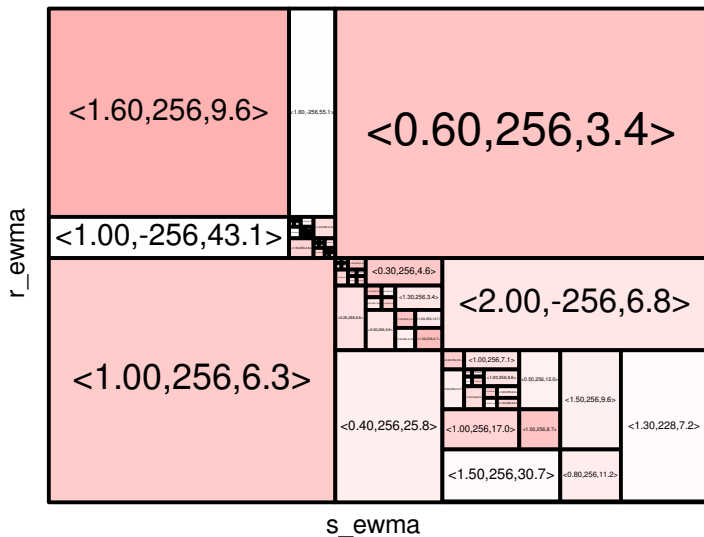
TCP ex Machina: Computer-Generated Congestion Control

RemyCC



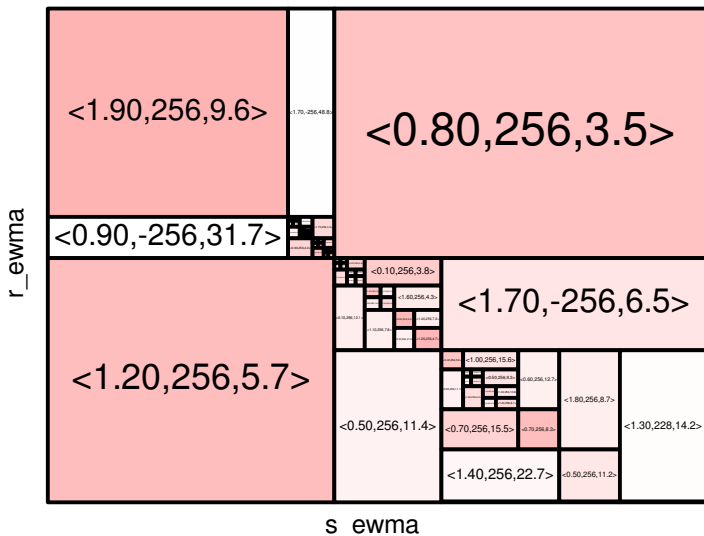


RemyCC

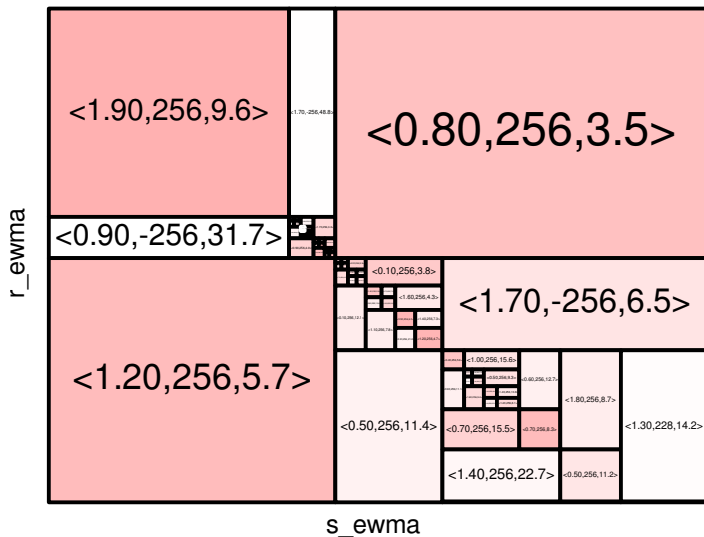


Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control



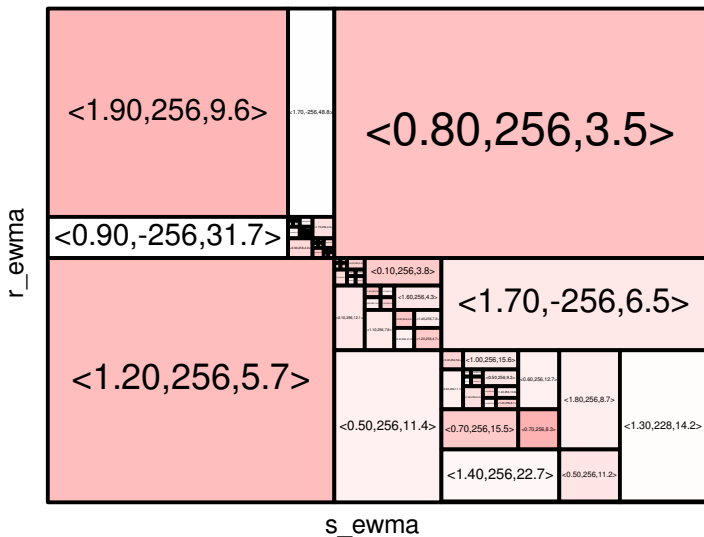
RemyCC



Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

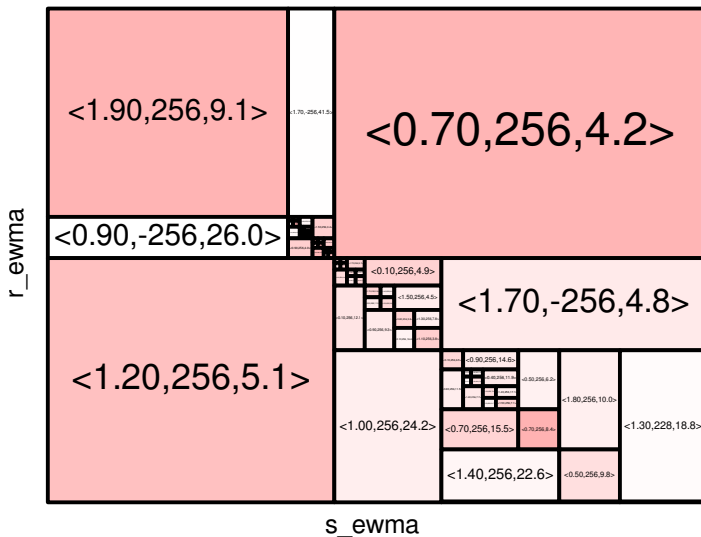
RemyCC



Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

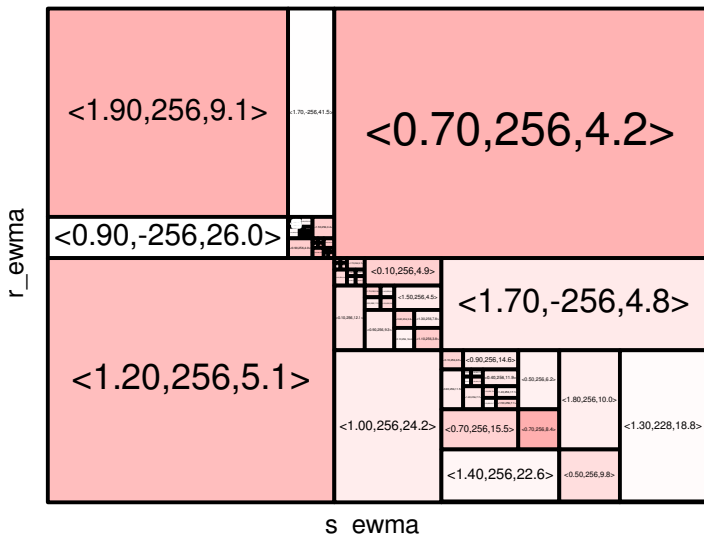
RemyCC



Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

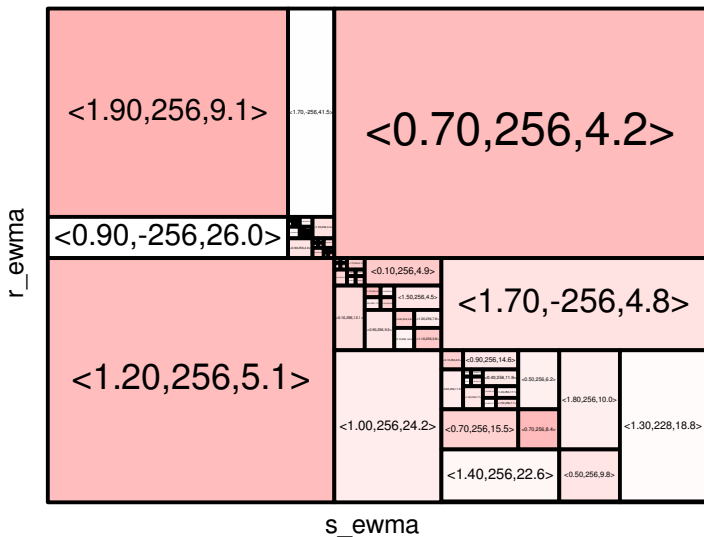
TCP ex Machina: Computer-Generated Congestion Control

RemyCC

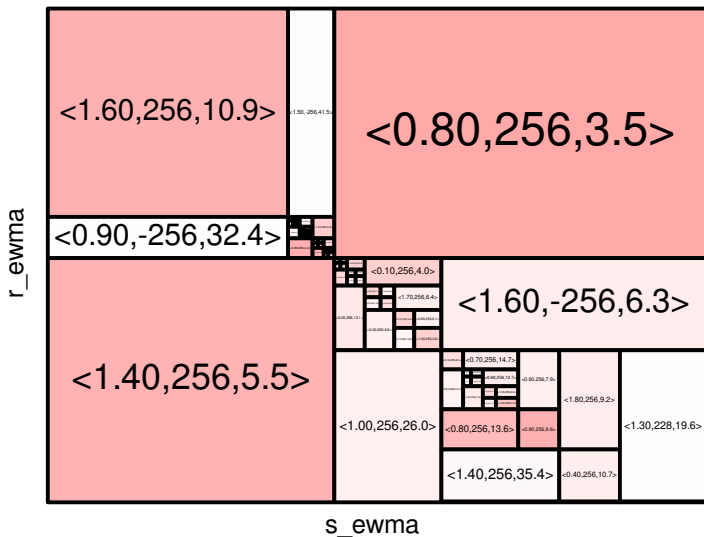


Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

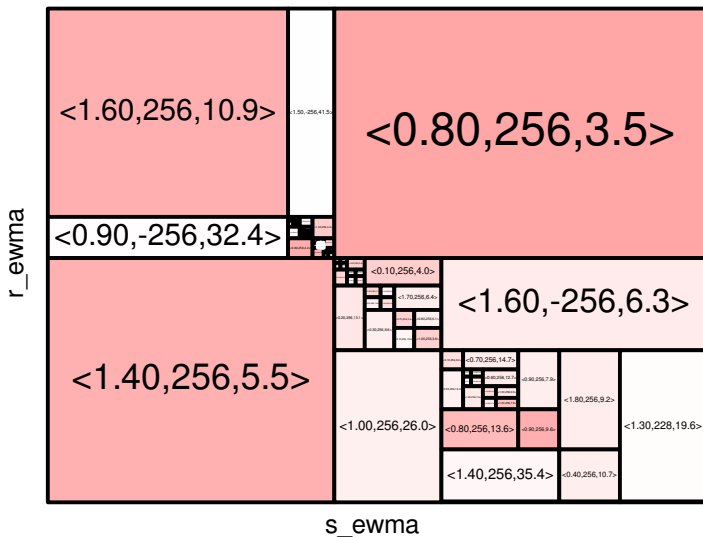
TCP ex Machina: Computer-Generated Congestion Control



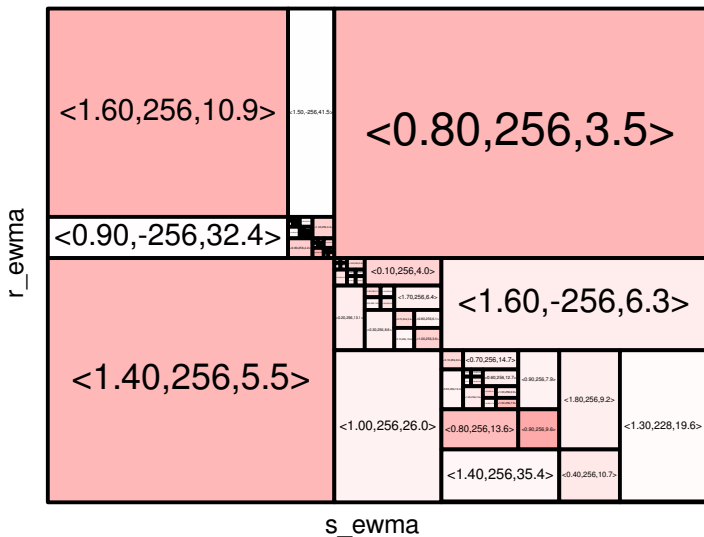
RemyCC



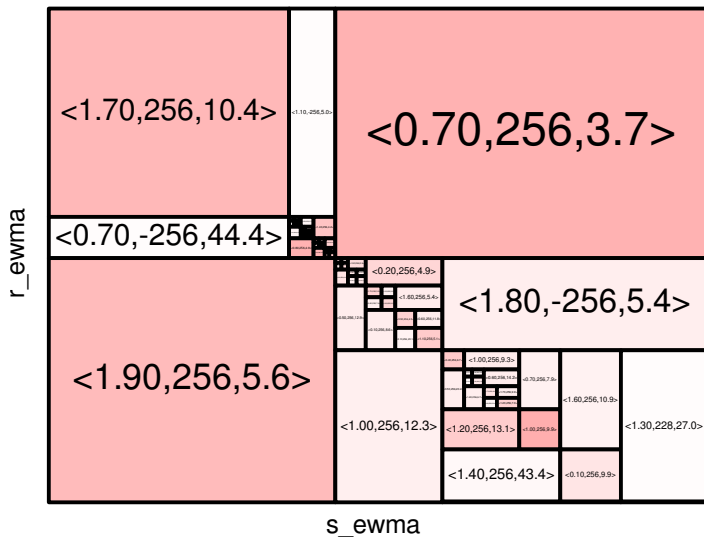
RemyCC



RemyCC



RemyCC

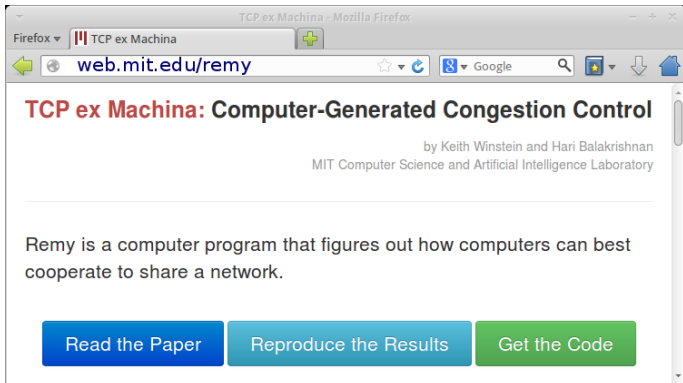


Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

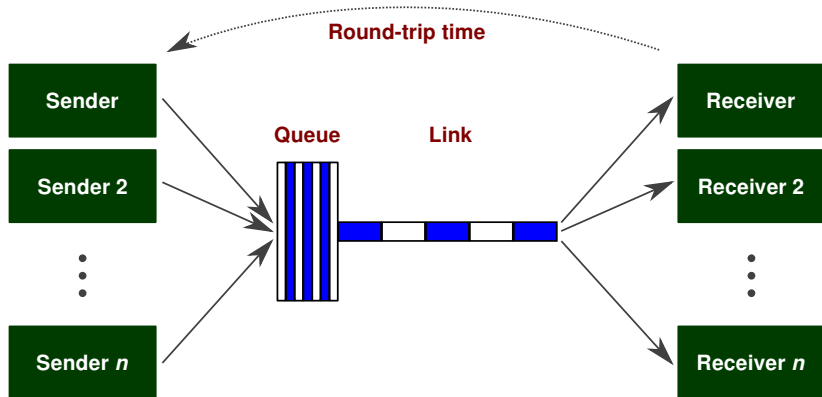
TCP ex Machina: Computer-Generated Congestion Control

Evaluation in ns-2

- ▶ End-to-end comparators: **NewReno, Cubic, Compound, Vegas**
- ▶ In-net comparators: **Cubic-over-sfqCoDel, XCP**
- ▶ Simulation setup published for replication



Scenario 1: fixed-rate network, homogenous senders



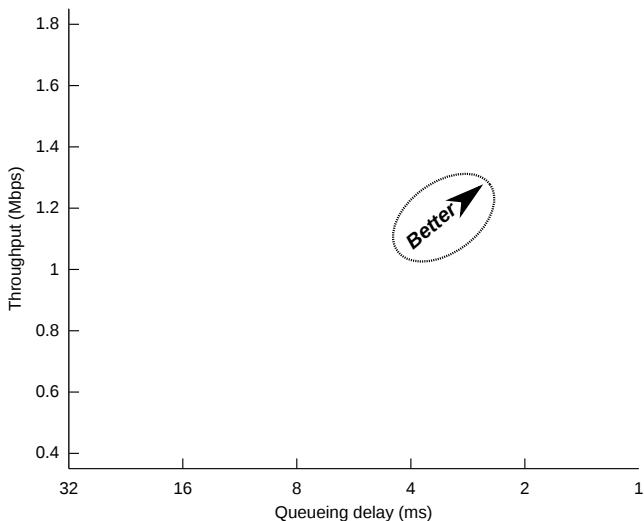
Scenario 1: details

Quantity	Simulation parameter	Remy assumptions
Link speed	15 Mbps	Uniform(10, 20) Mbps
RTT	150 ms	Uniform(100, 200) ms
n	8	Uniform(1, 16)
“On” process	$\exp[\mu = 100]$ kB	$\exp[\mu = 5]$ s
“Off” process	$\exp\left[\mu = \frac{1}{2}\right]$ s	$\exp[\mu = \mathbf{5}]$ s

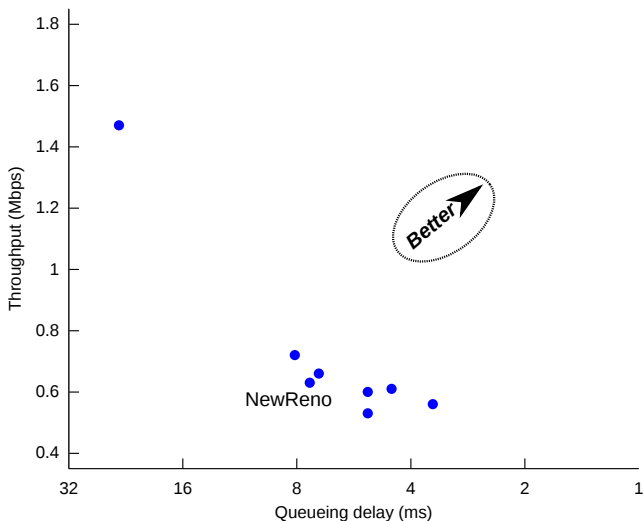
Remy objective:

$$\sum_i \log \left[\frac{\text{throughput}_i}{(\text{delay}_i)^\delta} \right]$$
$$\delta \in \left\{ \frac{1}{10}, 1, 10 \right\}$$

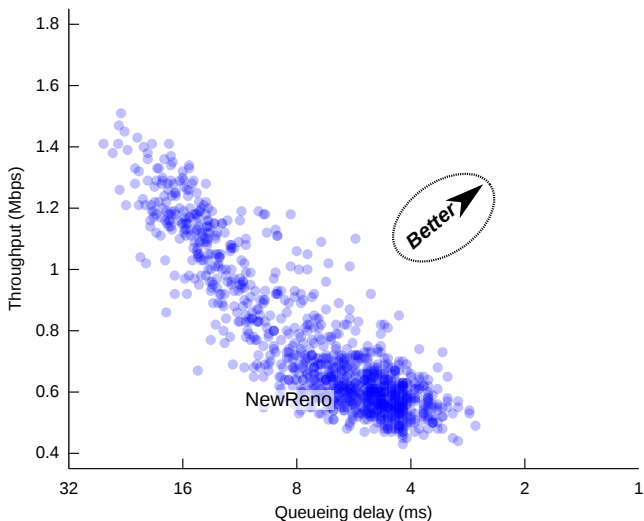
Scenario 1: throughput-delay plot



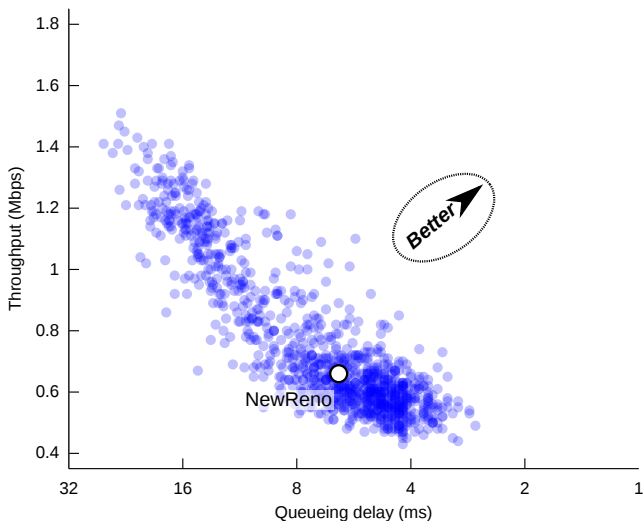
Scenario 1: throughput-delay plot



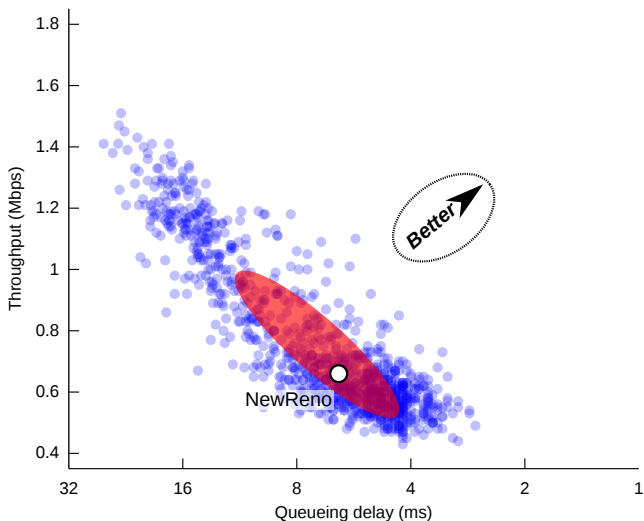
Scenario 1: throughput-delay plot



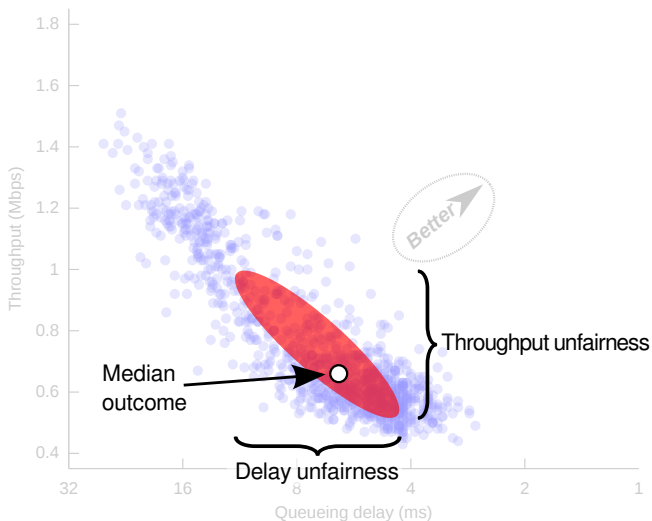
Scenario 1: throughput-delay plot



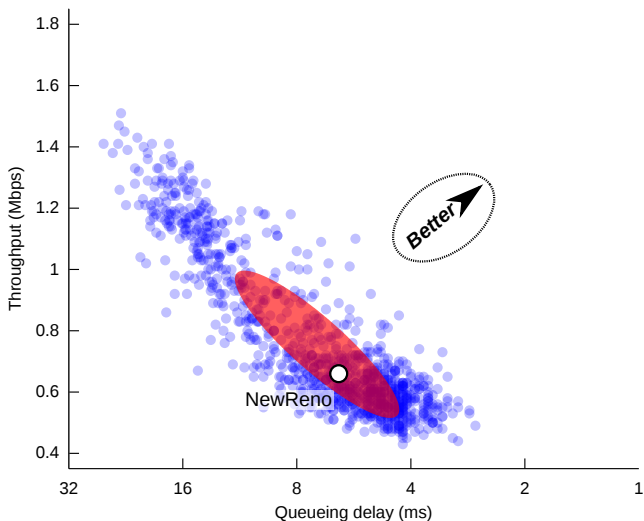
Scenario 1: throughput-delay plot



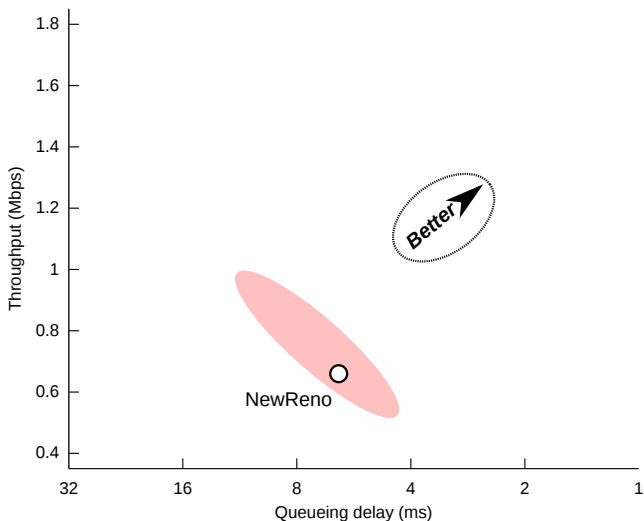
Scenario 1: throughput-delay plot



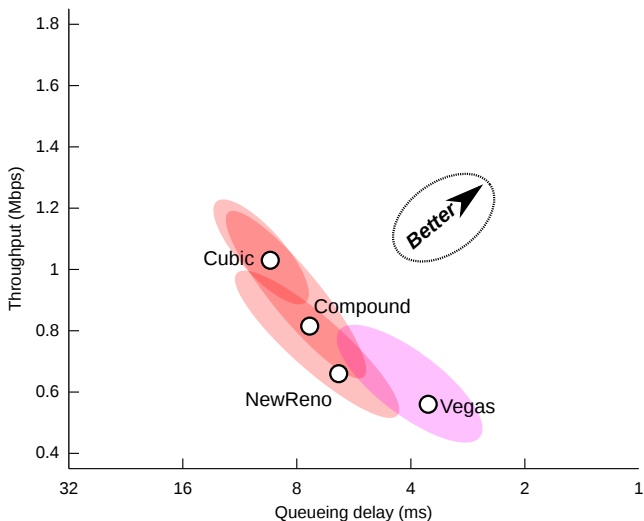
Scenario 1: throughput-delay plot



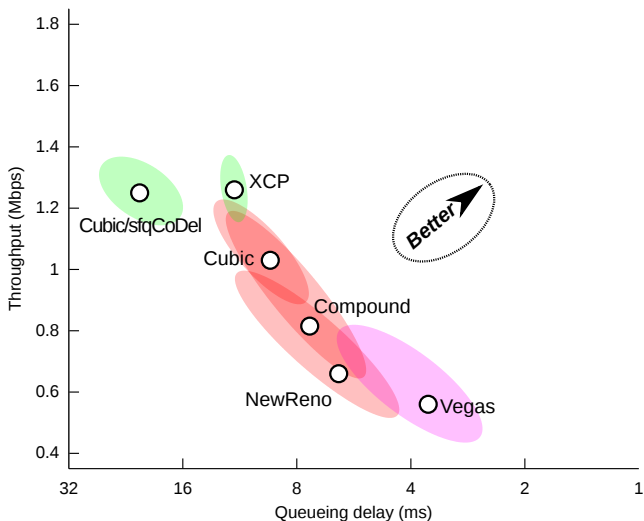
Scenario 1: throughput-delay plot



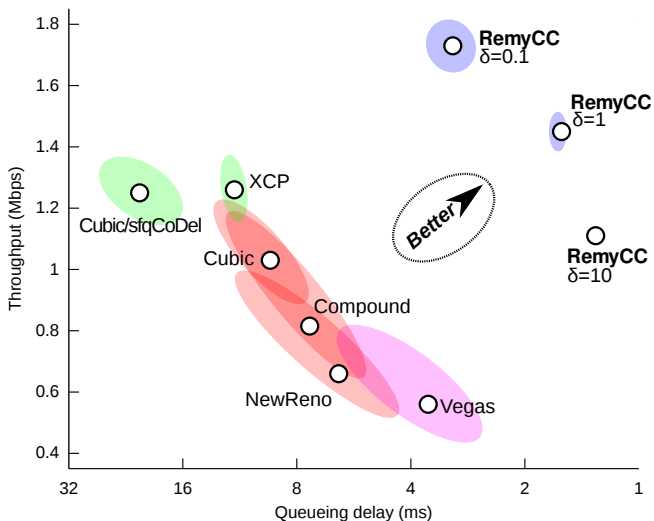
Scenario 1: throughput-delay plot



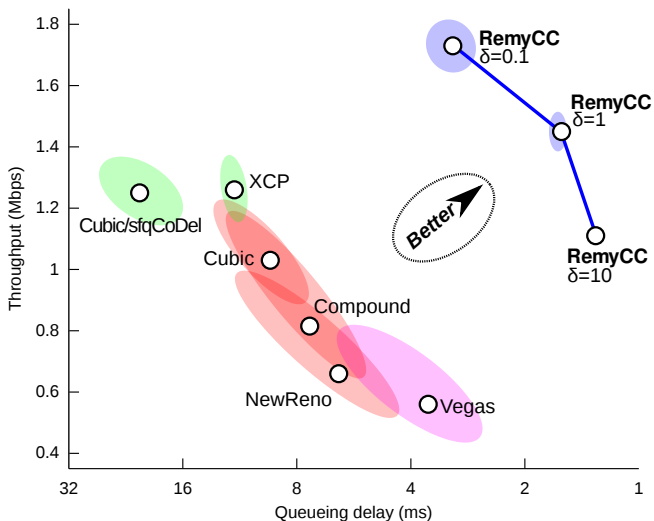
Scenario 1: throughput-delay plot



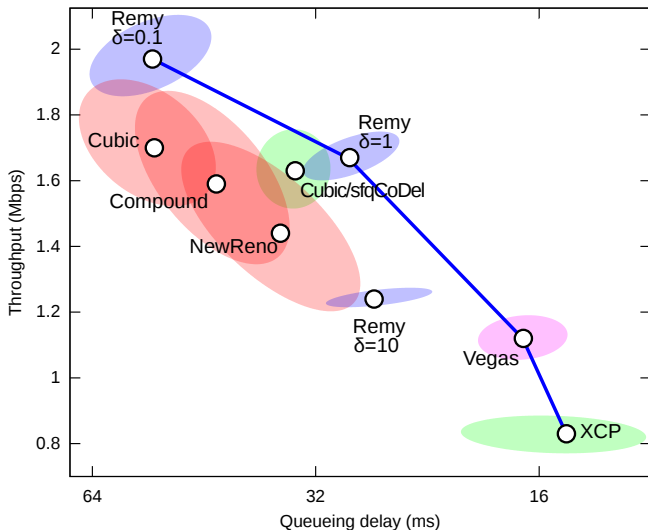
Scenario 1: throughput-delay plot



Scenario 1: throughput-delay plot



Scenario 2: Verizon LTE, $n = 8$

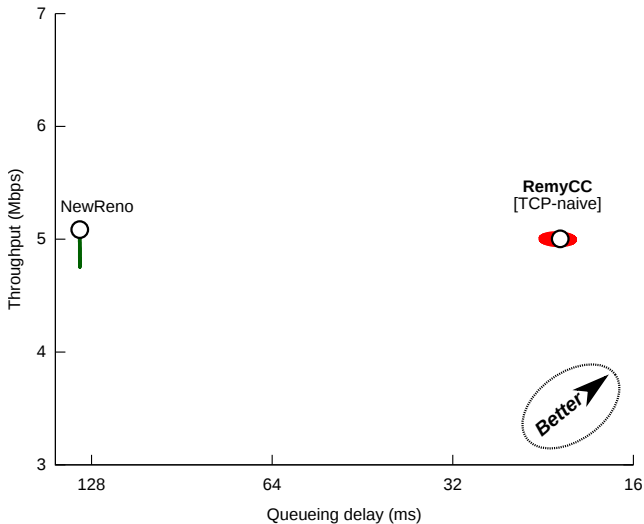


Remy as an instrument to study network science

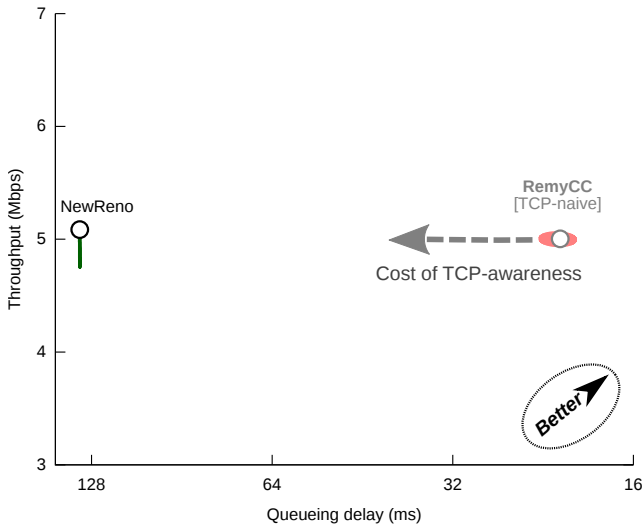
*From the perspective of a network endpoint,
what matters?*

*How difficult is it to learn a good protocol,
given an **imperfect** model of the network?*

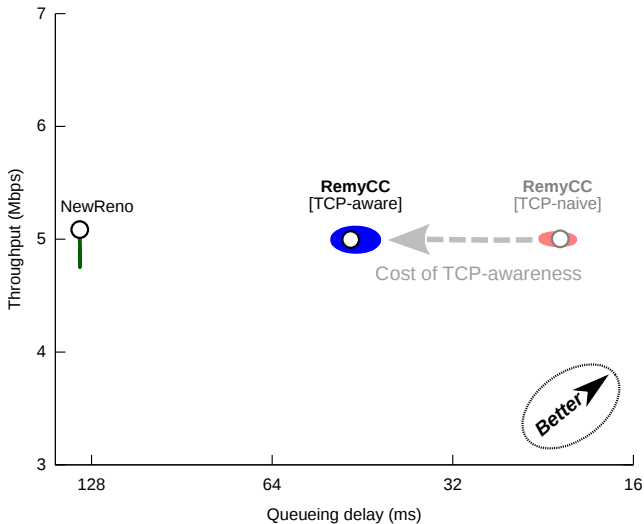
RemyCC competing against itself



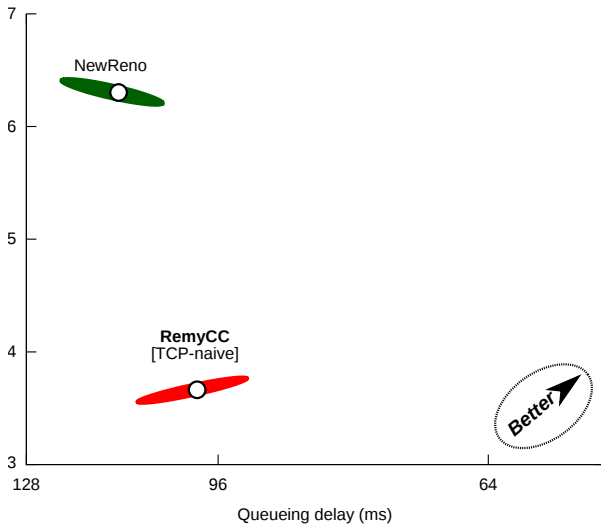
RemyCC competing against itself



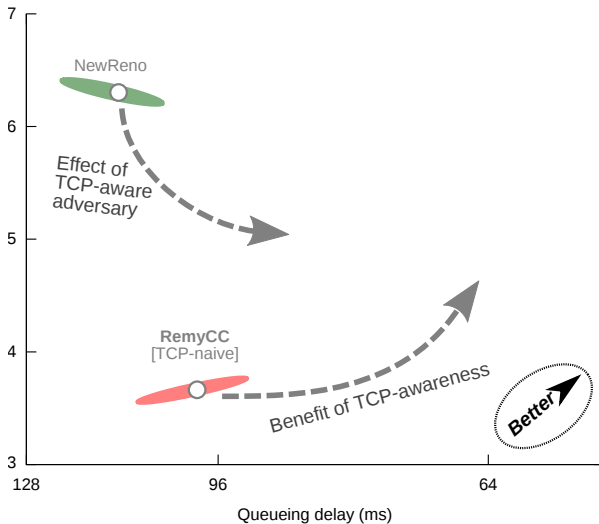
RemyCC competing against itself



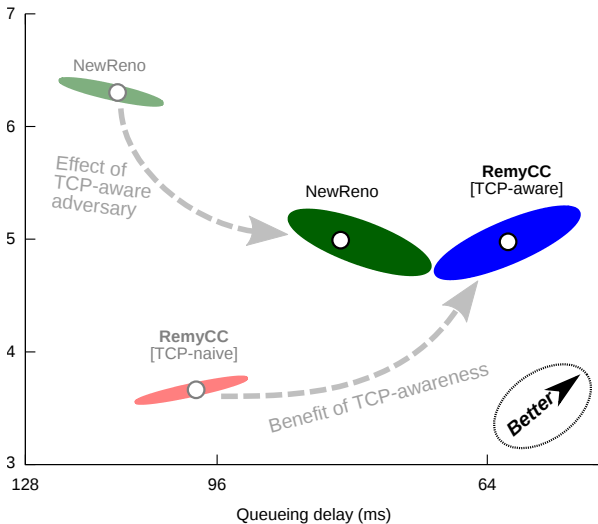
RemyCC competing against TCP NewReno



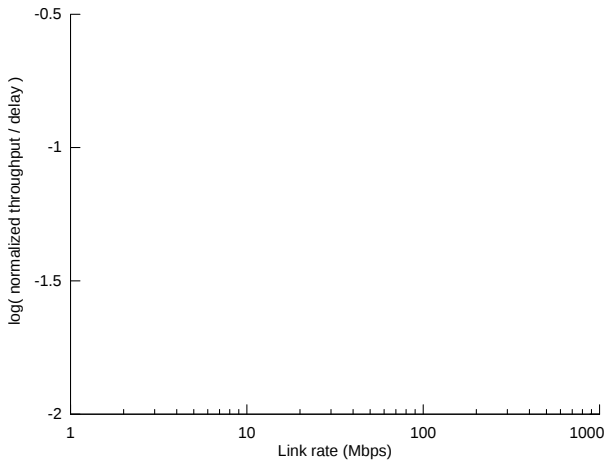
RemyCC competing against TCP NewReno



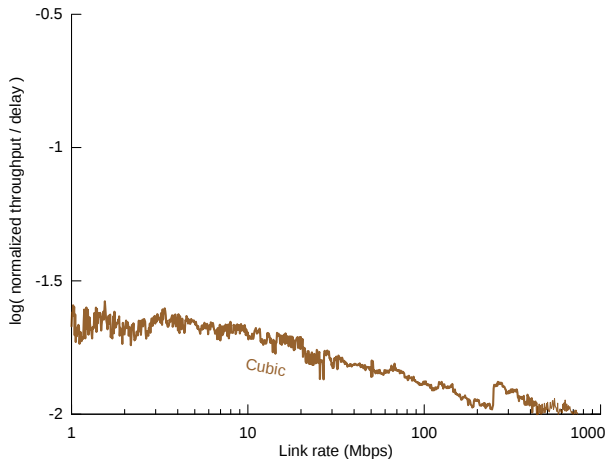
RemyCC competing against TCP NewReno



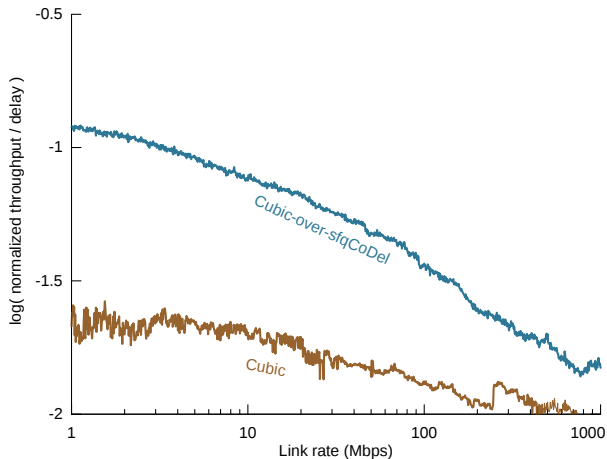
The cost of generality



The cost of generality



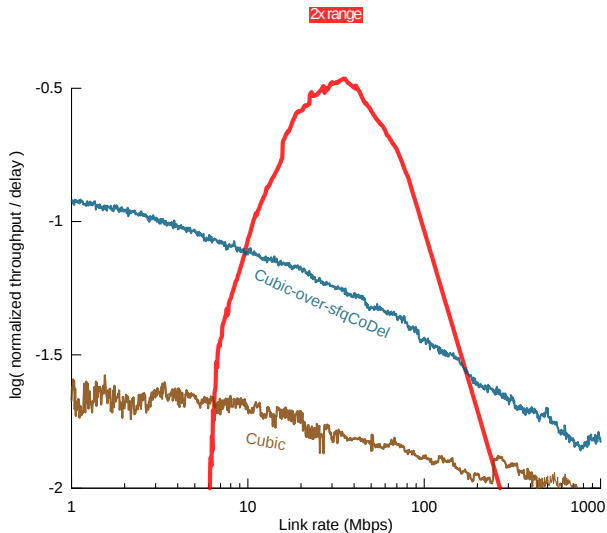
The cost of generality



Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

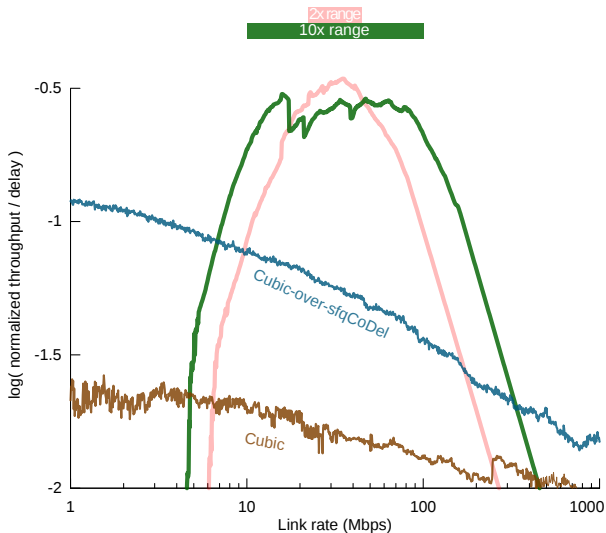
The cost of generality



Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

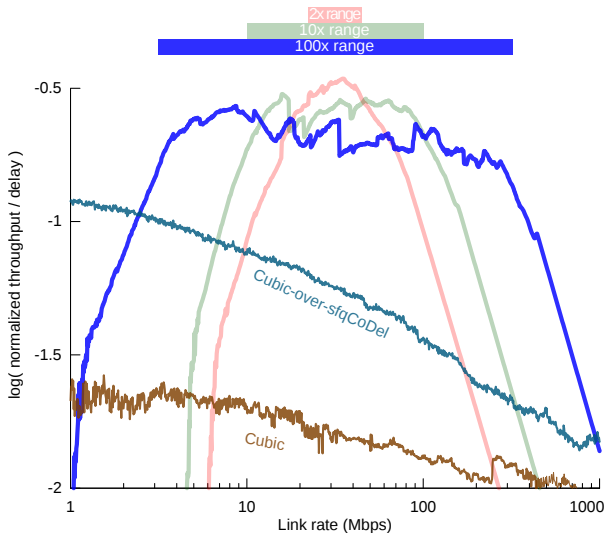
The cost of generality



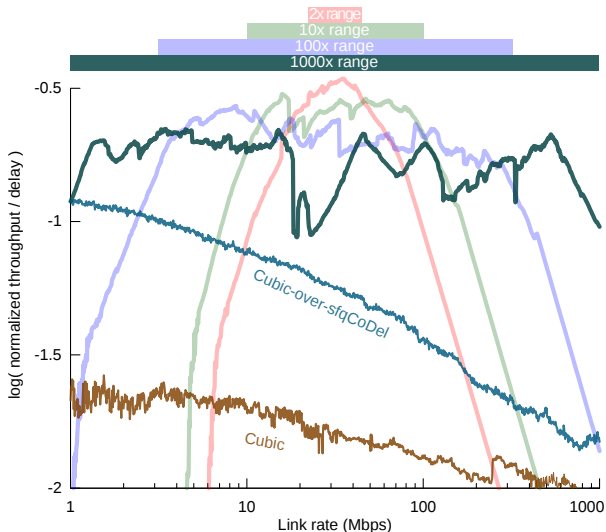
Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

TCP ex Machina: Computer-Generated Congestion Control

The cost of generality



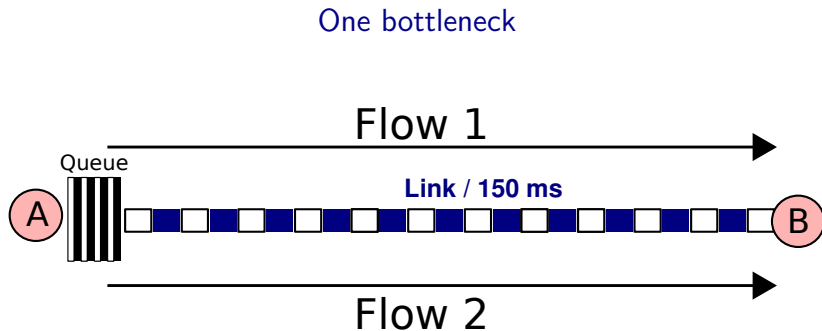
The cost of generality



Keith Winstein, Anirudh Sivaraman, Pratiksha Thaker, and Hari Balakrishnan

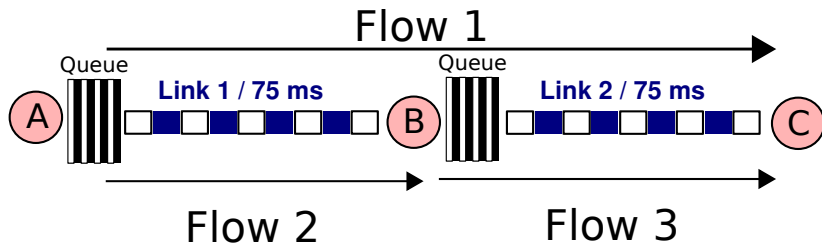
TCP ex Machina: Computer-Generated Congestion Control

When the model is wrong about the topology

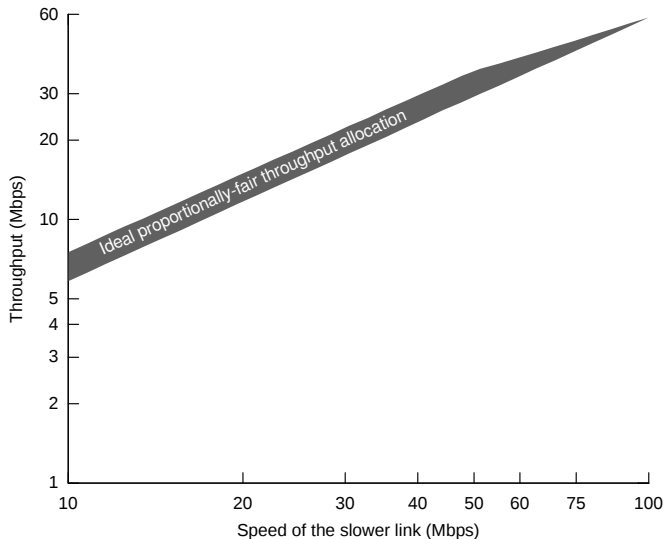


When the model is wrong about the topology

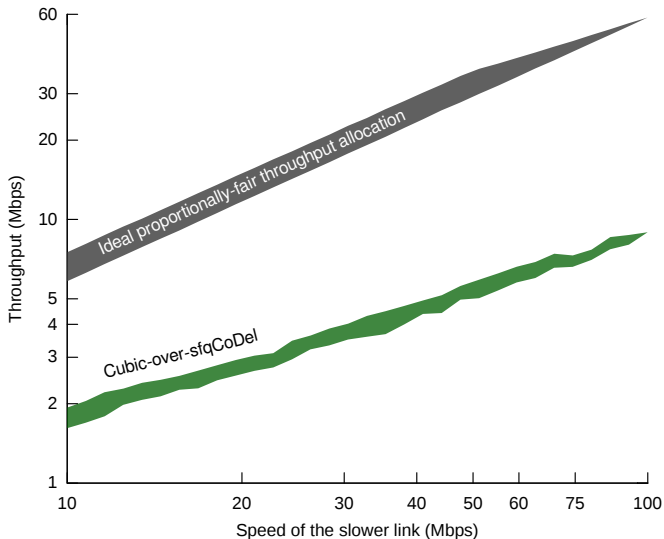
Two bottlenecks



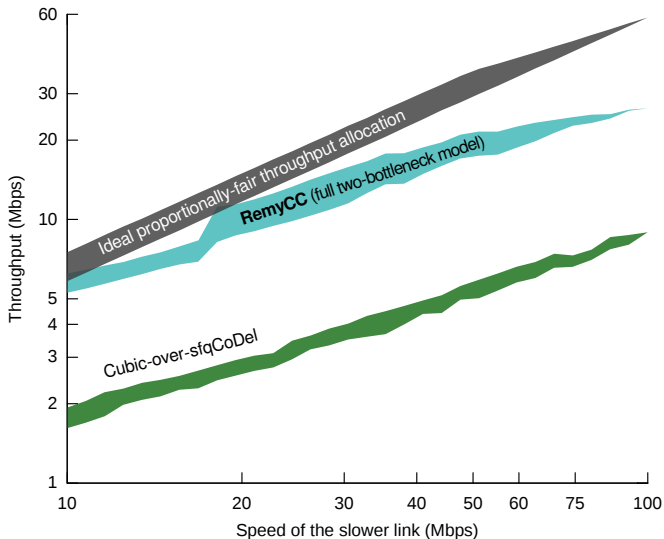
When the model is wrong about the topology



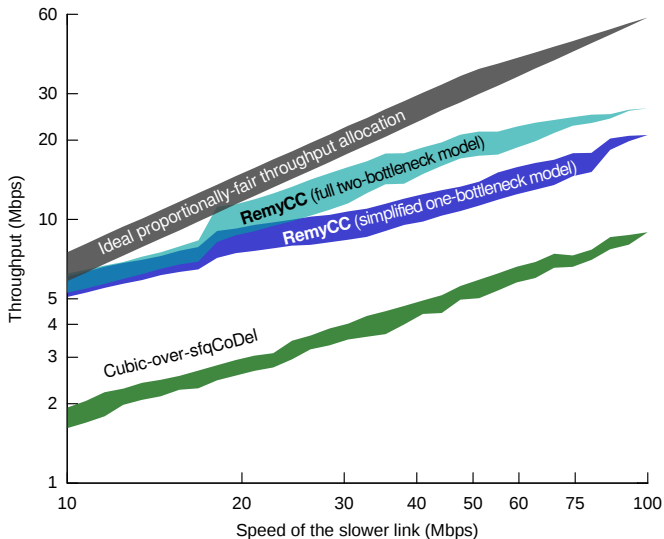
When the model is wrong about the topology



When the model is wrong about the topology



When the model is wrong about the topology



Systems ex Machina

- ▶ Explicit design considerations → **freedom to innovate**
- ▶ “If this protocol is the answer, what’s the question?”

keithw@mit.edu

<http://mit.edu/remy>