DATA CENTER TO THE HOME



koen.de_schepper@alcatel-lucent.com

March, 2015

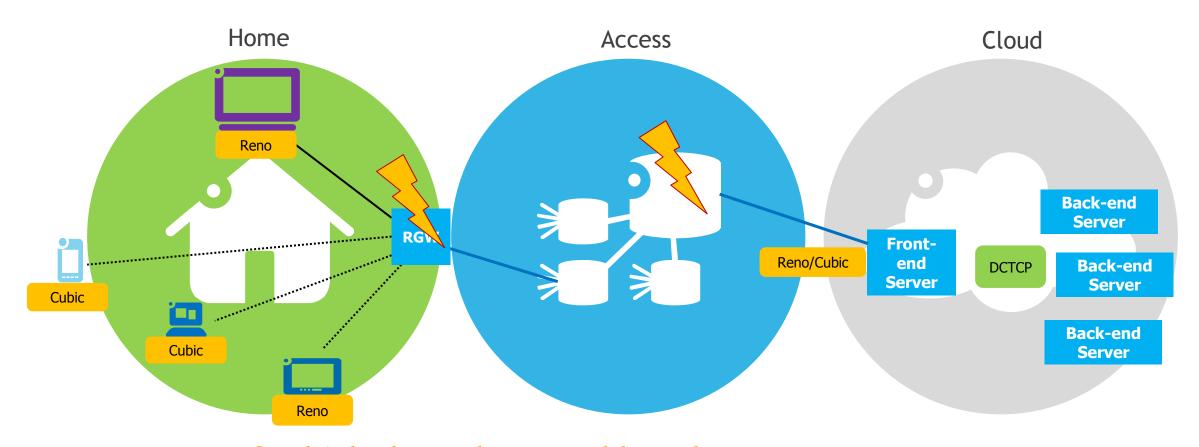


DCttH OBJECTIVE: UNIVERSAL SUPPORT FOR LOW LATENCY = SUPPORT FOR ADAPTIVE INTERACTIVE APPLICATIONS



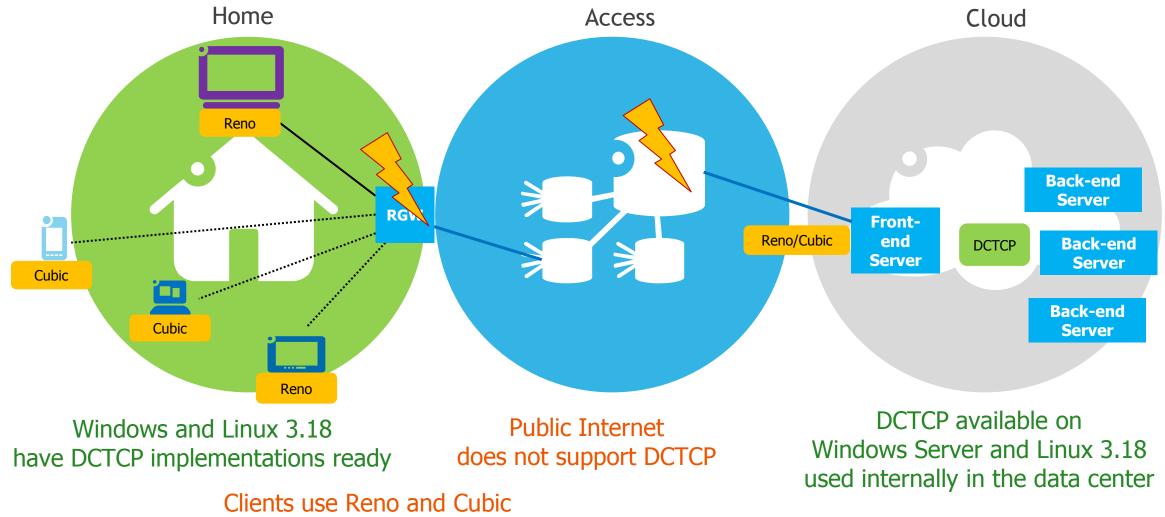


INTERACTIVE APPLICATIONS on the INTERNET ?



Large queues for high throughput and low drop ECN = No drop = Poor Latency ECN++ = Small queues = Bad for interactive applications = Low latency & High throughput

DATACENTER to the HOME ?

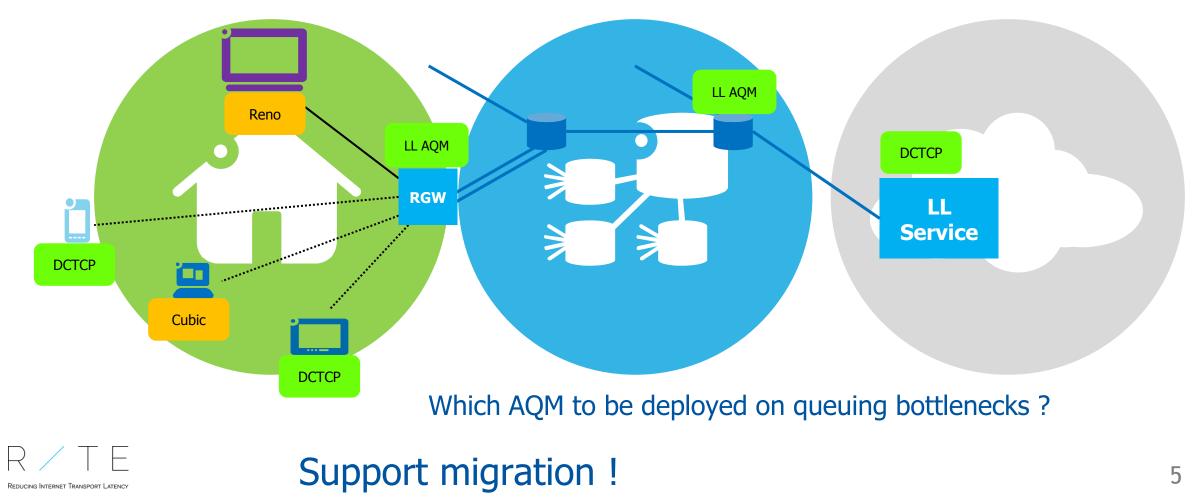


Can't use DCTCP without causing trouble

REDUCING INTERNET TRANSPORT LATENCY

MIGRATION OBJECTIVE: LOW LATENCY ACCESS TO THE CLOUD, EQUAL STEADY STATE THROUGHPUT TO RENO/CUBIC

Can DCTCP be used as Low Latency congestion controller ?



LOWER LATENCY BY SMARTER USE OF ECN DATA CENTER TCP

TCP (Reno)

←→

DCTCP

Response to congestion in sender

- Half the congestion window when drop detected in one RTT
- Reduce partially per marked packet; half if all marked in one RTT
 - ➔ React according to level of congestion

ECN feedback in receiver

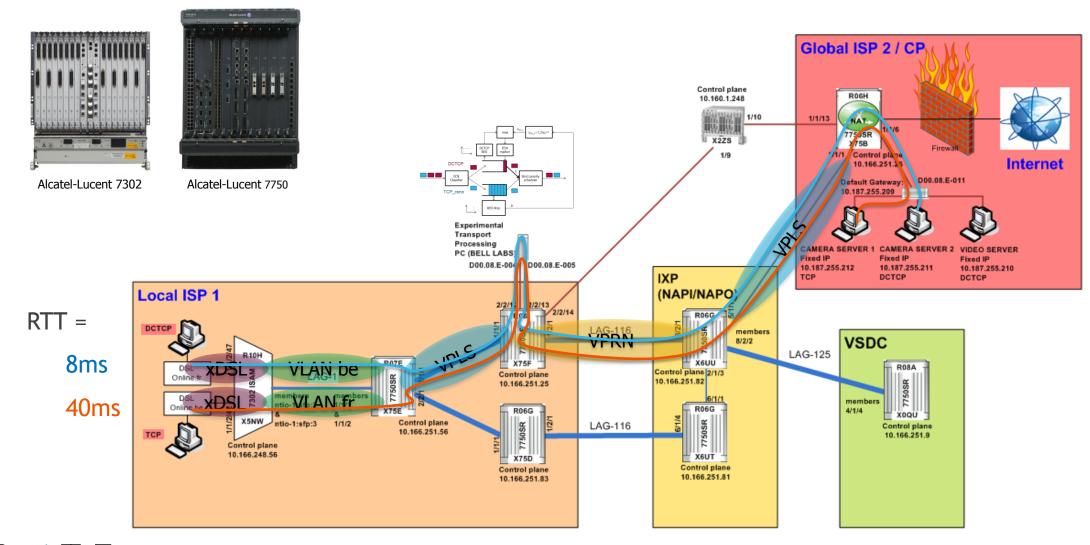
ECN marking in network

- Smooth and delay a drop or mark to allow bursts
- Don't smooth or delay queue size
- Shallower marking threshold

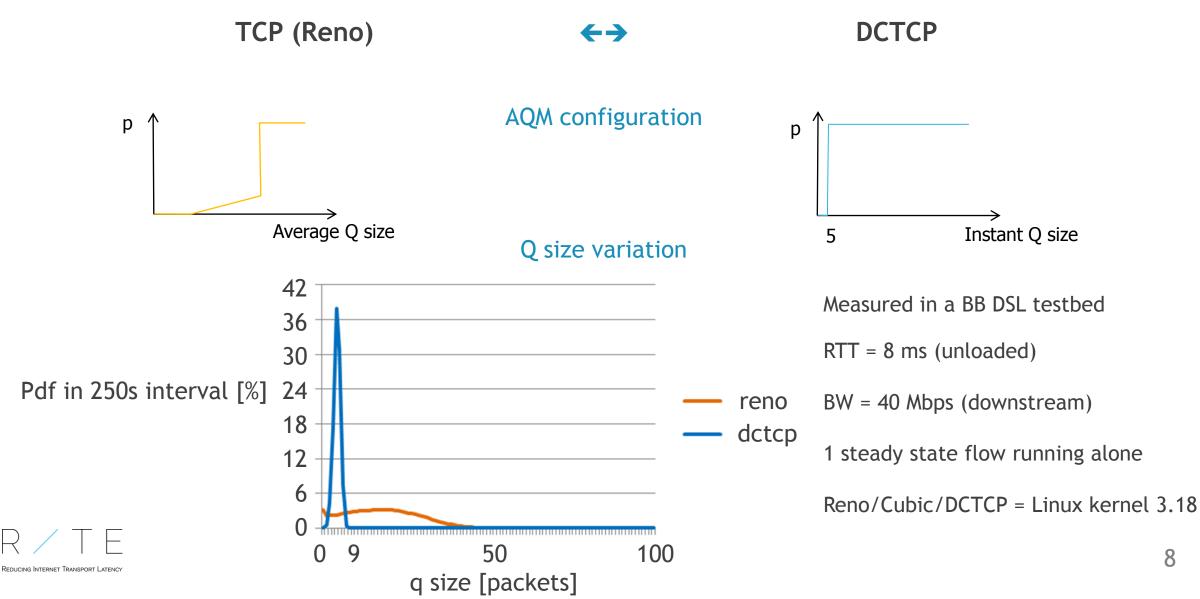
➔ immediate ECN marking



DEMONSTRATED ON A REAL BB RESIDENTIAL TESTBED

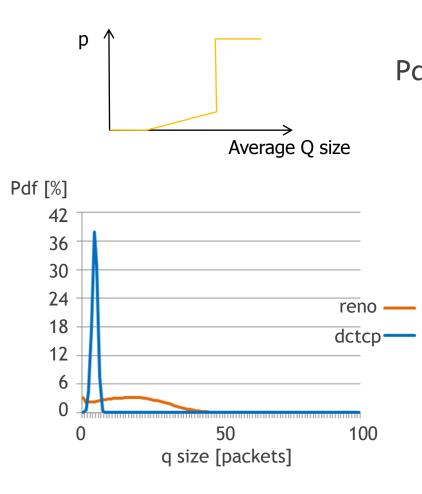


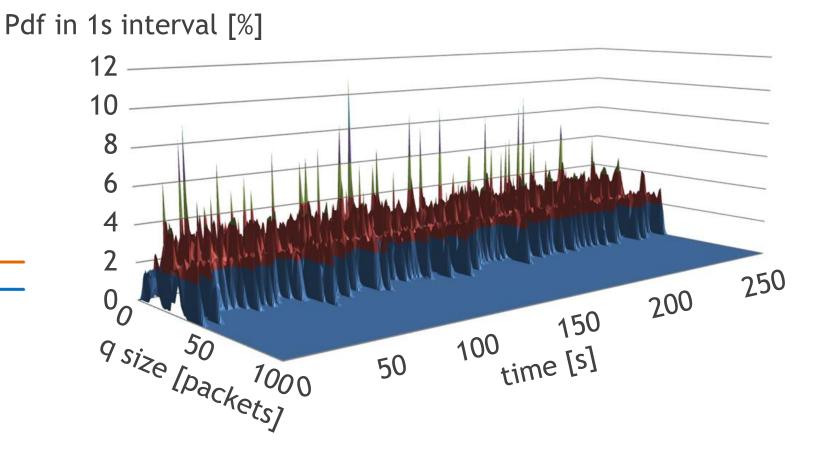
LOWER LATENCY BY SMARTER USE OF ECN DATA CENTER TCP



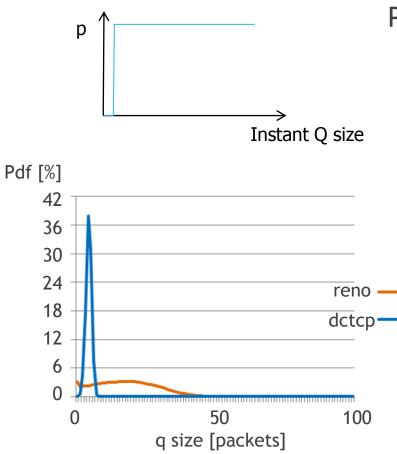
8

QUEUE SIZE AT DEQUEUE 1 TCP RENO FLOW (STEADY STATE)

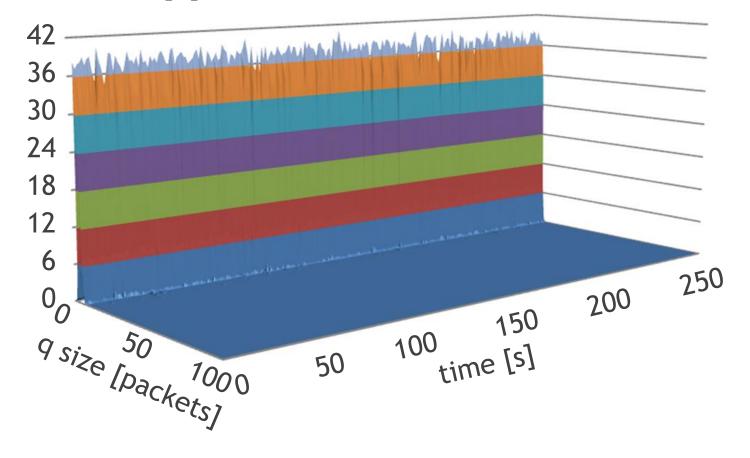




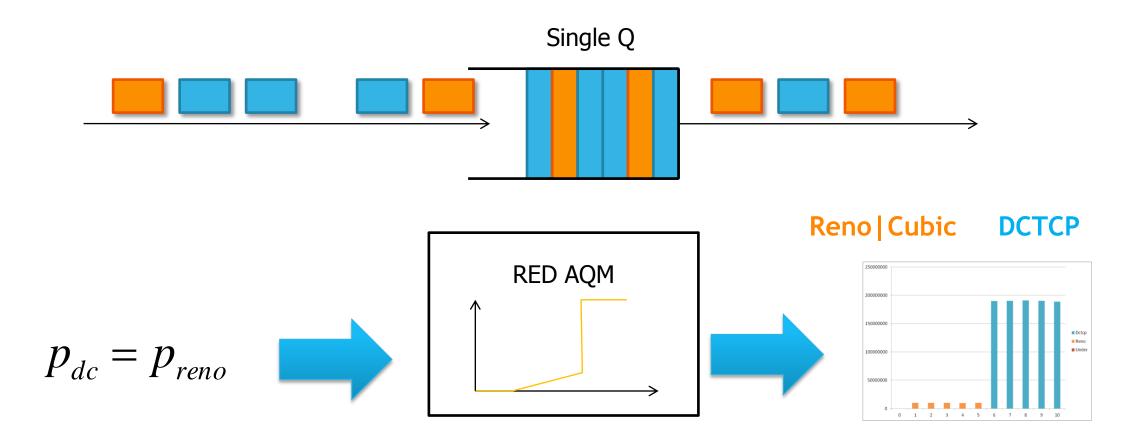
QUEUE SIZE AT DEQUEUE 1 DCTCP FLOW (STEADY STATE)



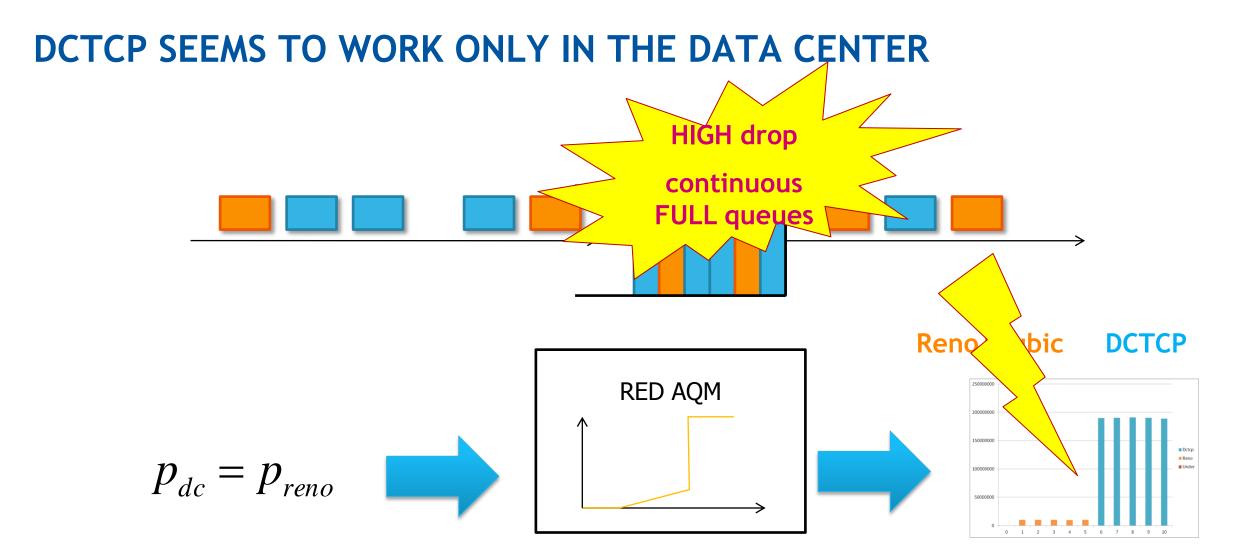
Pdf in 1s interval [%]



DCTCP DOES NOT WORK ON TRADITIONAL RED-ECN









THROUGHPUT:

DCTCP flows: 0

RTT = 8 ms (unloaded)

- BW = 40 Mbps (downstream)
- BDP = 27 full sized packets
- AQM = RED with recommended configuration*

X-axis: 0 - 250 sec

Y-axis: first row:

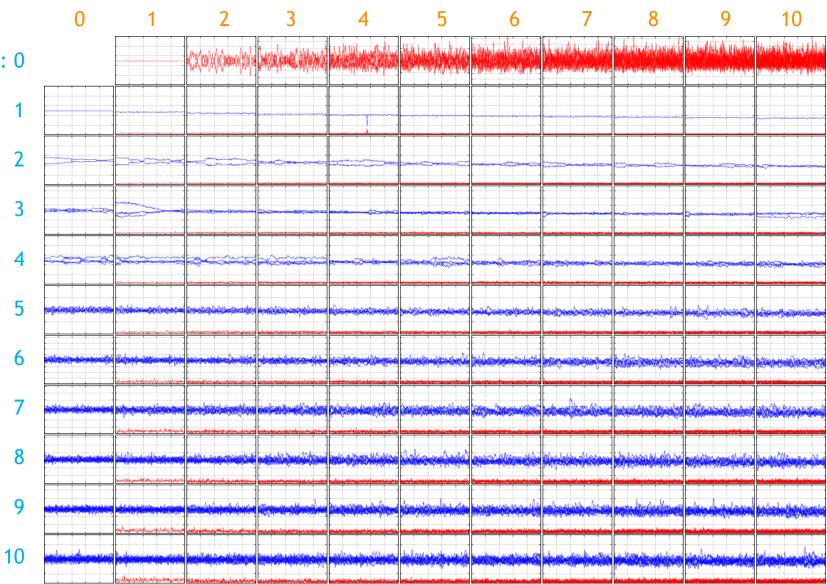
0 - (80 / <nbr_flows>) Mbps

Y-axis: other rows

REDUCING INTERNET TRANSPORT LATENCY

0 - (80 / <nbr_dctcp>) Mbps

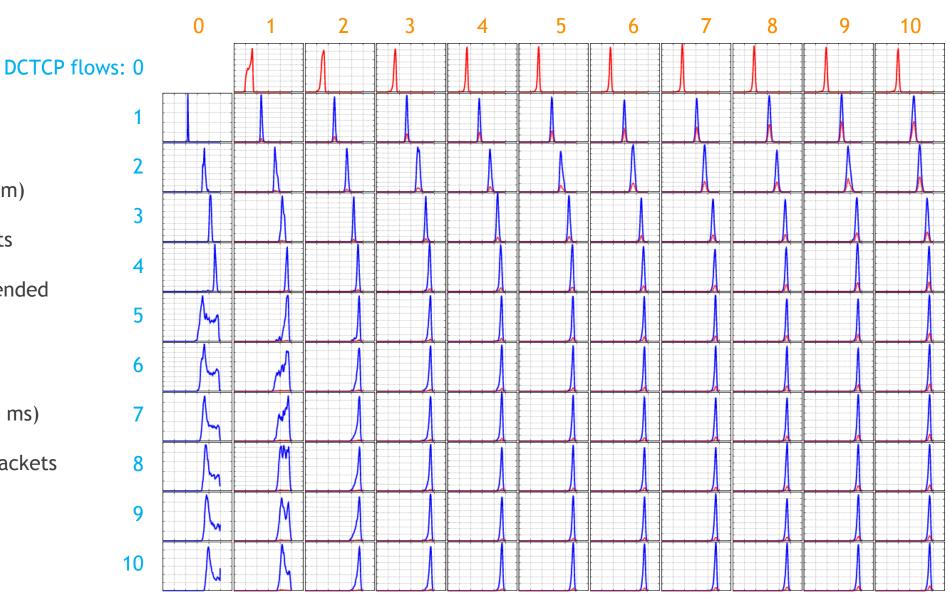
Cubic (= Reno) flows:



* tc qdisc add dev eth2 root red limit 1600000 min 120000 max 360000 avpkt 1000 burst 220 ecn bandwidth 40Mbit

Q SIZE PDF:

Cubic (= Reno) flows:



RTT = 8 ms (unloaded)

BW = 40 Mbps (downstream)

BDP = 27 full sized packets

AQM = RED with recommended configuration*

X-axis: 0 - 300 packets

(450 Kbytes, 90 ms)

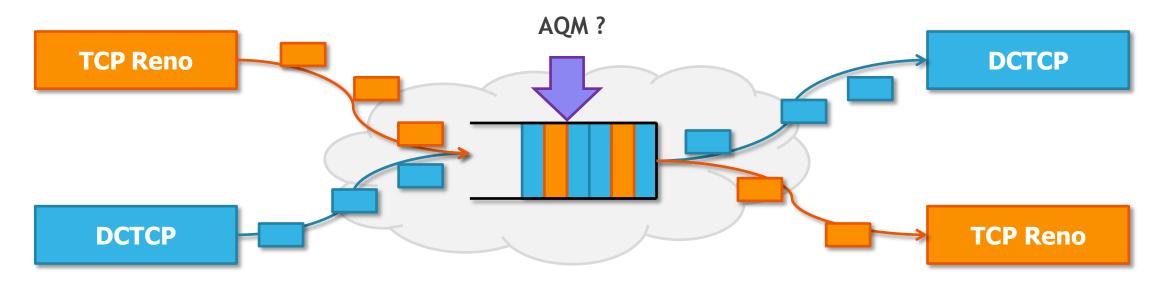
Y-axis: autoscale count packets

R T E

* tc qdisc add dev eth2 root red limit 1600000 min 120000 max 360000 avpkt 1000 burst 220 ecn bandwidth 40Mbit

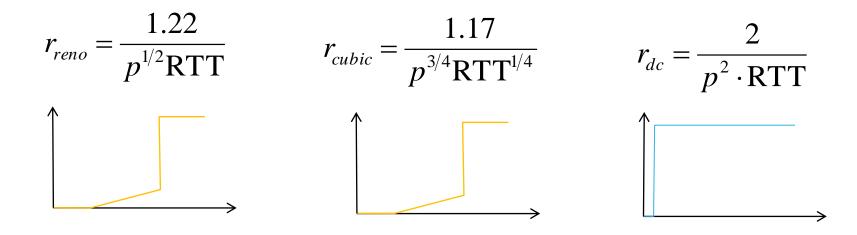
AQMS FOR EQUAL STEADY STATE RATE MIGRATION PATH FOR NEW CC SCHEMES

- How should an AQM guarantee an equal steady state rate for flows with different congestion control schemes
 - classify packets according to CC schemes
 - align the drop/mark probabilities



TCP CONGESTION CONTROL SCHEMES STEADY STATE RATE

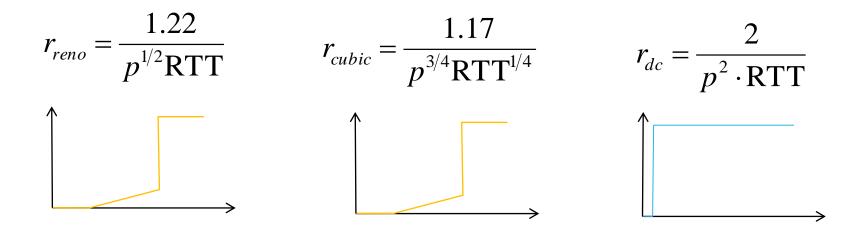
• Steady state rate has been calculated for existing CC schemes:





TCP CONGESTION CONTROL SCHEMES STEADY STATE RATE

• Steady state rate has been calculated for existing CC schemes:



• But we calculated that DCTCP running in non-on/off mode behaves as:

$$r_{dc_p} = \frac{2}{p \cdot RTT}$$



TCP CONGESTION CONTROL SCHEMES FAIRNESS BETWEEN DCTCP AND RENO

• Mark/drop probability relation for equal rate and RTT:

$$r_{reno} = r_{dc} \qquad \frac{1.22}{p_{reno}^{1/2} \text{RTT}_{reno}} = \frac{2}{p_{dc} \cdot \text{RTT}_{dc}}$$

$$p_{reno} = \left(\frac{p_{dc}}{1.63}\right)^2$$



TCP CONGESTION CONTROL SCHEMES FAIRNESS BETWEEN DCTCP AND RENO

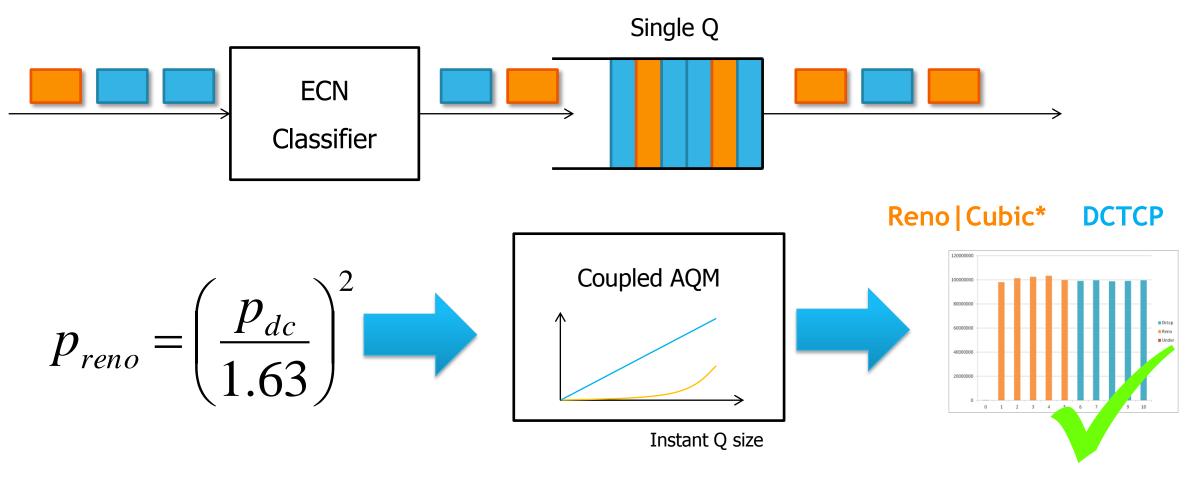
• Mark/drop probability relation for equal rate and RTT:

$$r_{reno} = r_{dc} \qquad \frac{1.22}{p_{reno}^{1/2} \text{RTT}_{reno}} = \frac{2}{p_{dc} \cdot \text{RTT}_{dc}}$$

 $p_{reno} = \left(\frac{p_{dc}}{1.63}\right)^2 \qquad \begin{array}{c} \text{Square is easy!} \\ \text{Compare Q size with 2 random variables} \\ P = f(Q) \qquad p \Rightarrow \text{Random}() < P \end{array}$ Square is easy! $p^2 \Rightarrow (\text{Random}() < P) \& \& (\text{Random}() < P)$ $p^2 \Rightarrow \max(\text{Random}(), \text{Random}()) < P$ 19

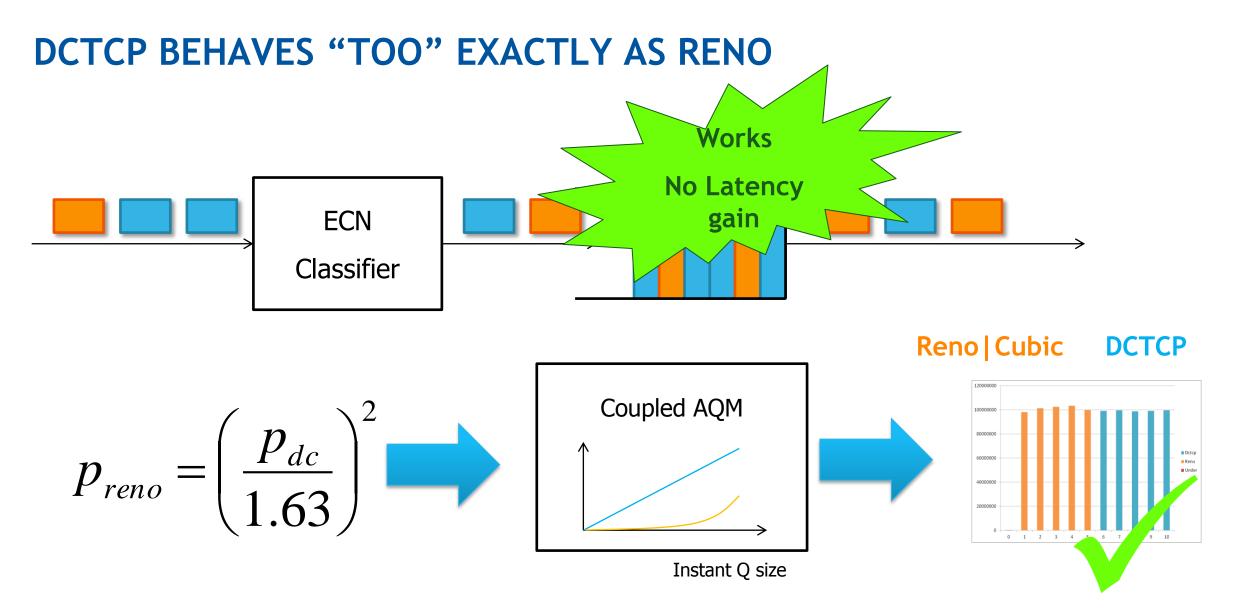


DCTCP BEHAVES EXACTLY AS RENO IF WE CORRECTLY CORRELATE MARKING AND DROPPING



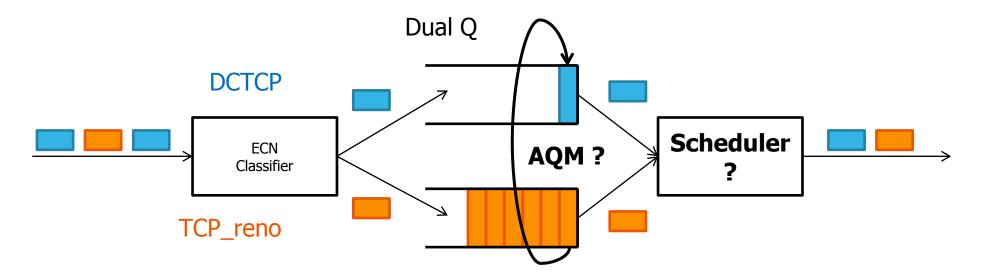
* Under local DC-access conditions (small BDP) Cubic behaves as Reno

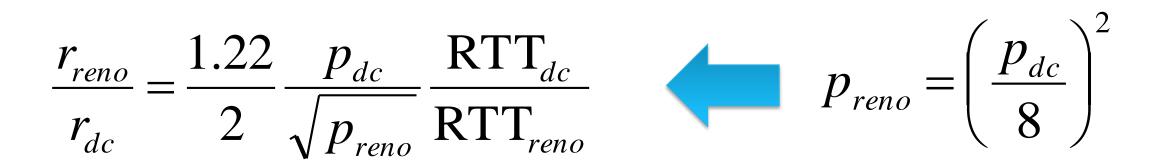
Slope starts from the origin to avoid ON/OFF behavior in steady state



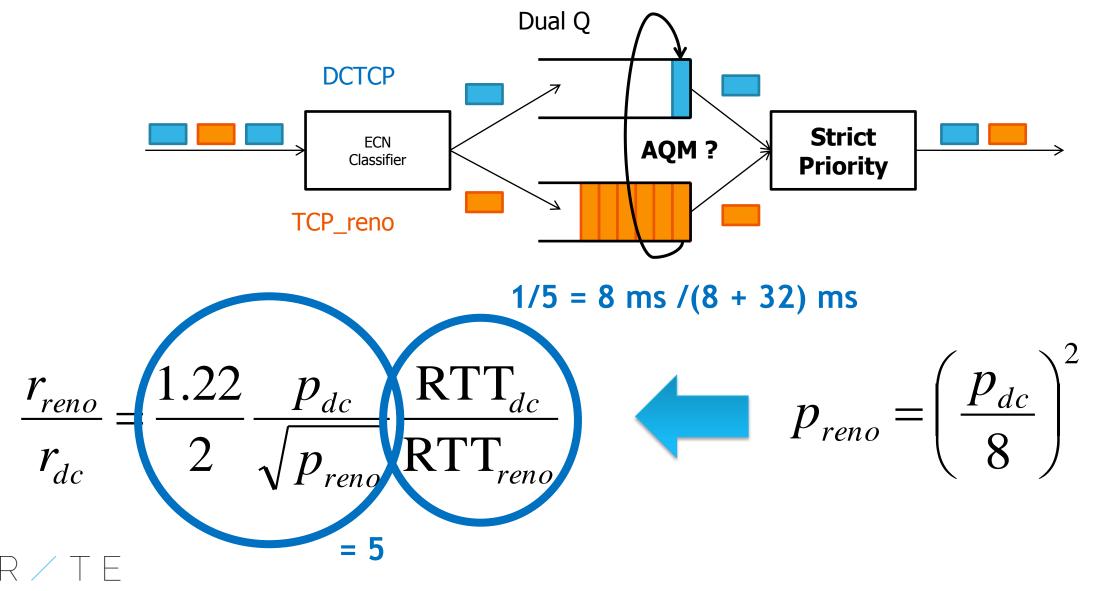


DUAL QUEUE - LOW LATENCY



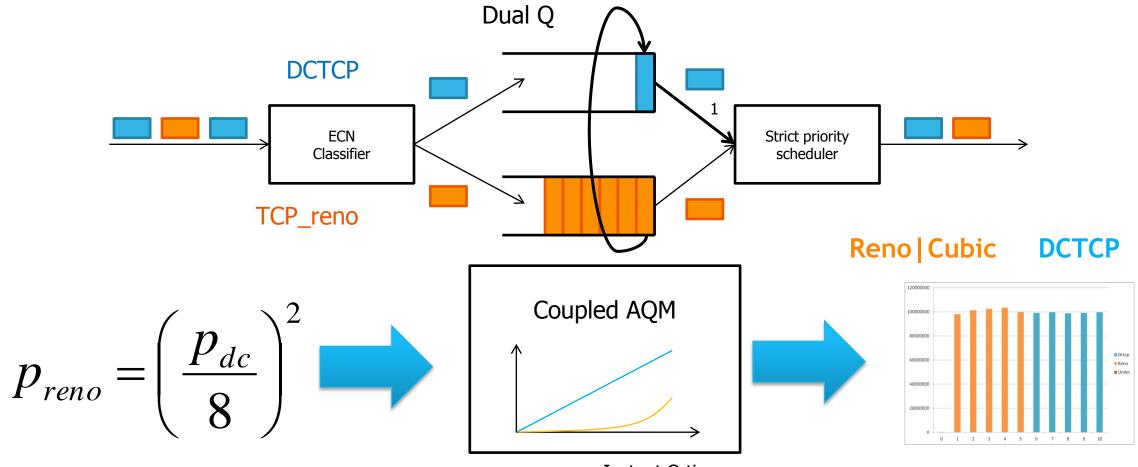


DUAL QUEUE - LOW LATENCY



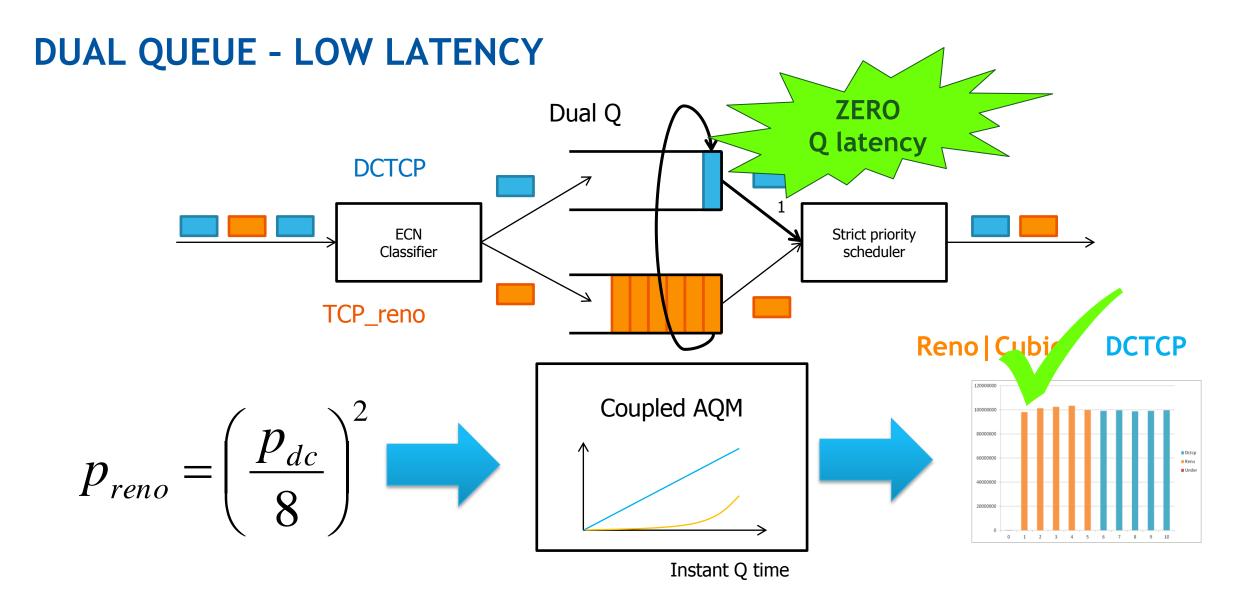
REDUCING INTERNET TRANSPORT LATENCY

DUAL QUEUE - LOW LATENCY



Instant Q time





Measure Q in time is important for optimal fairness !



THROUGHPUT:

DCTCP flows: (

RTT = 8 ms (unloaded)

BW = 40 Mbps (downstream)

BDP = 27 full sized packets

AQM = DualQ Coupled

X-axis: 0 - 250 sec

Y-axis: all rows:

0 - (80 / <nbr_flows>) Mbps

Cubic (= Reno) flows:

	0	1	2	3	4	5	6	7	8	9	10
s: 0			()000¢();() 000	ni) himpone	town have been a				Mildon ta certa trada nila nana nila internationality	With the base for the station were necessary memory and	e de la constant de la constant Constant de la constant de la consta
1		nophinikaciyaci	uto di coto me i li	www.							
2	(c);++++++++++++++++++++++++++++++++++++		a de la companya de l		<u>Kalidadidan</u>						
3	n afaising a haising an	Unecu prints	leiden ihreitereiten			Hariteteilikei				uttin has delanta et	
4	d de rether oa til en er er er	i Alensitu ulaitak M	manuterented	had Muga shi kekila	Natural and the first state of the first state of the second state of the second state of the second state of the			and addition of the date	ha hu and this works		
5	RATE AND THE REAL AND A	at an alway which	M. C. Carrier and State and State	lyst al der dame die	we stated a fight day	temperatur an and that the second second	naanne prannaa Iddaadhaa adddad	alilian laalik wulitad	y de la presenta conserva por serva por s En la la la porta de la port	in seal in an idea a	vaihadda.oodfied.eddod
-		Red Manares days be wh	legenerie generation of the second	ne totrollated by tota	romann ann an ann an ann ann 1 diùcht bha aile santainn	n ning og vinger og som	n e na stal de la stal stal stal stal stal stal stal de la stal de la stal stal stal stal stal stal stal s	ente de sala la la la com	fillen an	an in the second and a stand	an this for a dar with a
6							n diamateria de la constanta de				n sinterasiyasinin asi
7					u sanchu al-buru din badin da A pagarangi di pagarangi A pagarangi di pagarangi	Alexandron and a second se International second second International second	na an a				a fanal ever nader fri fransk
8	a lada abih ba da a bana ang ak	h defined a differ (forte		ershirikaan dhaladdik Tooroon gorooniyaa	alah di seria alah sa sahidi ya se Mana sa	ilanida della antesidado Antesida e della antesidado Antesida e della antesidado e pres	n beiteliseen allehan maangeren varianinge	leidlichthithithithithithithithithithithithithi	kaisleten ei dellende bisen. Annesissen statistissen so	la dat balan dasabash Marana Mana perena dangan perena Mana perena dangan perena	na na sta si di bi da si la septe Na timpi propona esta si na s
9	el en las fatas de la construir par alterna espe a construir de la construir par la construir de	addillardlijder Veldelard	ka Alilia ulk alk	kalati kanikata teleta. 1995-yani memperanta	i di tin di	de des stations de las de recto propositions de las des	te, en le définité de la des recommendant parties). Se statile te a did se di General synthesis de provisione	and the standard of the And these many second standards and the second standards and the second standard standards and the second standard standard	Hard Lade a dised on the second s	tele ana ana ana ana ana ana ana ana ana an
10		Wesserfel Walk And	all have be all h	in and his bold to a side complete registration of a	in blande, eibin dwar o rafi arstegner werder voor		tettetetetetetetetetetetetetetetetetet			Here da farre i da ance entre Here da farre i da ance entre Histori i strand relation i ta	erne di del stan per contra constan Sette de constante de la constante de la constante

Q SIZE PDF:

DCTCP flows: 0

Cubic (= Reno) flows:

RTT = 8 ms (unloaded)

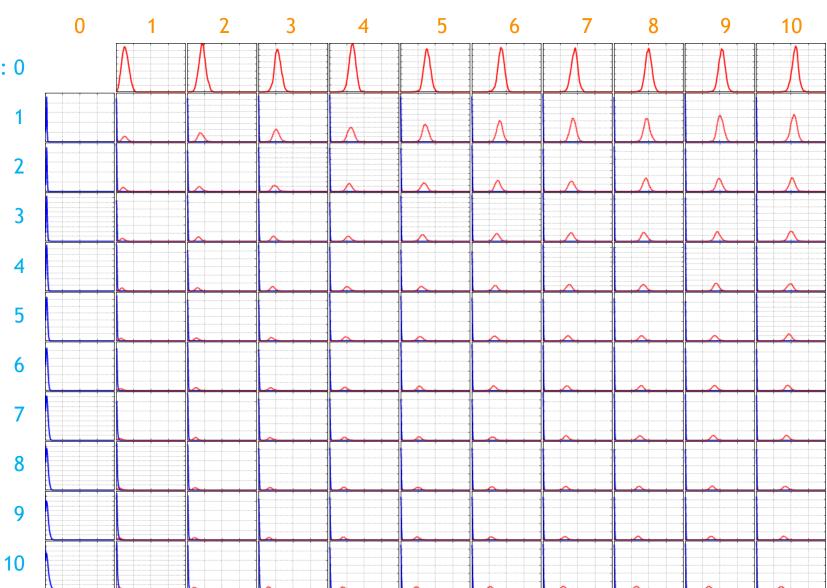
- BW = 40 Mbps (downstream)
- BDP = 27 full sized packets

AQM = DualQ Coupled

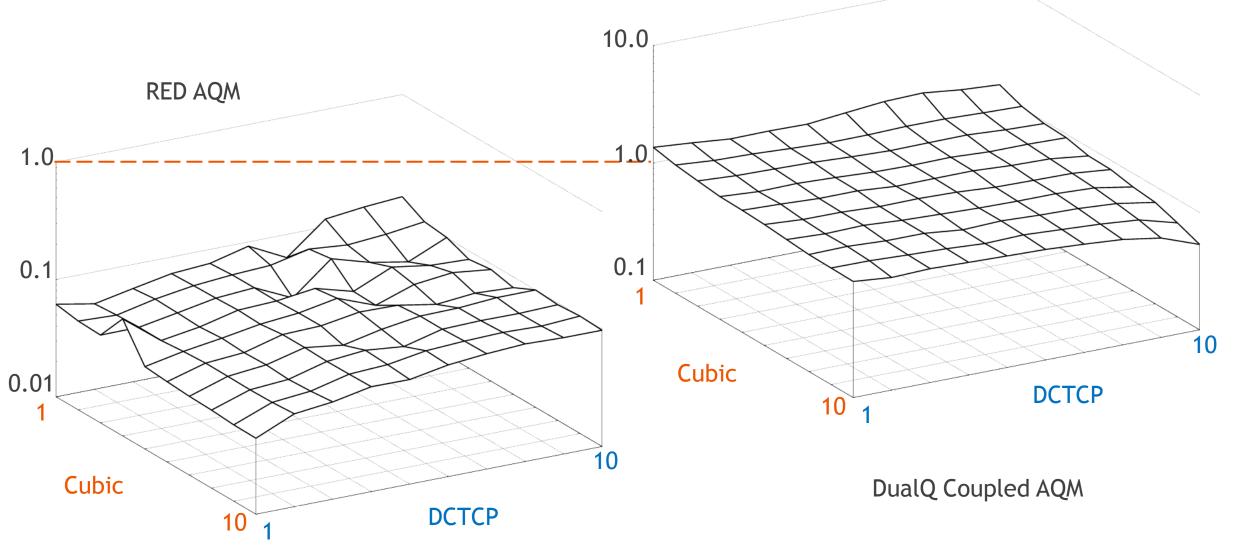
X-axis: 0 - 300 packets

(450 Kbytes, 90/w ms)

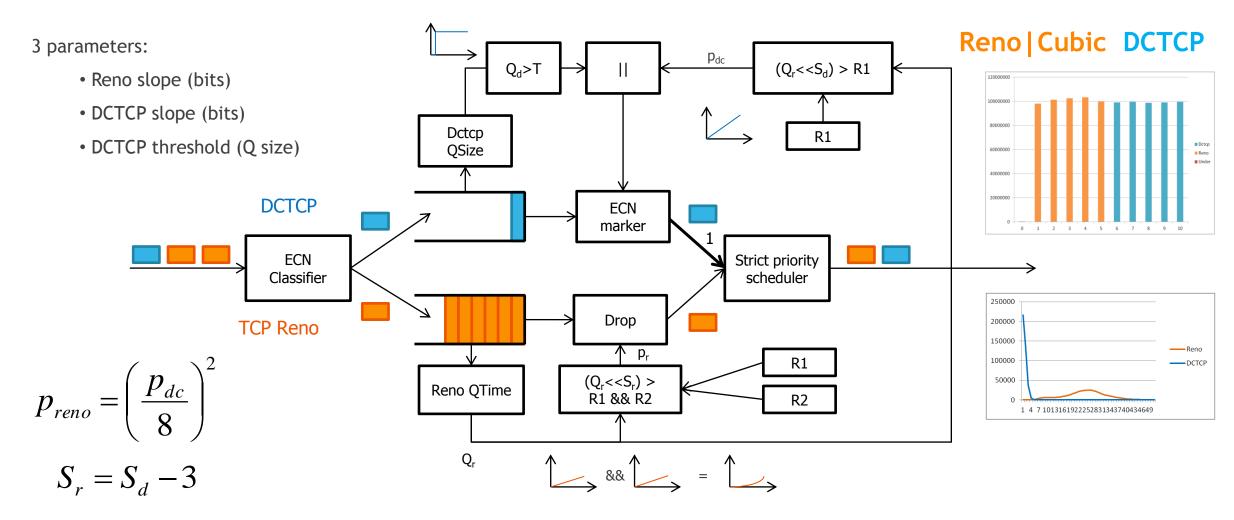
Y-axis: autoscale count packets



THROUGHPUT RATIO (CUBIC / DCTCP)



DETAILED IMPLEMENTATION



ADAPTIVE INTERACTIVE APPLICATIONS

Panoramic interactive video



Video/Voice conferencing



• Remote control,

FUTURE WORK & CONCLUSIONS ?

- Dynamic behaviour to be investigated (expected to be 5x better due to 5x latency reduction)
- Unmanaged Low Level network Service → Native support for Adaptive Interactive Applications
- Better usage for ECN (marking can be more often than dropping)
 → x/p relation for ECN based congestion controller (x determining the marking rate)
 → p² relation between mark and drop in AQM
- Backwards compatible: DCTCP should respond to drop as Reno
 - Currently 3.18 Linux implementation fails on this aspect
- Steady state throughput fairness between DCTCP and Reno|Cubic
 - Only if DCTCP flows are terminated to nearby (local) datacenter
 - If longer RTT, DCTCP flows are getting lower throughput than Reno|Cubic
 - Reno|Cubic fallback if throughput is too low and base RTT is too long ?
 - > Define a TCP congestion controller which is less (/not) dependent on the RTT ?

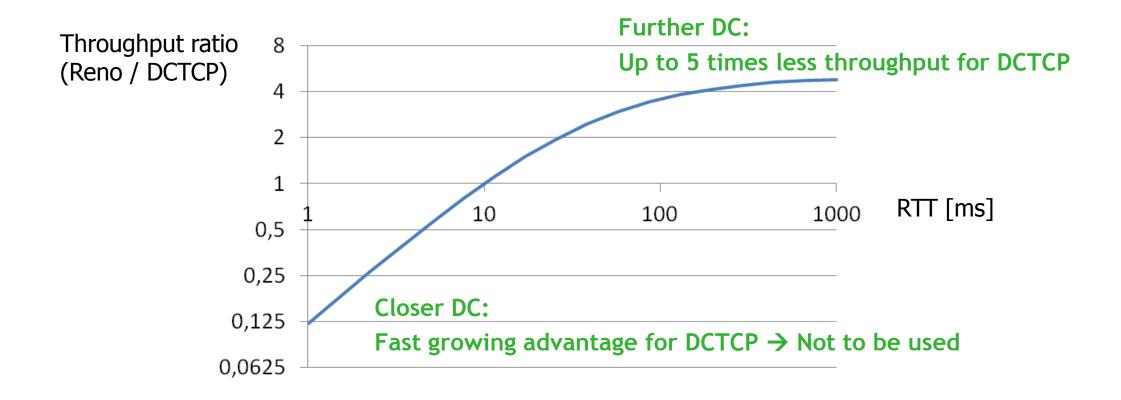


Questions

koen.de_schepper@alcatel-lucent.com



BASE RTT FAIRNESS WHAT IF THE DATACENTER IS FURTHER OR CLOSER



Coupled AQM configured for 10ms base RTT and 40ms Reno queue time (1/5 RTT ratio)



DCTCP STEADY STATE THROUGHPUT WITH SLOPE-RED

Per "long" RTT: $W \leftarrow W + 1$ (1)

And also per RTT:
$$W \leftarrow W \left(1 - \frac{\alpha}{2} \right)$$
 (2)

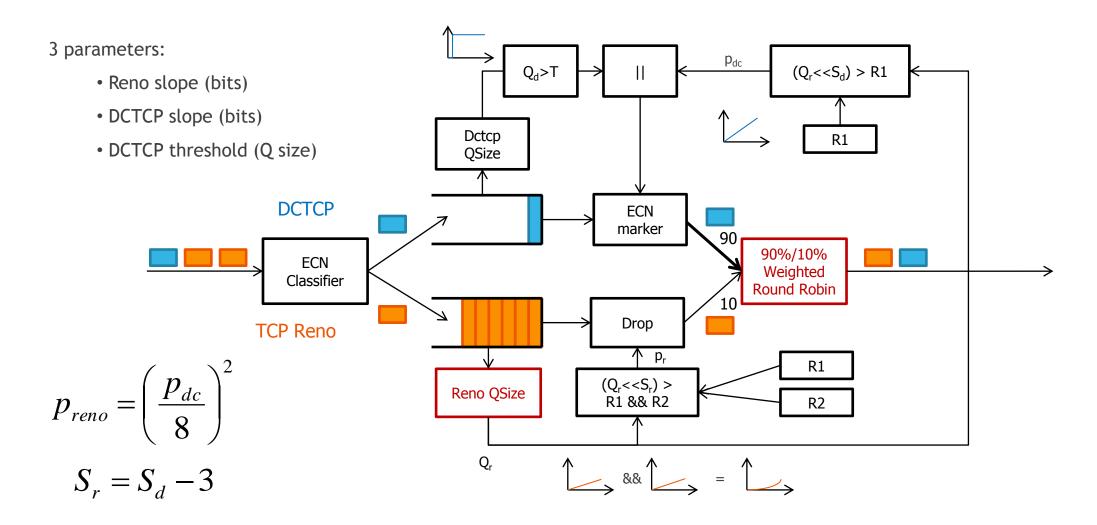
In steady state if (1) is compensated $W \leftarrow W - 1 = W \left(1 - \frac{1}{W} \right)$ so from (2) if $\frac{\alpha}{2} = \frac{1}{W}$ (3)

As $\alpha \leftarrow (1-g)\alpha + gp$, if *p* is stable in steady state $\alpha = p$ (4)

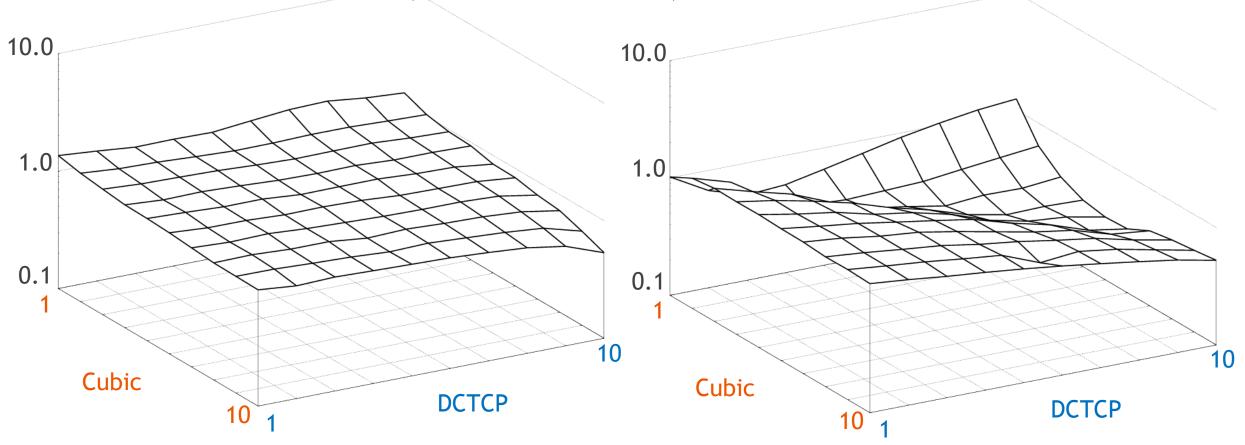
The instantaneous rate
$$r = \frac{W}{rtt}$$
 (5) thus (3,4,5) $r = \frac{2}{p \cdot rtt}$



QUEUE SIZE BASED COUPLED AQM



QUEUE SIZE BASED COUPLED AQM THROUGHPUT RATIO (CUBIC / DCTCP)



Qtime - DualQ Coupled AQM

Qsize - DualQ Coupled AQM

REDUCING INTERNET TRANSPORT LATENCY