Using TCP/IP Traffic shaping to achieve iSCSI service predictability

Paper presentation

Jarle Bjørgeengen

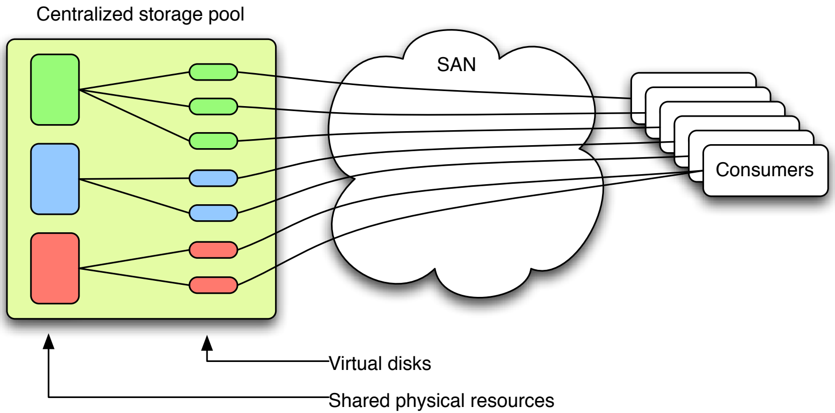
University of Oslo / USIT

November 11, 2010

Outline

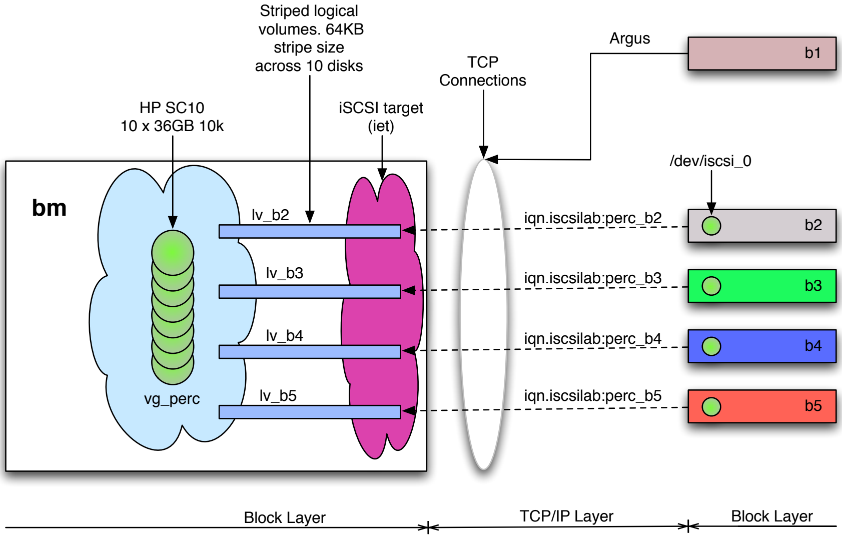
- About resource sharing in storage devices
- Lab setup / job setup
- Experiment illustrating the problem
- One half of the solution: the throttle
- Live demo
  - The throttle
  - Part two of the solution: the controller
- How the controller works
- Conclusion and future work

General problem of sharing resources

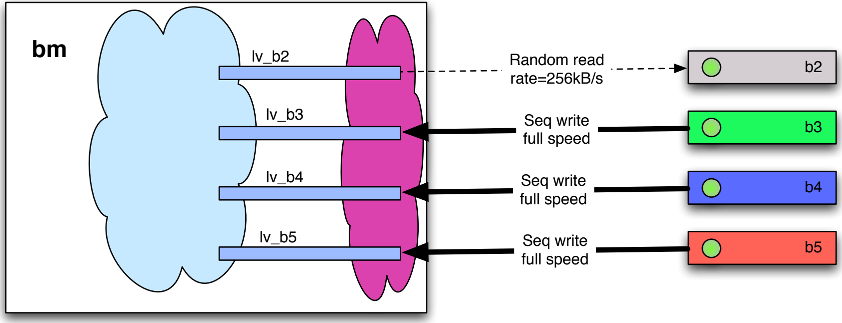


Free competition causes unpredictable I/O performance for any given consumer.

Lab setup

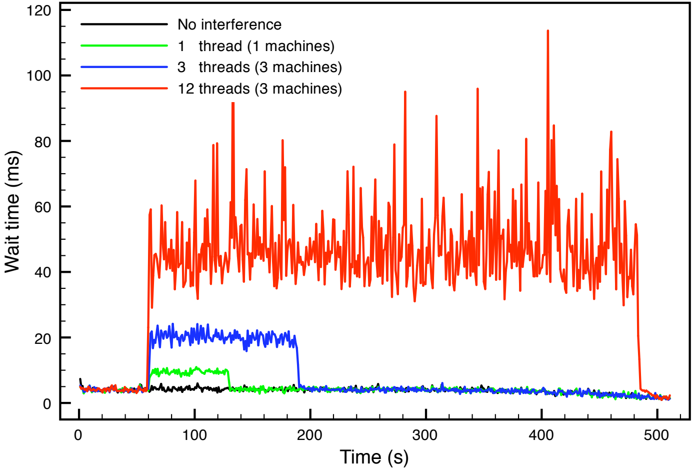

Striped logical volumes. 64KB stripe size across 10 disks
Argus
b1
TCP Connections
HP SC10
10 x 36GB 10k
iSCSI target (iet)
/dev/iscsi_0
bm
lv_b2
iqn.iscsilab:perc_b2
b2
lv_b3
iqn.iscsilab:perc_b3
b3
lv_b4
iqn.iscsilab:perc_b4
b4
vg_perc
lv_b5
iqn.iscsilab:perc_b5
b5
Block Layer
TCP/IP Layer
Block Layer

Is read response time affected by write activity ?

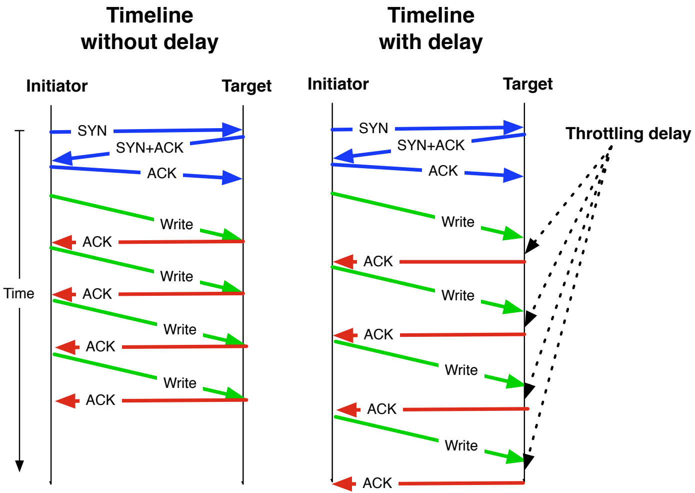


The Answer is yes

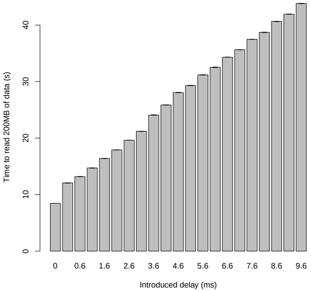
- Long response times adversely affect application service availability.

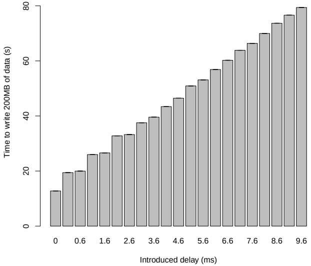


Throttling method



Relation between packet delay and average rate





- Write rate 15 MB/s - 2.5 MB/s
- Read rate 22 MB/s - 5 MB/s

Managing consumers

- Need to operate on sets of consumers (throttlable={10.0.0.243,10.0.0.244})
- Ipset: One rule to match them all

```
ipset -N $throttlable  ipmap --network 10.0.0.0/24
ipset -A $throttlable 10.0.0.243
ipset -A $throttlable 10.0.0.244
iptables --match-set $throttlable dst -j MARK --set-mark $mark
```

- The mark is a step in the range of available packet delays

Live demonstration

- Manual throttling and QoS specification
- An automatic QoS policy and automated throttling

Dynamic throttling decision

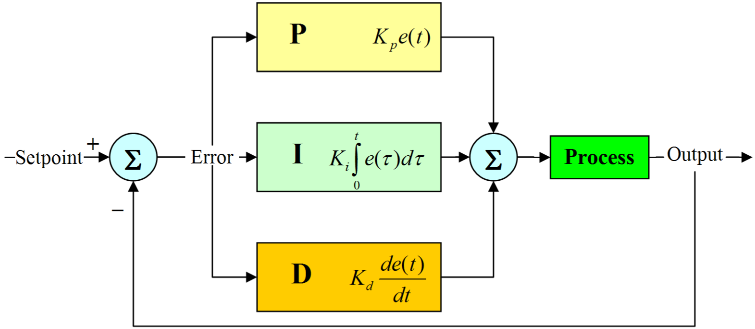
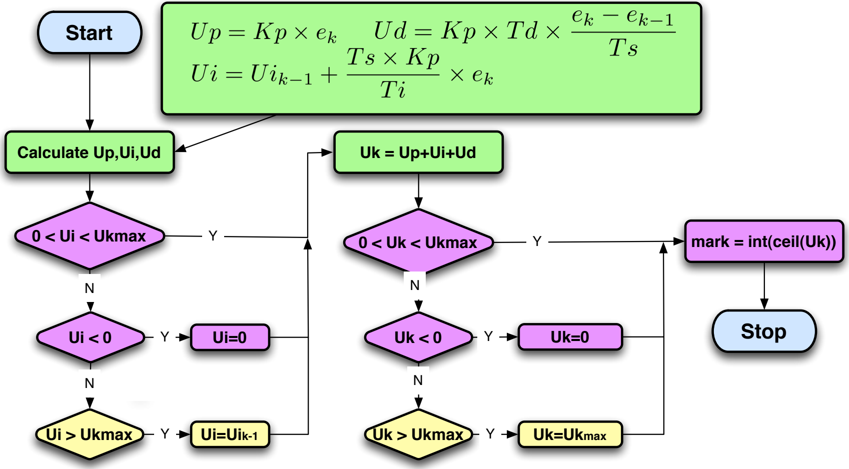


Figure: Block diagram of a PID controller. Created by SilverStar(at)en.wikipedia. Licensed under the terms of Creative Commons Attribution 2.5 Generic.

Modified PID function



Start

$$Up = Kp \times e_k \qquad Ud = Kp \times Td \times \frac{e_k - e_{k-1}}{Ts}$$

$$Ui = Ui_{k-1} + \frac{Ts \times Kp}{Ti} \times e_k$$

Calculate Up,Ui,Ud

0 < Ui < Ukmax

Ui < 0

Ui=0

Ui > Ukmax

Ui=Ui$_{k-1}$

Uk = Up+Ui+Ud

0 < Uk < Ukmax

Uk < 0

Uk=0

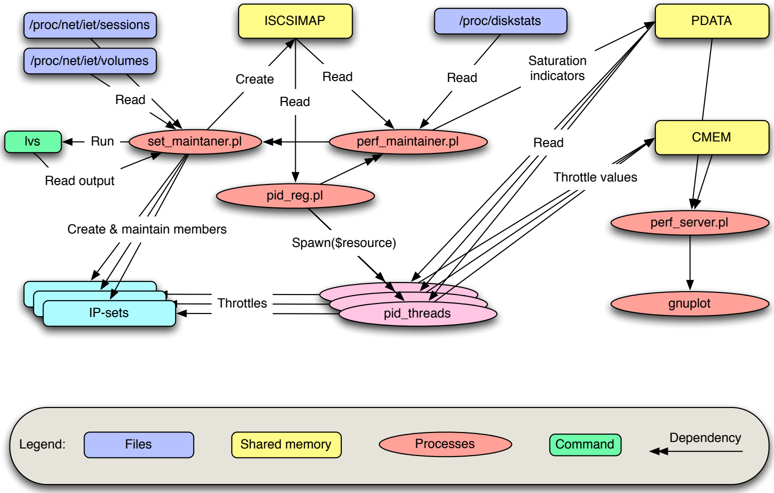
Uk > Ukmax

Uk=Uk$_{max}$

mark = int(ceil(Uk))

Stop

The completely automated approach



Impact

- The packet delay throttle is very efficient
  - Solves the throttling need completely for iSCSI (likely other TCP based storage networks too)
- The modified PID controller is consistently keeping response time low in spite of rapidly changing load interference.
- The concept is widely applicable

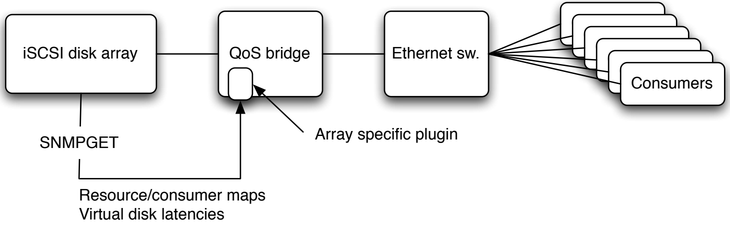
Impact

- The packet delay throttle is very efficient
  - Solves the throttling need completely for iSCSI (likely other TCP based storage networks too)
- The modified PID controller is consistently keeping response time low in spite of rapidly changing load interference.
- The concept is widely applicable

Impact

- The packet delay throttle is very efficient
  - Solves the throttling need completely for iSCSI (likely other TCP based storage networks too)
- The modified PID controller is consistently keeping response time low in spite of rapidly changing load interference.
- The concept is widely applicable

Future work



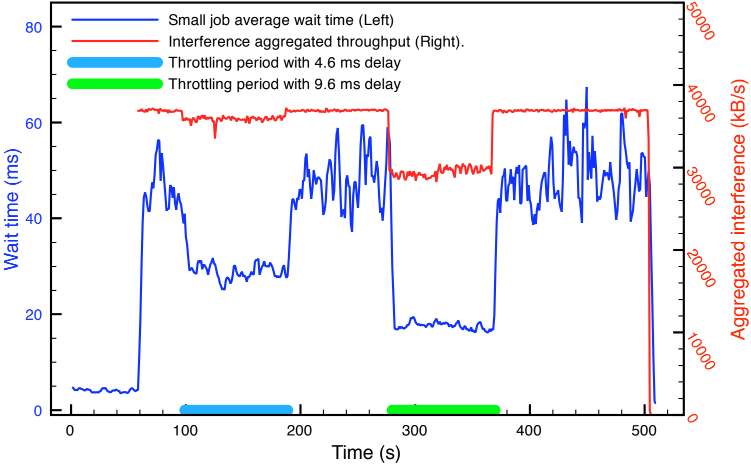
- Packet delay throttle with other algorithms
- PID controller with other throttles

Thanks for the attention !

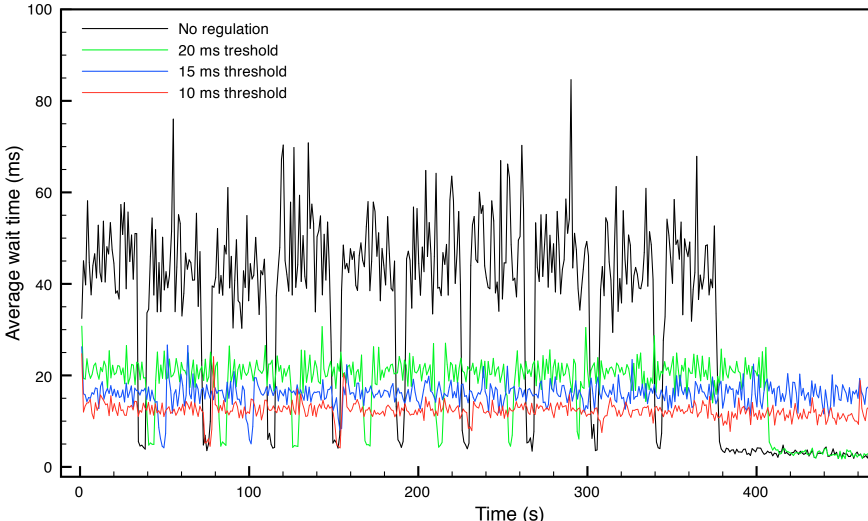
Overhead

- Negligeble overhead introduced by TC filters
- Differences measured 20 times
- t-test 99% confidence shows 0.4% / 1.7 %• overhead for read/write (worst case)

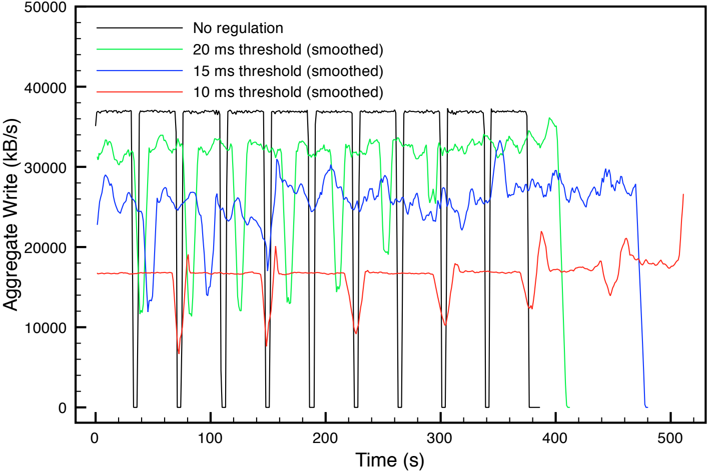
Is response time improved by throttling ?



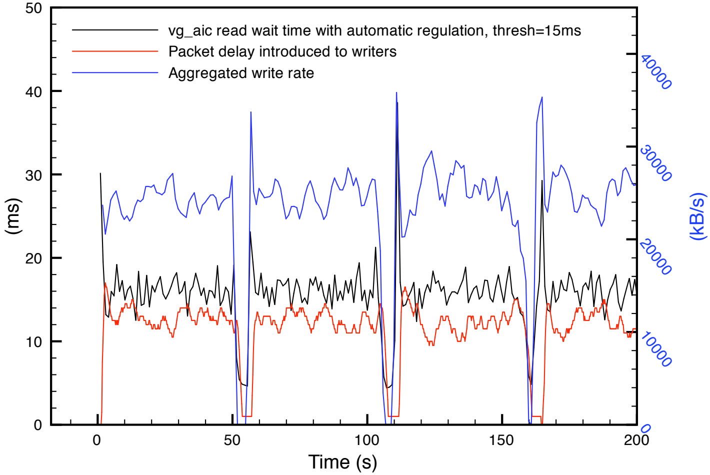
Automatically controlled wait time



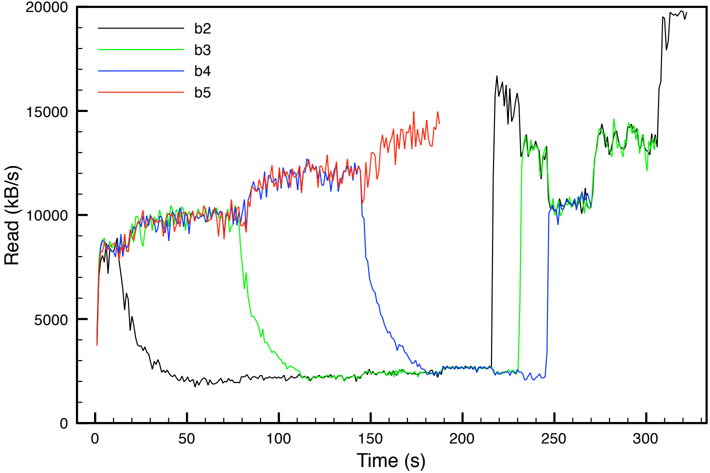
The throttled rates



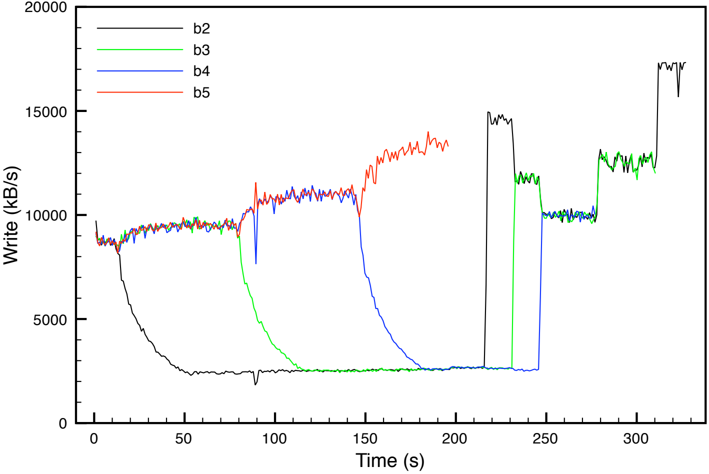
Exposing the throttling value



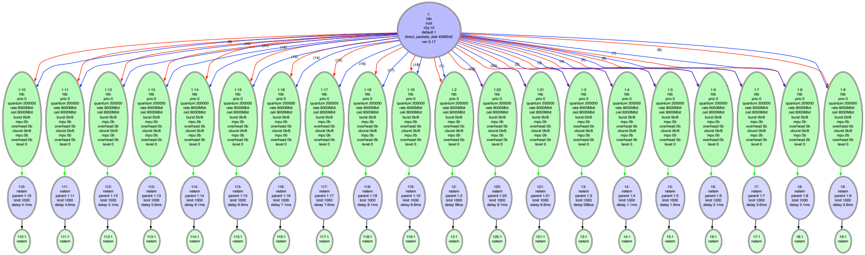
Effect of the packet delay throttle: Reads



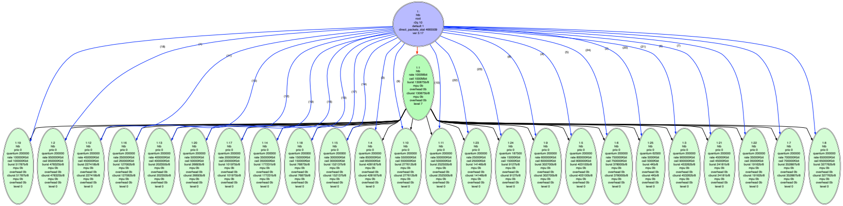
Effect of the packet delay throttle: Writes



The tc delay queues


1: htb root r2q 10 default 1 direct_packets_stat 4399042 ver 3.17
(9)
(10)
(11)
(12)
(13)
(14)
(15)
(16)
(17)
(18)
(1)
(19)
(20)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
1:10 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:11 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:12 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:13 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:14 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:15 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:16 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:17 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:18 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:19 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:2 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:20 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:21 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:3 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:4 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:5 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:6 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:7 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:8 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
1:9 htb prio 0 quantum 200000 rate 8000Mbit ceil 8000Mbit burst 0b/8 mpu 0b overhead 0b cburst 0b/8 mpu 0b overhead 0b level 0
110: netem parent 1:10 limit 1000 delay 4.1ms
111: netem parent 1:11 limit 1000 delay 4.6ms
112: netem parent 1:12 limit 1000 delay 5.1ms
113: netem parent 1:13 limit 1000 delay 5.6ms
114: netem parent 1:14 limit 1000 delay 6.1ms
115: netem parent 1:15 limit 1000 delay 6.6ms
116: netem parent 1:16 limit 1000 delay 7.1ms
117: netem parent 1:17 limit 1000 delay 7.6ms
118: netem parent 1:18 limit 1000 delay 8.1ms
119: netem parent 1:19 limit 1000 delay 8.6ms
12: netem parent 1:2 limit 1000 delay 99us
120: netem parent 1:20 limit 1000 delay 9.1ms
121: netem parent 1:21 limit 1000 delay 9.6ms
13: netem parent 1:3 limit 1000 delay 598us
14: netem parent 1:4 limit 1000 delay 1.1ms
15: netem parent 1:5 limit 1000 delay 1.6ms
16: netem parent 1:6 limit 1000 delay 2.1ms
17: netem parent 1:7 limit 1000 delay 2.6ms
18: netem parent 1:8 limit 1000 delay 3.1ms
19: netem parent 1:9 limit 1000 delay 3.6ms
110:1 netem
111:1 netem
112:1 netem
113:1 netem
114:1 netem
115:1 netem
116:1 netem
117:1 netem
118:1 netem
119:1 netem
12:1 netem
120:1 netem
121:1 netem
13:1 netem
14:1 netem
15:1 netem
16:1 netem
17:1 netem
18:1 netem
19:1 netem

The tc bandwidth queues



1:
htb
root
r2q 10
default 1
direct_packets_stat 4665509
ver 3.17

1:1
htb
rate 1000Mbit
ceil 1000Mbit
burst 130875b/8
mpu 0b
overhead 0b
cburst 130875b/8
mpu 0b
overhead 0b
level 7

(18) (1) (11) (15) (12) (19) (16) (13) (17) (14) (3) (9) (10) (22) (23) (8) (4) (5) (24) (2) (20) (21) (6) (7)

1:19
htb
prio 0
quantum 200000
rate 100000Kbit
ceil 100000Kbit
burst 51787b/8
mpu 0b
overhead 0b
cburst 51787b/8
mpu 0b
overhead 0b
level 0

1:2
htb
prio 0
quantum 200000
rate 950000Kbit
ceil 950000Kbit
burst 478325b/8
mpu 0b
overhead 0b
cburst 478325b/8
mpu 0b
overhead 0b
level 0

1:12
htb
prio 0
quantum 200000
rate 450000Kbit
ceil 450000Kbit
burst 227418b/8
mpu 0b
overhead 0b
cburst 227418b/8
mpu 0b
overhead 0b
level 0

1:16
htb
prio 0
quantum 200000
rate 250000Kbit
ceil 250000Kbit
burst 127062b/8
mpu 0b
overhead 0b
cburst 127062b/8
mpu 0b
overhead 0b
level 0

1:13
htb
prio 0
quantum 200000
rate 400000Kbit
ceil 400000Kbit
burst 202350b/8
mpu 0b
overhead 0b
cburst 202350b/8
mpu 0b
overhead 0b
level 0

1:20
htb
prio 0
quantum 200000
rate 50000Kbit
ceil 50000Kbit
burst 26893b/8
mpu 0b
overhead 0b
cburst 26893b/8
mpu 0b
overhead 0b
level 0

1:17
htb
prio 0
quantum 200000
rate 200000Kbit
ceil 200000Kbit
burst 101975b/8
mpu 0b
overhead 0b
cburst 101975b/8
mpu 0b
overhead 0b
level 0

1:14
htb
prio 0
quantum 200000
rate 350000Kbit
ceil 350000Kbit
burst 177231b/8
mpu 0b
overhead 0b
cburst 177231b/8
mpu 0b
overhead 0b
level 0

1:18
htb
prio 0
quantum 200000
rate 150000Kbit
ceil 150000Kbit
burst 76875b/8
mpu 0b
overhead 0b
cburst 76875b/8
mpu 0b
overhead 0b
level 0

1:15
htb
prio 0
quantum 200000
rate 300000Kbit
ceil 300000Kbit
burst 152137b/8
mpu 0b
overhead 0b
cburst 152137b/8
mpu 0b
overhead 0b
level 0

1:4
htb
prio 0
quantum 200000
rate 850000Kbit
ceil 850000Kbit
burst 428187b/8
mpu 0b
overhead 0b
cburst 428187b/8
mpu 0b
overhead 0b
level 0

1:10
htb
prio 0
quantum 200000
rate 550000Kbit
ceil 550000Kbit
burst 277612b/8
mpu 0b
overhead 0b
cburst 277612b/8
mpu 0b
overhead 0b
level 0

1:11
htb
prio 0
quantum 200000
rate 500000Kbit
ceil 500000Kbit
burst 252500b/8
mpu 0b
overhead 0b
cburst 252500b/8
mpu 0b
overhead 0b
level 0

1:23
htb
prio 0
quantum 200000
rate 25000Kbit
ceil 25000Kbit
burst 14146b/8
mpu 0b
overhead 0b
cburst 14146b/8
mpu 0b
overhead 0b
level 0

1:24
htb
prio 0
quantum 187500
rate 15000Kbit
ceil 15000Kbit
burst 9127b/8
mpu 0b
overhead 0b
cburst 9127b/8
mpu 0b
overhead 0b
level 0

1:9
htb
prio 0
quantum 200000
rate 600000Kbit
ceil 600000Kbit
burst 302700b/8
mpu 0b
overhead 0b
cburst 302700b/8
mpu 0b
overhead 0b
level 0

1:5
htb
prio 0
quantum 200000
rate 800000Kbit
ceil 800000Kbit
burst 403100b/8
mpu 0b
overhead 0b
cburst 403100b/8
mpu 0b
overhead 0b
level 0

1:6
htb
prio 0
quantum 200000
rate 750000Kbit
ceil 750000Kbit
burst 378000b/8
mpu 0b
overhead 0b
cburst 378000b/8
mpu 0b
overhead 0b
level 0

1:25
htb
prio 0
quantum 62500
rate 5000Kbit
ceil 5000Kbit
burst 4Kb/8
mpu 0b
overhead 0b
cburst 4Kb/8
mpu 0b
overhead 0b
level 0

1:3
htb
prio 0
quantum 200000
rate 900000Kbit
ceil 900000Kbit
burst 453262b/8
mpu 0b
overhead 0b
cburst 453262b/8
mpu 0b
overhead 0b
level 0

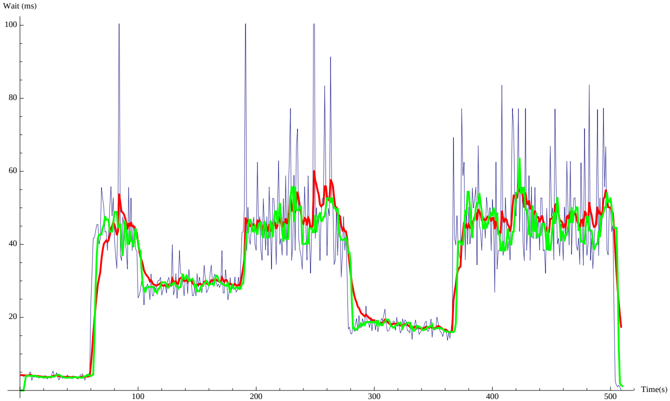
1:21
htb
prio 0
quantum 200000
rate 45000Kbit
ceil 45000Kbit
burst 24181b/8
mpu 0b
overhead 0b
cburst 24181b/8
mpu 0b
overhead 0b
level 0

1:22
htb
prio 0
quantum 200000
rate 35000Kbit
ceil 35000Kbit
burst 19162b/8
mpu 0b
overhead 0b
cburst 19162b/8
mpu 0b
overhead 0b
level 0

1:7
htb
prio 0
quantum 200000
rate 700000Kbit
ceil 700000Kbit
burst 352887b/8
mpu 0b
overhead 0b
cburst 352887b/8
mpu 0b
overhead 0b
level 0

1:8
htb
prio 0
quantum 200000
rate 650000Kbit
ceil 650000Kbit
burst 327762b/8
mpu 0b
overhead 0b
cburst 327762b/8
mpu 0b
overhead 0b
level 0

Input signal



- Red: Exponential Weighted Moving Average (EWMA)
- Green: Moving median
- $L_{(t)} = l_{(t)}\alpha + L_{(t-1)}(1 - \alpha)$
- EWMA, also called low pass filter

$$u(t) = \overbrace{\underbrace{K_p e(t)}_{\text{Proportional}} + \underbrace{\frac{K_p}{T_i} \int_0^t e(\tau) d\tau}_{\text{Integral}} + \underbrace{K_p T_d e'(t)}_{\text{Derivative}}}^{\textit{Continous}}$$

$$u_k = \underbrace{\underbrace{u_{k-1}}_{\textit{Previous}} + \underbrace{K_p(1 + \frac{T}{T_i})e_k - K_p e_{k-1} + \frac{K_p T_d}{T}(e_k - 2e_{k-1} + e_{k-2})}_{\textit{Delta}}}_{\textit{Incremental form}}$$

$$u_k = \underbrace{\underbrace{K_p e_k}_{\textit{Proportional}} + \underbrace{u_{i(k-1)} + \frac{K_p T}{Ti} e_k}_{\textit{Integral}} + \underbrace{\frac{K_p T_d}{T}(e_k - e_{k-1})}_{\textit{Derivative}}}_{\textit{Absolute form}}$$