My experience, results and remarks to BW estimation and CT measurement tools

Ji_í Navrátil
SLAC

My interest is concerned on

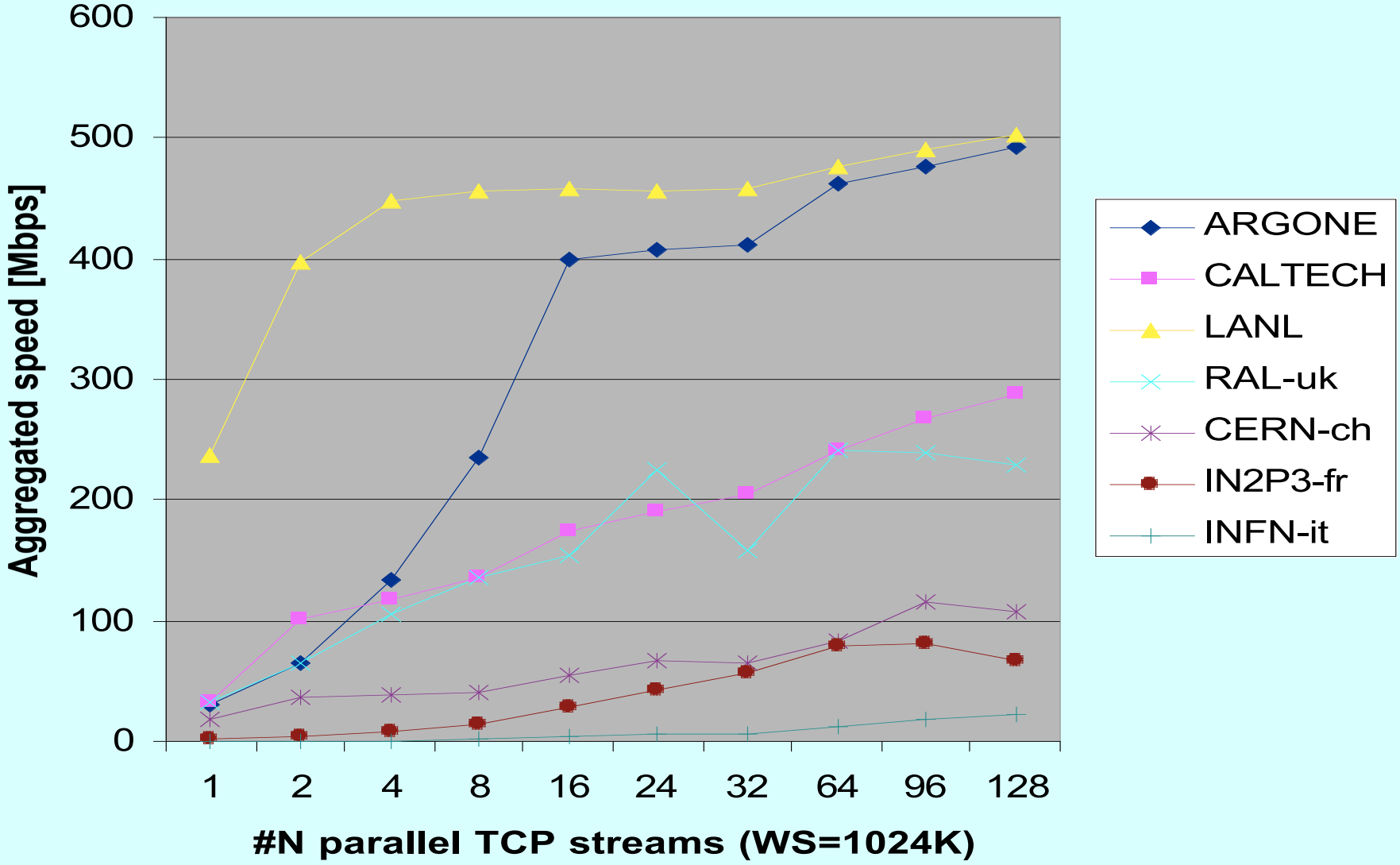
- Full BW utilization on the path
- Tools for BW estimation
- Improving of CT detection

The main goal of our daily effort is

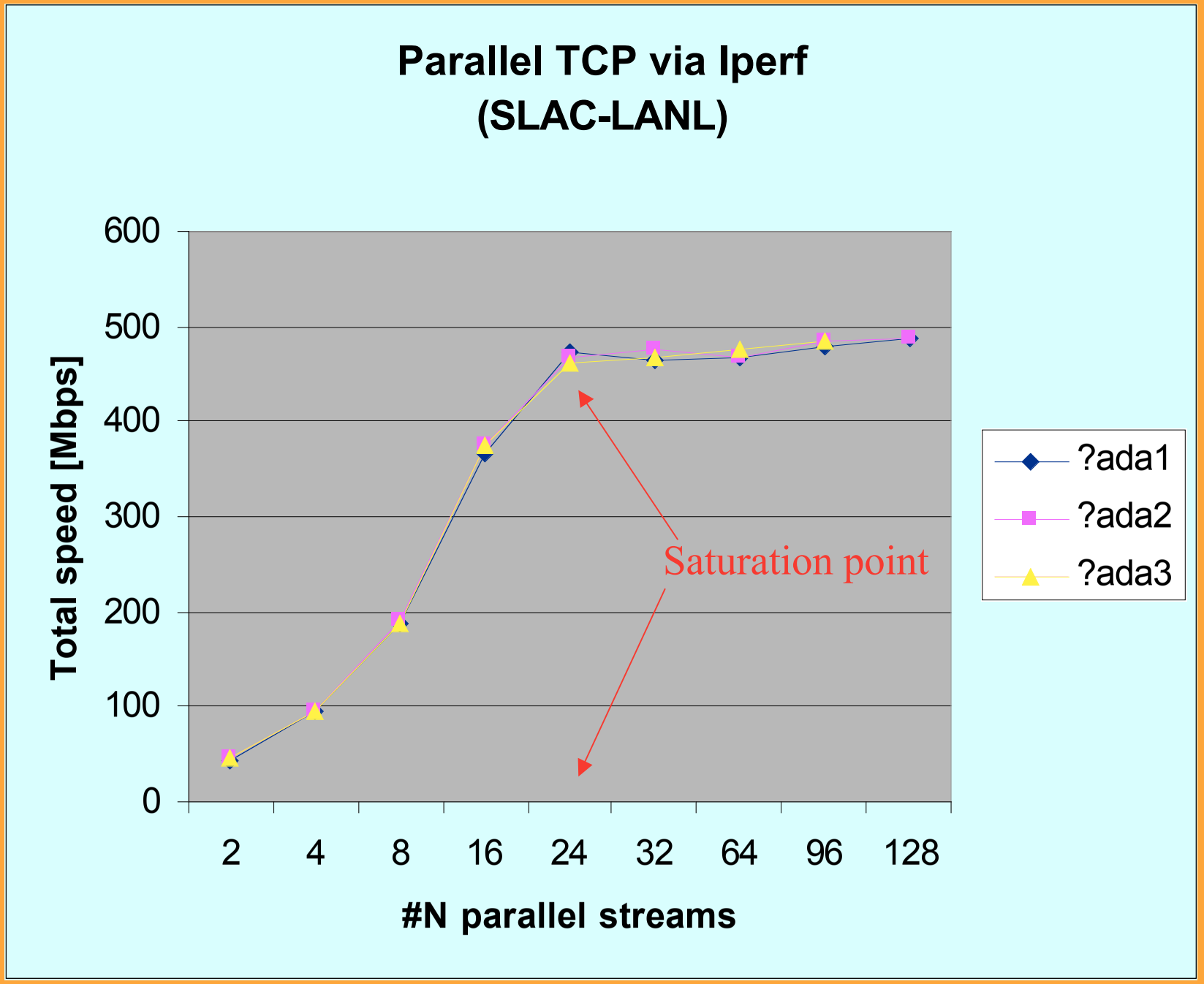
Maximum BW utilization (inspiration)

- **Les Cottrell IEPM metrics**
- **Jin Guojun netest-2 recommendation**
- **Feng Wu-Chun simulations**

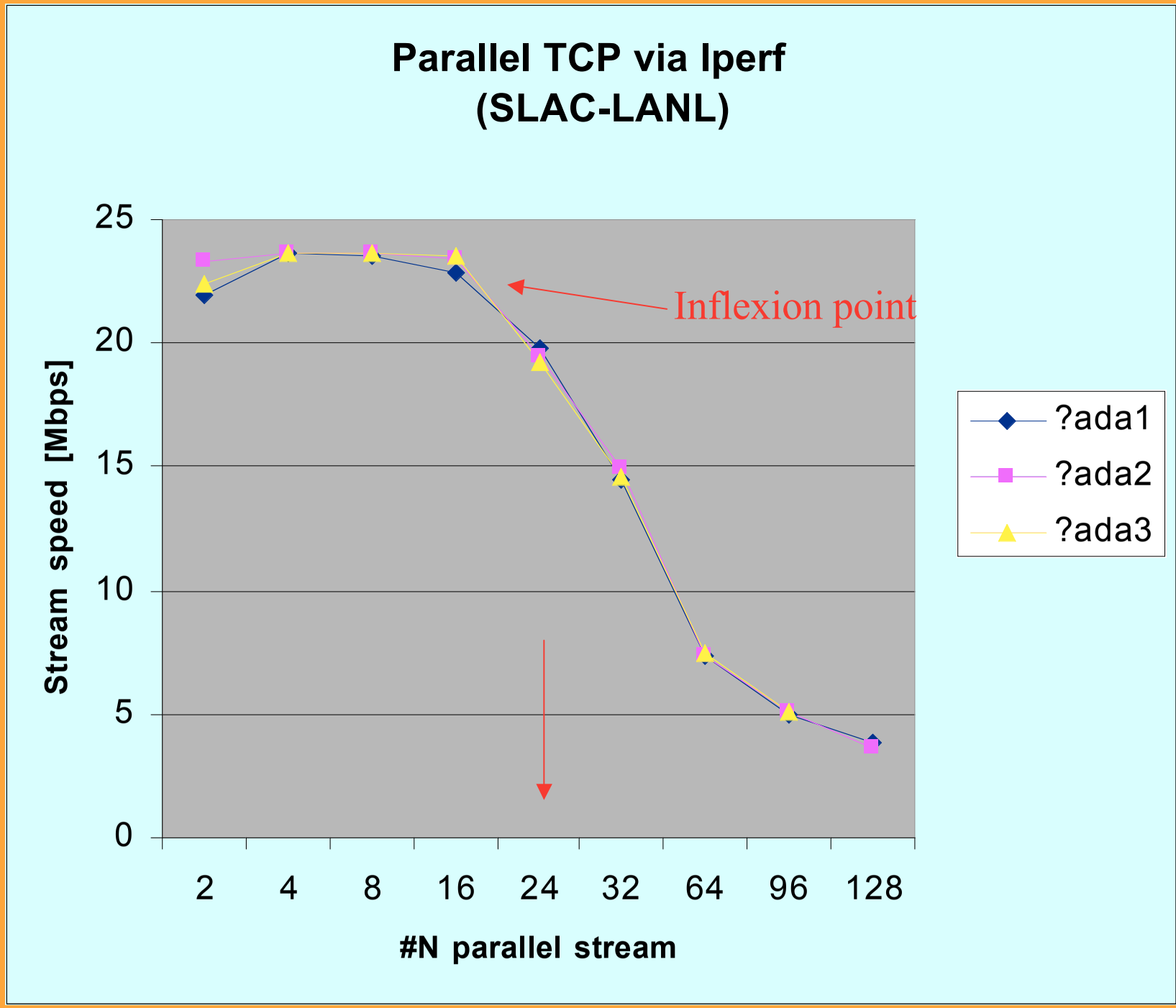
Parallel TCP via Iperf (SLAC to LABS)

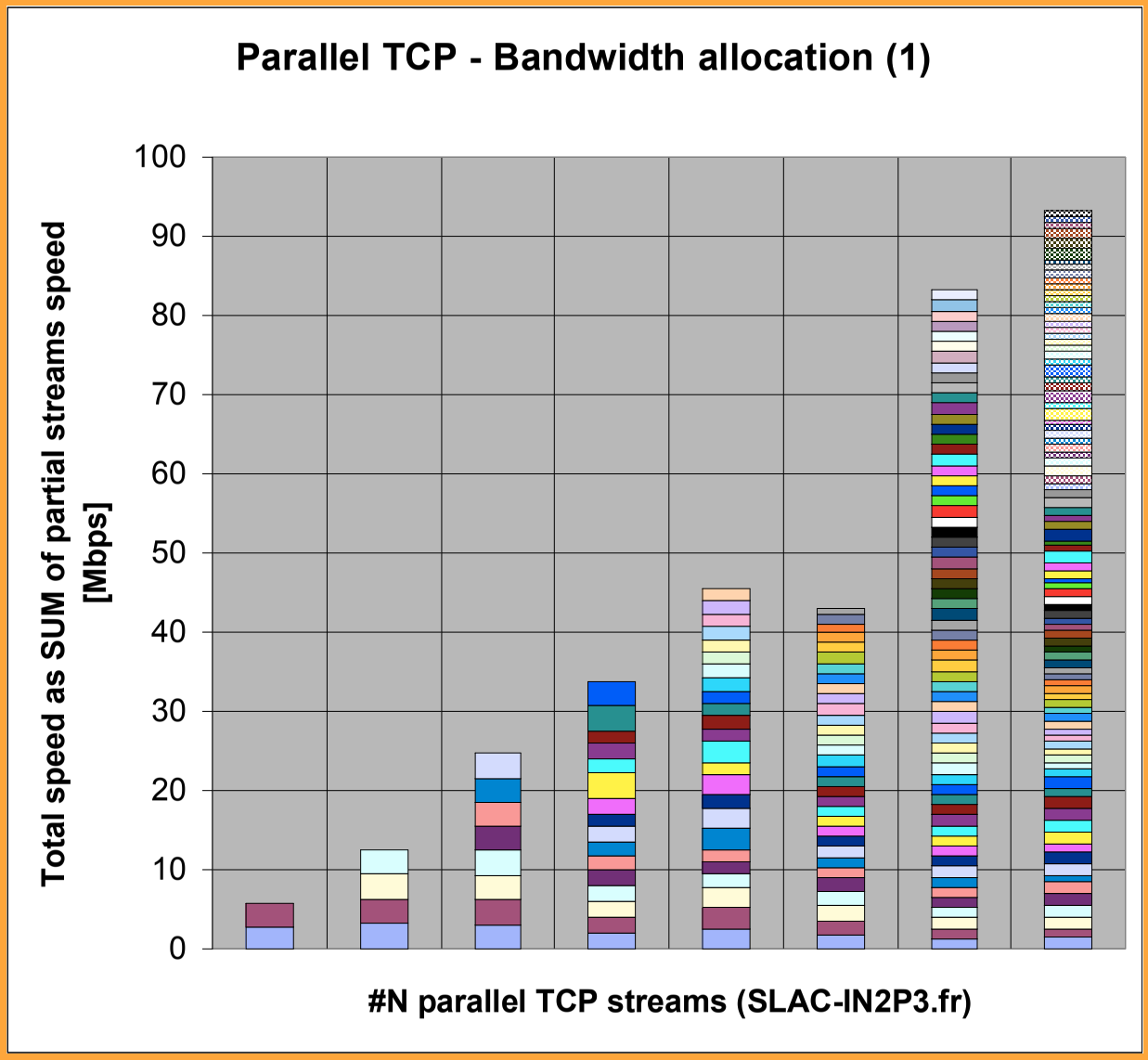


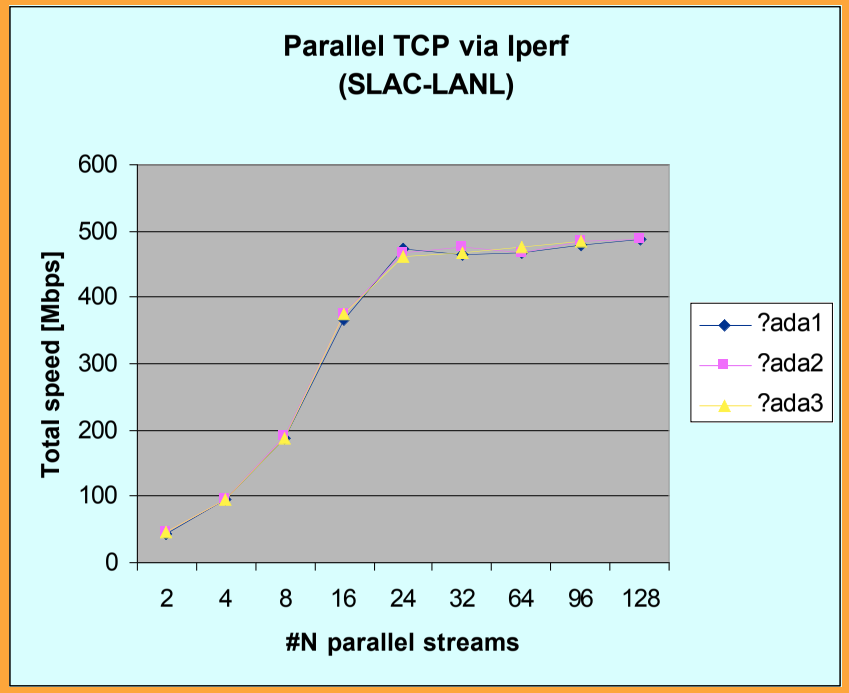
- **Area of saturation = area of FULL utilization**
  - **Parallel streams**
  - **TCP windows size (not sufficient for HSL)**
  - **Combination of both methods**

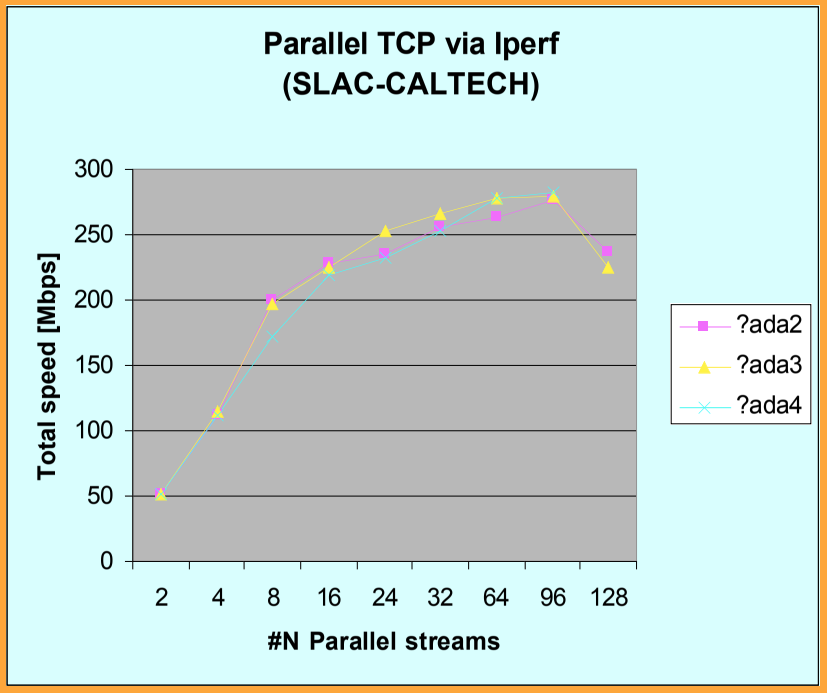

Parallel TCP via Iperf
(SLAC-LANL)
Total speed [Mbps]
600
500
400
300
200
100
0
2
4
8
16
24
32
64
96
128
#N parallel streams
Saturation point
?ada1
?ada2
?ada3

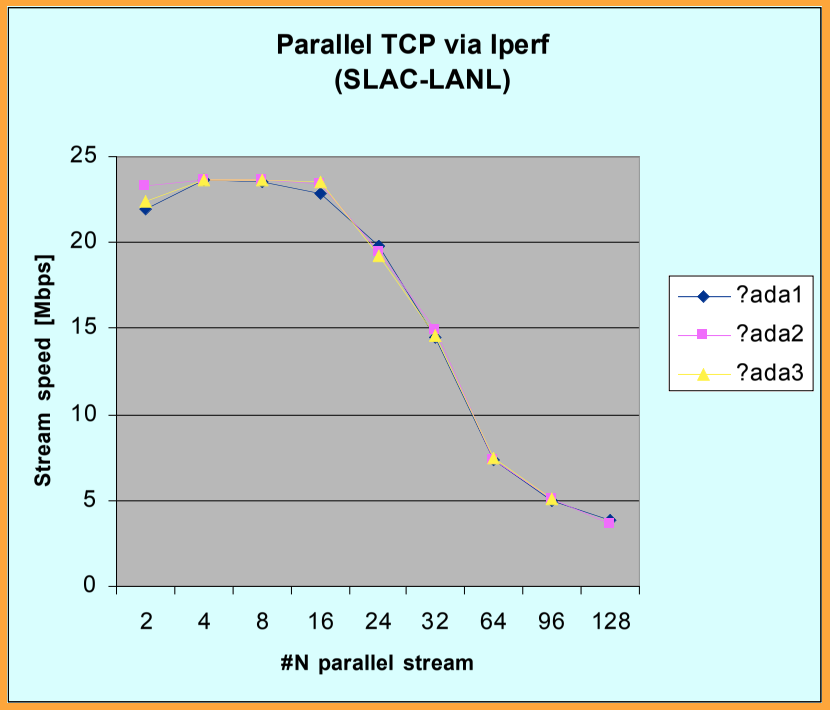
0	Mbytes	Mbits	Median	Average	Sdev
1	26,4	22			
2	26,3	22			
0	**52,7**	**44**	**22**	**22**	**0**
1	28,3	23,6			
2	28,3	23,6			
3	28,4	23,7			
4	28,4	23,7			
0	**113,4**	**94,6**	**23,65**	**23,65**	**0,057735**
1	28,3	23,6			
2	28,6	23,9			
3	27,7	23,2			
4	28,6	23,9			
5	28,3	23,7			
6	28	23,4			
7	26,8	22,4			
8	28,4	23,7			
0	**224,7**	**188**	**23,65**	**23,475**	**0,494975**

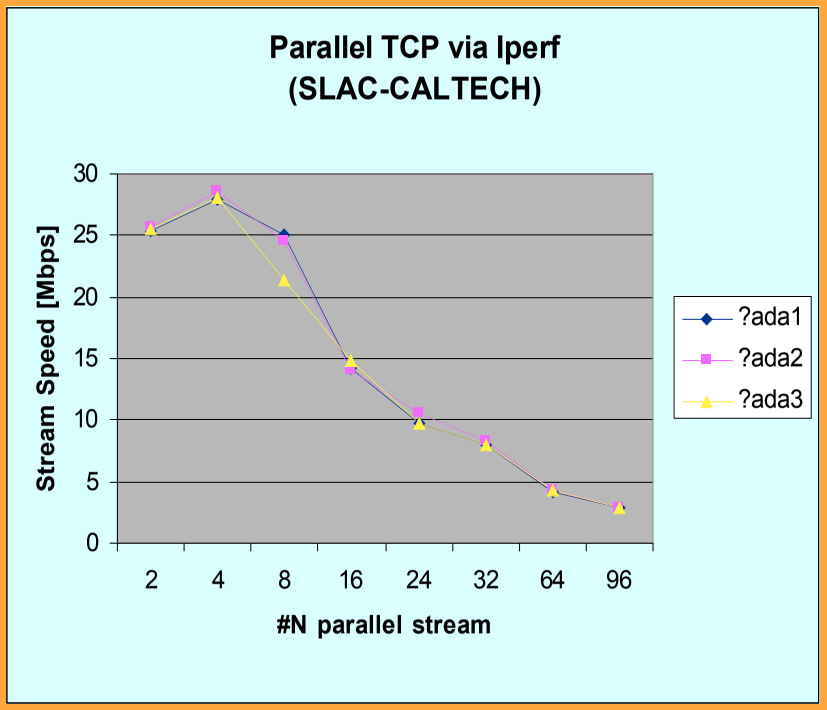

Parallel TCP via Iperf
(SLAC-LANL)
Stream speed [Mbps]
25
20
15
10
5
0
2
4
8
16
24
32
64
96
128
#N parallel stream
Inflexion point
?ada1
?ada2
?ada3


Parallel TCP - Bandwidth allocation (1)
Total speed as SUM of partial streams speed [Mbps]
100
90
80
70
60
50
40
30
20
10
0
#N parallel TCP streams (SLAC-IN2P3.fr)


Parallel TCP via Iperf
(SLAC-LANL)
Total speed [Mbps]
600
500
400
300
200
100
0
2
4
8
16
24
32
64
96
128
#N parallel streams
?ada1
?ada2
?ada3


Parallel TCP via Iperf
(SLAC-CALTECH)
Total speed [Mbps]
300
250
200
150
100
50
0
2
4
8
16
24
32
64
96
128
#N Parallel streams
?ada2
?ada3
?ada4


Parallel TCP via Iperf
(SLAC-LANL)
Stream speed [Mbps]
25
20
15
10
5
0
2
4
8
16
24
32
64
96
128
#N parallel stream
?ada1
?ada2
?ada3

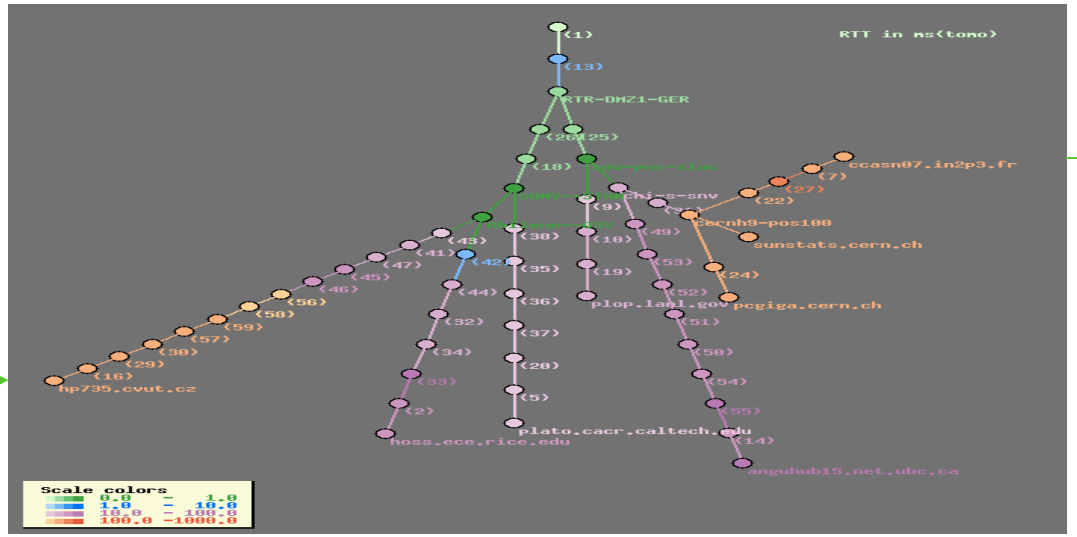

Parallel TCP via Iperf
(SLAC-CALTECH)
Stream Speed [Mbps]
30
25
20
15
10
5
0
2
4
8
16
24
32
64
96
#N parallel stream
?ada1
?ada2
?ada3

Why such distribution of throughput of individual streams

?

Internet path

In



Out

Virtual queue

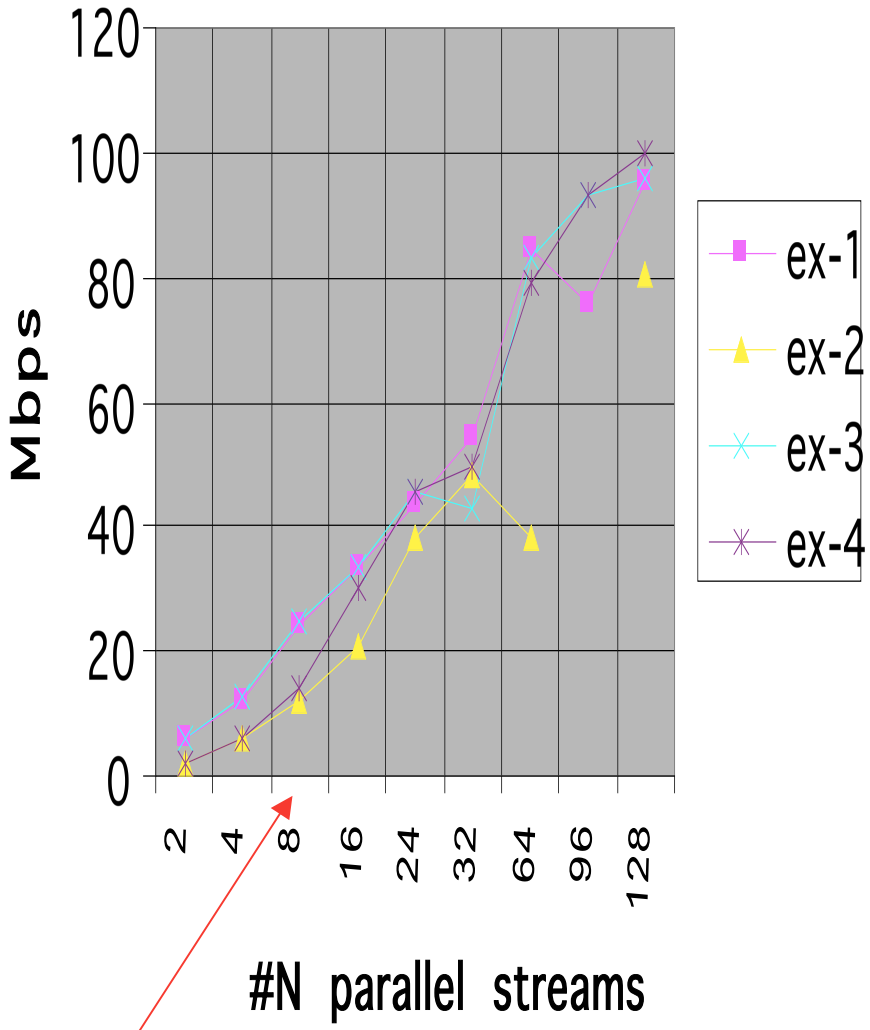
- All parallel streams share same “virtual queue”
- All “my traffic” share same queue with outside CT

Problem on slow and high loaded lines is

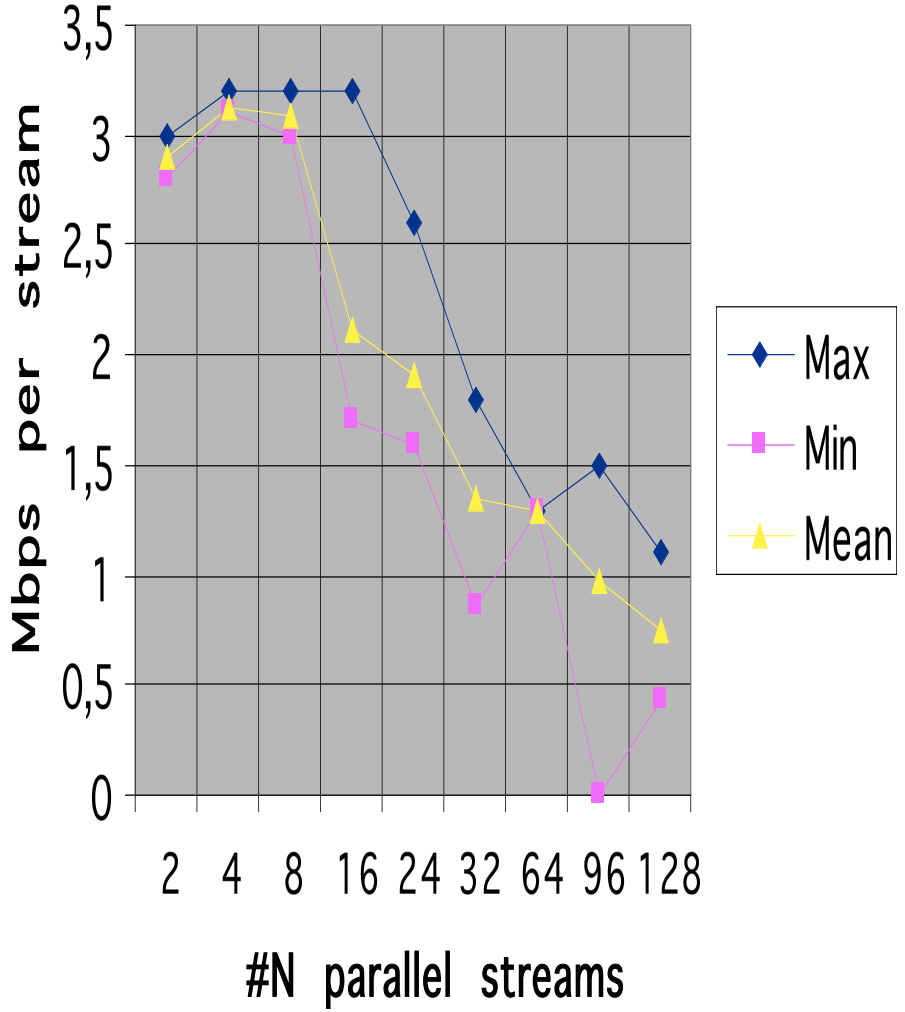
visible in time reports (In iperf)
visible in statistics for “streams speed”

- 10 second in Iperf means 10 second of sending ! Not receiving.

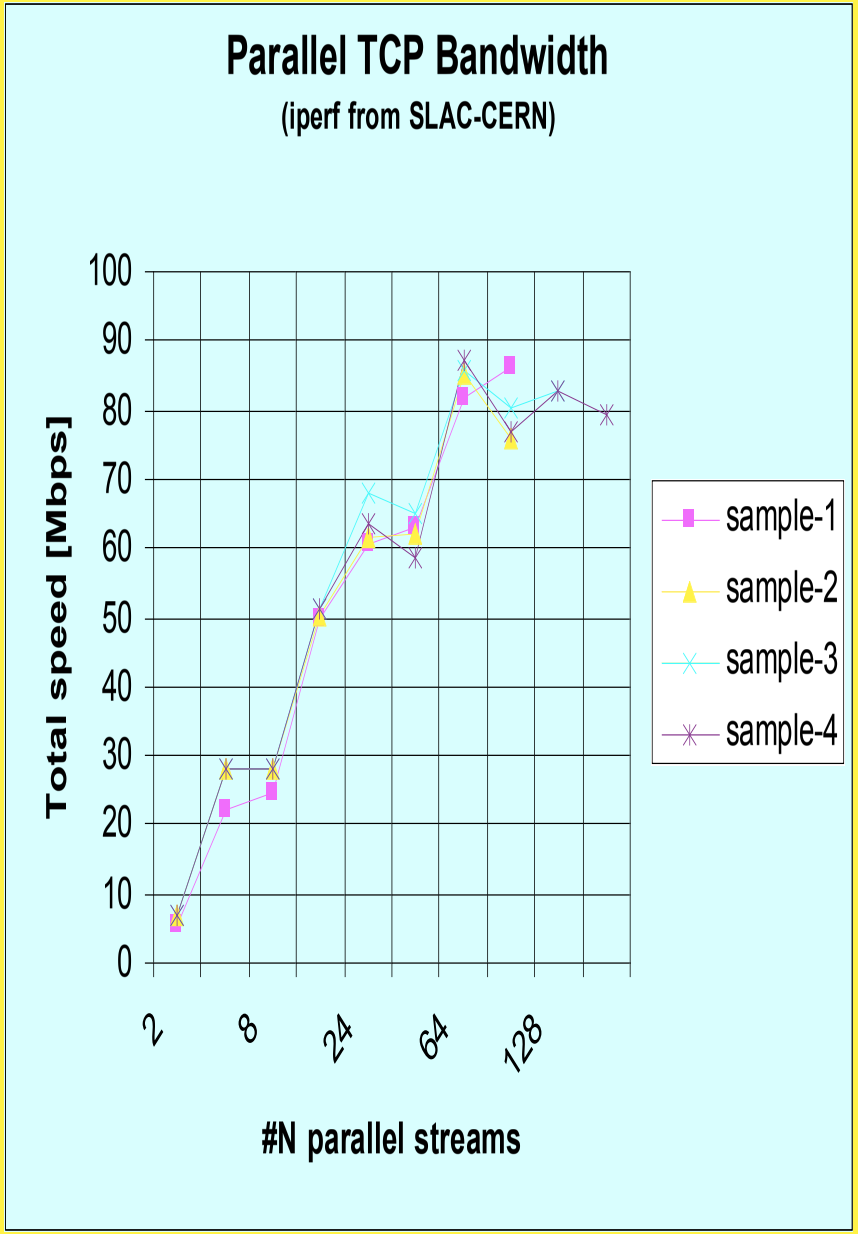
Aggregated speed (SLAC-IN2P3)

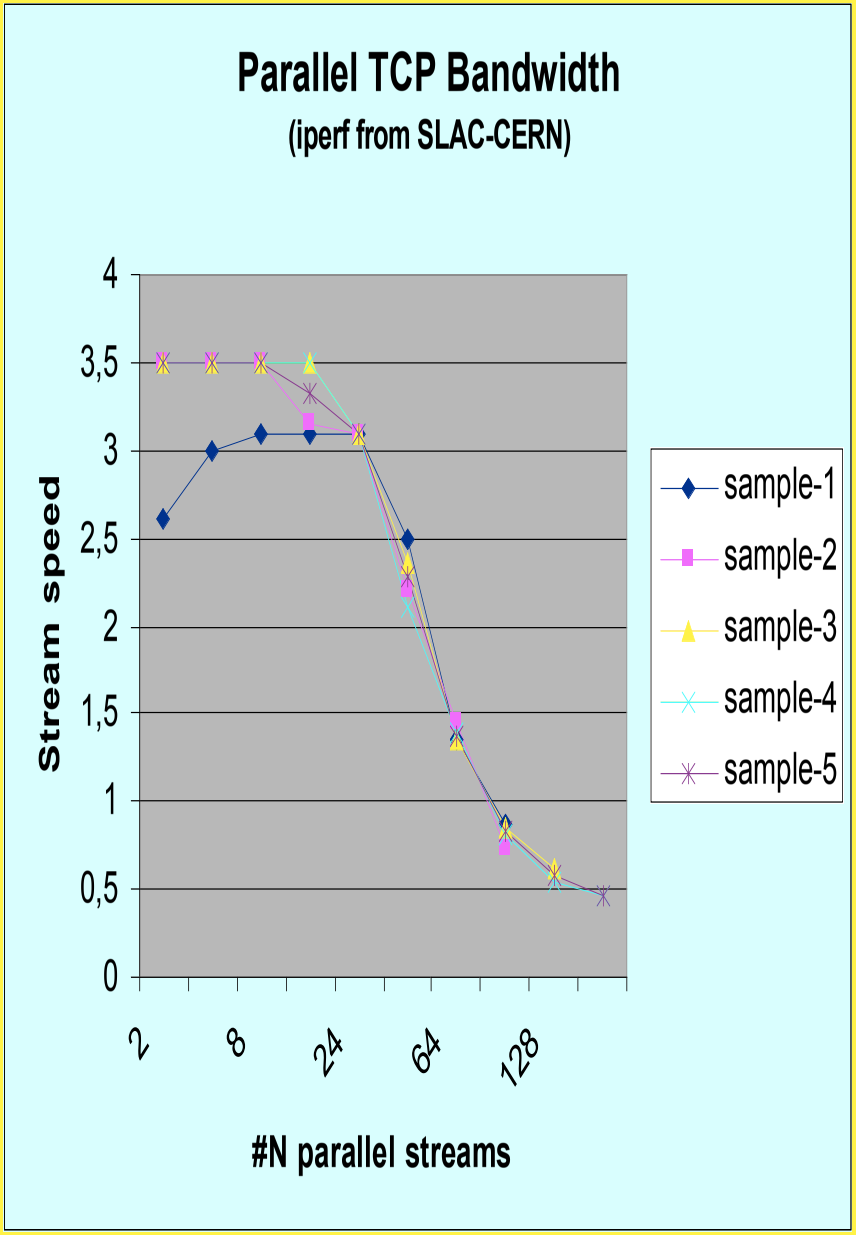


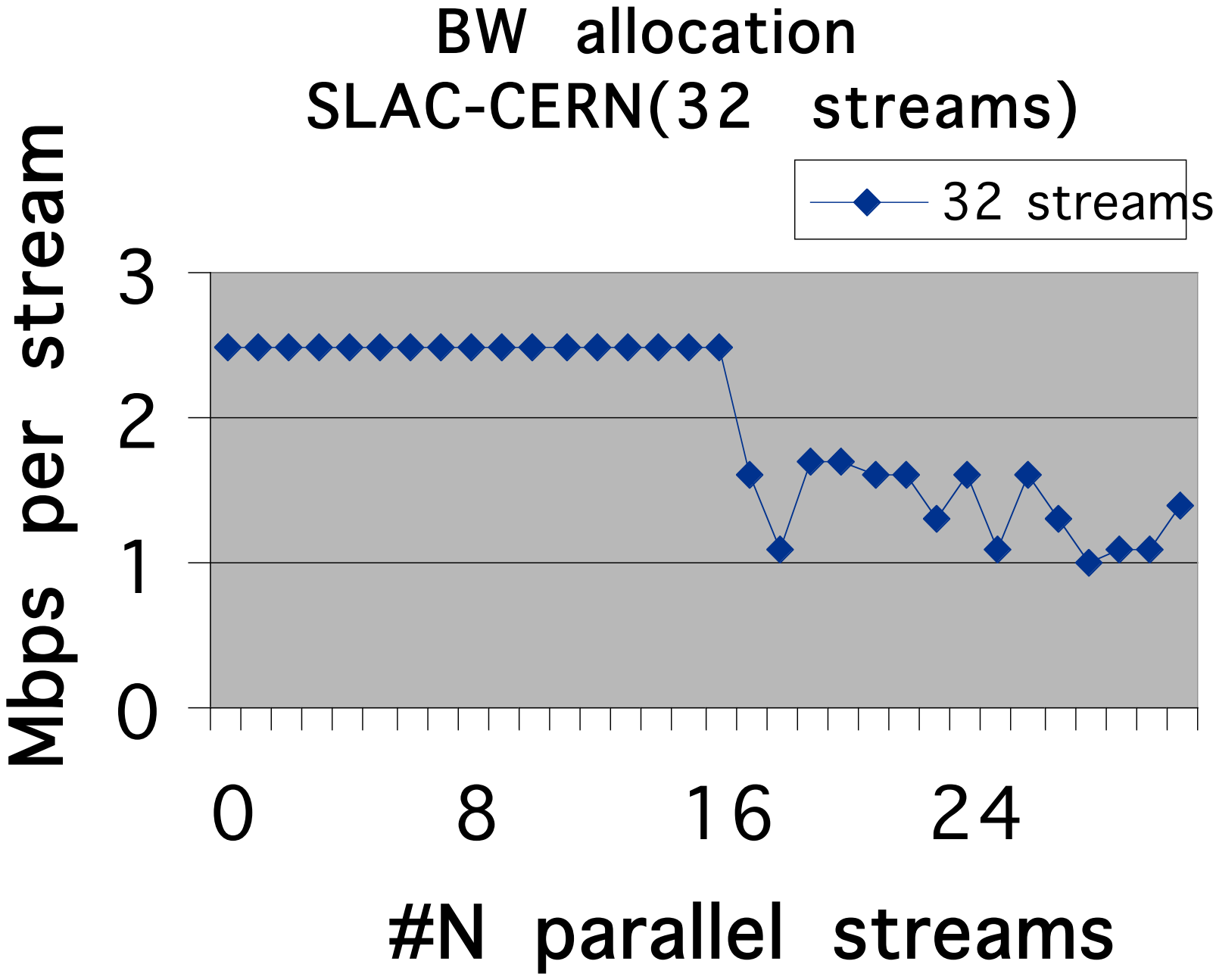
Bandwidth allocation (SLAC-IN2P2)

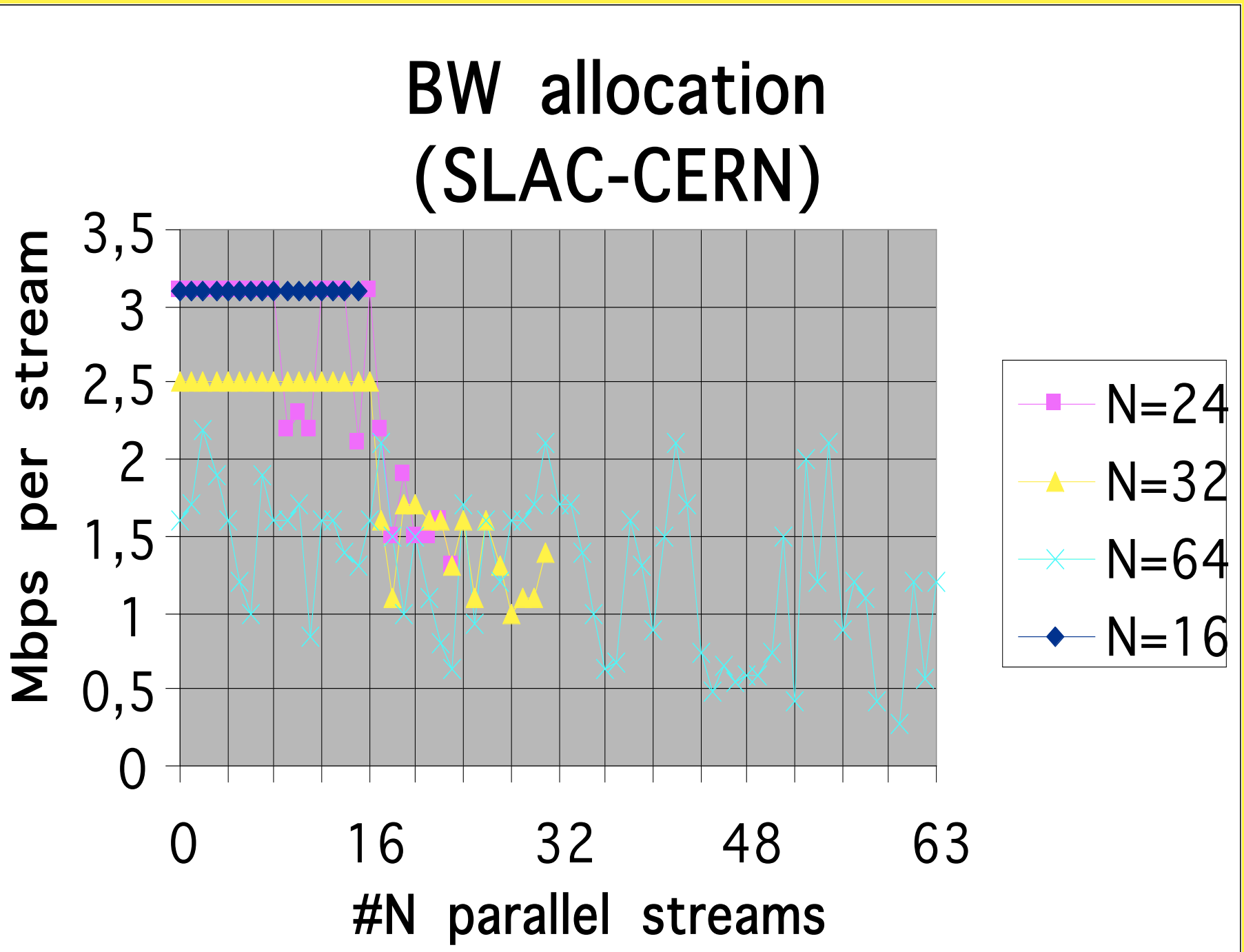


Metric with 8 parallel streams doesn't fit to the BW estimation

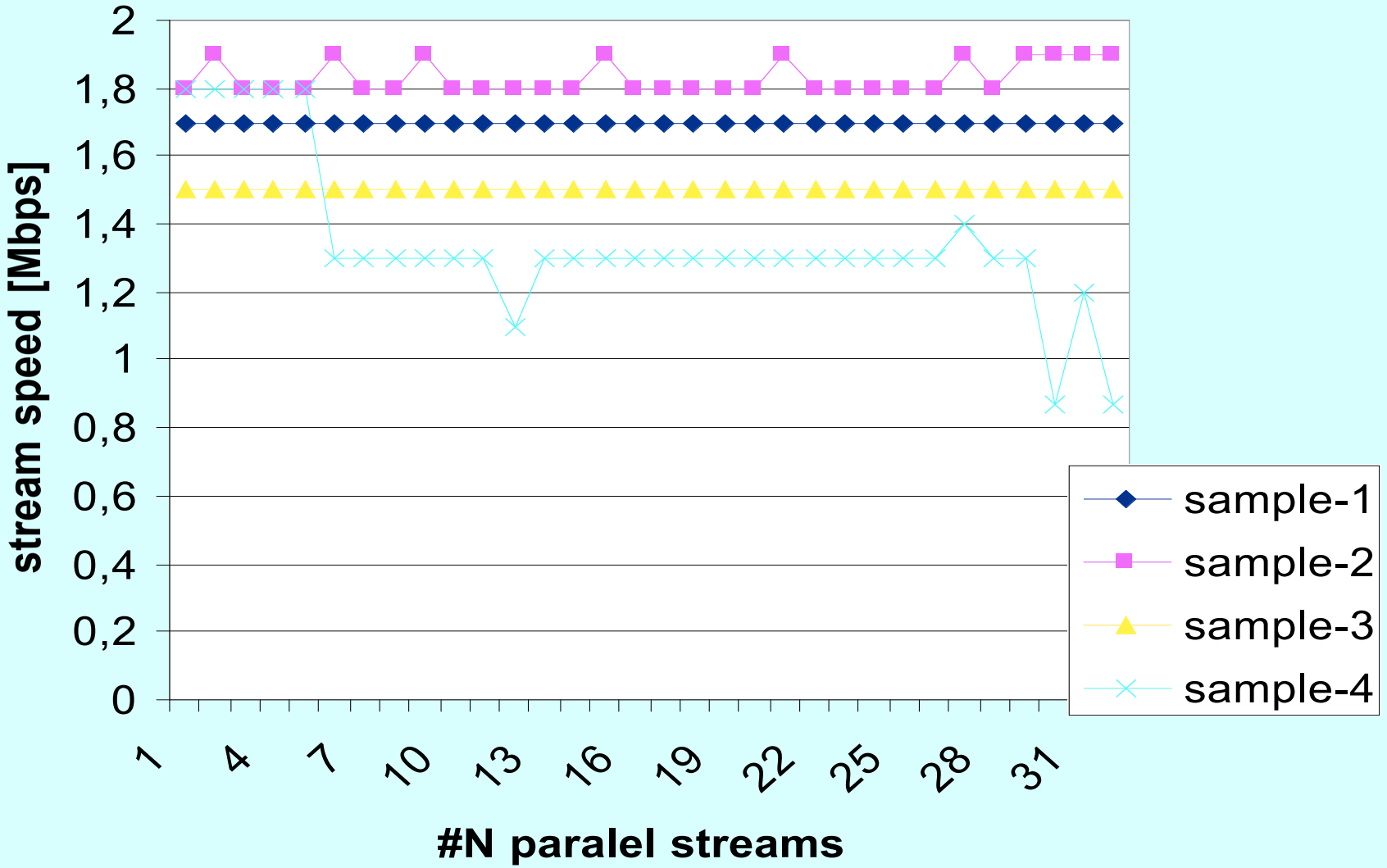

Parallel TCP Bandwidth
(iperf from SLAC-CERN)
Total speed [Mbps]
100
90
80
70
60
50
40
30
20
10
0
2
8
24
64
128
#N parallel streams
sample-1
sample-2
sample-3
sample-4


Parallel TCP Bandwidth
(iperf from SLAC-CERN)
Stream speed
4
3,5
3
2,5
2
1,5
1
0,5
0
2
8
24
64
128
#N parallel streams
sample-1
sample-2
sample-3
sample-4
sample-5

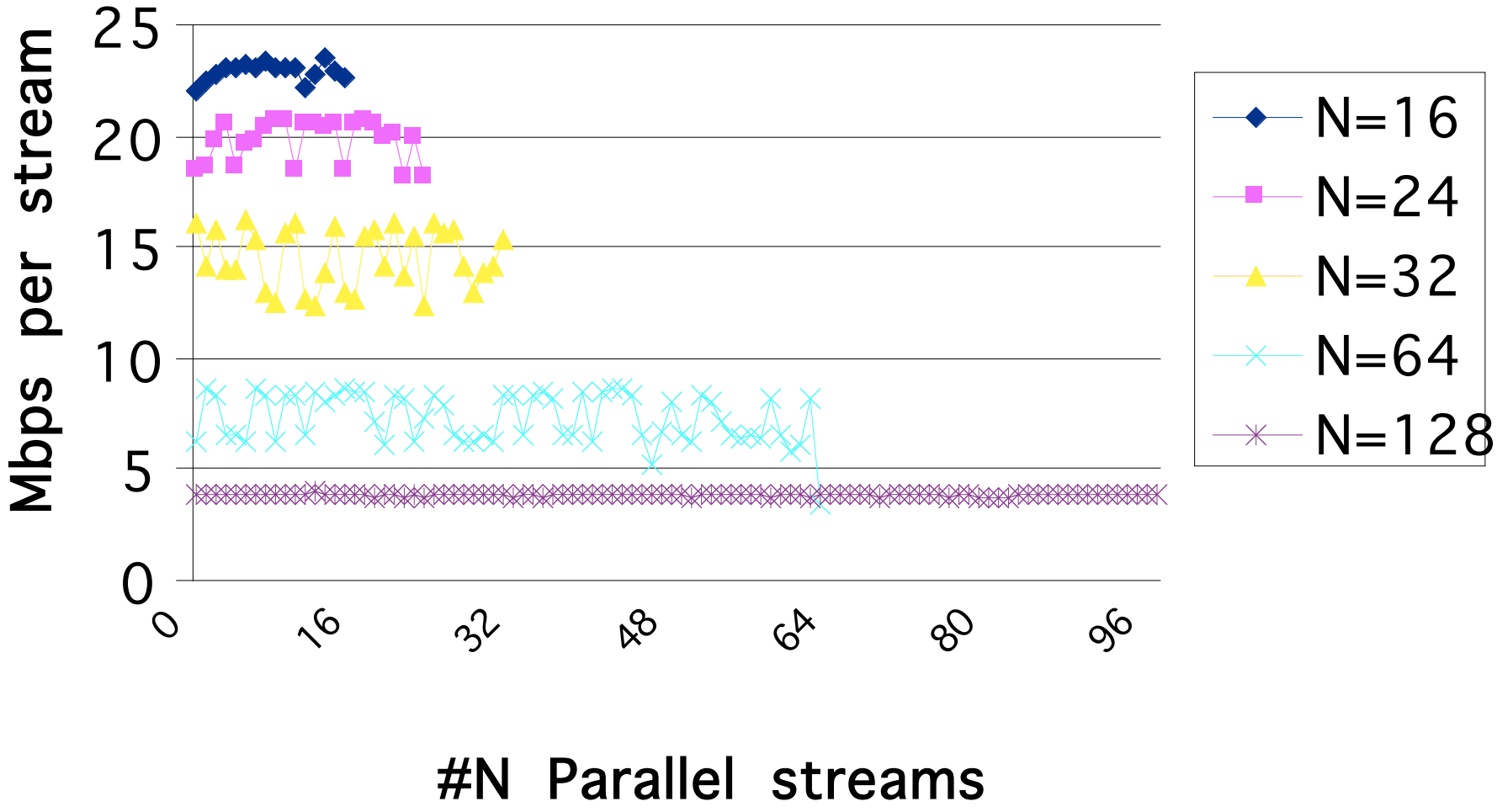

BW allocation
SLAC-CERN(32 streams)
32 streams
Mbps per stream
3
2
1
0
0
8
16
24
#N parallel streams

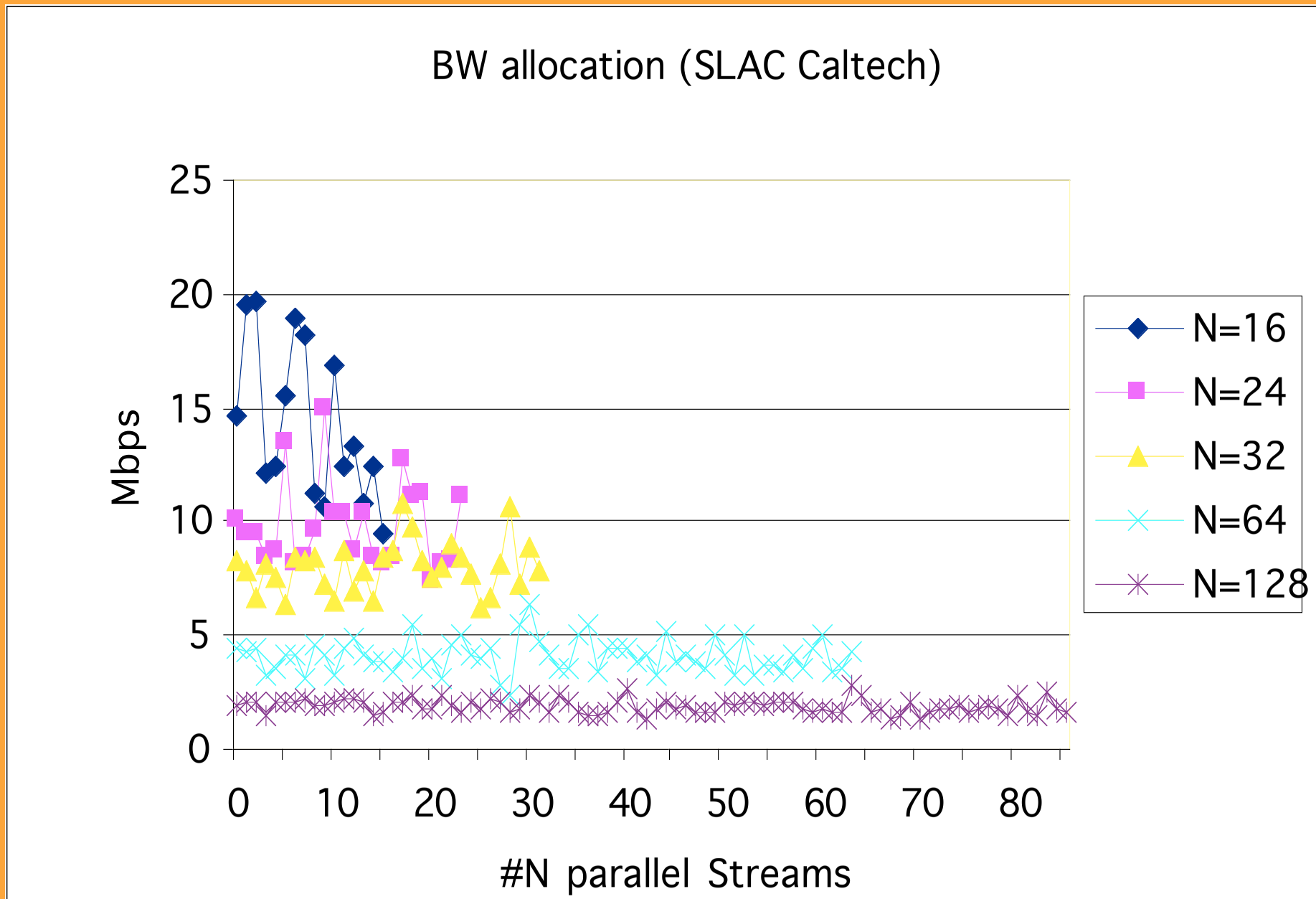

BW allocation
(SLAC-CERN)
Mbps per stream
3,5
3
2,5
2
1,5
1
0,5
0
0
16
32
48
63
#N parallel streams
N=24
N=32
N=64
N=16

Parallel TCP via Iperf
(32 streams SLAC-IN2P3)



BW allocation (SLAC LANL)




BW allocation (SLAC Caltech)
Mbps
25
20
15
10
5
0
0
10
20
30
40
50
60
70
80
#N parallel Streams
N=16
N=24
N=32
N=64
N=128

Used tools for BW estimations

- Pathrate
- Pathload
- Iperf
- Netest-2
- Incite BWe
- UDPmon

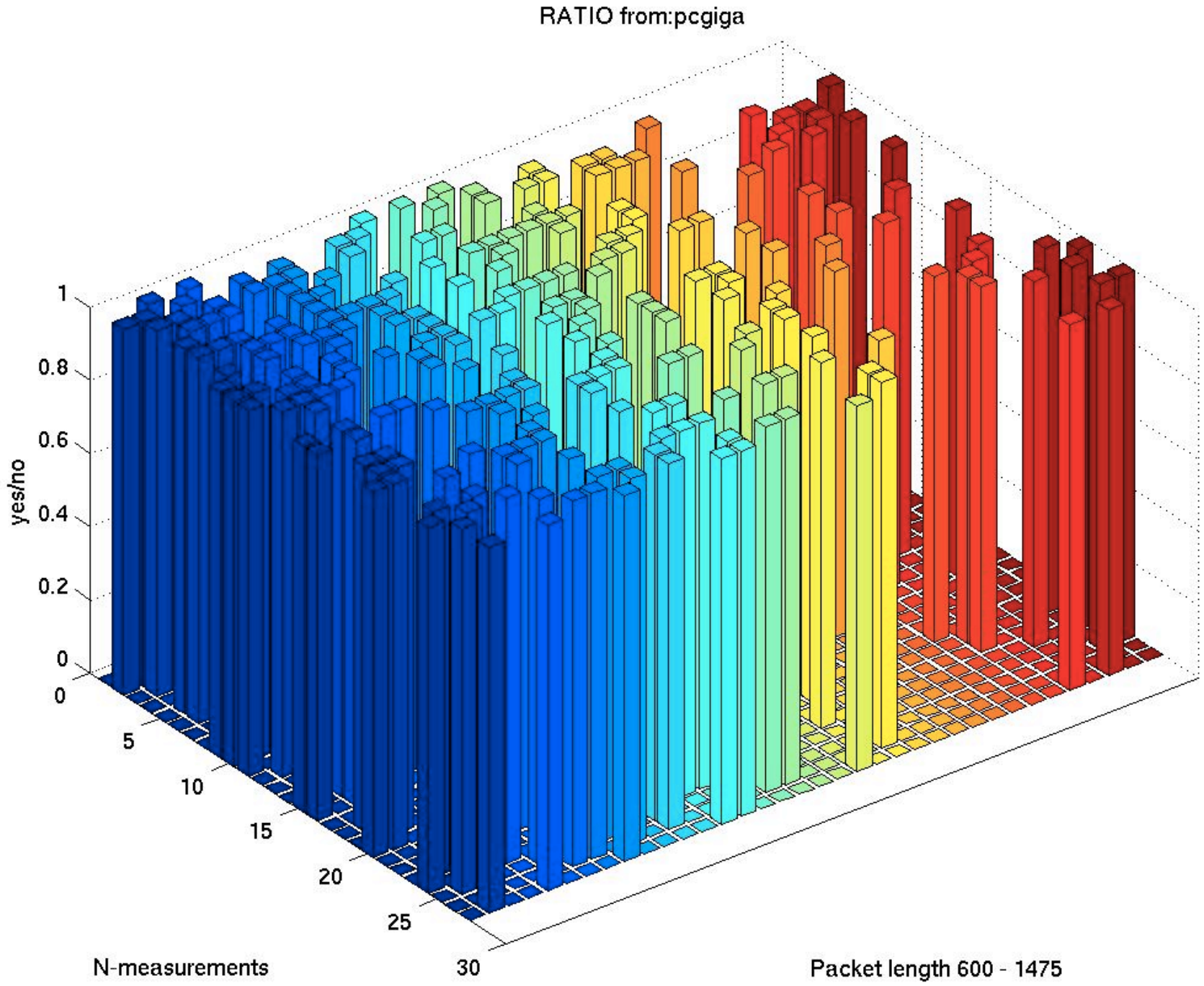
Short characteristic of methods

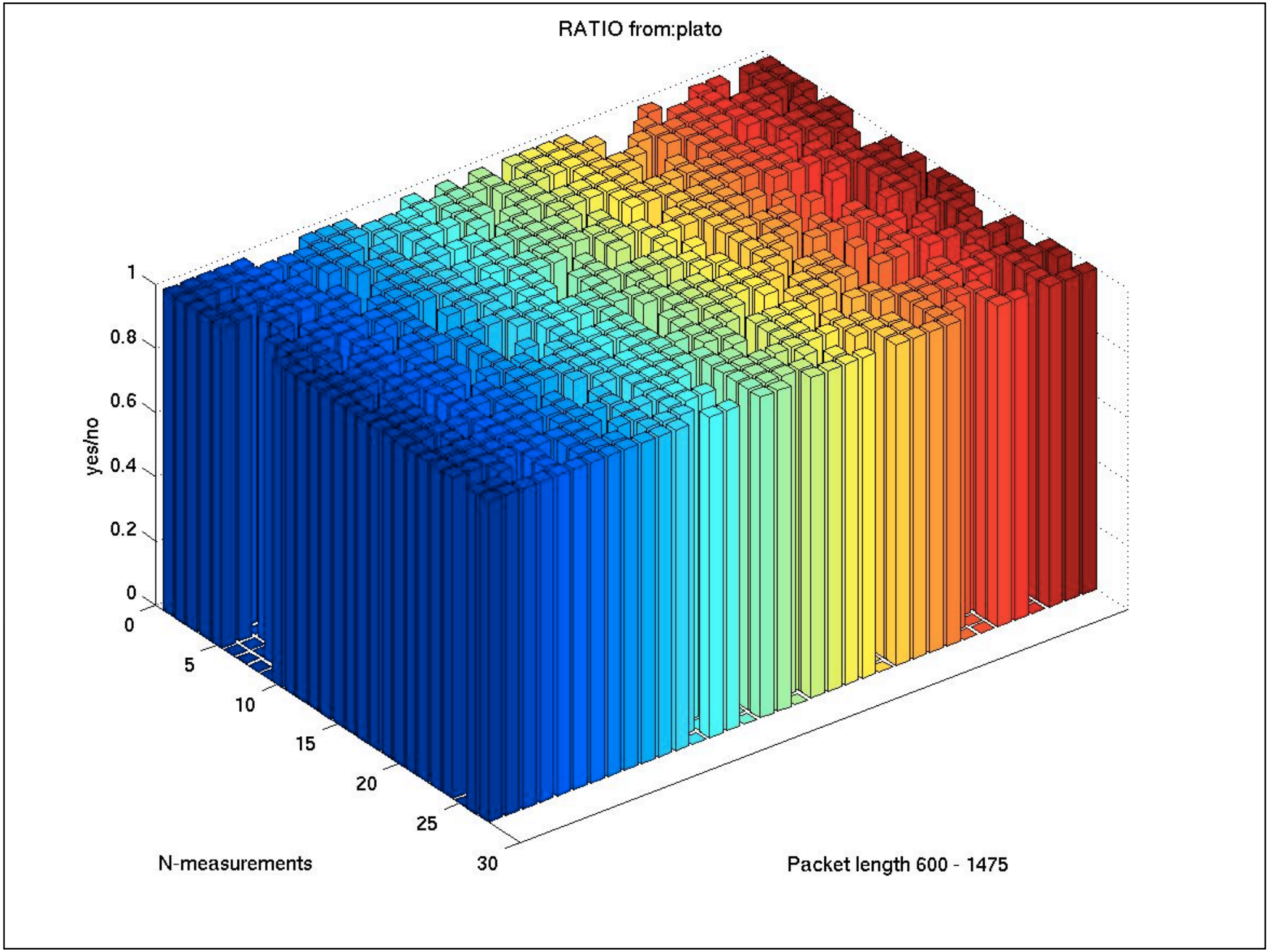
Patharate

+ accurate
- first version 2.1 not very reliable (reject more than 60%)
- long run time
- report of results (too many text info)

Pathload

+ accurate, fast, light
- limited range of operation (< 155 Mbps)
- sensitive to heavy loaded lines


RATIO from:pcgiga
yes/no
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
30
N-measurements
Packet length 600 - 1475


RATIO from:plato
yes/no
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
30
N-measurements
Packet length 600 - 1475

Short characteristic of methods

Iperf

+ reliable

- must be configured to use full BW and post processed

- create heavy load on lines during operation

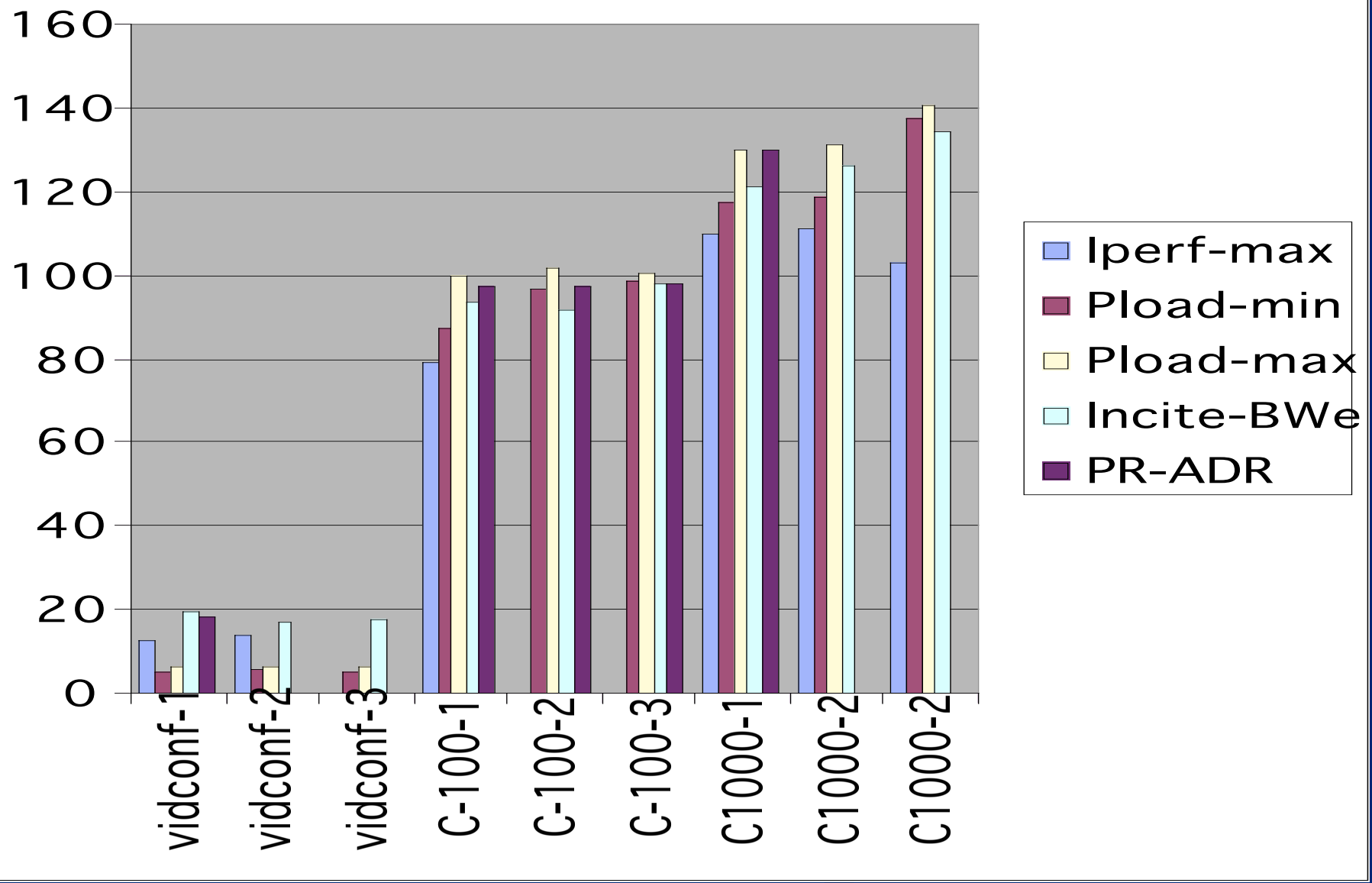
Netest-2

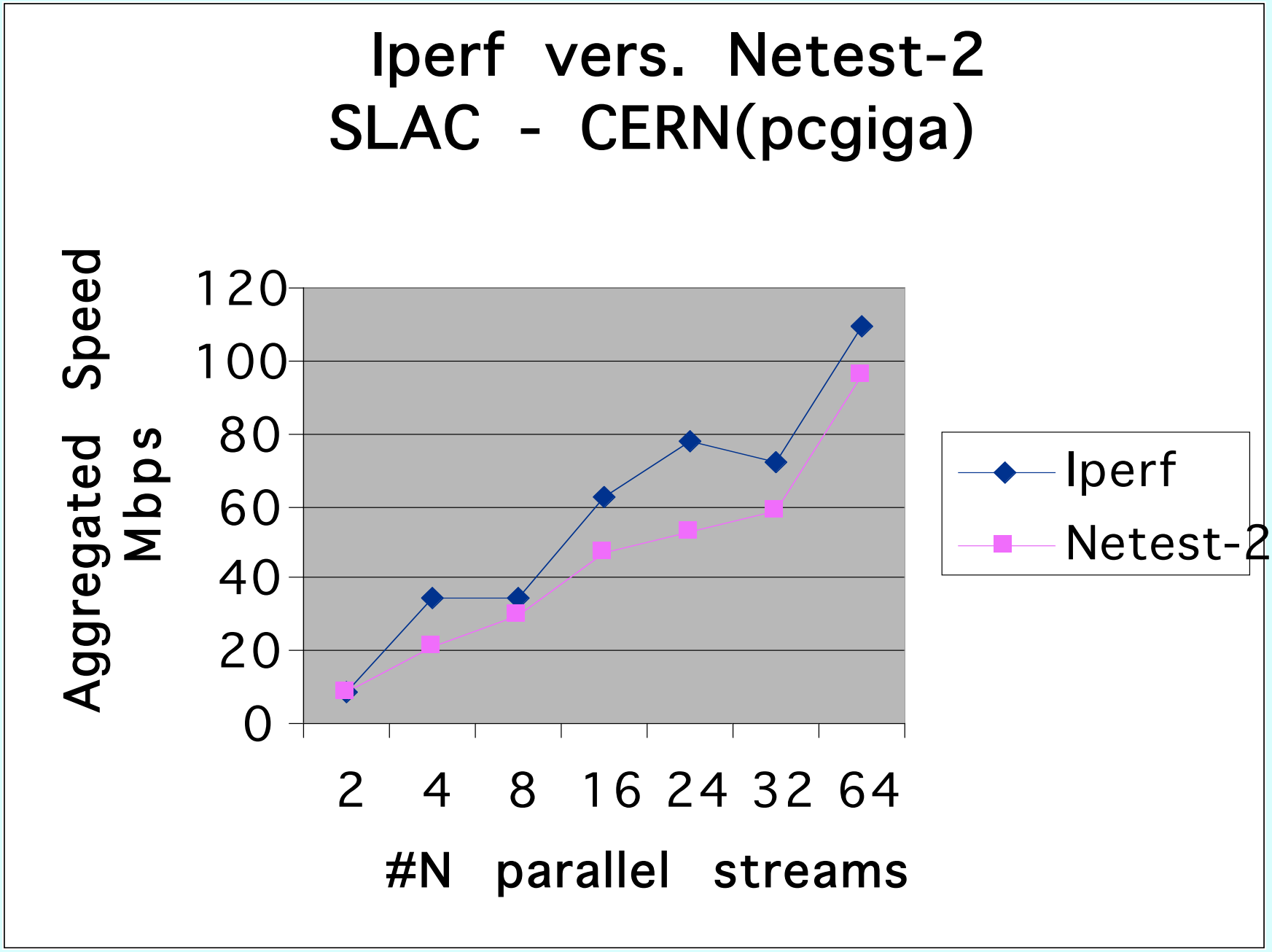
+ reliable, gives recommendations of WS

- must be configured to use full BW,

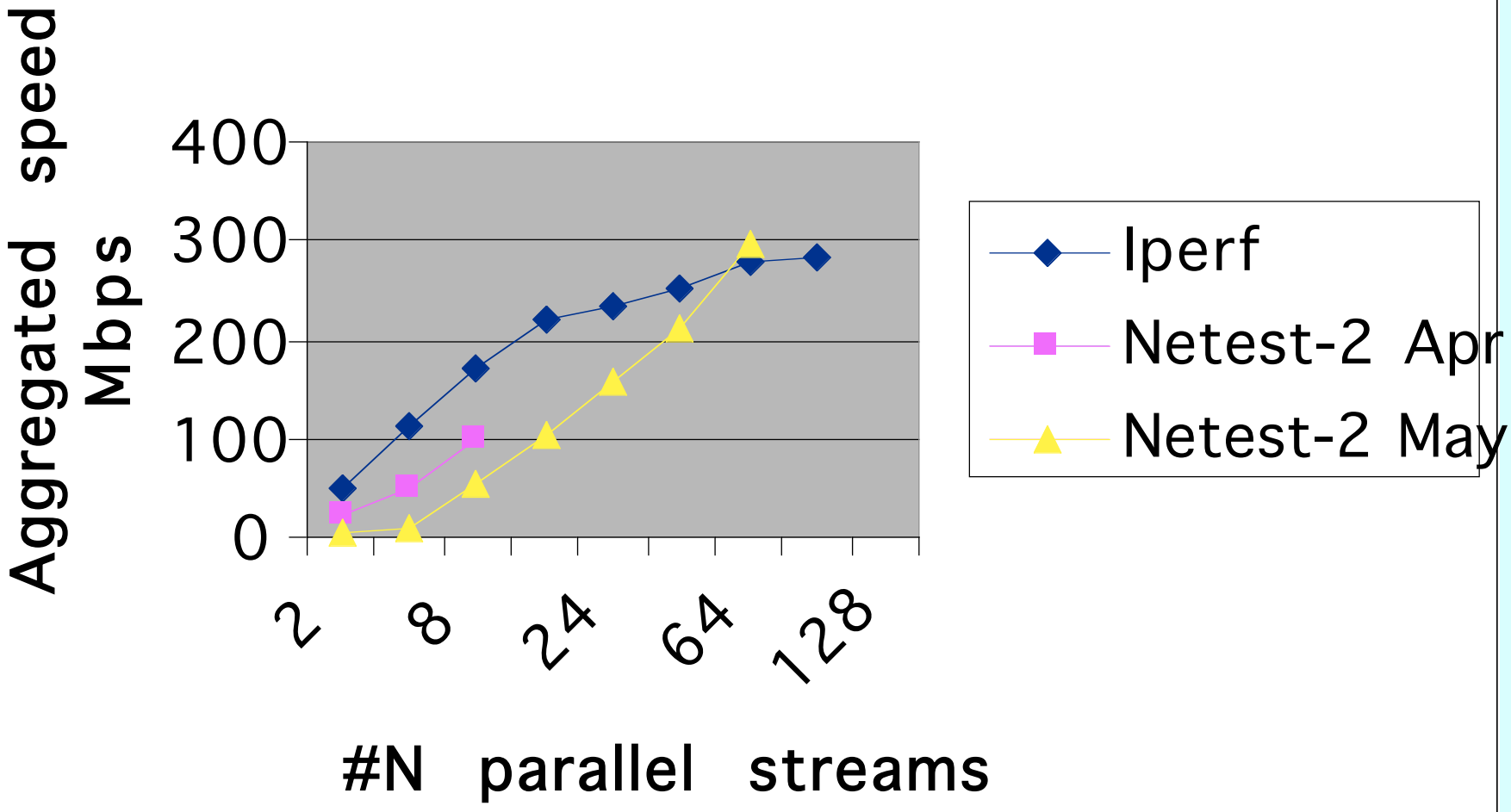
- different timing scheme, problematic accuracy

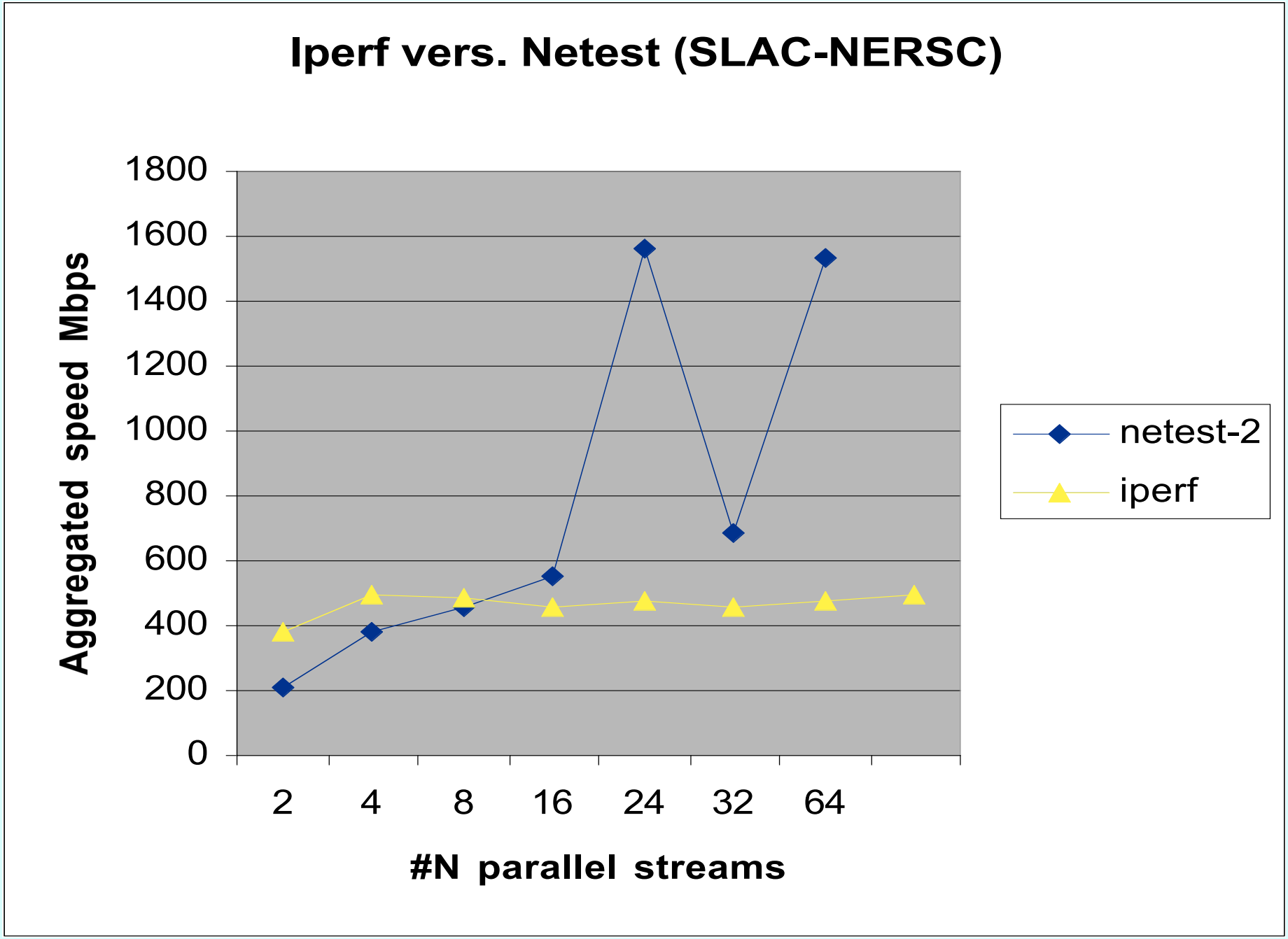
Small comparisons

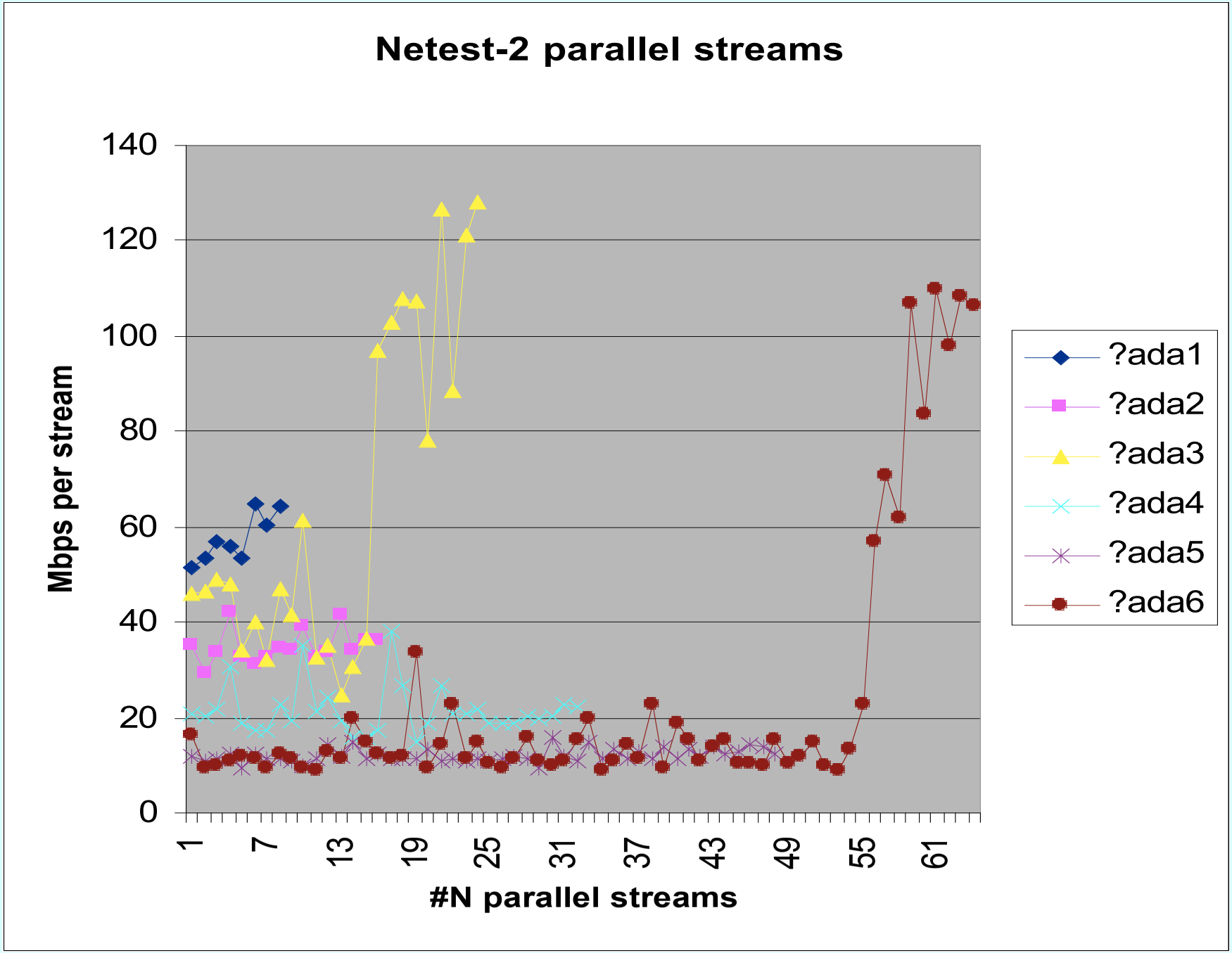



Iperf vers. Netest-2
SLAC - CERN(pcgiga)
Aggregated Speed Mbps
120
100
80
60
40
20
0
2
4
8
16
24
32
64
#N parallel streams
Iperf
Netest-2

Iperf vers. Netetst-2
SLAC-Caltech




Iperf vers. Netest (SLAC-NERSC)
Aggregated speed Mbps
1800
1600
1400
1200
1000
800
600
400
200
0
2
4
8
16
24
32
64
#N parallel streams
netest-2
iperf

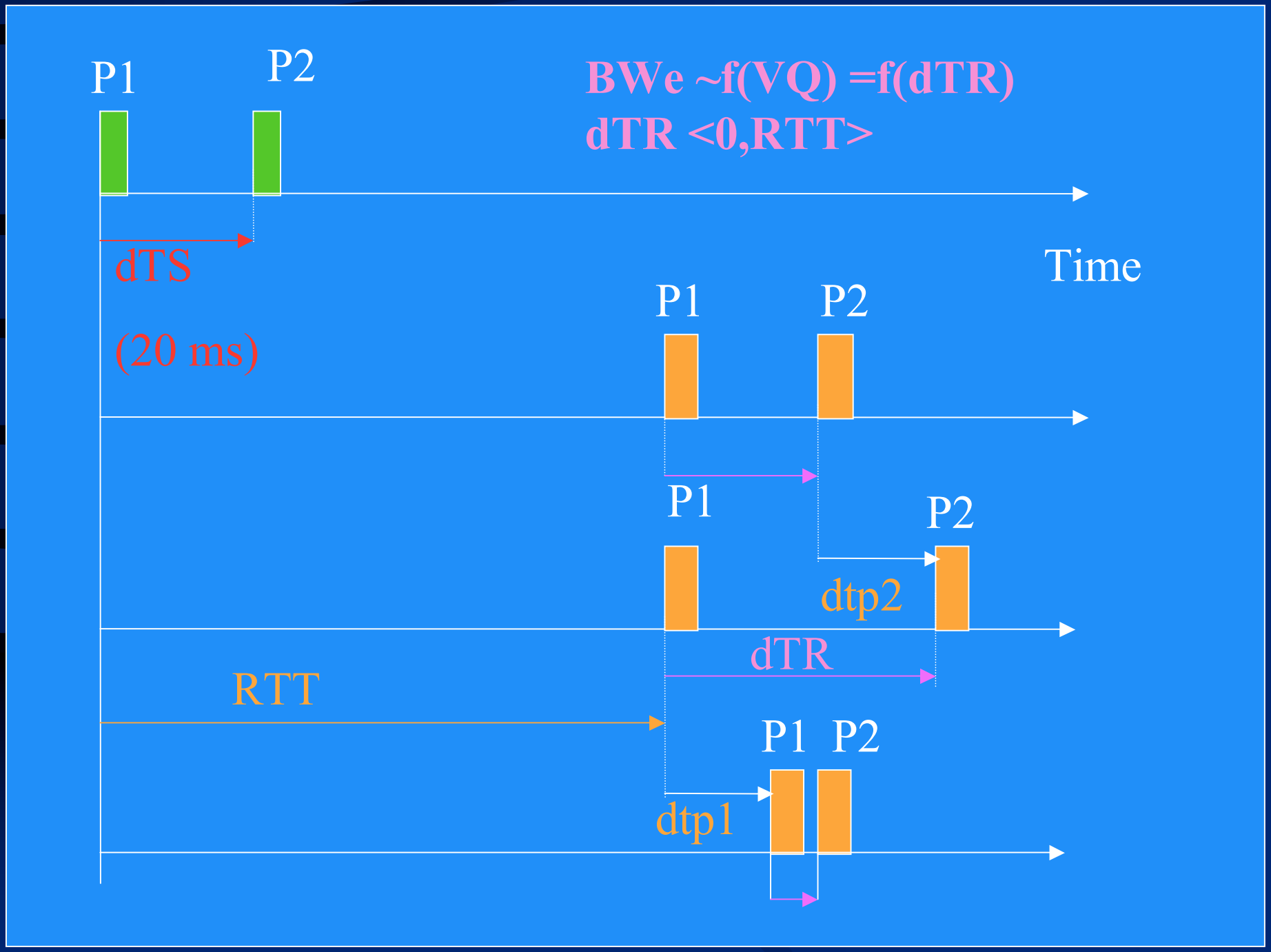

Netest-2 parallel streams
Mbps per stream
140
120
100
80
60
40
20
0
1
7
13
19
25
31
37
43
49
55
61
#N parallel streams
?ada1
?ada2
?ada3
?ada4
?ada5
?ada6

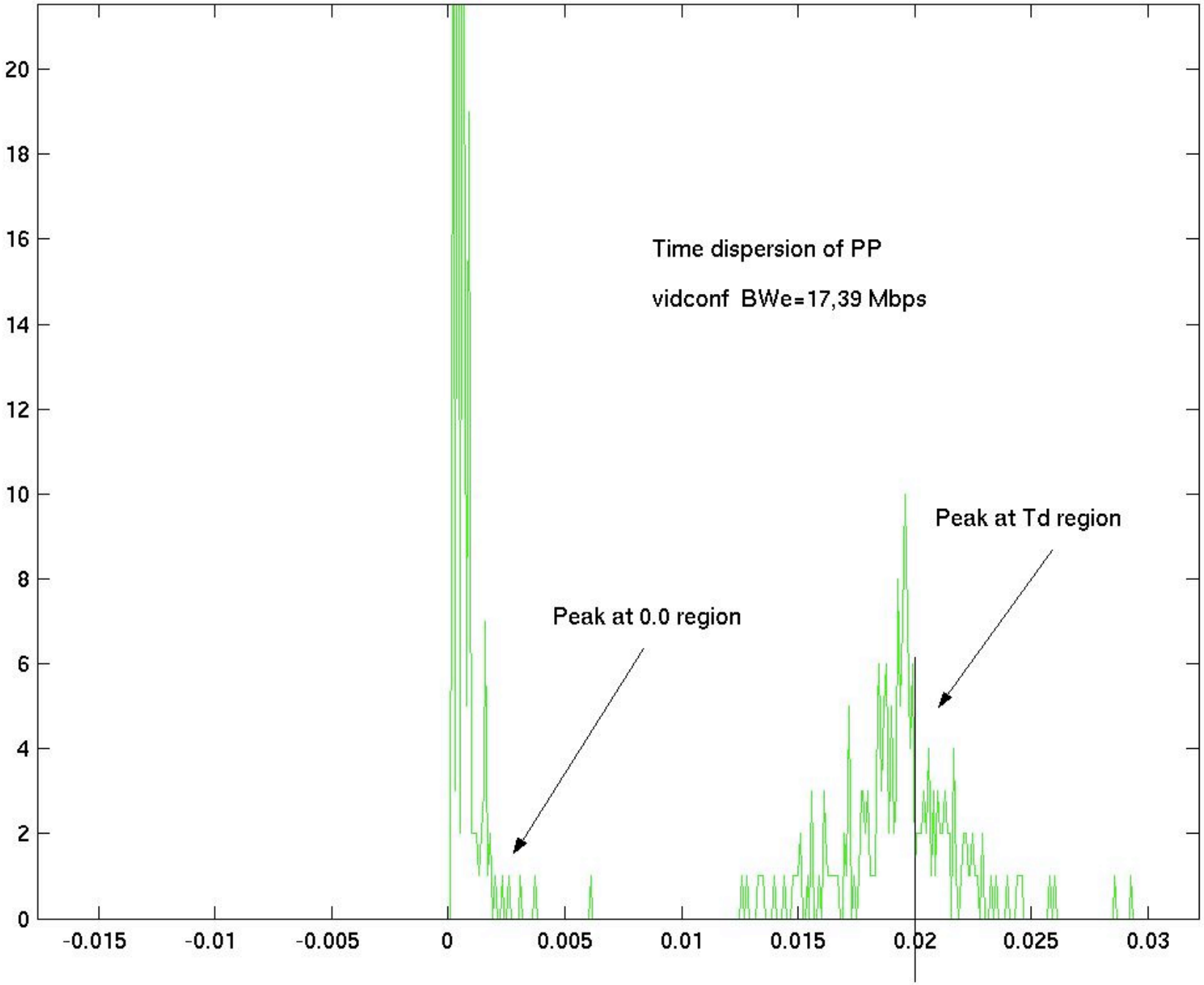
INCITE: Edge-based Traffic Processing and Service Inference for High-Performance Networks

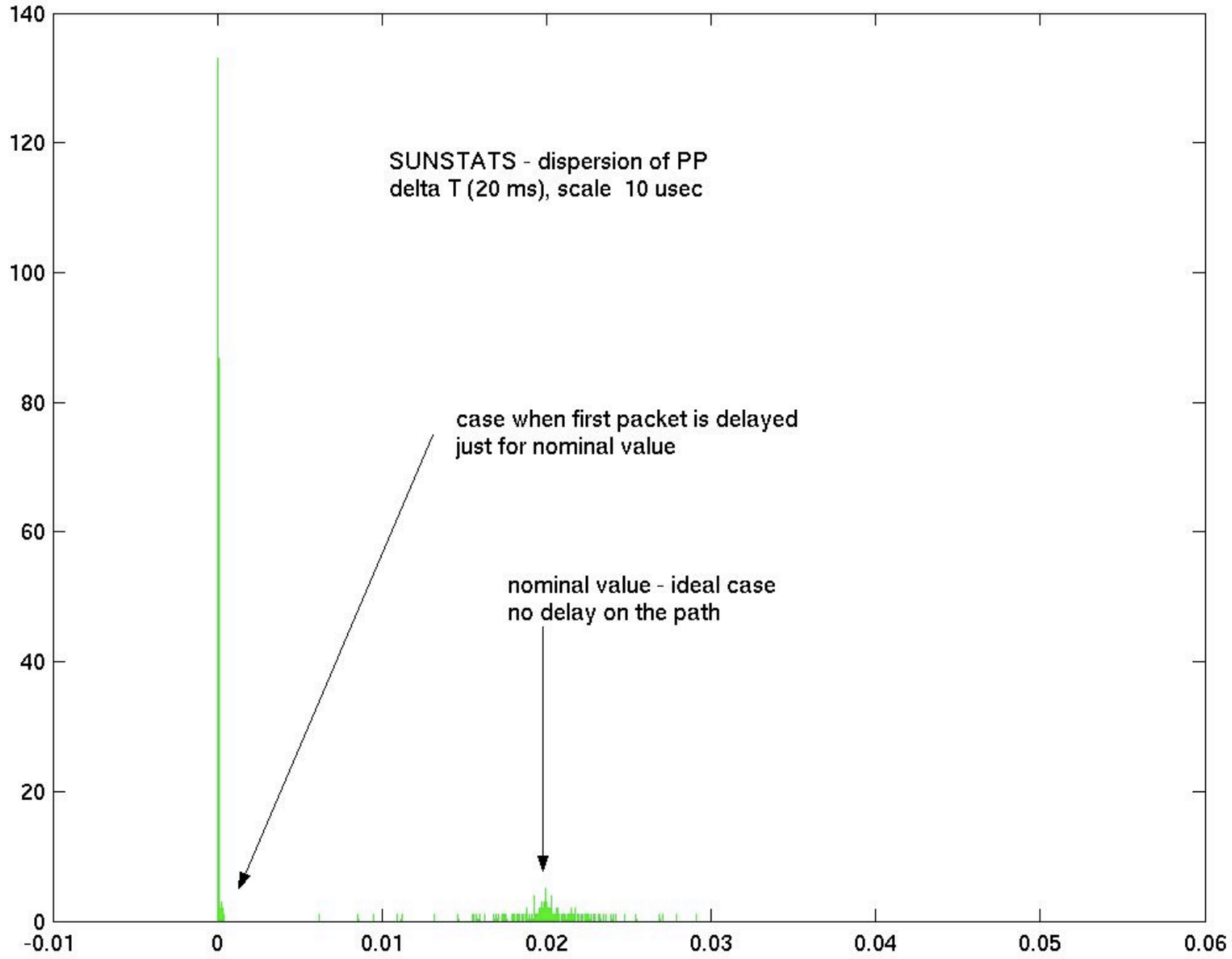
Richard Baraniuk, Rice University; Les Cottrell, SLAC; Wu-chun Feng, LANL

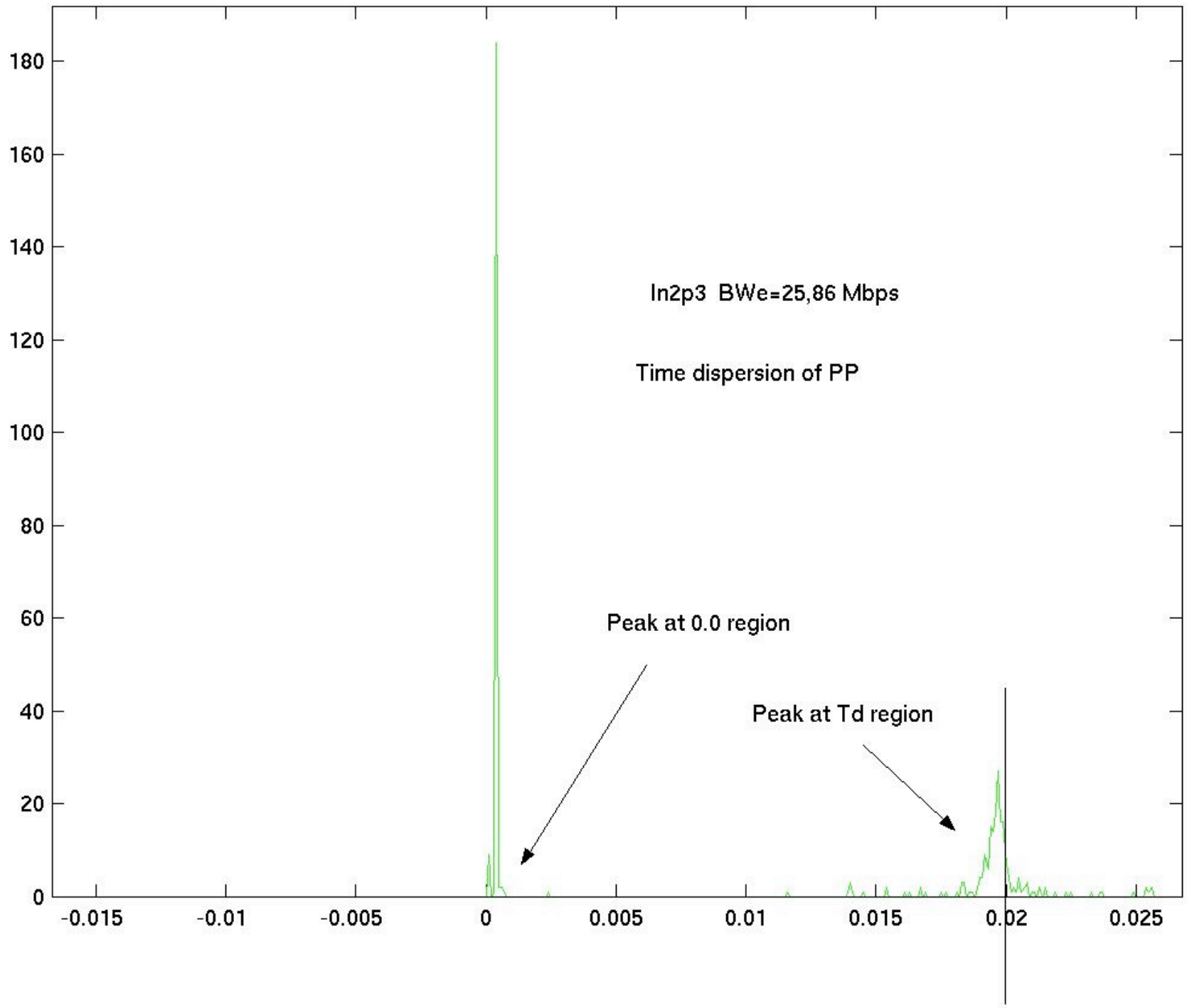
Incite

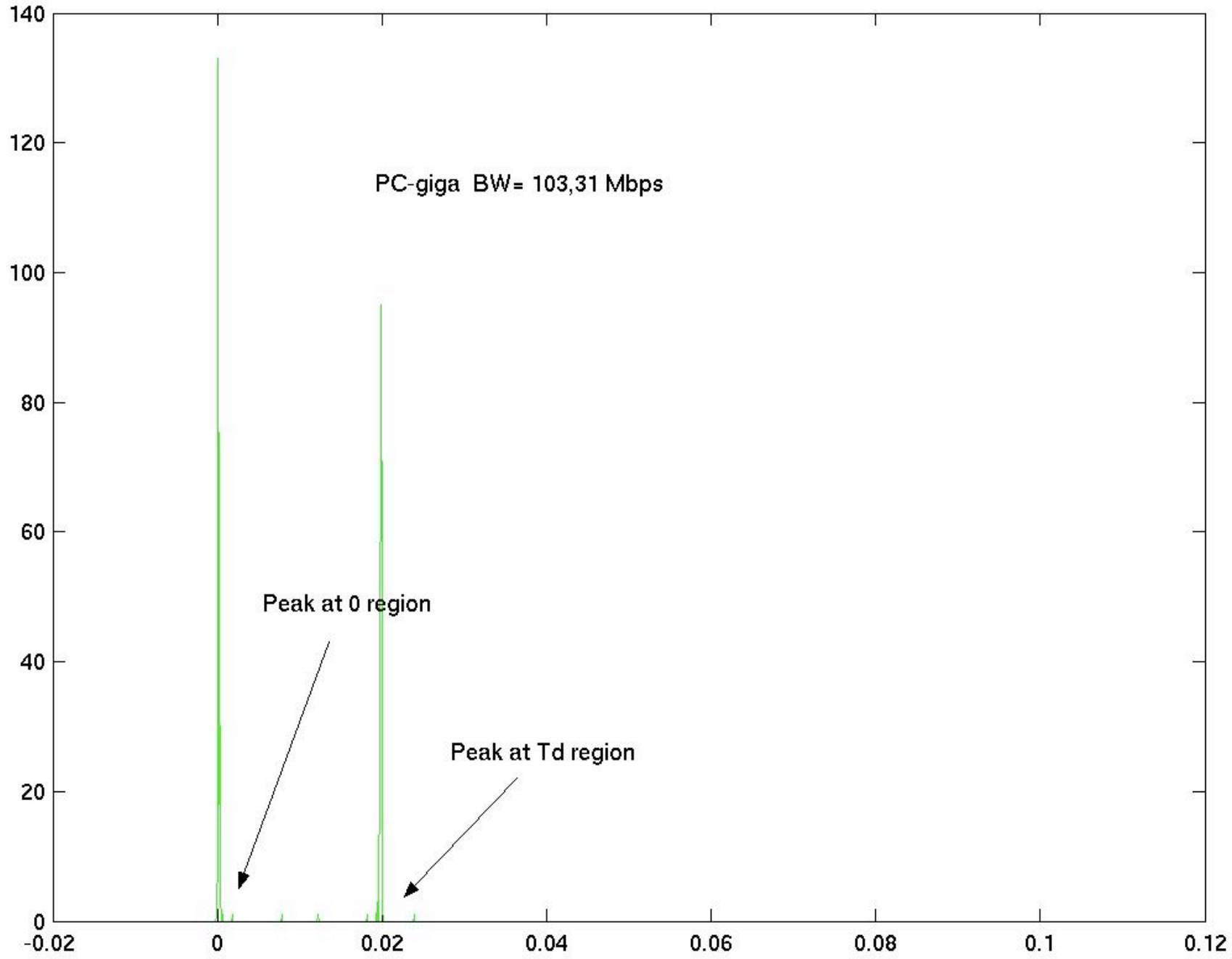
BW estimator

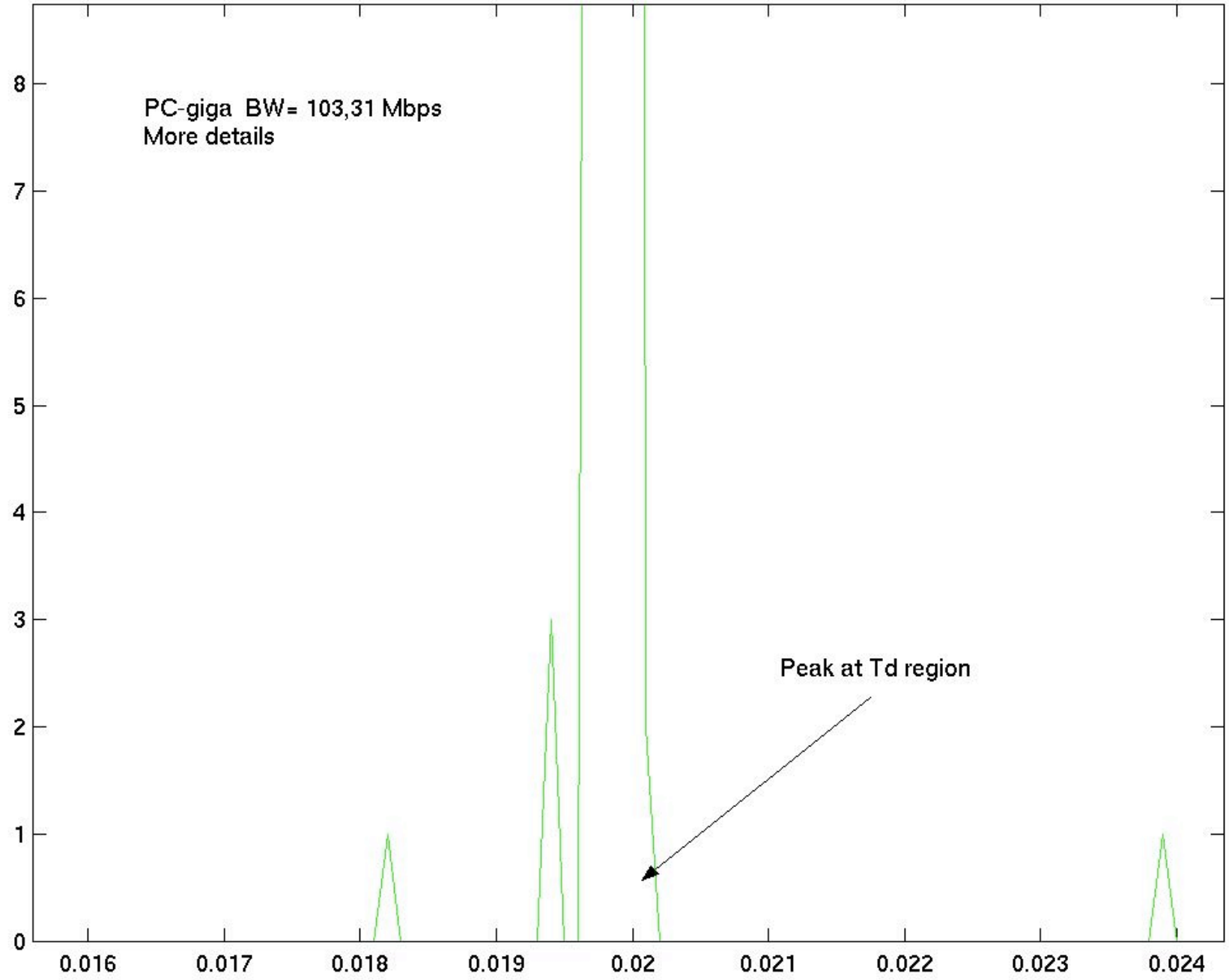

P1
P2
BWe ~f(VQ) =f(dTR)
dTR <0,RTT>
dTS
(20 ms)
Time
P1
P2
P1
P2
dtp2
dTR
RTT
P1
P2
dtp1

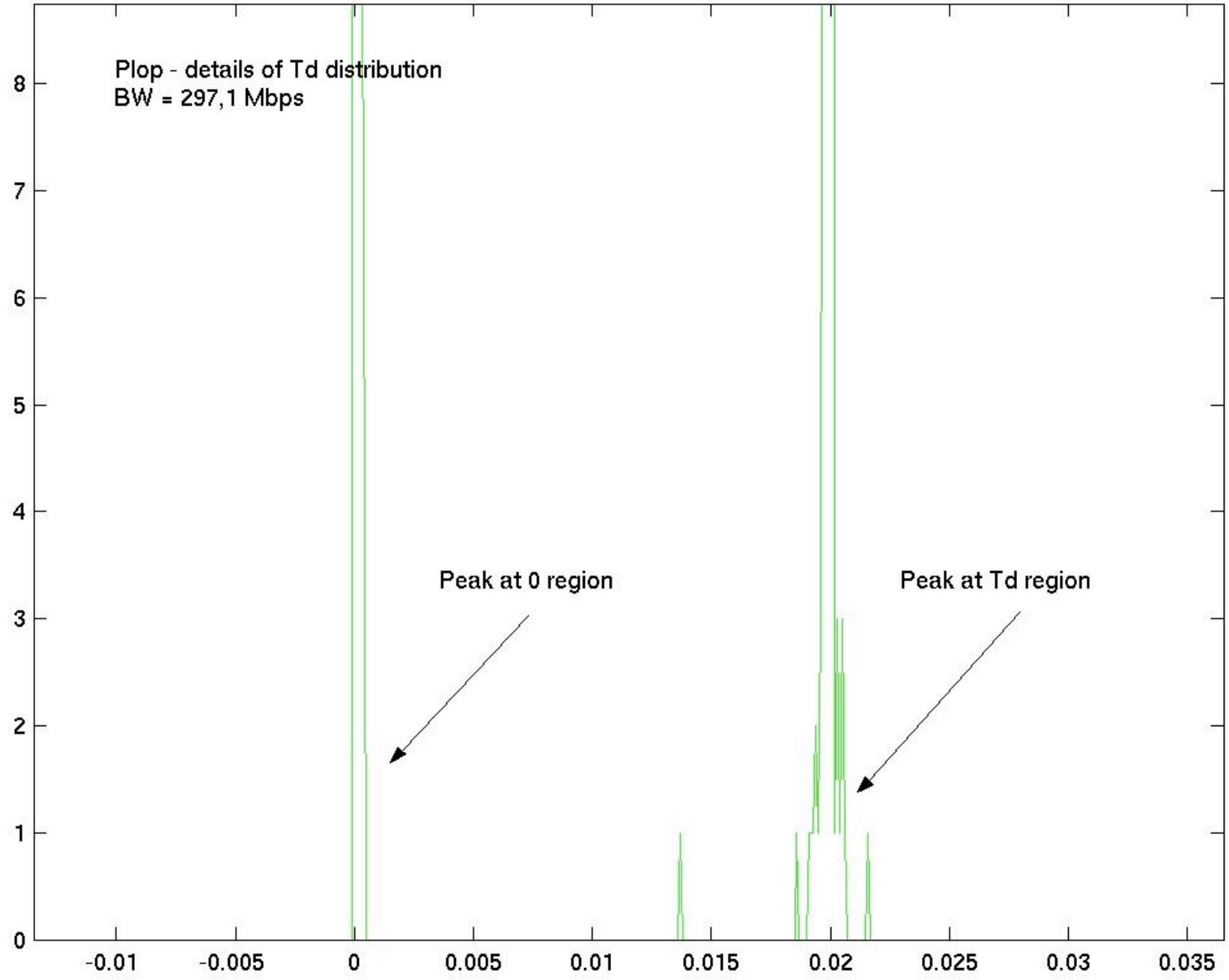

Time dispersion of PP
vidconf BWe=17,39 Mbps
Peak at 0.0 region
Peak at Td region
20
18
16
14
12
10
8
6
4
2
0
-0.015
-0.01
-0.005
0
0.005
0.01
0.015
0.02
0.025
0.03


SUNSTATS - dispersion of PP
delta T (20 ms), scale 10 usec
case when first packet is delayed
just for nominal value
nominal value - ideal case
no delay on the path
140
120
100
80
60
40
20
0
-0.01
0
0.01
0.02
0.03
0.04
0.05
0.06


180
160
140
120
100
80
60
40
20
0
-0.015
-0.01
-0.005
0
0.005
0.01
0.015
0.02
0.025
In2p3 BWe=25,86 Mbps
Time dispersion of PP
Peak at 0.0 region
Peak at Td region


PC-giga BW= 103,31 Mbps
Peak at 0 region
Peak at Td region
140
120
100
80
60
40
20
0
-0.02
0
0.02
0.04
0.06
0.08
0.1
0.12


PC-giga BW= 103,31 Mbps
More details
Peak at Td region
8
7
6
5
4
3
2
1
0
0.016
0.017
0.018
0.019
0.02
0.021
0.022
0.023
0.024


Plop - details of Td distribution
BW = 297,1 Mbps
Peak at 0 region
Peak at Td region
0
1
2
3
4
5
6
7
8
-0.01
-0.005
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035

freq= dv/dt

data_volume ~ pkt_length*8

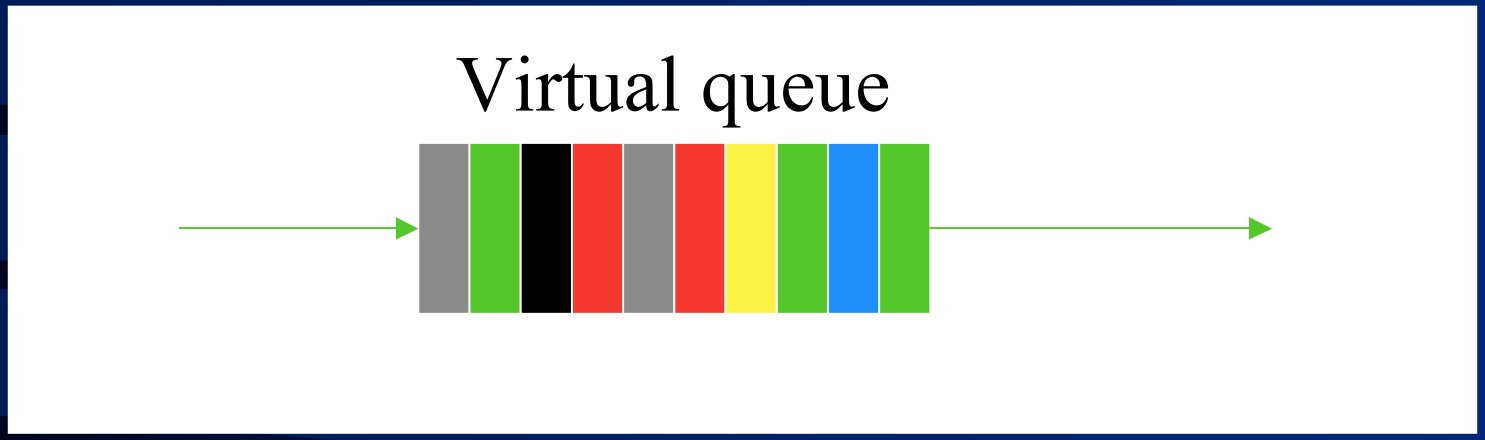
dt ~ F(VQ) ~ F(Internet-path)

BWe=Mean(freq)

Open Problem

What is pkt_length?

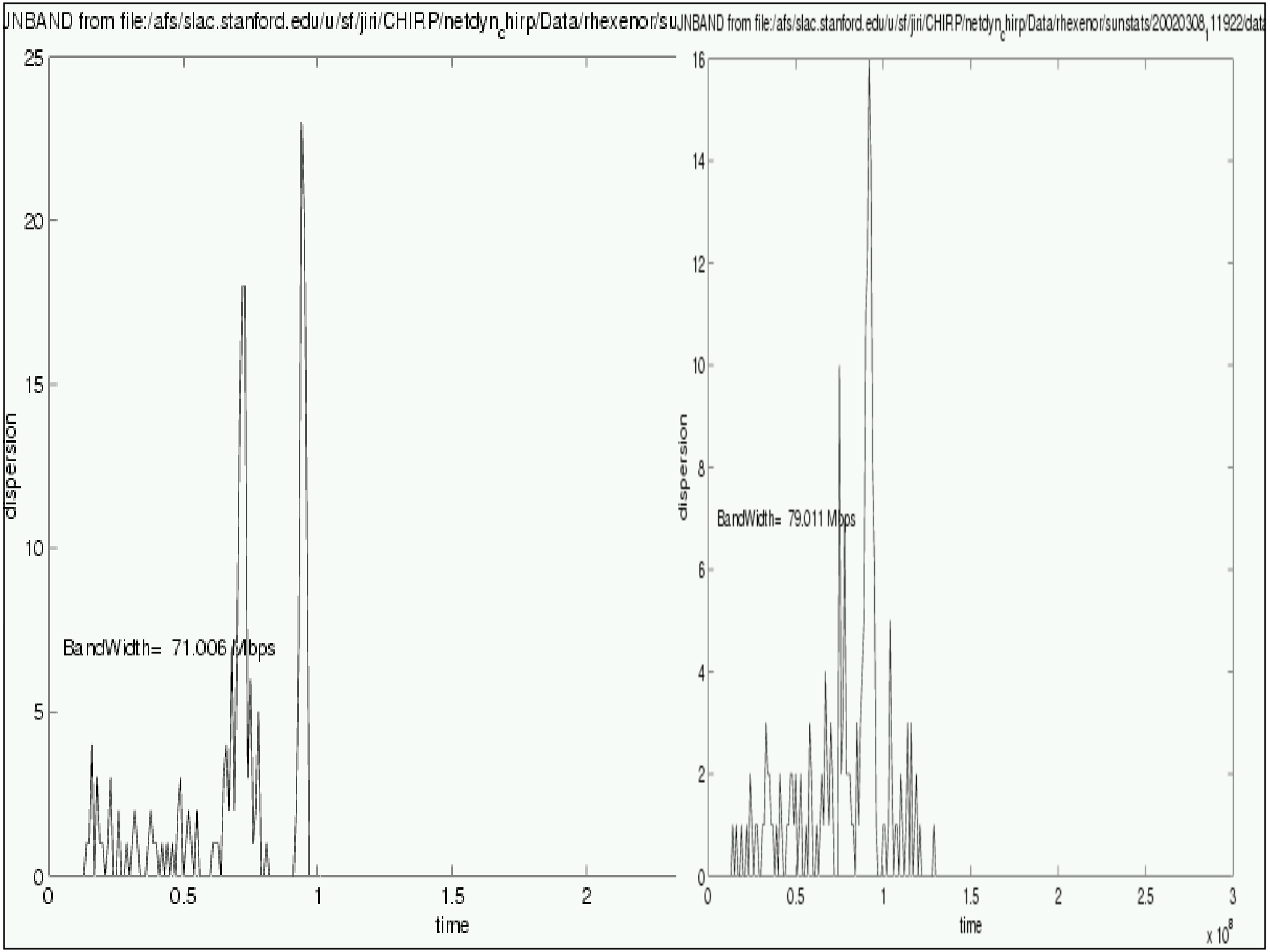
- **All "my test traffic" share same queue with outside CT**

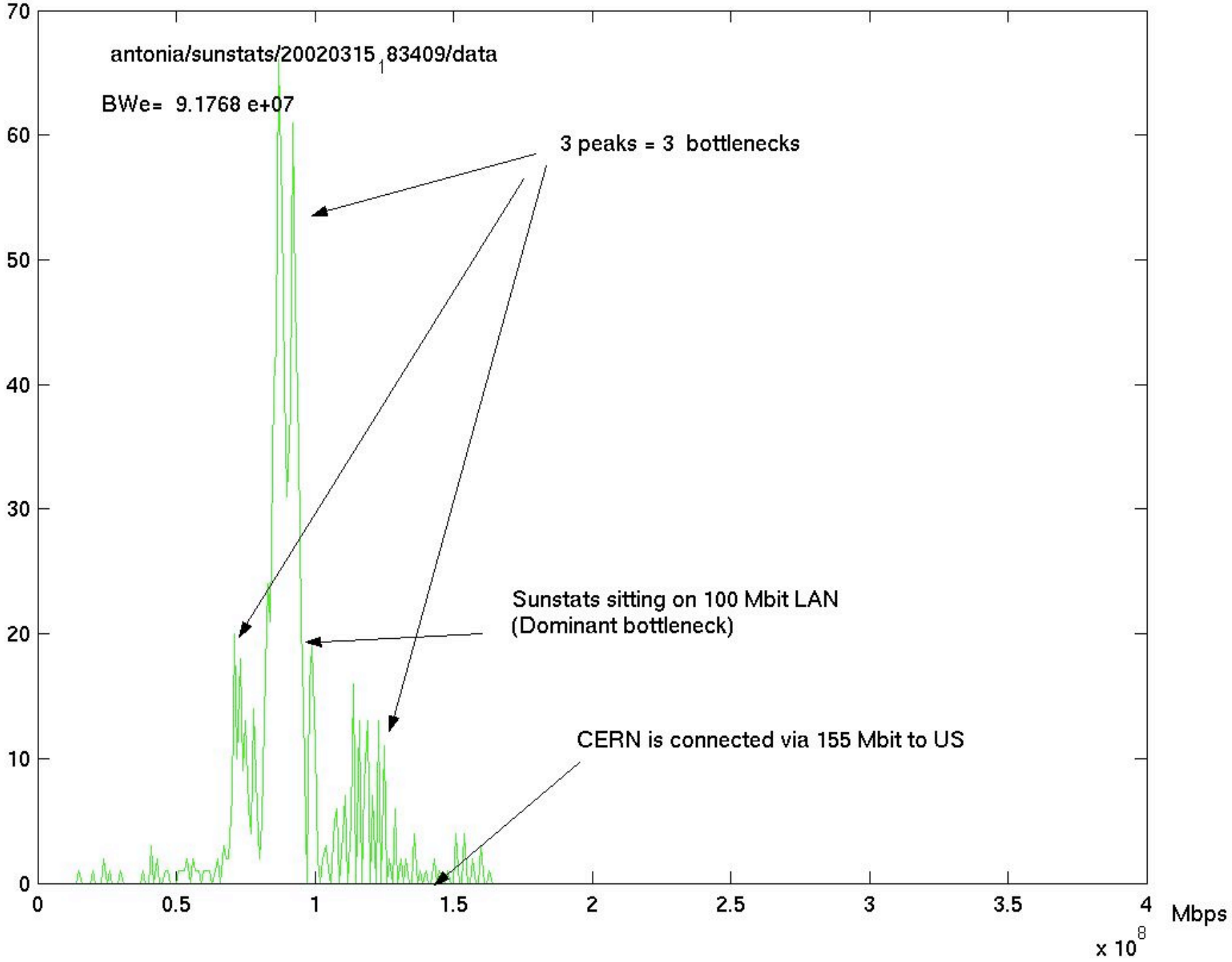


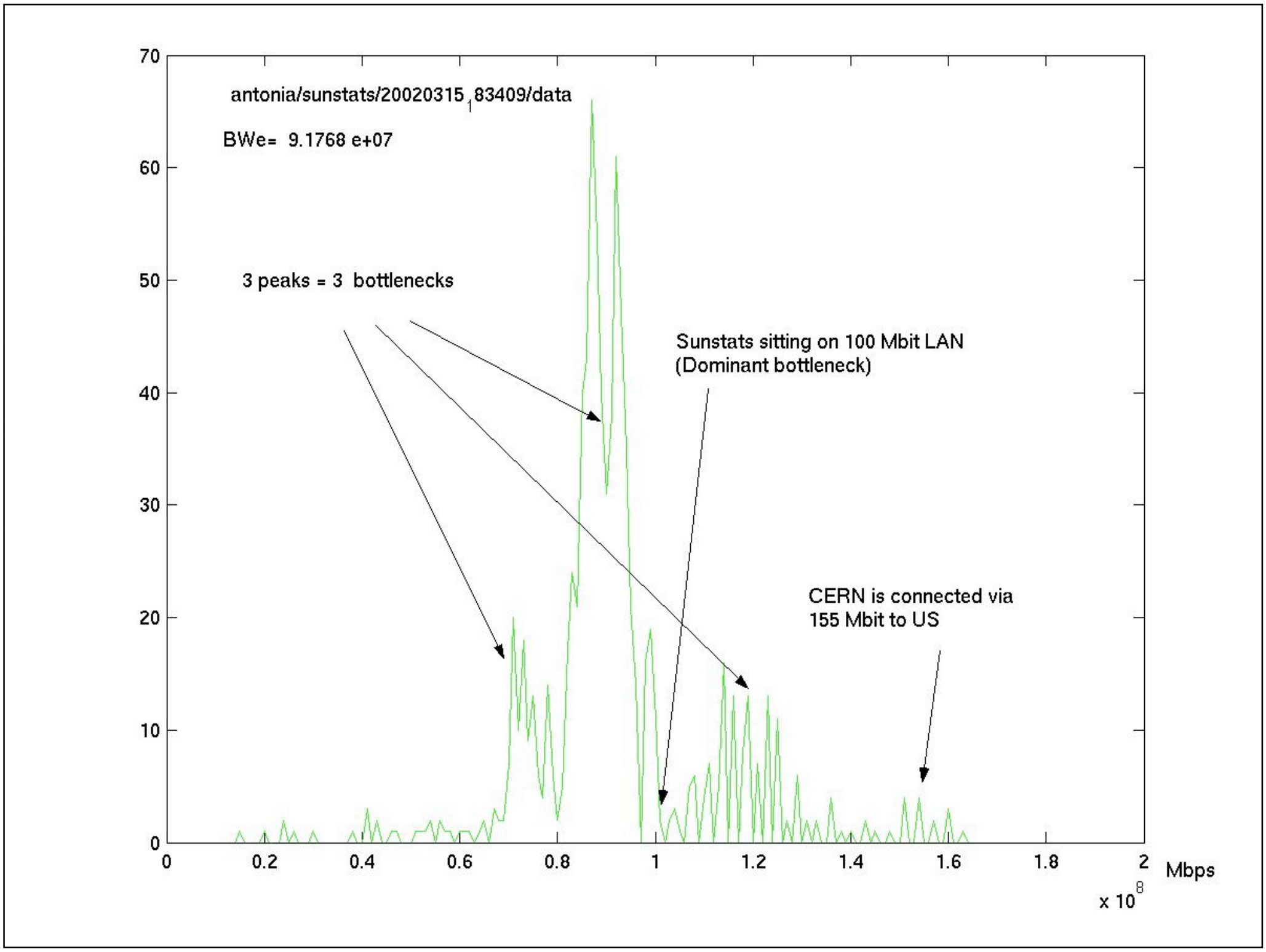
(It means that delay caused to my pkts by VQ is not dependent only on my pkt_lengths !)

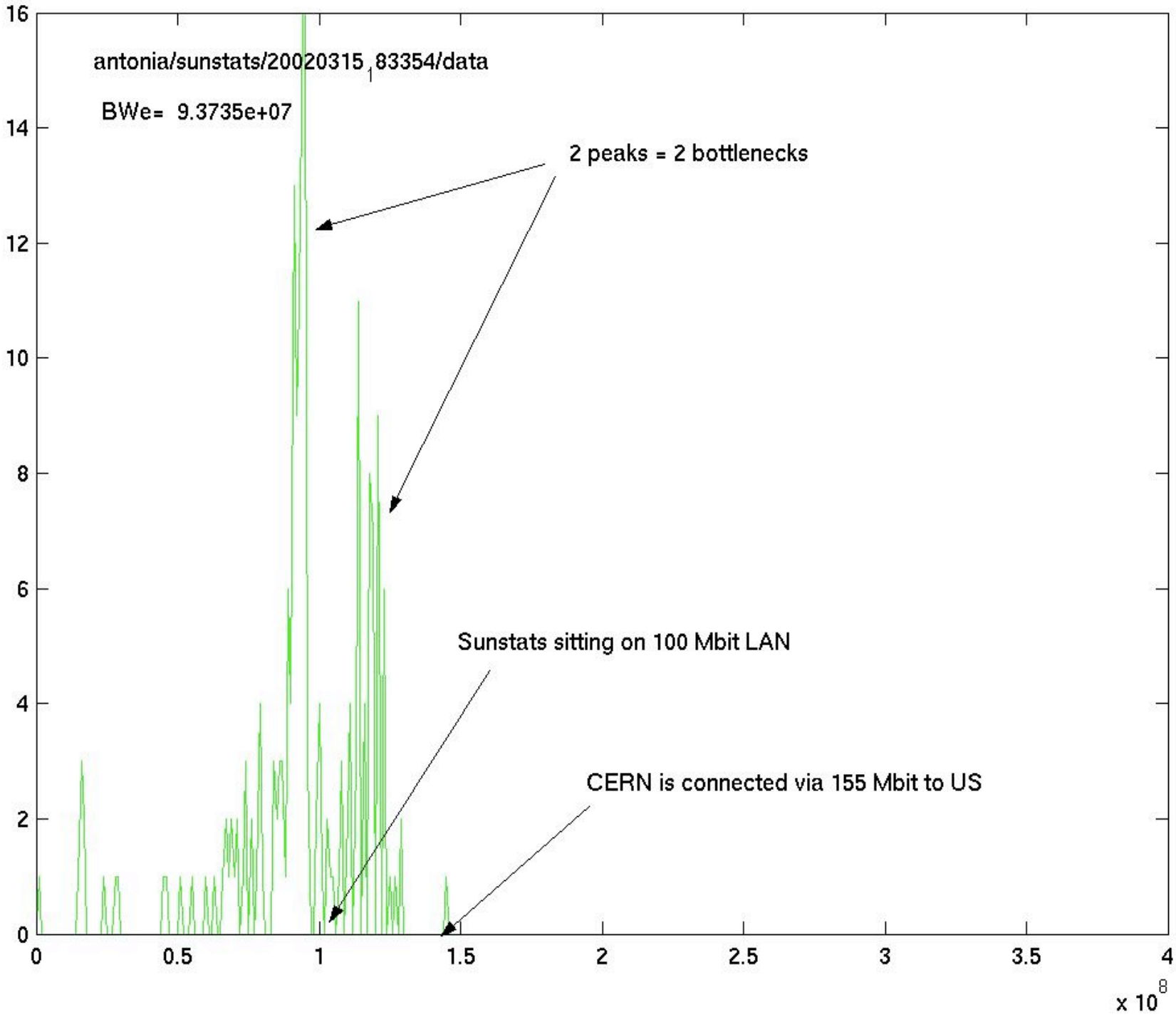
The accuracy is dependent on the knowlegde of packet distribution on particular path.

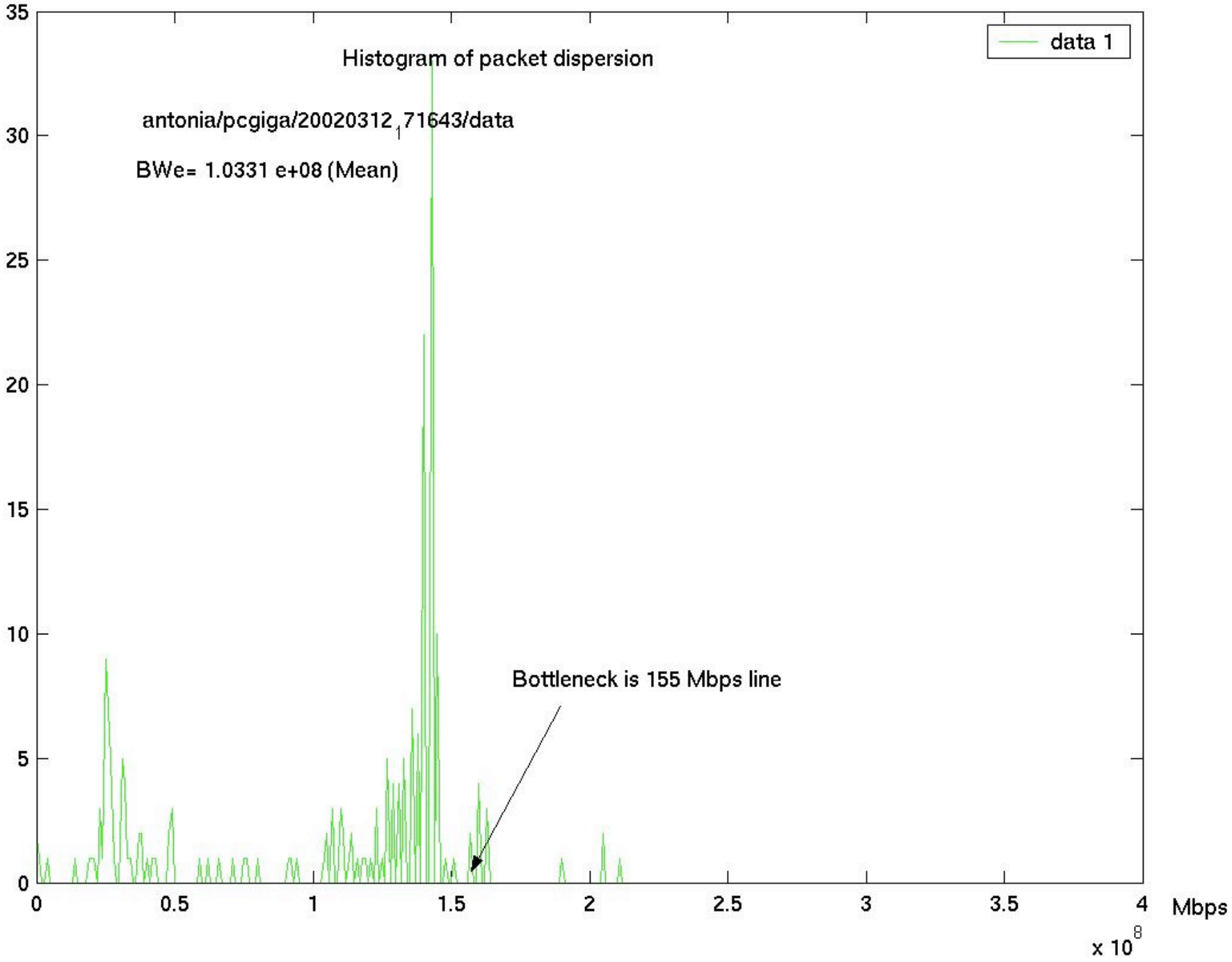
The average packet length ~1000 bytes gives reasonable results.

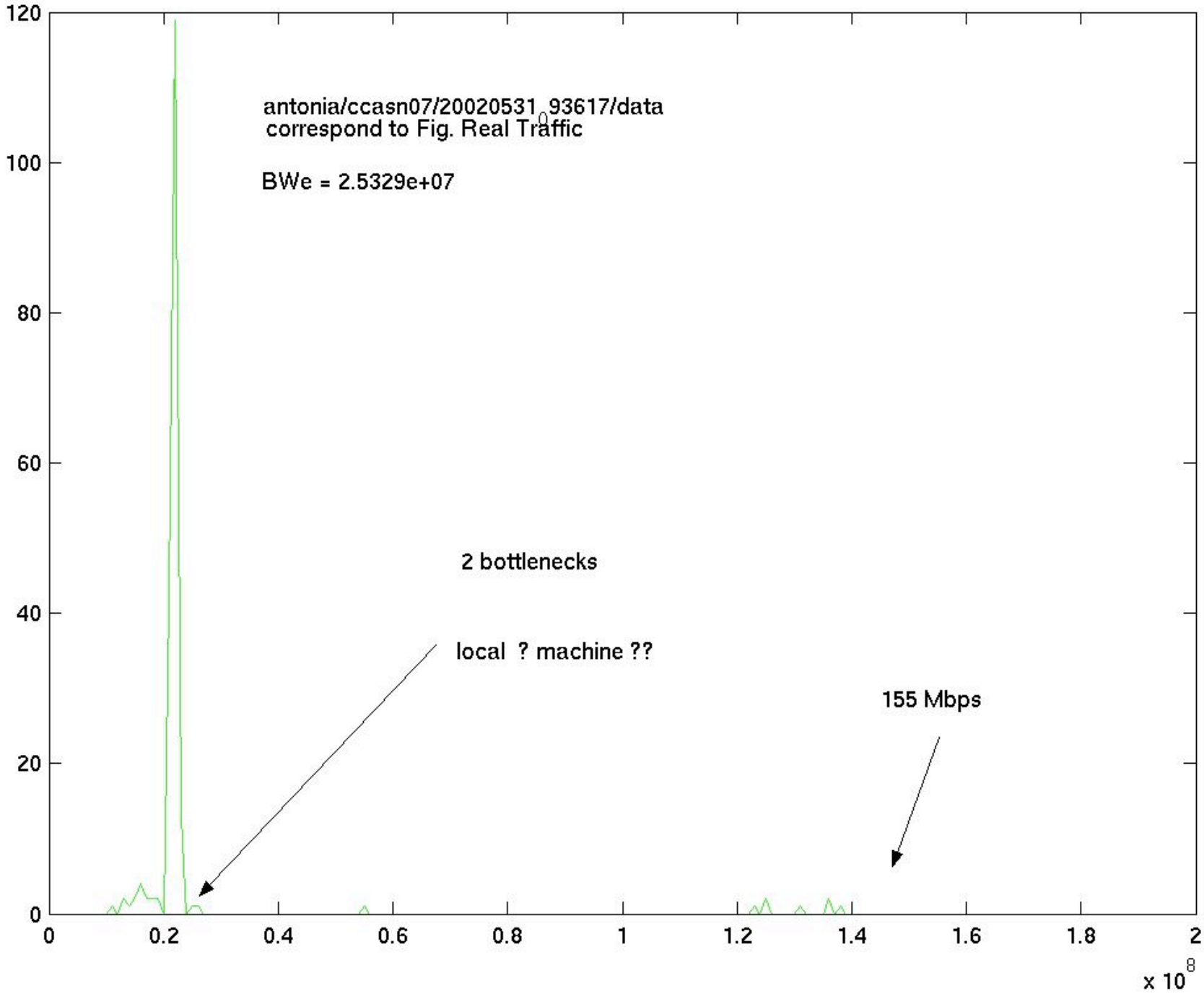

JNBAND from file:/afs/slac.stanford.edu/u/sf/jiri/CHIRP/netdyn_chirp/Data/rhexenor/su
JNBAND from file:/afs/slac.stanford.edu/u/sf/jiri/CHIRP/netdyn_chirp/Data/rhexenor/sunstats/20020308_11922/data
dispersion
time
BandWidth= 71.006 Mbps
dispersion
time
BandWidth= 79.011 Mbps
x 10^8

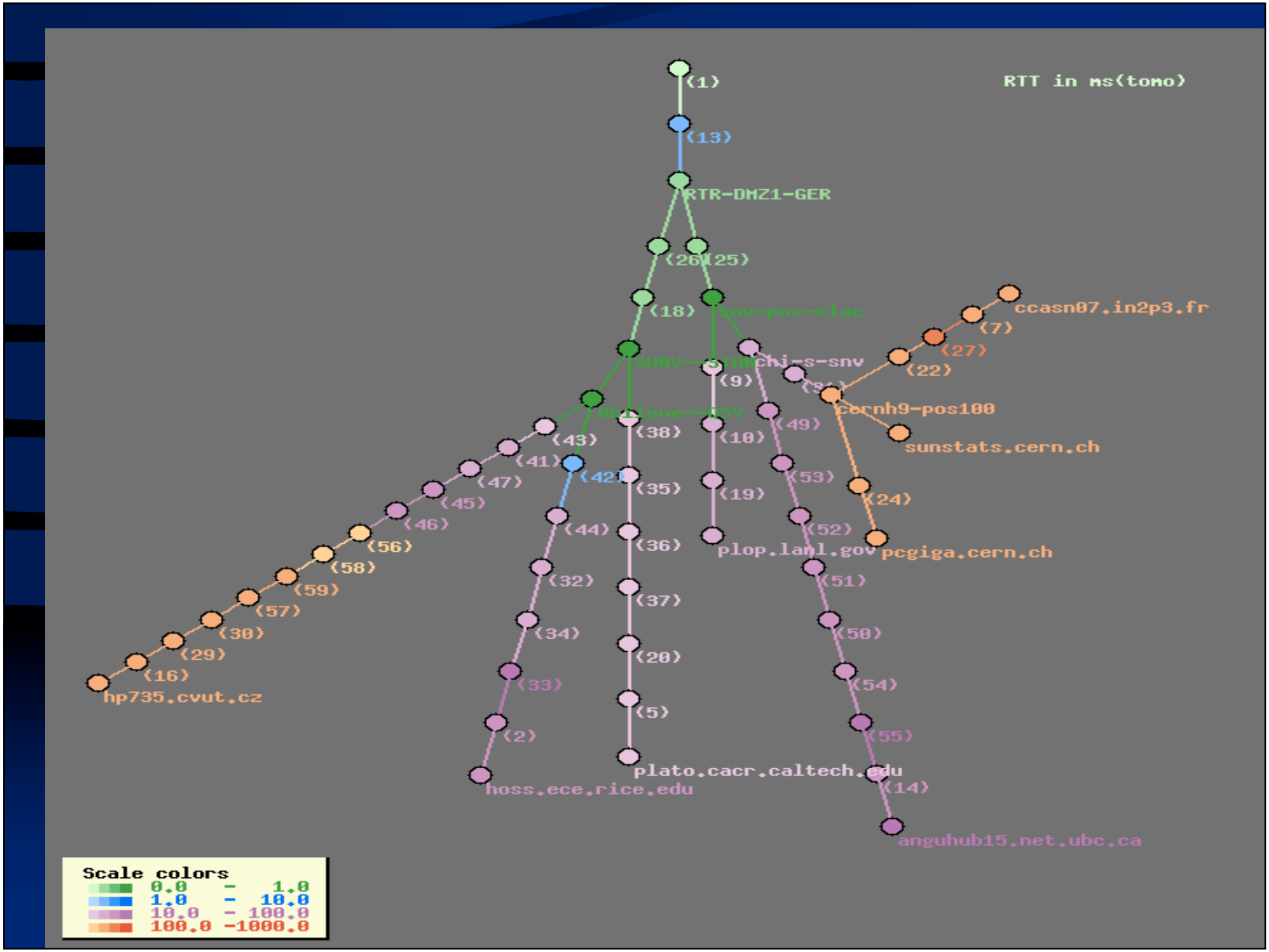

antonia/sunstats/20020315_183409/data
BWe= 9.1768 e+07
3 peaks = 3 bottlenecks
Sunstats sitting on 100 Mbit LAN
(Dominant bottleneck)
CERN is connected via 155 Mbit to US
70
60
50
40
30
20
10
0
0
0.5
1
1.5
2
2.5
3
3.5
4
Mbps
x 10^8

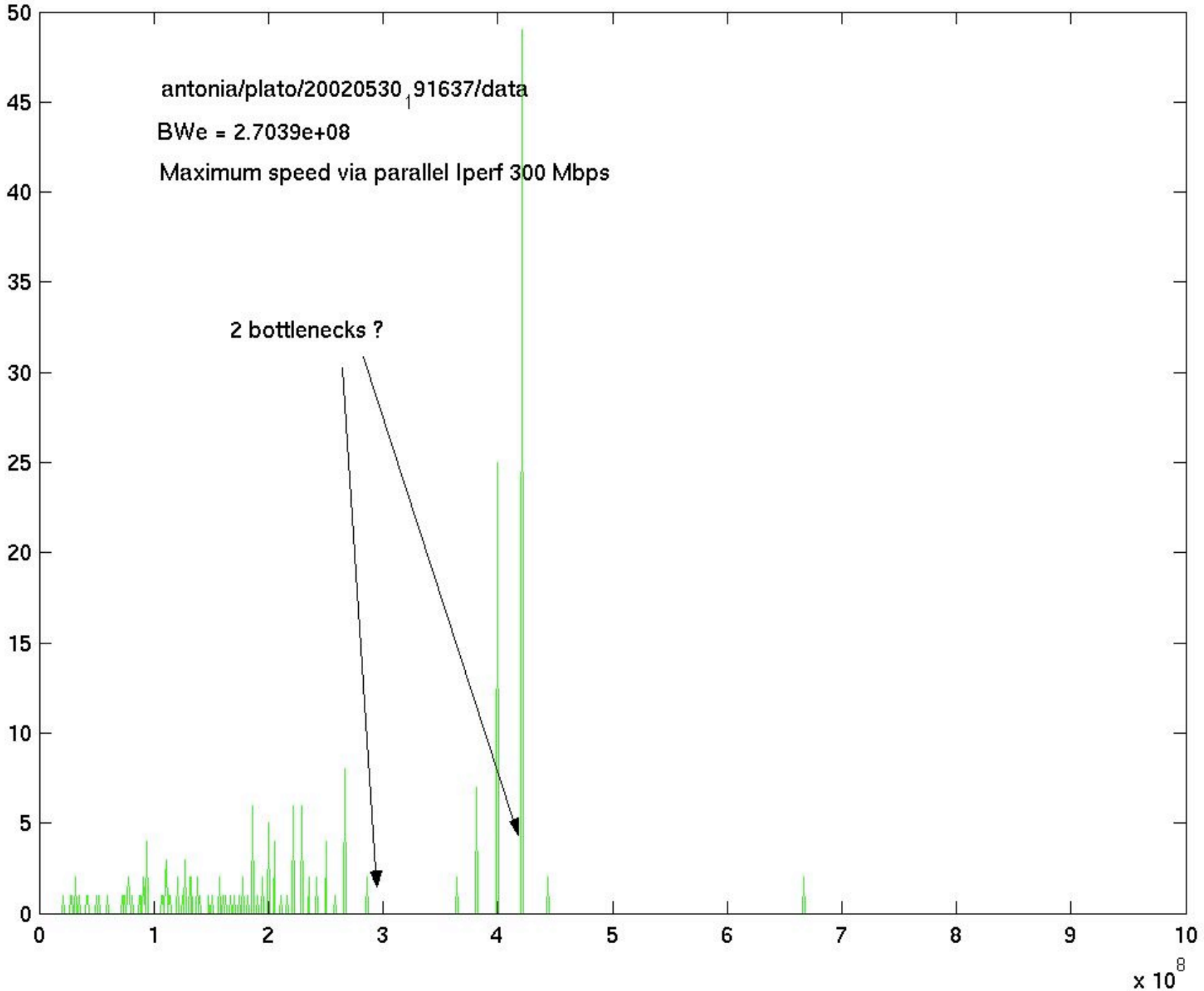

antonia/sunstats/20020315_183409/data
BWe= 9.1768 e+07
3 peaks = 3 bottlenecks
Sunstats sitting on 100 Mbit LAN
(Dominant bottleneck)
CERN is connected via
155 Mbit to US
70
60
50
40
30
20
10
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x 10^8
Mbps

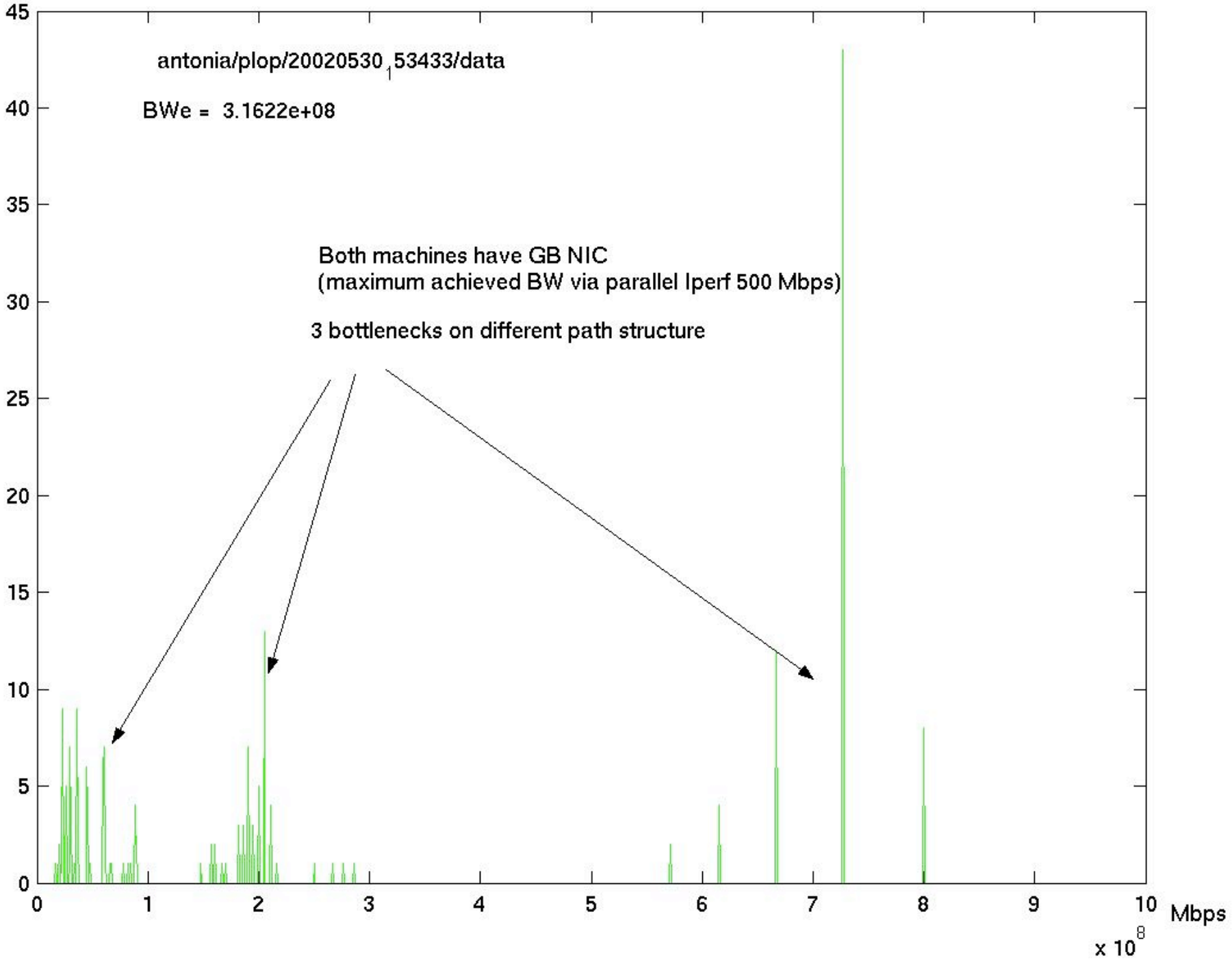

antonia/sunstats/20020315_183354/data
BWe= 9.3735e+07
2 peaks = 2 bottlenecks
Sunstats sitting on 100 Mbit LAN
CERN is connected via 155 Mbit to US
16
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3
3.5
4
x 10^8

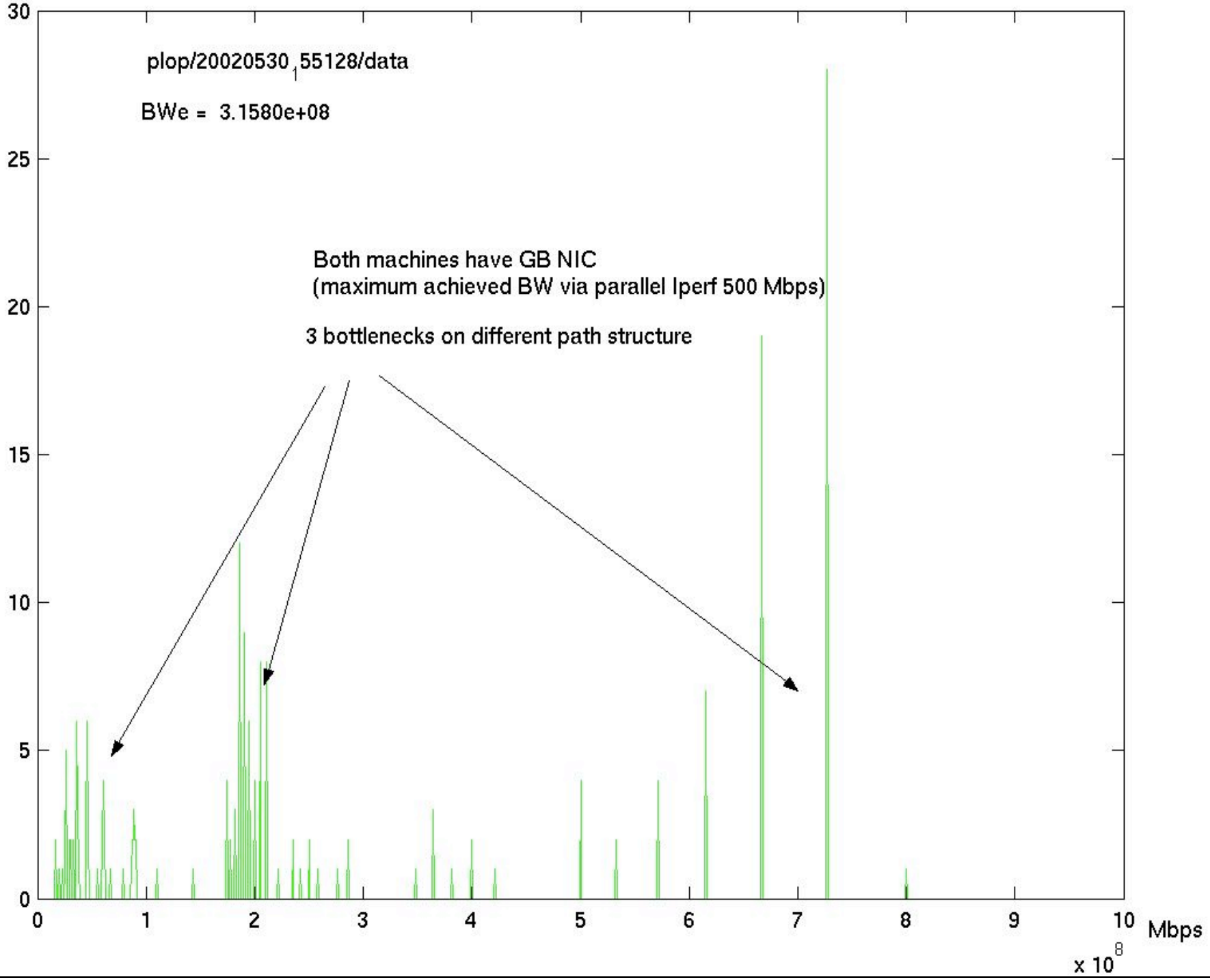

Histogram of packet dispersion
antonia/pcgiga/20020312_171643/data
BWe= 1.0331 e+08 (Mean)
data 1
Bottleneck is 155 Mbps line
35
30
25
20
15
10
5
0
0
0.5
1
1.5
2
2.5
3
3.5
4
x 10^8
Mbps

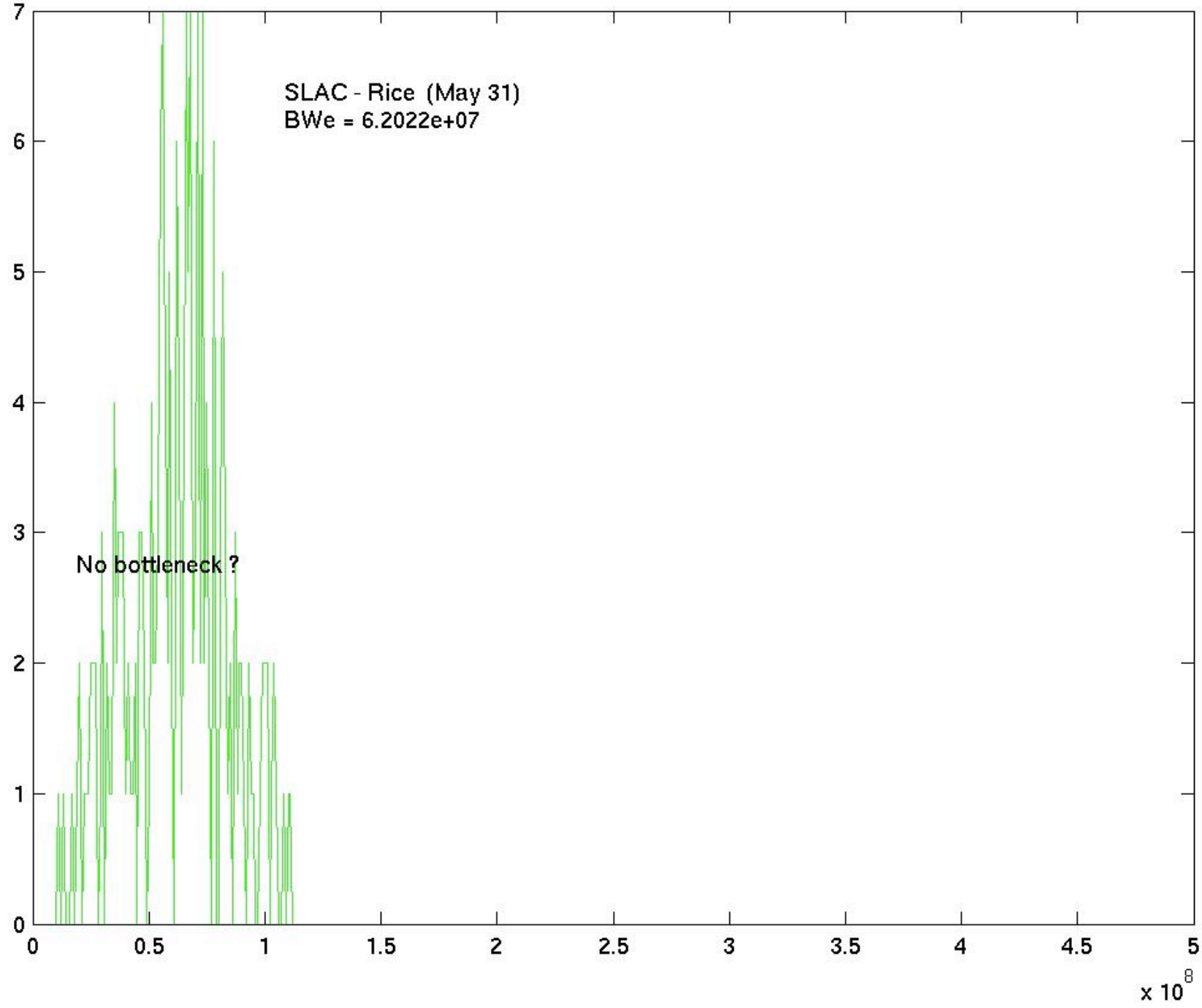

antonia/ccasn07/20020531_93617/data
correspond to Fig. Real Traffic
BWe = 2.5329e+07
2 bottlenecks
local ? machine ??
155 Mbps
120
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x 10^8

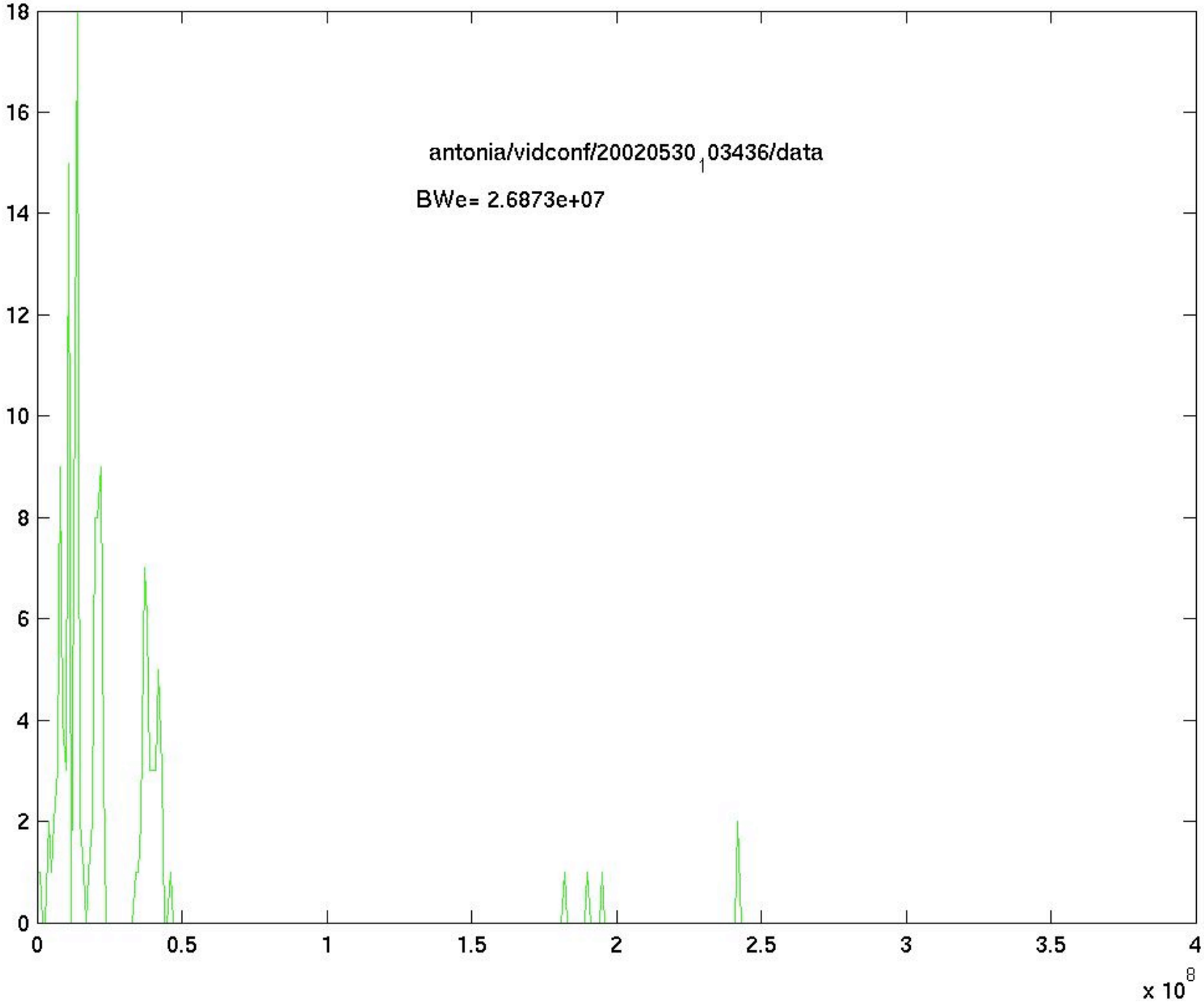

RTT in ms(tomo)
(1)
(13)
RTR-DMZ1-GER
(26)
(25)
(18)
snv-pos-slac
ccasn07.in2p3.fr
(7)
(27)
(22)
chi-s-snv
(9)
cernh9-pos100
sunstats.cern.ch
(49)
(10)
(43)
(38)
(41)
(42)
(47)
(35)
(45)
(19)
(53)
(24)
(46)
(44)
(36)
(56)
plop.lanl.gov
pcgiga.cern.ch
(58)
(32)
(51)
(59)
(37)
(57)
(34)
(50)
(30)
(20)
(29)
(33)
(54)
(16)
(5)
hp735.cvut.cz
(2)
(55)
plato.cacr.caltech.edu
(14)
hoss.ece.rice.edu
anguhub15.net.ubc.ca
Scale colors
0.0 - 1.0
1.0 - 10.0
10.0 - 100.0
100.0 -1000.0


antonia/plato/20020530_191637/data
BWe = 2.7039e+08
Maximum speed via parallel Iperf 300 Mbps
2 bottlenecks ?
50
45
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
x 10^8

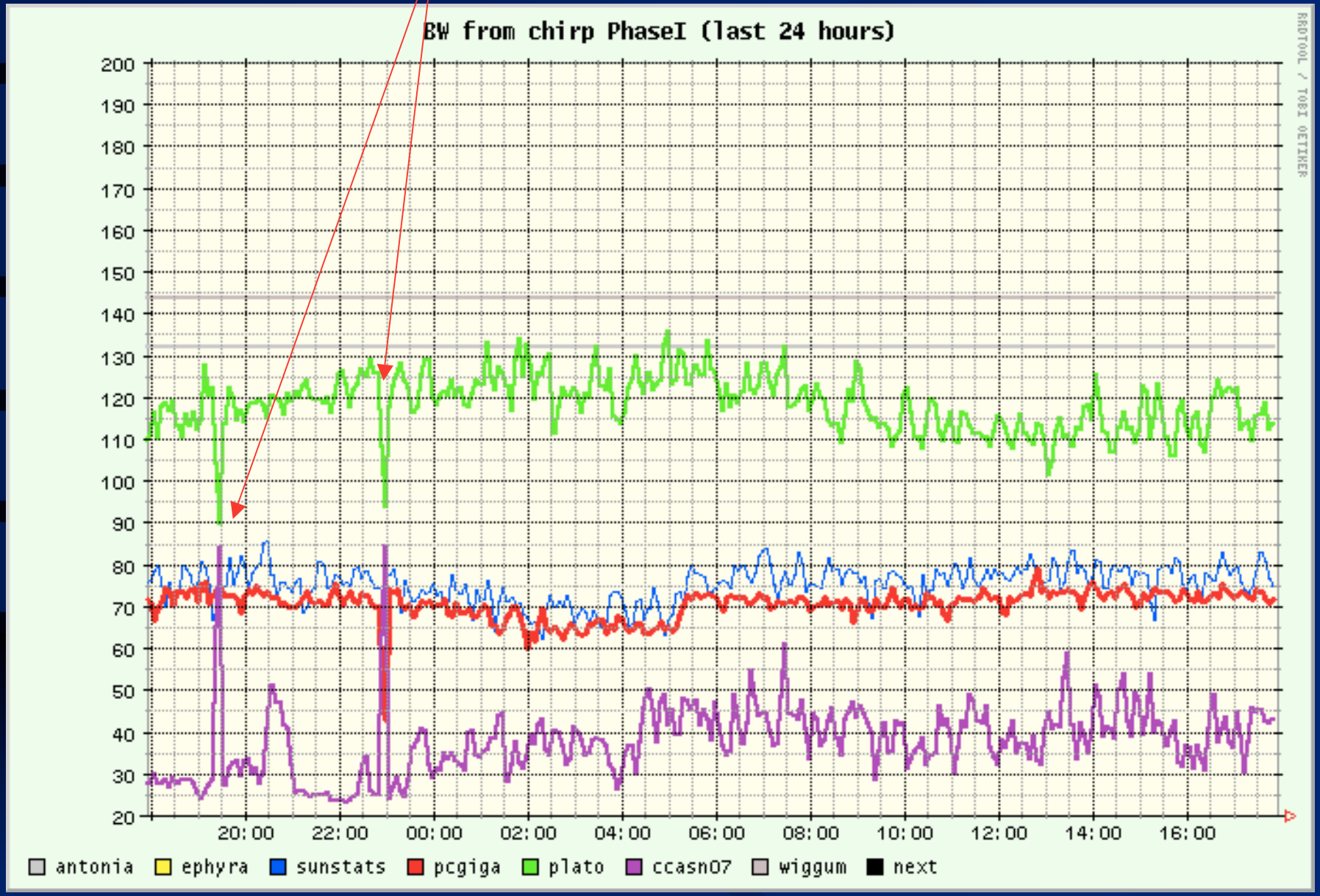

antonia/plop/20020530_153433/data
BWe = 3.1622e+08
Both machines have GB NIC
(maximum achieved BW via parallel Iperf 500 Mbps)
3 bottlenecks on different path structure
45
40
35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
Mbps
x 10^8

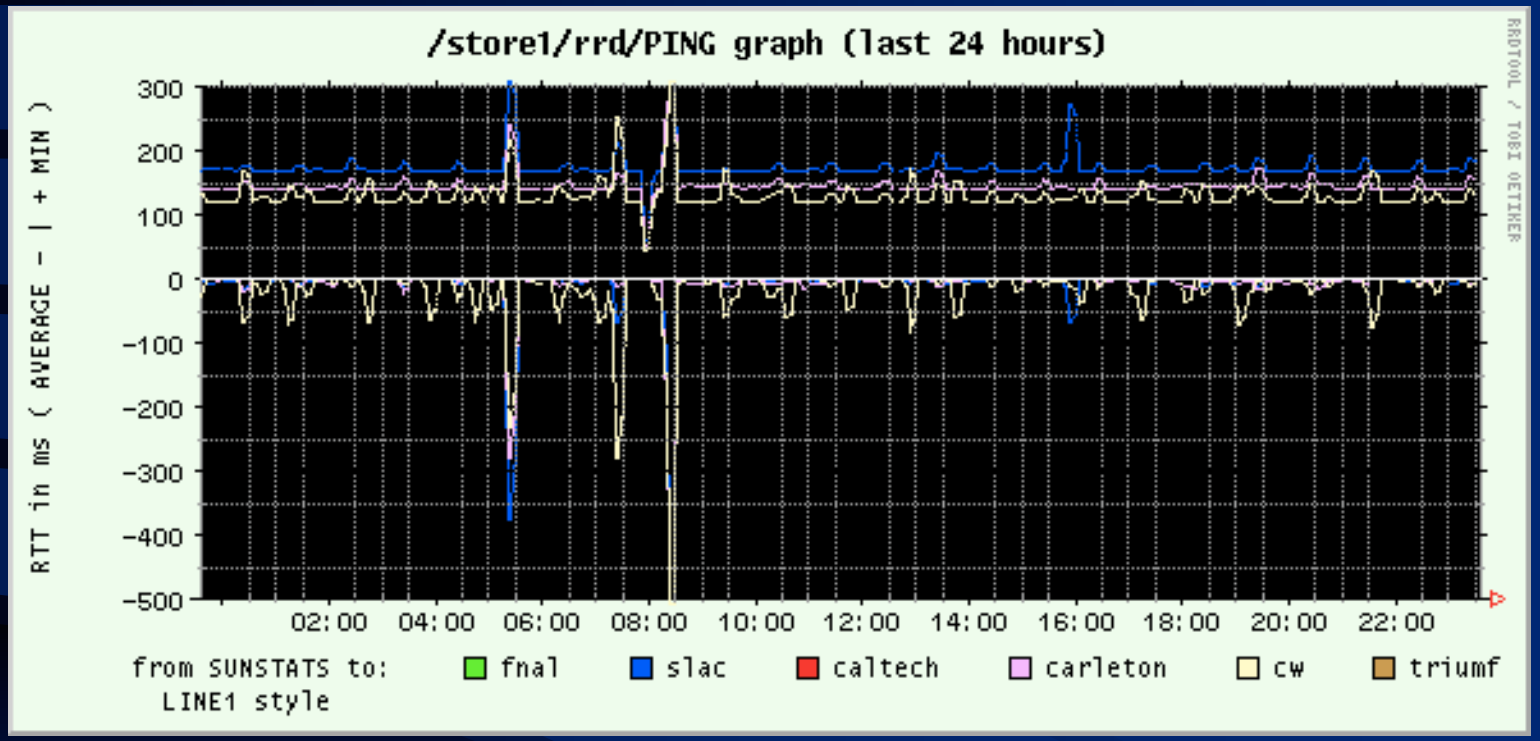

plop/20020530_155128/data
BWe = 3.1580e+08
Both machines have GB NIC
(maximum achieved BW via parallel Iperf 500 Mbps)
3 bottlenecks on different path structure
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10
Mbps
x 10^8


SLAC - Rice (May 31)
BWe = 6.2022e+07
No bottleneck ?
0
1
2
3
4
5
6
7
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
x 10^8


antonia/vidconf/20020530_103436/data
BWe= 2.6873e+07
18
16
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3
3.5
4
x 10^8

Reaction to the network problems




/store1/rrd/PING graph (last 24 hours)
RRDTOOL / TOBI OETIKER
RTT in ms (AVERAGE - | + MIN)
300
200
100
0
-100
-200
-300
-400
-500
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
from SUNSTATS to:
LINE1 style
fnal
slac
caltech
carleton
cw
triumf

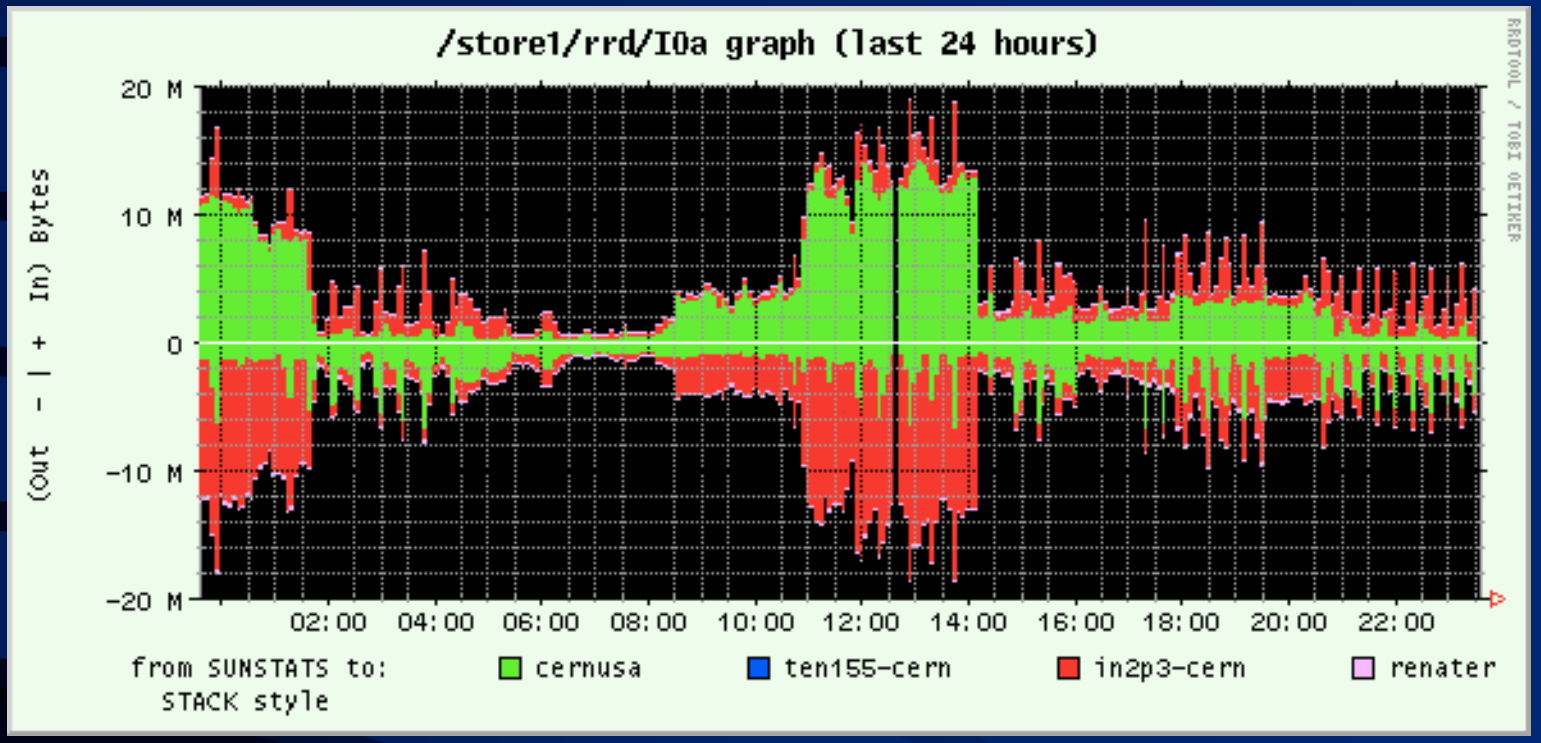
INCITE: Edge-based Traffic Processing and Service Inference for High-Performance Networks

Richard Baraniuk, Rice University; Les Cottrell, SLAC; Wu-chun Feng, LANL

Incite

CT detection

$$CT = f(dTR)$$

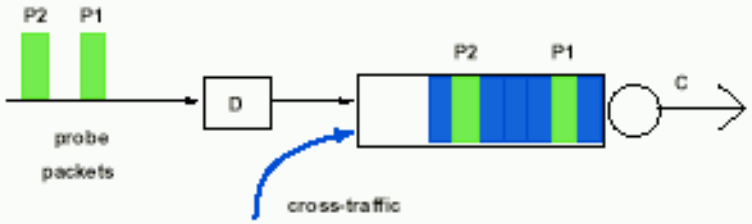

/store1/rrd/I0a graph (last 24 hours)
(Out - | + In) Bytes
20 M
10 M
0
-10 M
-20 M
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
from SUNSTATS to:
STACK style
cernusa
ten155-cern
in2p3-cern
renater
RRDTOOL / TOBI OETIKER

MF-CT Features and benefits

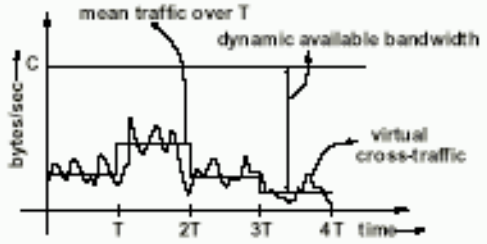
- No need access to routers !
  - Current monitoring systems for Load of traffic are based on SNMP or Flows (needs access to routers)
- Low cost:
  - Allows permanent monitoring (20 pkts/sec ~ overhead 10 Kbytes/sec)
  - Can be used as data provider for ABW prediction (**ABW=BW-CT**)
- We have 2 data points of CT per second !!
- Weak point for common use

  MATLAB code

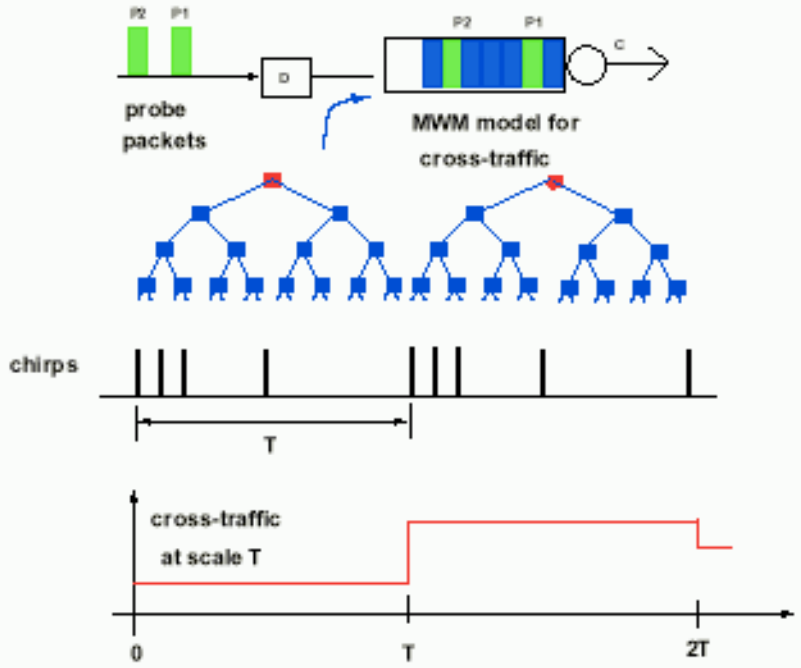
Probing

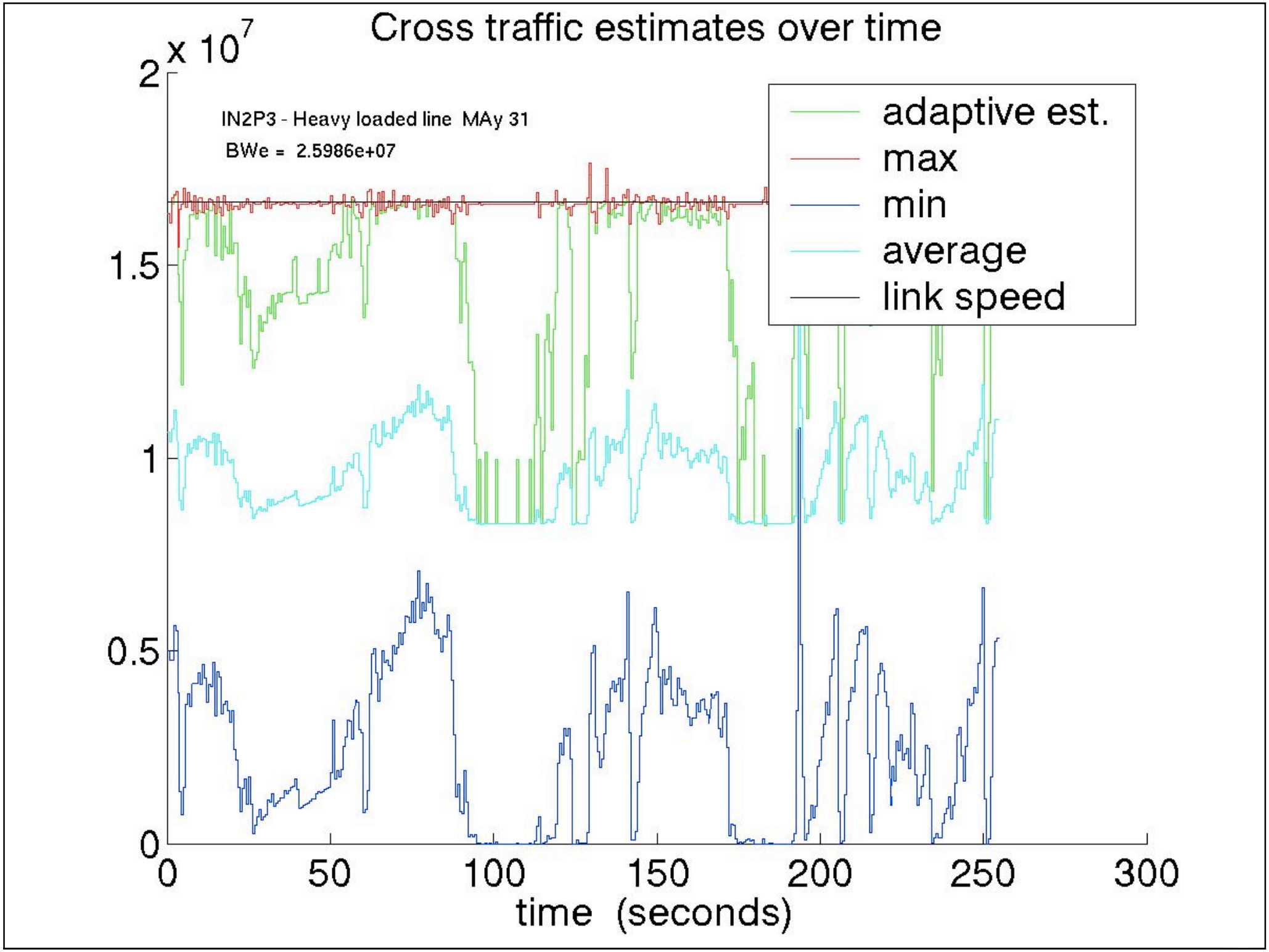


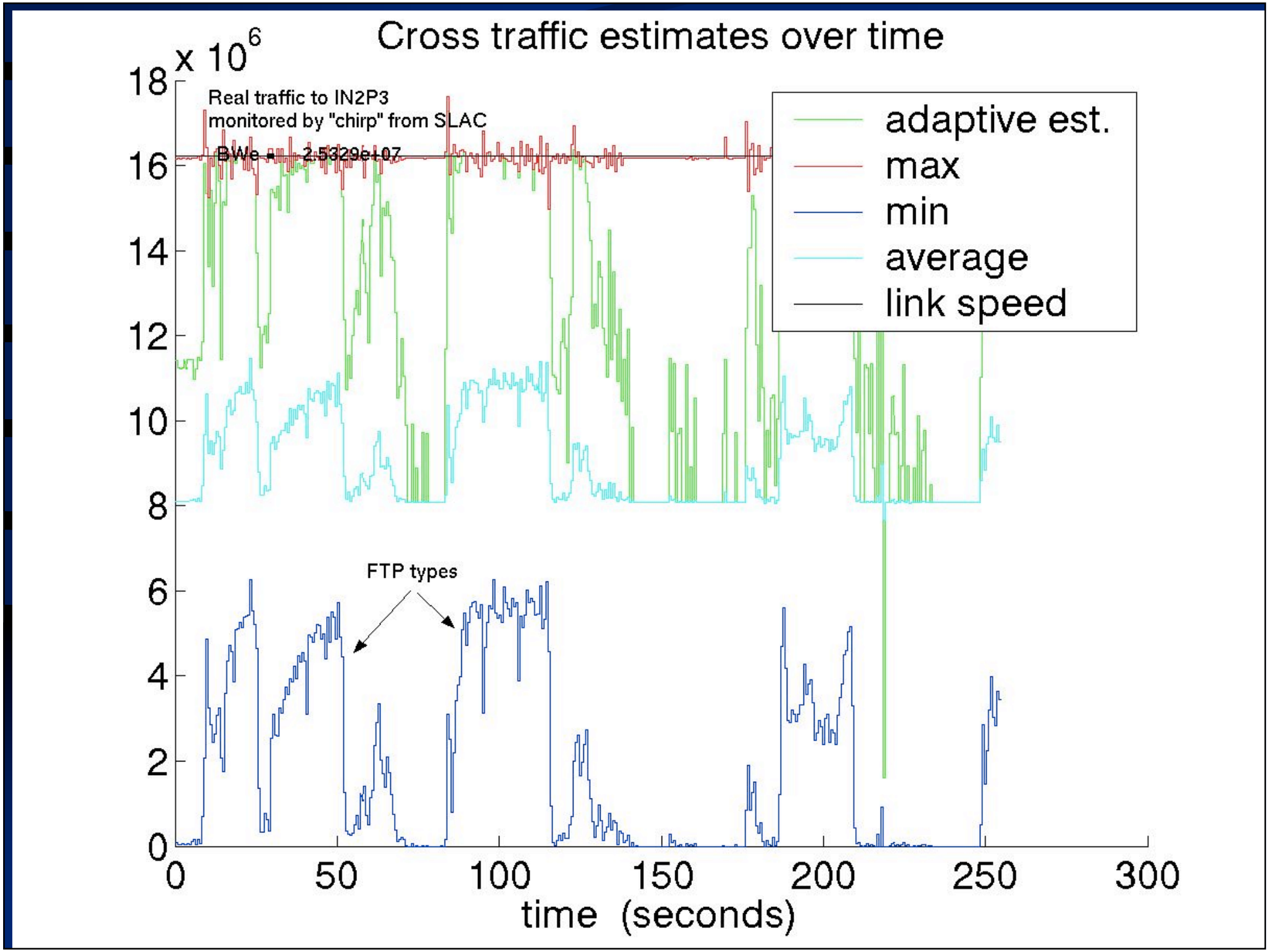
- delay ⇒ queue length ⇒ volume of cross-traffic

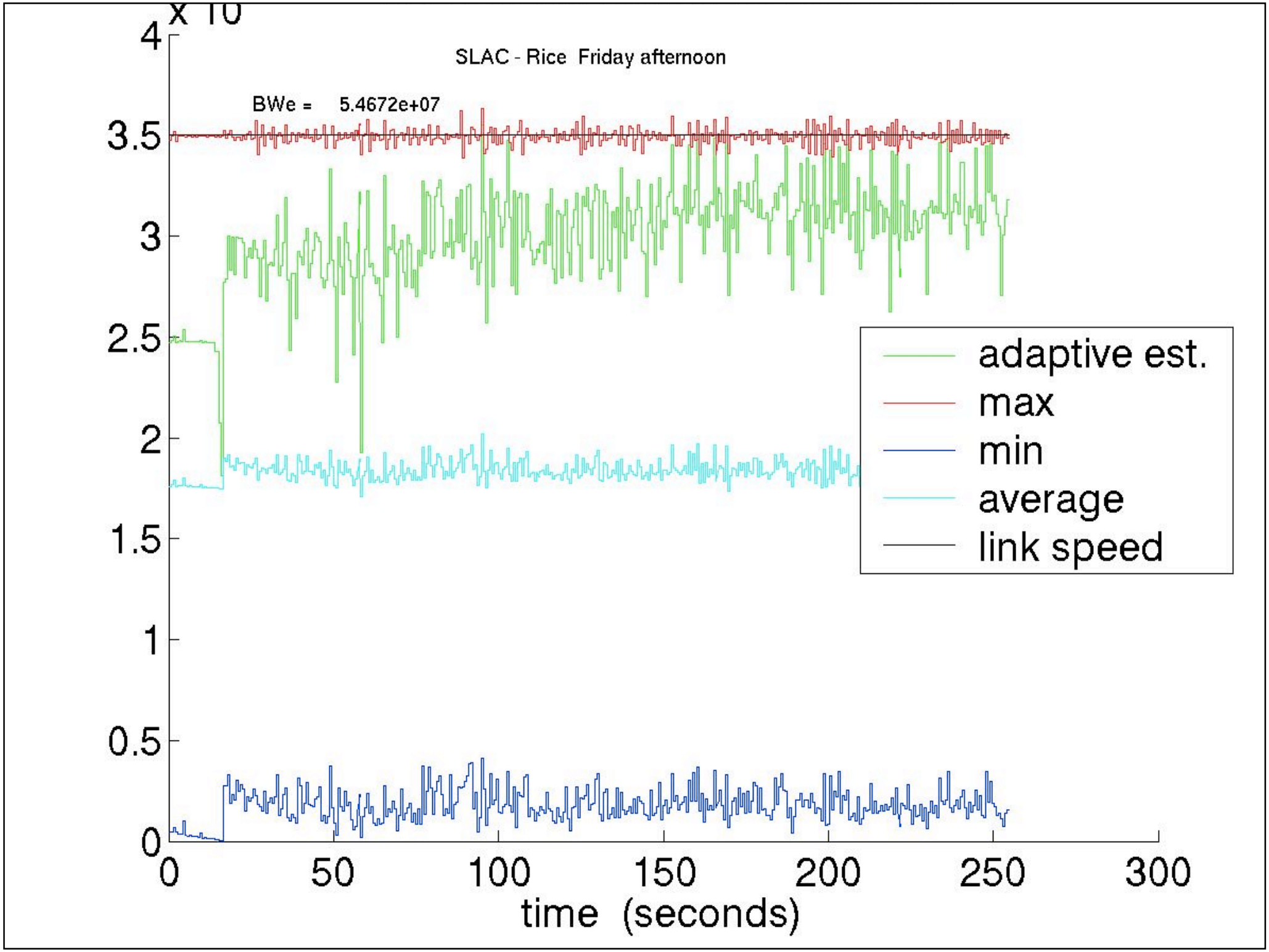


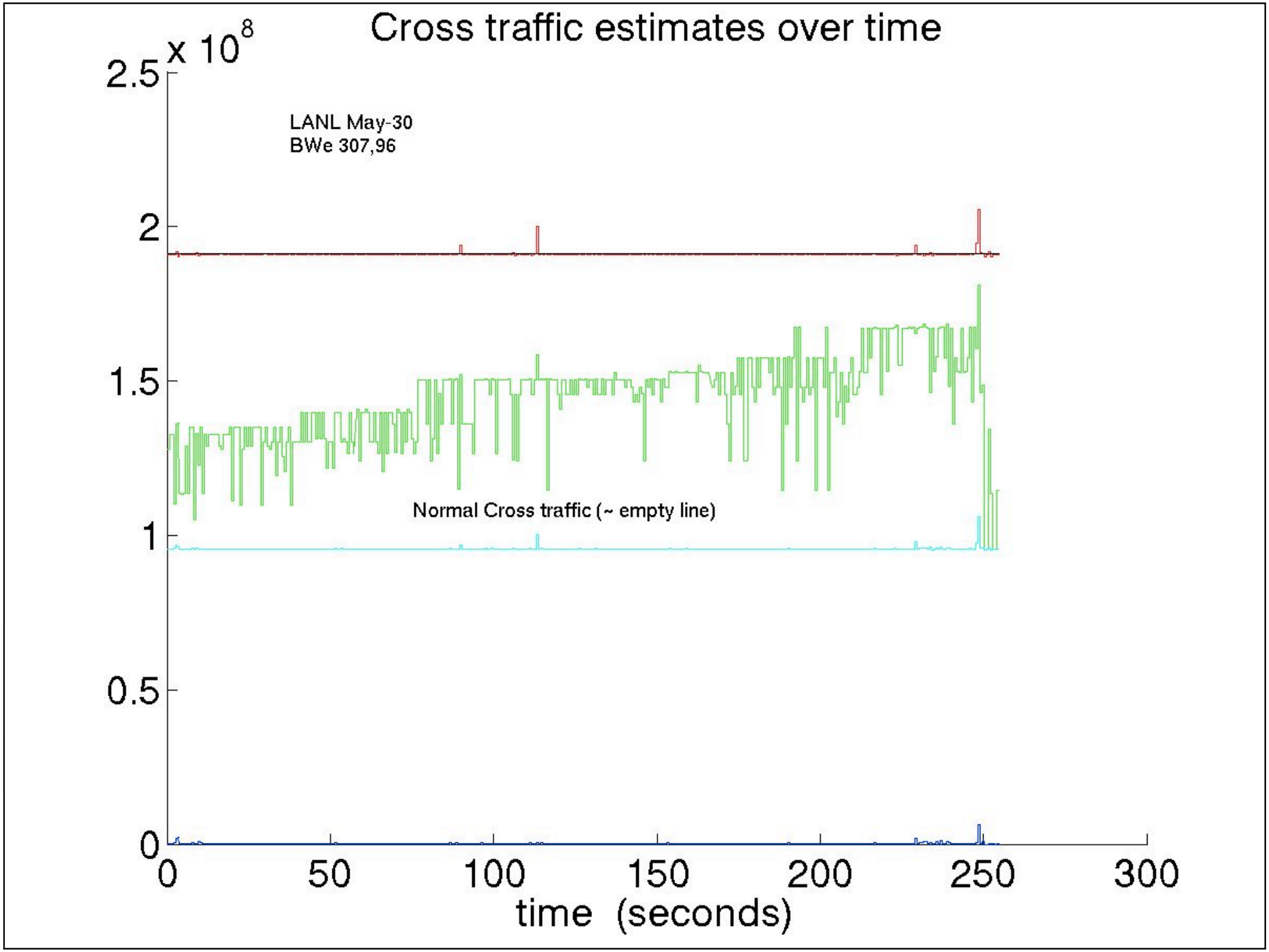
Cross-Traffic Estimation

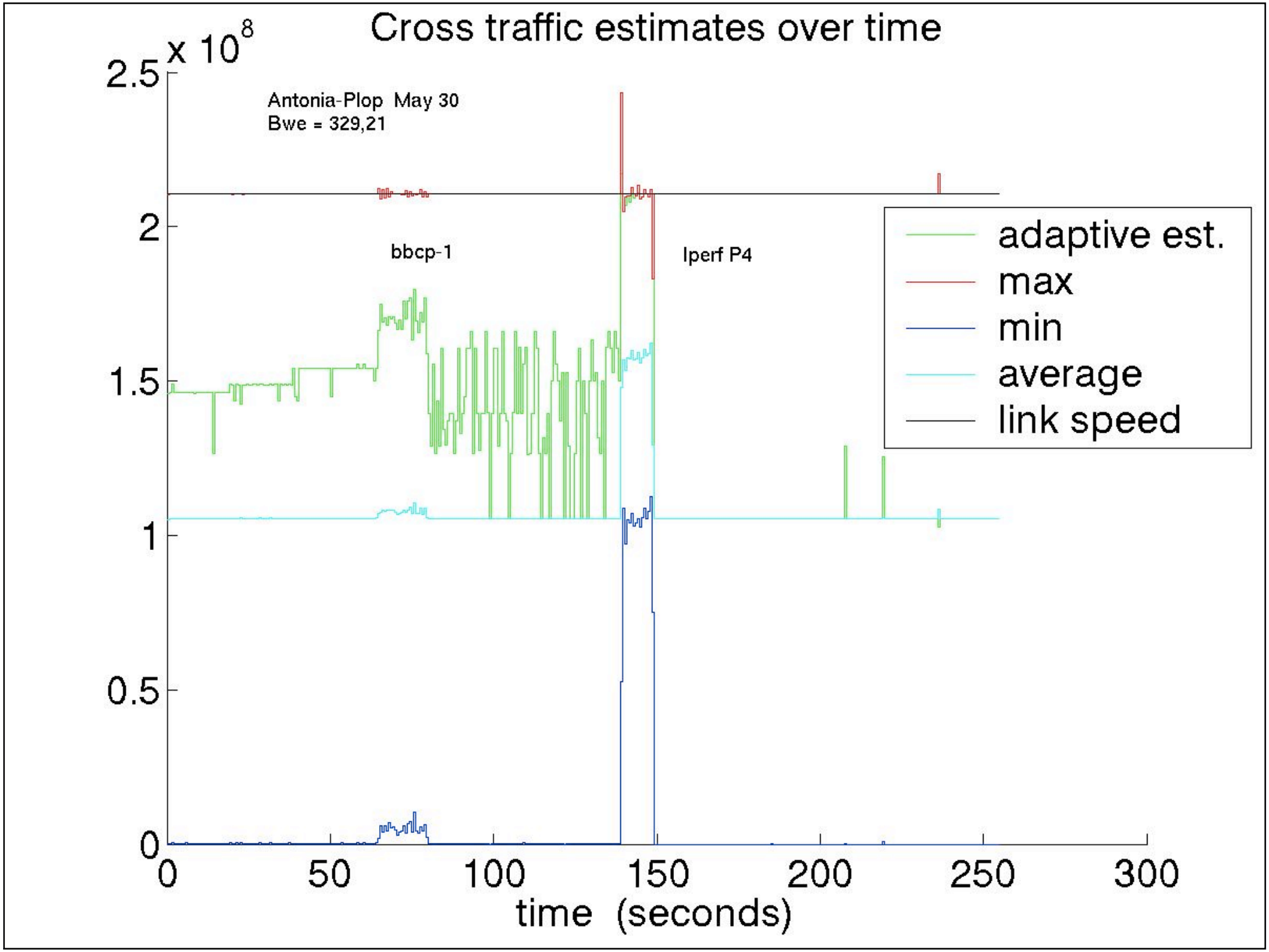


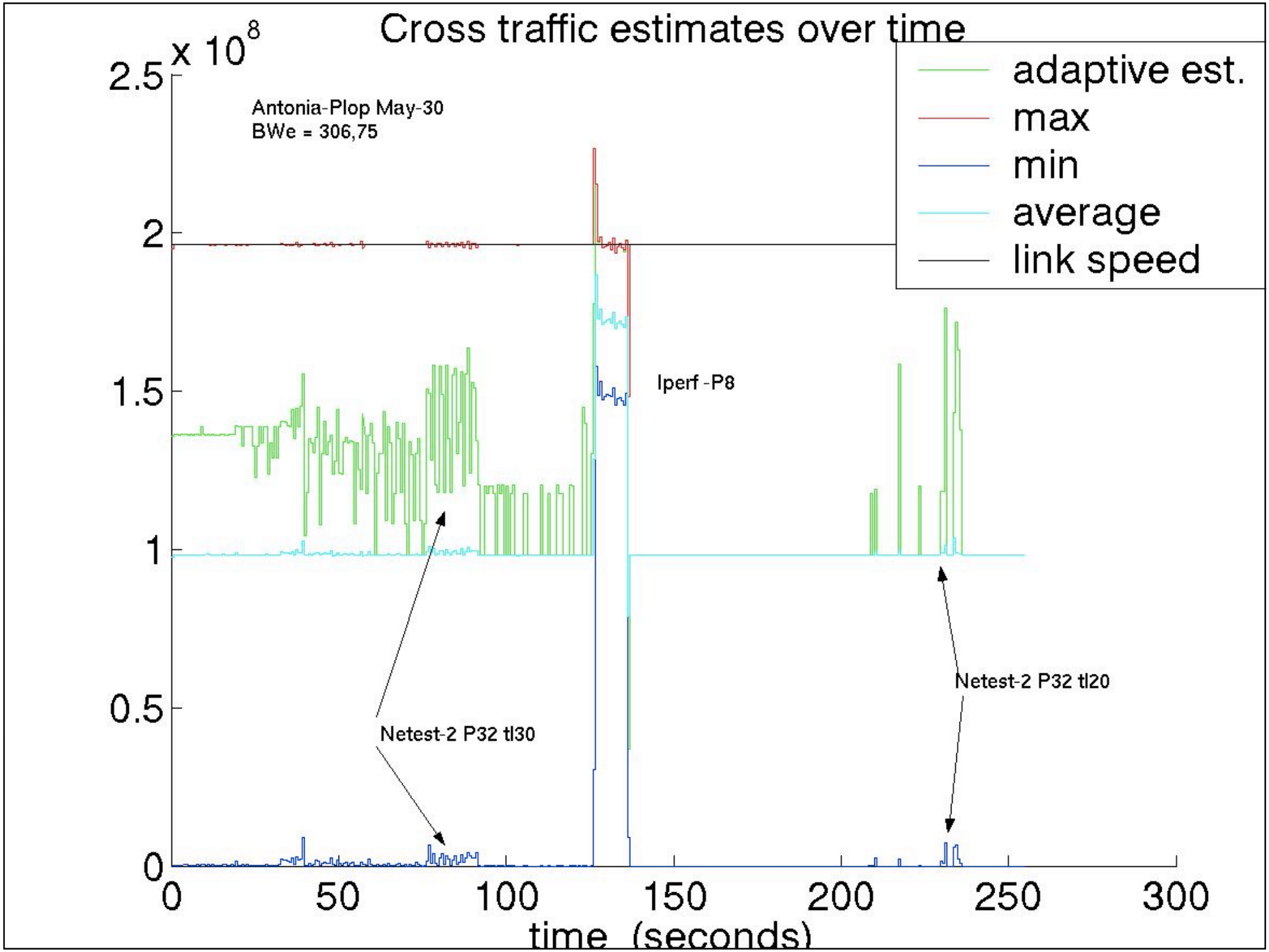

Cross traffic estimates over time
x 10^7
2
1.5
1
0.5
0
IN2P3 - Heavy loaded line MAy 31
BWe = 2.5986e+07
adaptive est.
max
min
average
link speed
0
50
100
150
200
250
300
time (seconds)

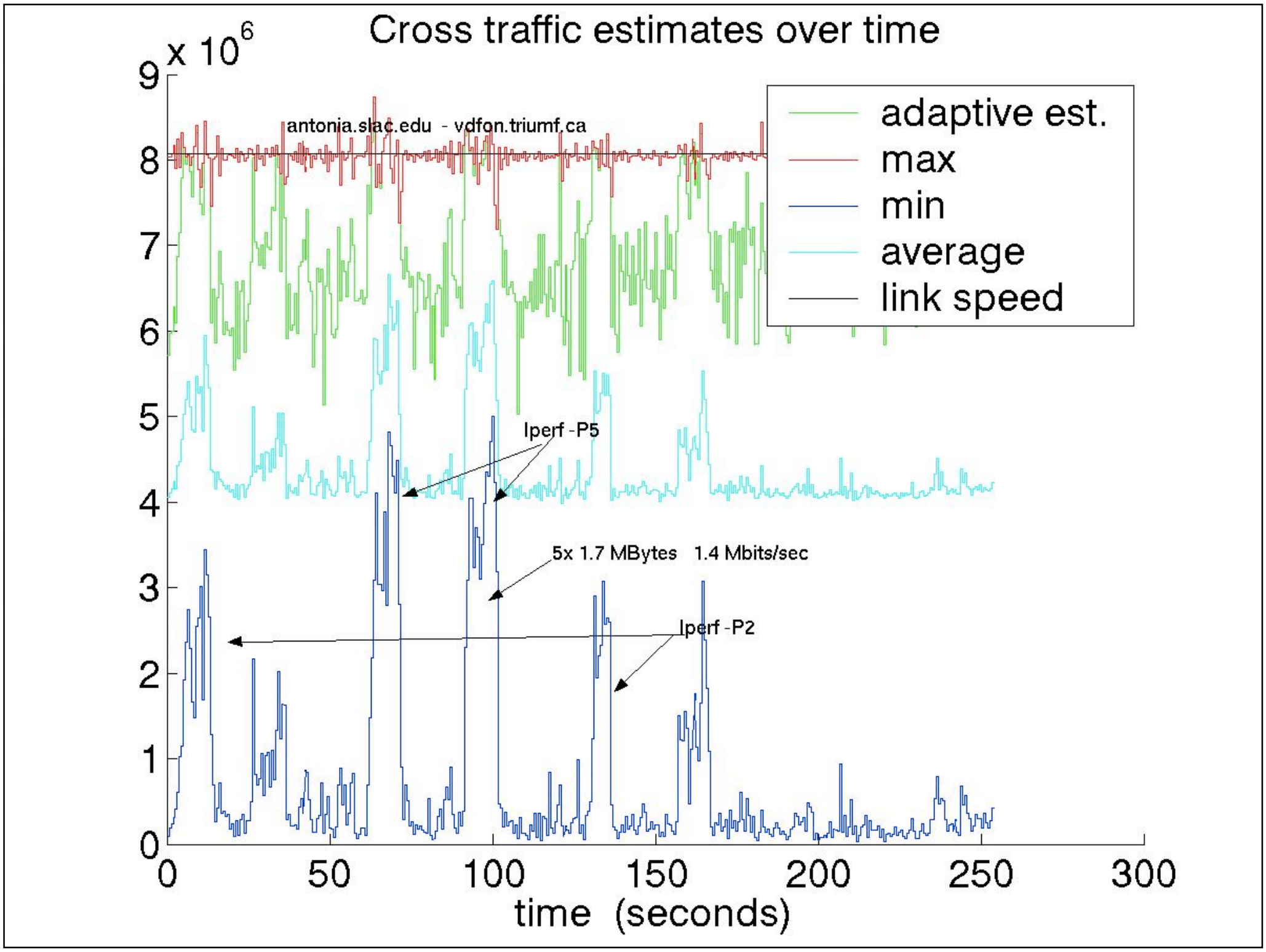

Cross traffic estimates over time
x 10^6
Real traffic to IN2P3
monitored by "chirp" from SLAC
BWe = 2.5329e+07
adaptive est.
max
min
average
link speed
FTP types
18
16
14
12
10
8
6
4
2
0
0
50
100
150
200
250
300
time (seconds)

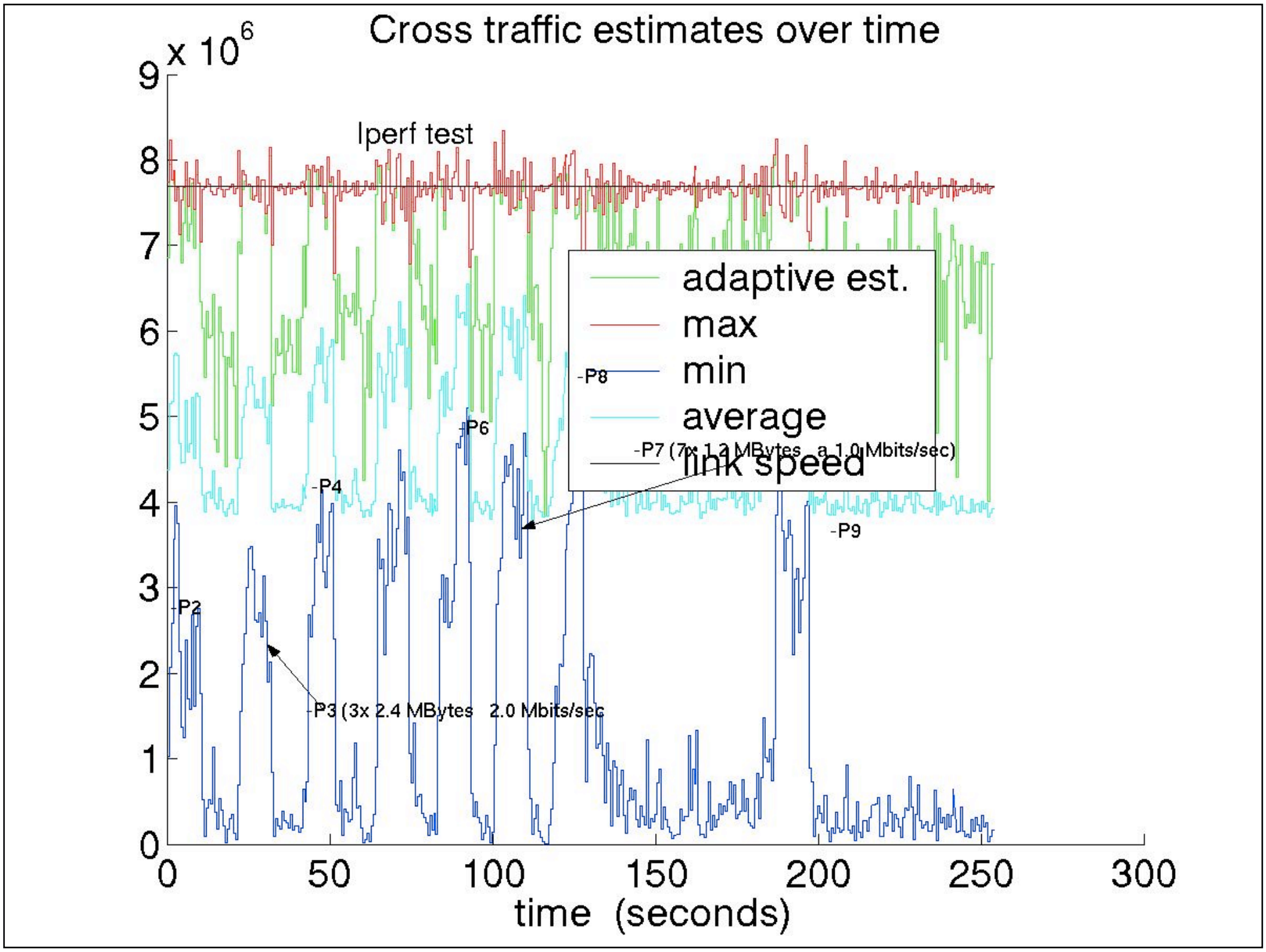

x 10
4
SLAC - Rice Friday afternoon
BWe = 5.4672e+07
3.5
3
2.5
2
1.5
1
0.5
0
0
50
100
150
200
250
300
time (seconds)
adaptive est.
max
min
average
link speed

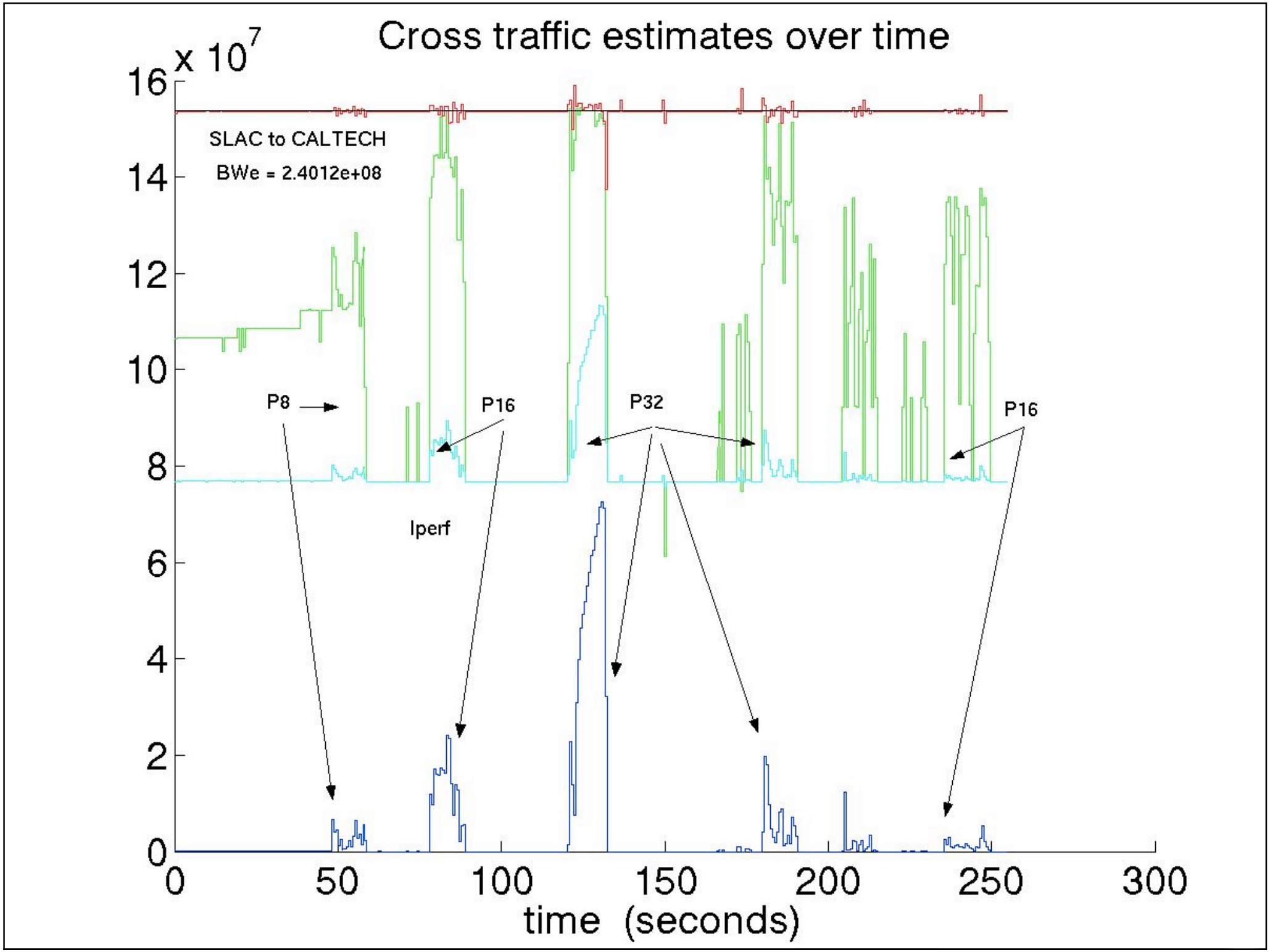

Cross traffic estimates over time
x 10^8
2.5
2
1.5
1
0.5
0
LANL May-30
BWe 307,96
Normal Cross traffic (~ empty line)
0
50
100
150
200
250
300
time (seconds)

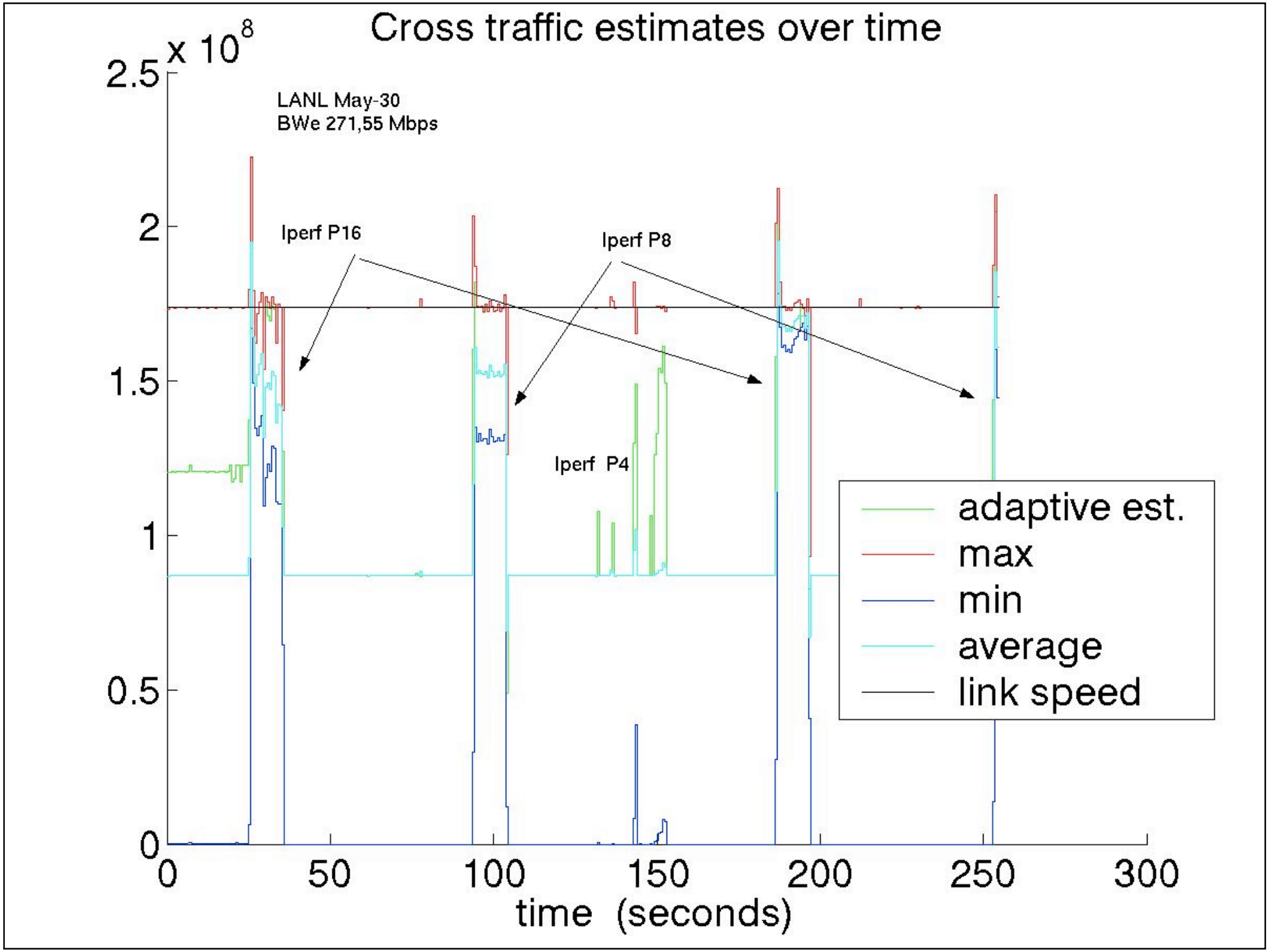

Cross traffic estimates over time
x 10^8
2.5
2
1.5
1
0.5
0
Antonia-Plop May 30
Bwe = 329,21
bbcp-1
Iperf P4
adaptive est.
max
min
average
link speed
0
50
100
150
200
250
300
time (seconds)


Cross traffic estimates over time
x 10^8
2.5
2
1.5
1
0.5
0
Antonia-Plop May-30
BWe = 306,75
adaptive est.
max
min
average
link speed
Iperf -P8
Netest-2 P32 tl30
Netest-2 P32 tl20
0
50
100
150
200
250
300
time (seconds)


Cross traffic estimates over time
x 10^6
antonia.slac.edu - vdfon.triumf.ca
adaptive est.
max
min
average
link speed
Iperf -P5
5x 1.7 MBytes 1.4 Mbits/sec
Iperf -P2
time (seconds)
0
1
2
3
4
5
6
7
8
9
0
50
100
150
200
250
300


Cross traffic estimates over time
x 10^6
9
8
7
6
5
4
3
2
1
0
Iperf test
adaptive est.
max
min
average
link speed
-P8
-P6
-P7 (7x 1.2 MBytes a 1.0 Mbits/sec)
-P4
-P9
-P2
-P3 (3x 2.4 MBytes 2.0 Mbits/sec
0
50
100
150
200
250
300
time (seconds)

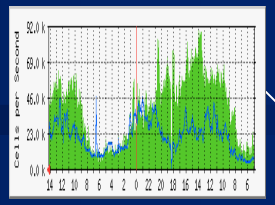

Cross traffic estimates over time
x 10^7
16
14
12
10
8
6
4
2
0
SLAC to CALTECH
BWe = 2.4012e+08
P8
P16
P32
P16
Iperf
0
50
100
150
200
250
300
time (seconds)


Cross traffic estimates over time
x 10^8
2.5
2
1.5
1
0.5
0
LANL May-30
BWe 271,55 Mbps
Iperf P16
Iperf P8
Iperf P4
adaptive est.
max
min
average
link speed
0
50
100
150
200
250
300
time (seconds)

MF-CT verification model

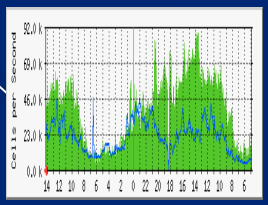
UDP echo

SNMP counter



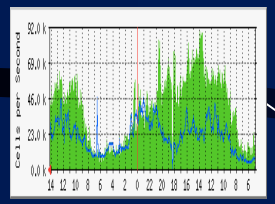
CERN

SNMP counter



internet

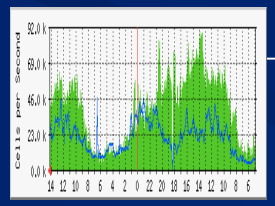
SNMP counter



SLAC

IN2P3

SNMP counter



MF-CT Simulator

UDP echo