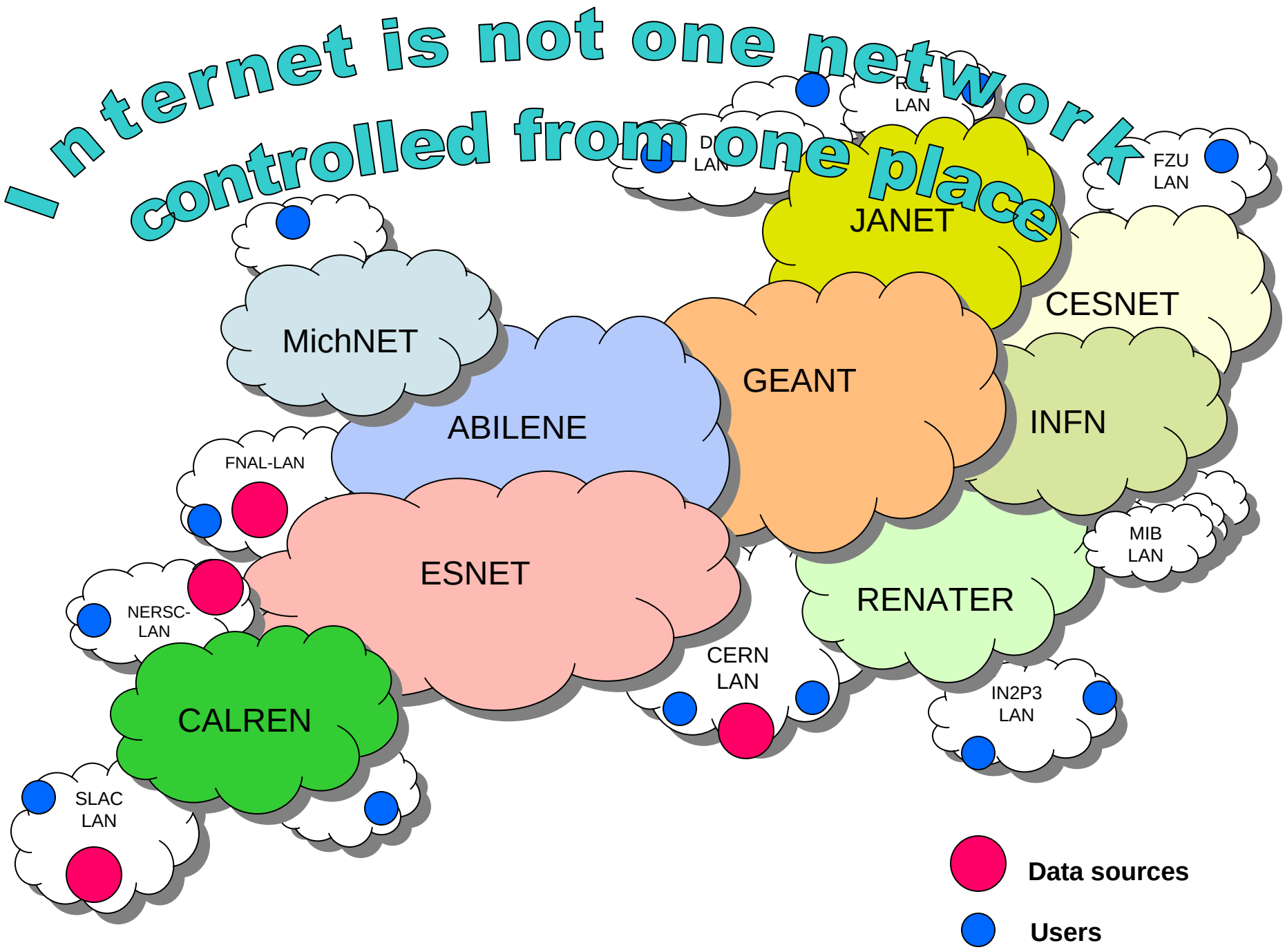
What we have learned from developing and running ABwE

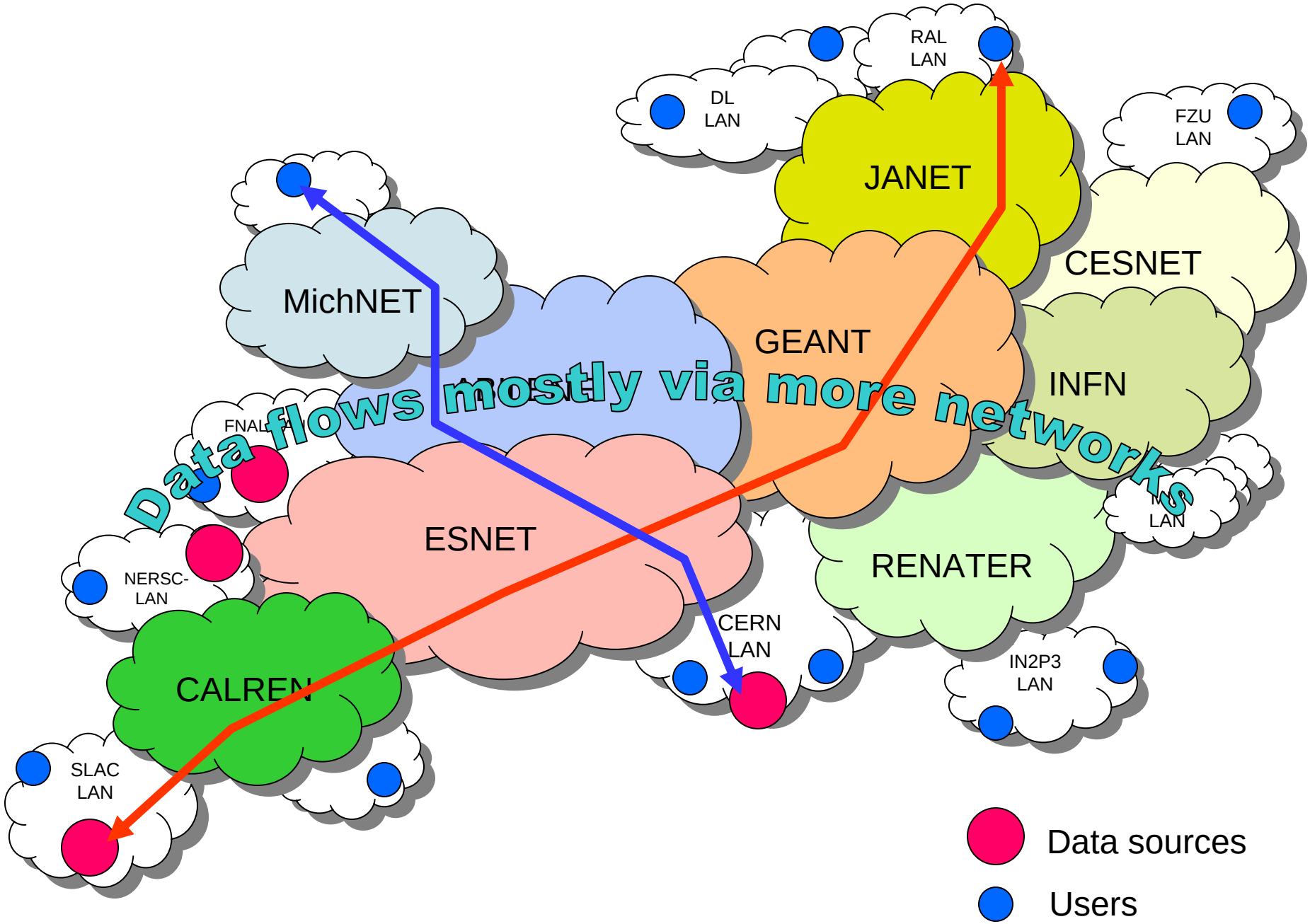
Jiri Navratil, Les R.Cottrell
(SLAC)

Why E2E tools are needed

- **The scientific community is increasingly dependent on networking as international cooperation grows.** HEP users (needs transfer huge amount of data between experimental sites as **SLAC, FNAL, CERN**, etc. (where data is created) and home institutes spread over the world)
- What ISPs (as Abilene,Esnet,Geant..) can offer to the users for getting information?

  (Not too much because they are only in the middle of the path and they don't cover all parts of connections)


Internet is not one network
controlled from one place
JANET
CESNET
MichNET
GEANT
ABILENE
INFN
FNAL-LAN
ESNET
MIB
LAN
RENATER
NERSC-
LAN
CERN
LAN
CALREN
IN2P3
LAN
SLAC
LAN
FZU
LAN
Data sources
Users


RAL
LAN
DL
LAN
JANET
FZU
LAN
CESNET
MichNET
GEANT
INFN
Data flows mostly via more networks
FNAL
ESNET
NERSC-
LAN
RENATER
CERN
LAN
IN2P3
LAN
CALREN
SLAC
LAN
Data sources
Users

- There must be always **somebody who gives complex information** to the **users of the community**

  or

  the **users** have to have **a tool which give them such information**

- **How fast I can transfer 20 GB from my experimental site (SLAC,CERN) to my home institute?**
- **Can I run graphical 3D visualization program with data located 1000 miles away?**
- **How stable is line ? (Can I use it in the same conditions for 5 minutes or 2 hours or whole day ?)**

  All such questions must be **replied in few seconds** doesn't matter if for **individual user** **or for Grid brokers**

- **Global science has no day and night.**

To reply this we needed the tools that could be used in continuous mode 24 hours a day 7 days a week which can non intrusively detect changes on multiple path or on demand by any user

ABwE: Basic terminology:

- ***Generally:***

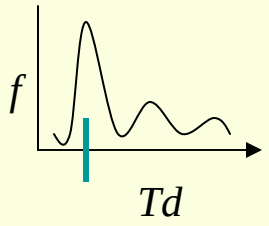
  Available bandwidth* = *Capacity* – *Load

- ***ABwE* measure *Td* – *Time dispersion P1-P2*** *(20x PP)*

  We are trying to distinguish two basic states in our results:

  - "***Dominate (free)***" – when *Td* ~= *const*

  -"***loaded***" with Td = other value



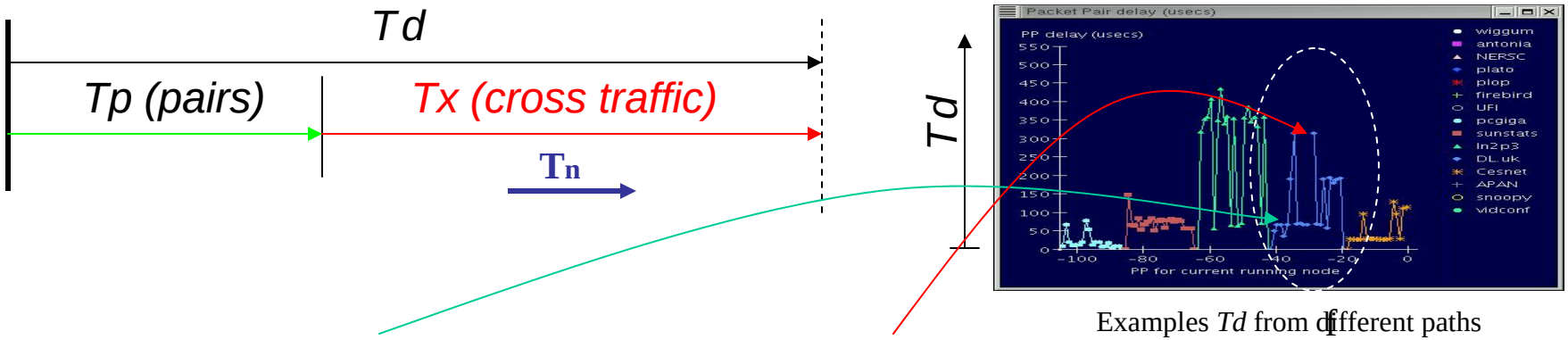
Td results from "***Dominate***" state are used to estimate

DBC - Dynamic Bottleneck Capacity

Td measured during the "***loaded***" state is used to estimate the level of **XTR (cross traffic)**

ABw* = *DBC* – *XTR

Abing: Estimation principles:

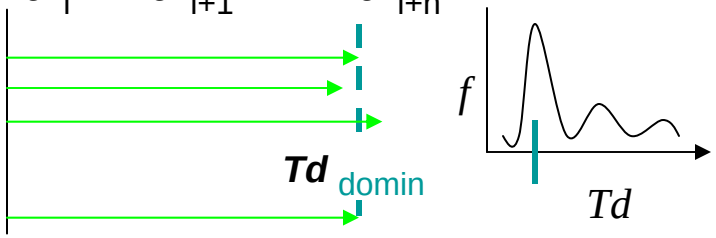


Examples *Td* from different paths

”Dominating state”

(when sustained load or no load)

$Td_i = Td_{i+1} = .. Td_{i+n}$



$Dbc = L_{pp}/Td_{domin}$

“Load state”

(when load is changing)

$q = Tx/Tn$ *(Tx=Td –Tp)*

Tx – *busy time (transmit time for cross trafic)*

Tn – *transmit time for average packet*

q – ***relative queue increment (QDF)*** *during decision interval Td (h-1)*

$u = q/(q+1)$

$CT = u * Dbc$

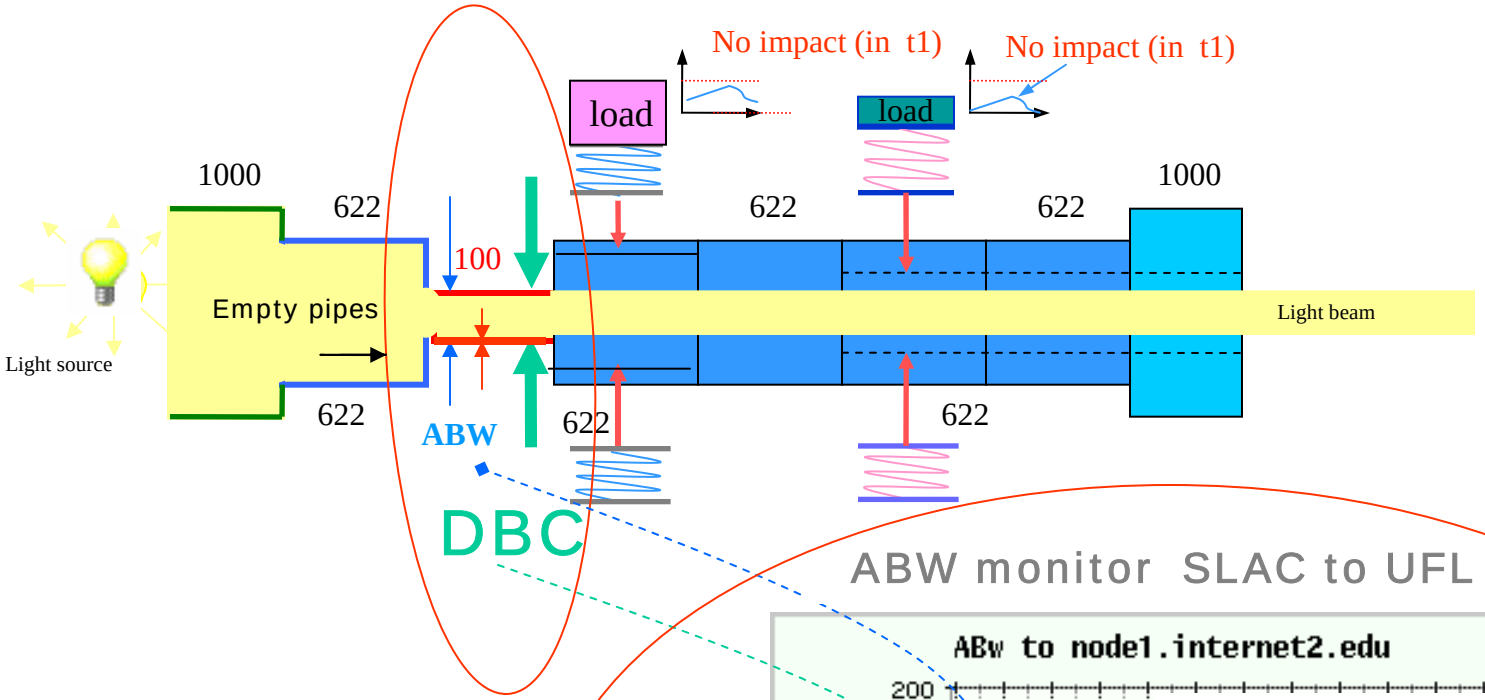
$Abw = Dbc - CT$

What is DBC

- DBC *characterize instant high capacity bottleneck that* ***DOMINATE*** *on the path*
- *It covers situations when* ***routers*** *in the path are* ***overloaded*** *and sending* ***packets back to back*** *with its maximal rates*
- *We discovered that in most cases* **only one node dominates** *in the instant of our measurements* (in our *decision interval*)

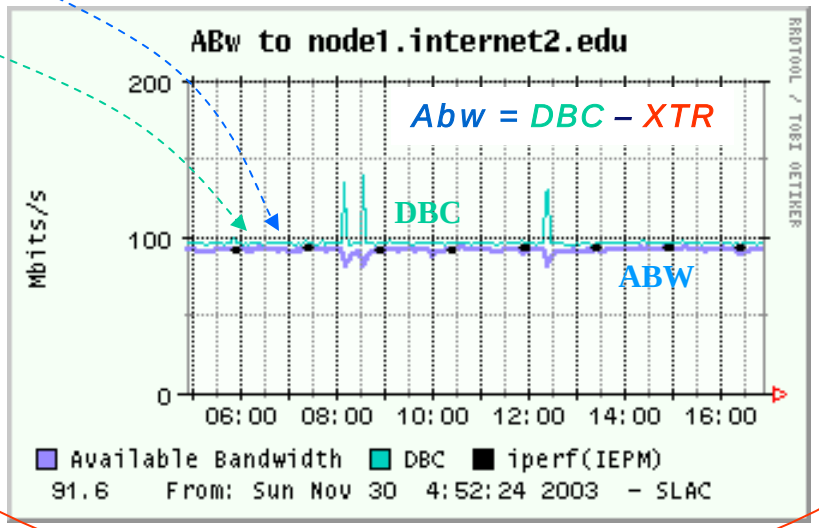
ABwE: Example of narrow link in the path

(Pipes analogy with different diameter and aperture)



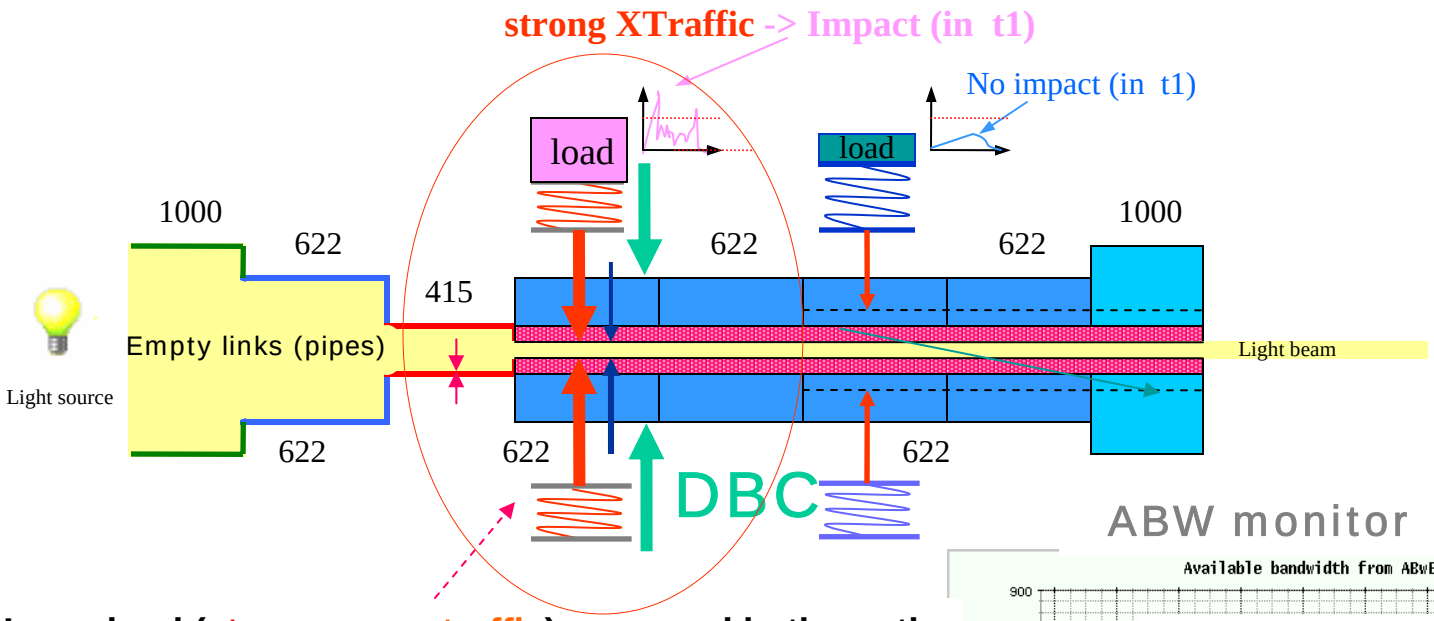
link that has domination effect on bandwidth

ABW monitor SLAC to UFL

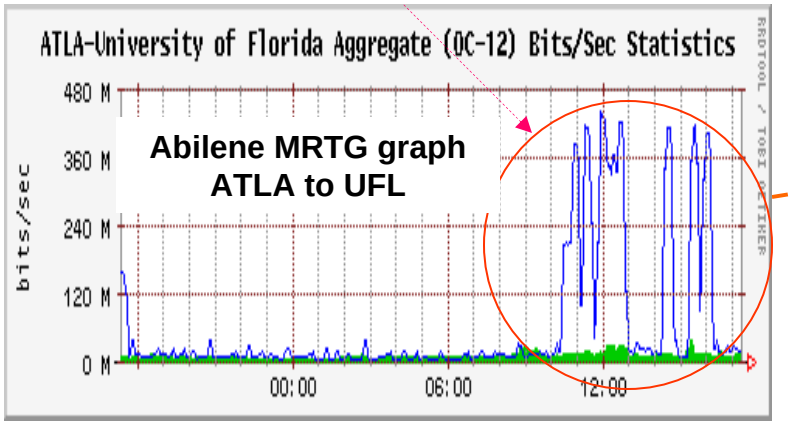


Example of heavy loaded link in the path

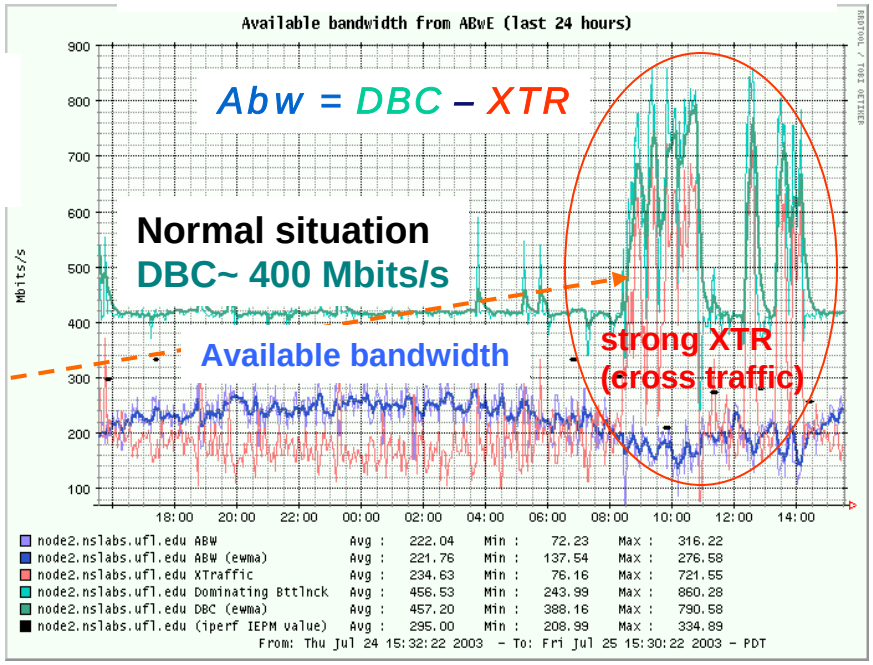
(Pipes analogy with different diameter and aperture)



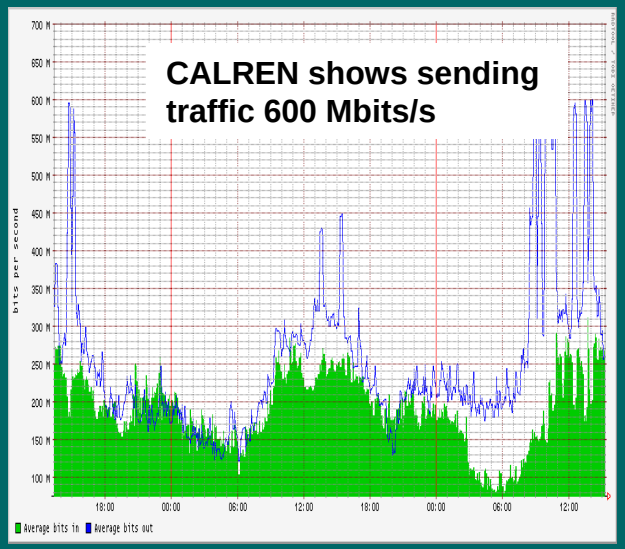
Heavy load (strong cross traffic) appeared in the path
It shows new DBC **in the path because this load** dominates **in whole path !**

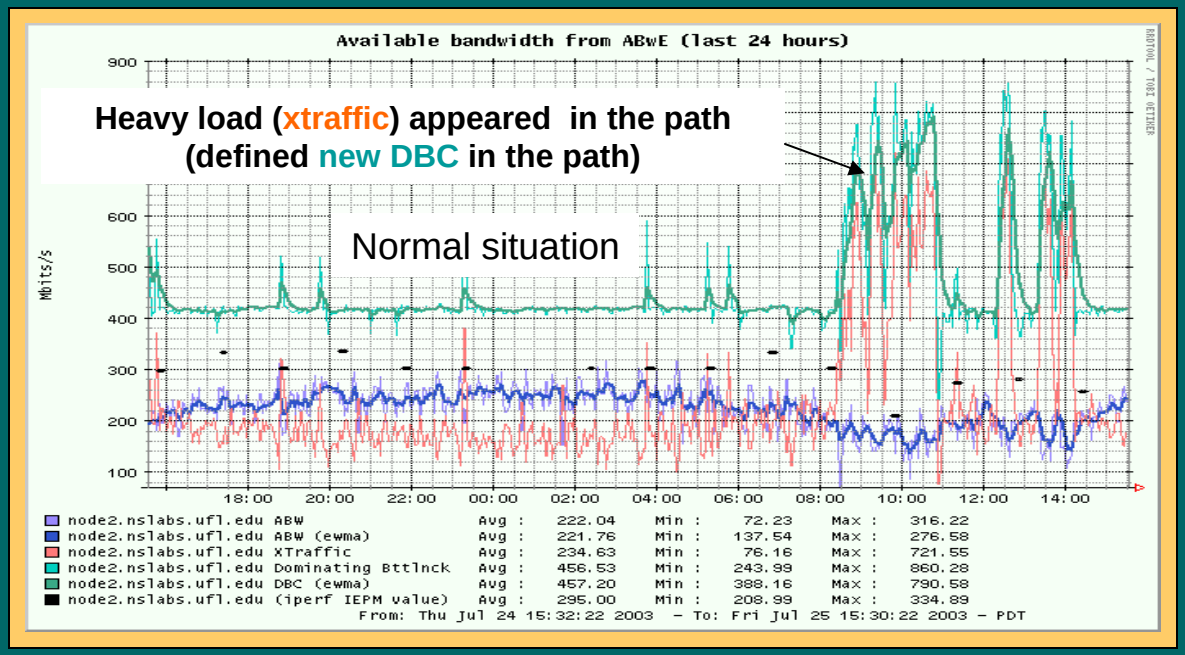


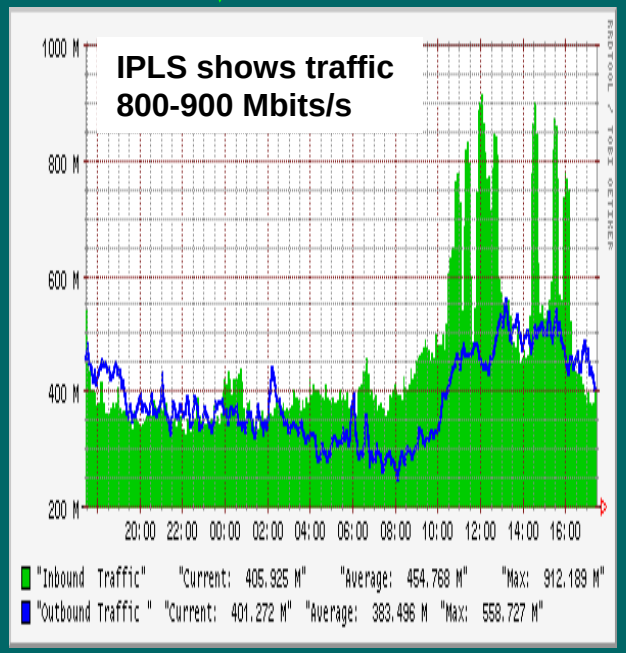
ABW monitor SLAC to UFL



ABwE / MRTG match: TCP test to UFL





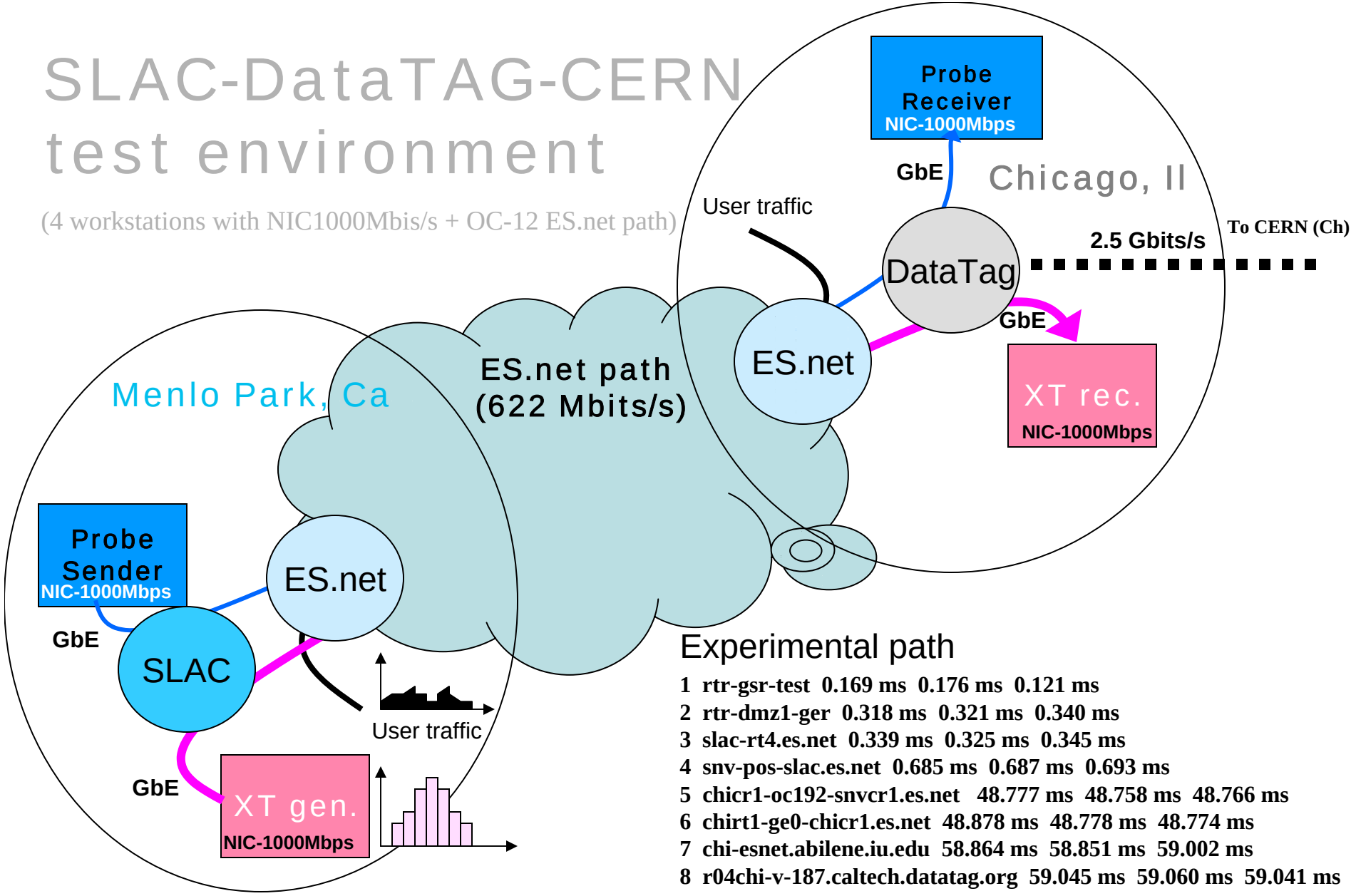


Confront ABwE results with other tools

Iperf,Pathload,Pathchirp

SLAC-DataTAG-CERN test environment

(4 workstations with NIC1000Mbis/s + OC-12 ES.net path)



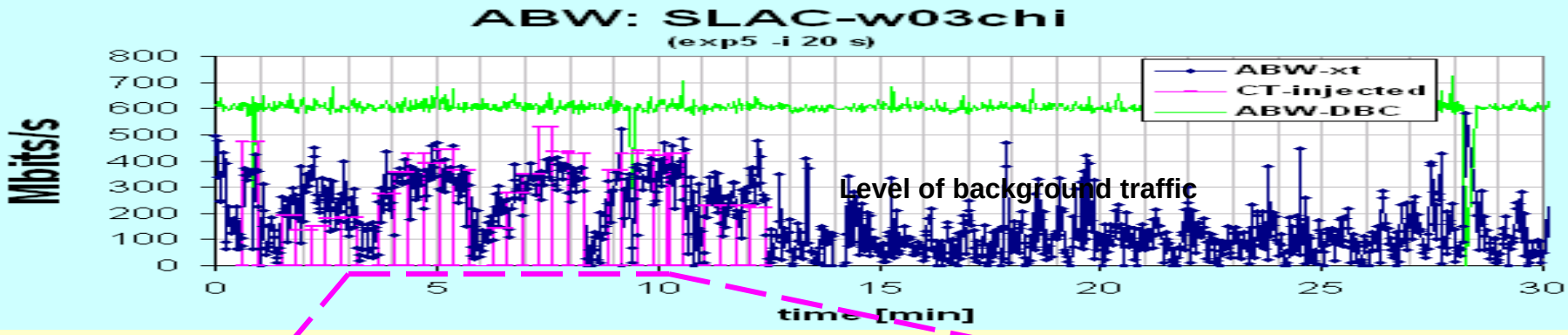
Experimental path

1 rtr-gsr-test 0.169 ms 0.176 ms 0.121 ms
2 rtr-dmz1-ger 0.318 ms 0.321 ms 0.340 ms
3 slac-rt4.es.net 0.339 ms 0.325 ms 0.345 ms
4 snv-pos-slac.es.net 0.685 ms 0.687 ms 0.693 ms
5 chicr1-oc192-snvcr1.es.net 48.777 ms 48.758 ms 48.766 ms
6 chirt1-ge0-chicr1.es.net 48.878 ms 48.778 ms 48.774 ms
7 chi-esnet.abilene.iu.edu 58.864 ms 58.851 ms 59.002 ms
8 r04chi-v-187.caltech.datatag.org 59.045 ms 59.060 ms 59.041 ms

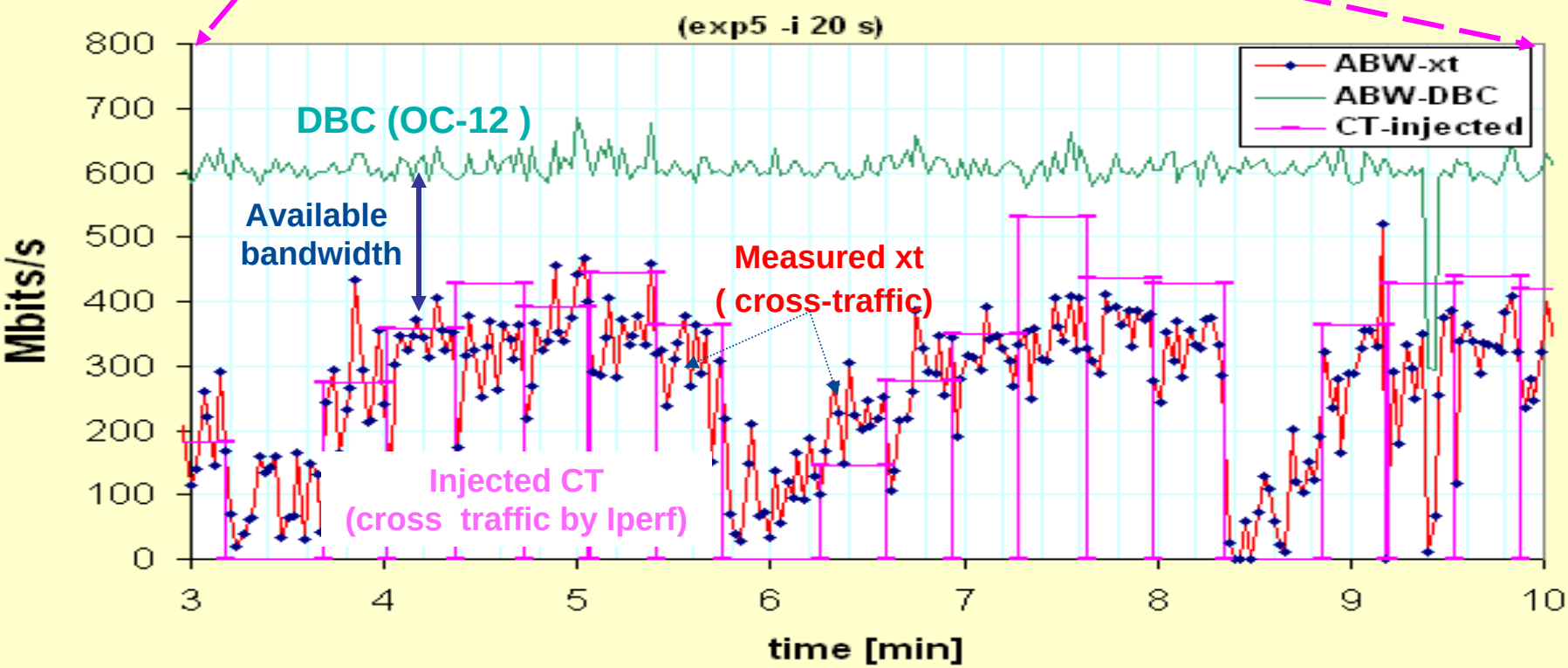
Probing packets
Injected Cross traffic
User traffic (background)

The match of the cross traffic

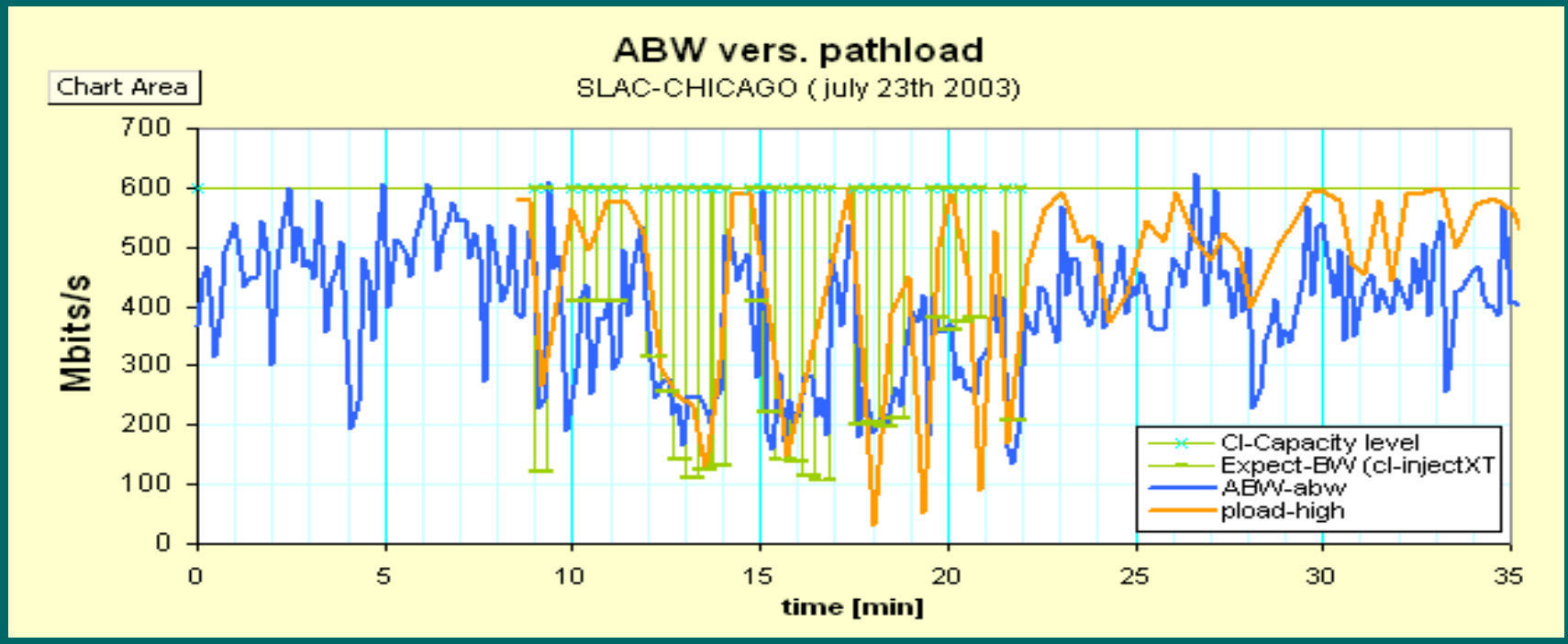
(ABW – XT compare to injection traffic generated by Iperf)

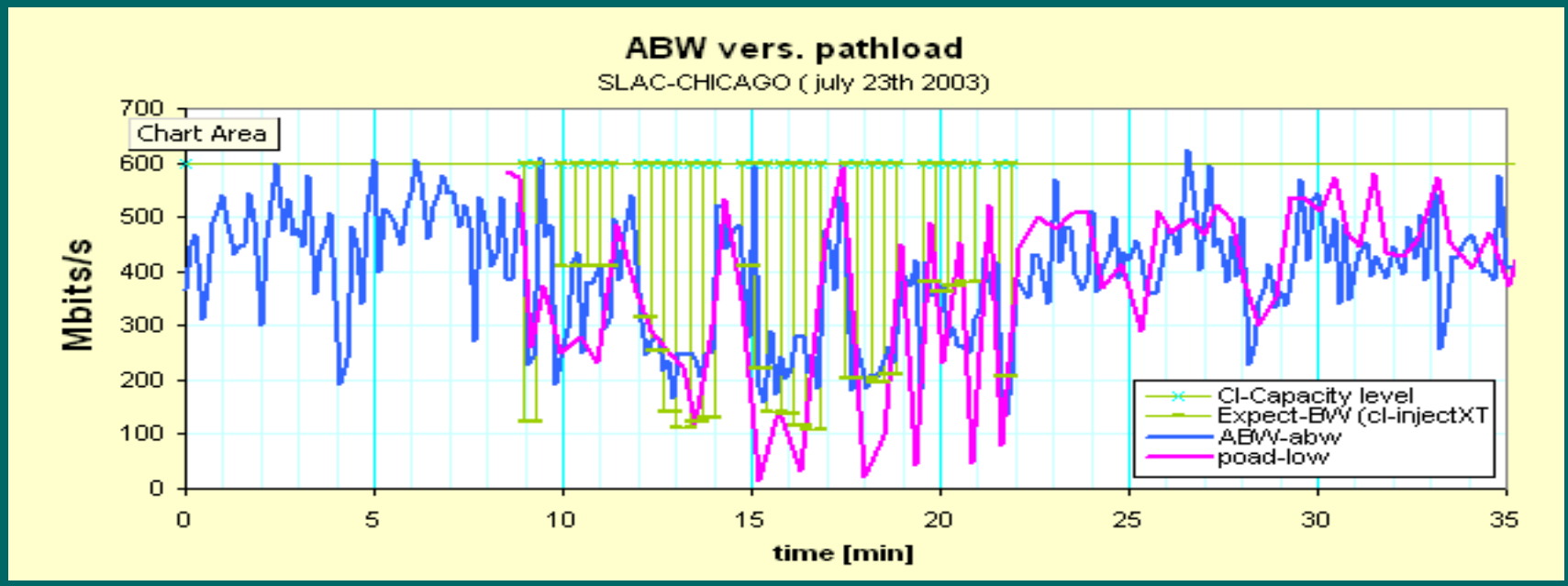


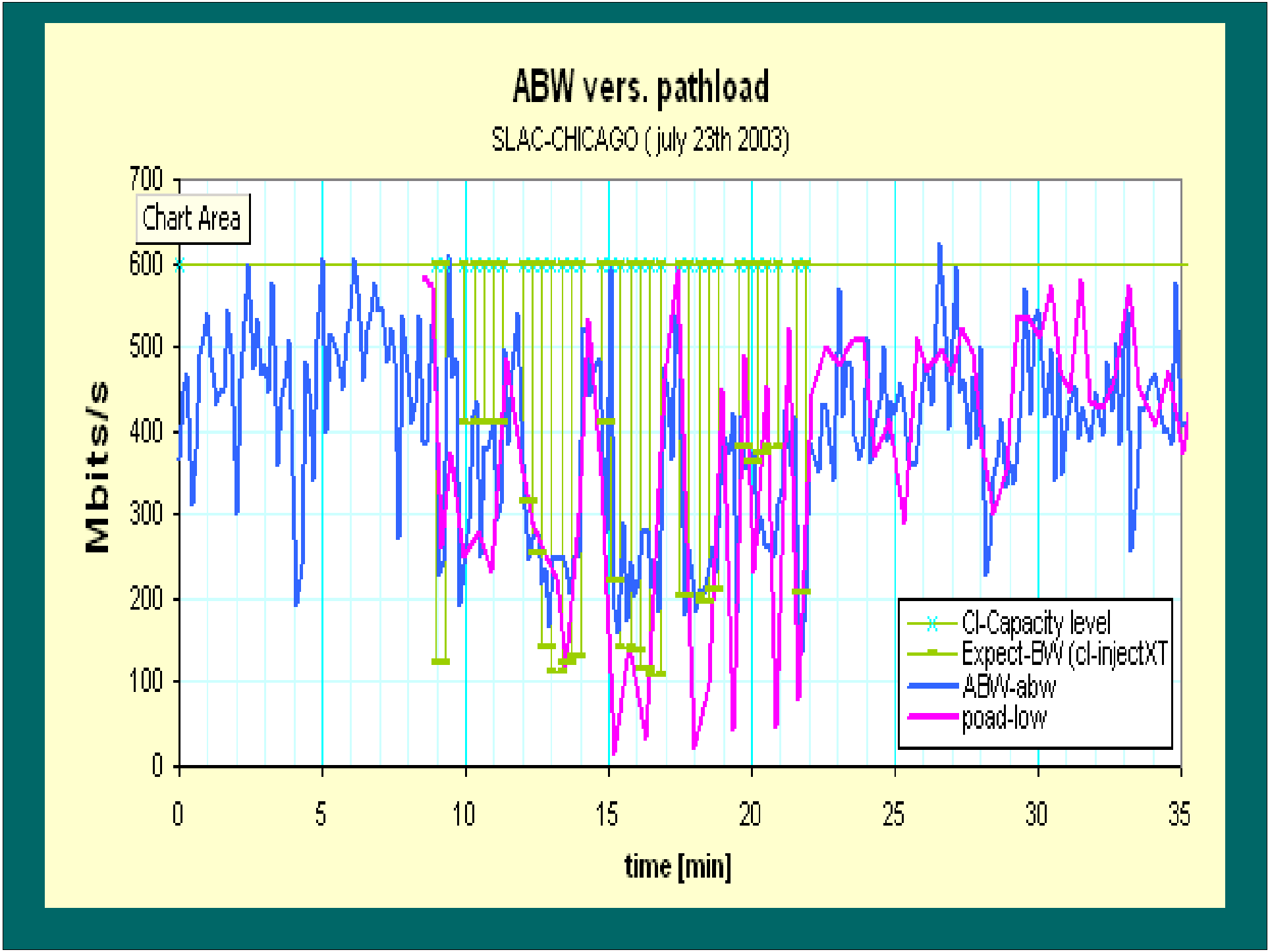
Zoom



Conlusion: Iperf measure own performance which can approach DBC (in best case)

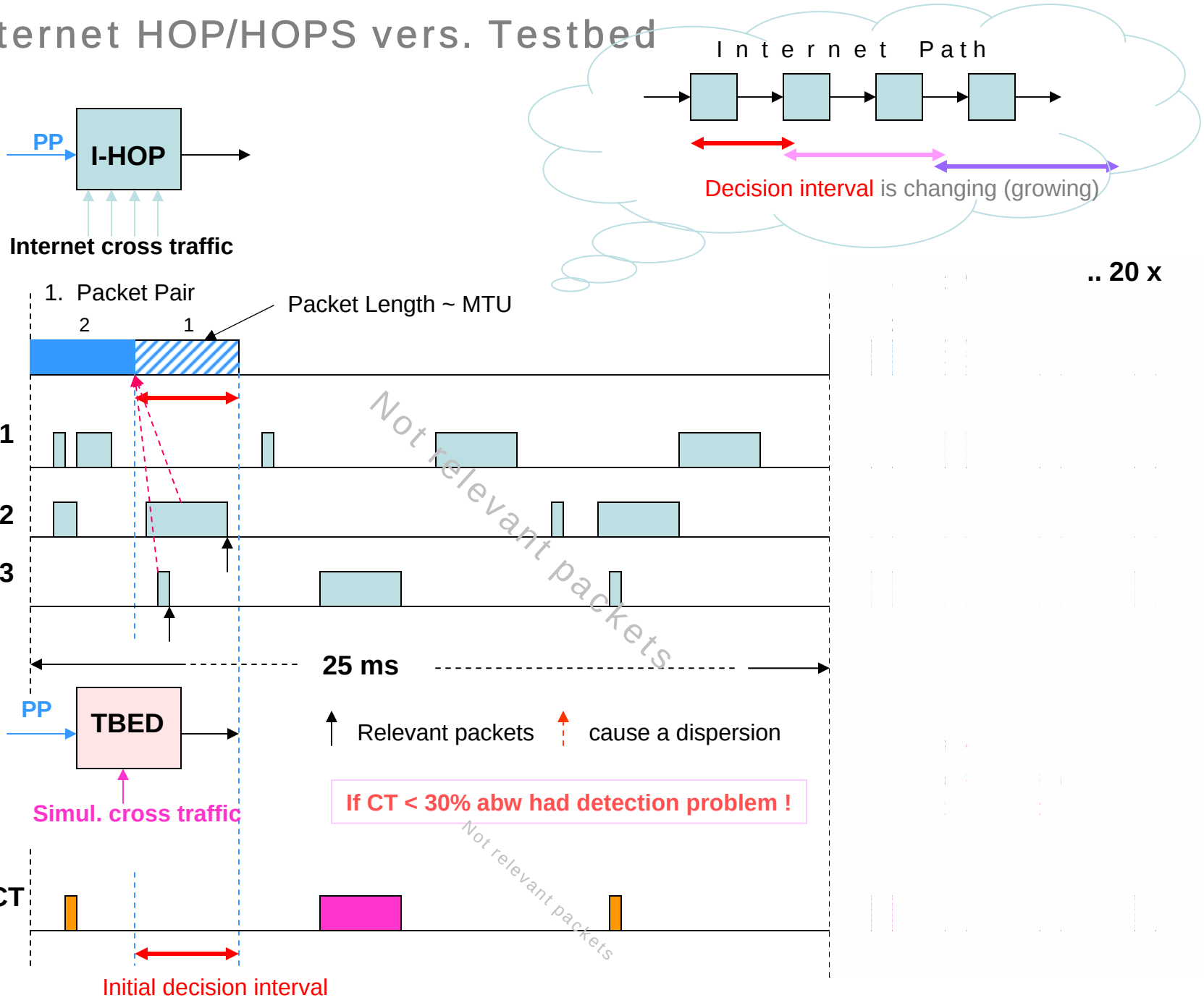

ABW vers. pathload
SLAC-CHICAGO (july 23th 2003)
Chart Area
Mbits/s
700
600
500
400
300
200
100
0
0
5
10
15
20
25
30
35
time [min]
CI-Capacity level
Expect-BW (cl-injectXT
ABW-abw
pload-high


ABW vers. pathload
SLAC-CHICAGO (july 23th 2003)
Chart Area
Mbits/s
700
600
500
400
300
200
100
0
0
5
10
15
20
25
30
35
time [min]
CI-Capacity level
Expect-BW (cl-injectXT
ABW-abw
poad-low

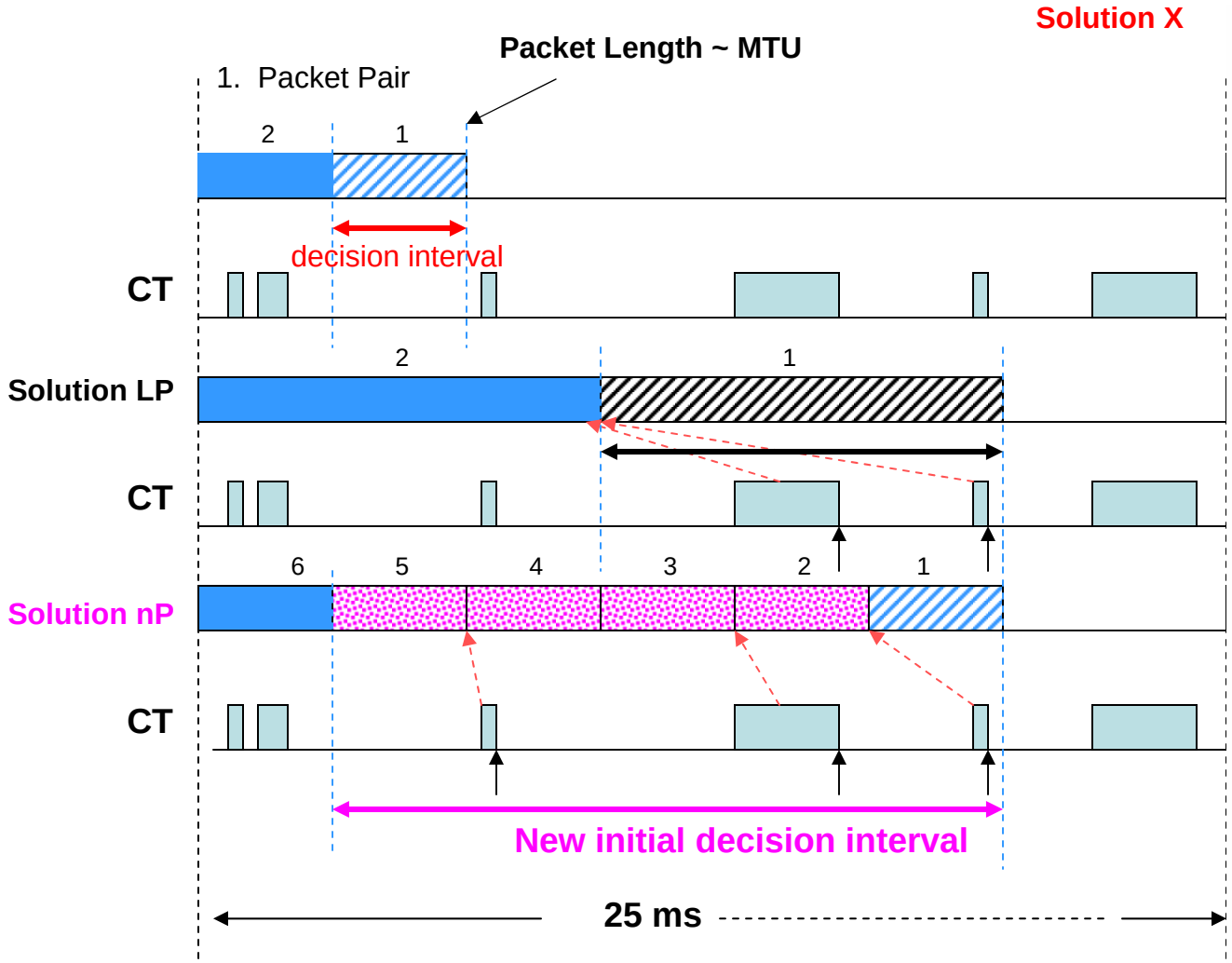

ABW vers. pathload
SLAC-CHICAGO (july 23th 2003)
Chart Area
Mbits/s
700
600
500
400
300
200
100
0
0
5
10
15
20
25
30
35
time [min]
CI-Capacity level
Expect-BW (cl-injectXT
ABW-abw
poad-low

What we learned from CAIDA testbed

Internet HOP/HOPS vers. Testbed



How to improve “detection effectiveness”

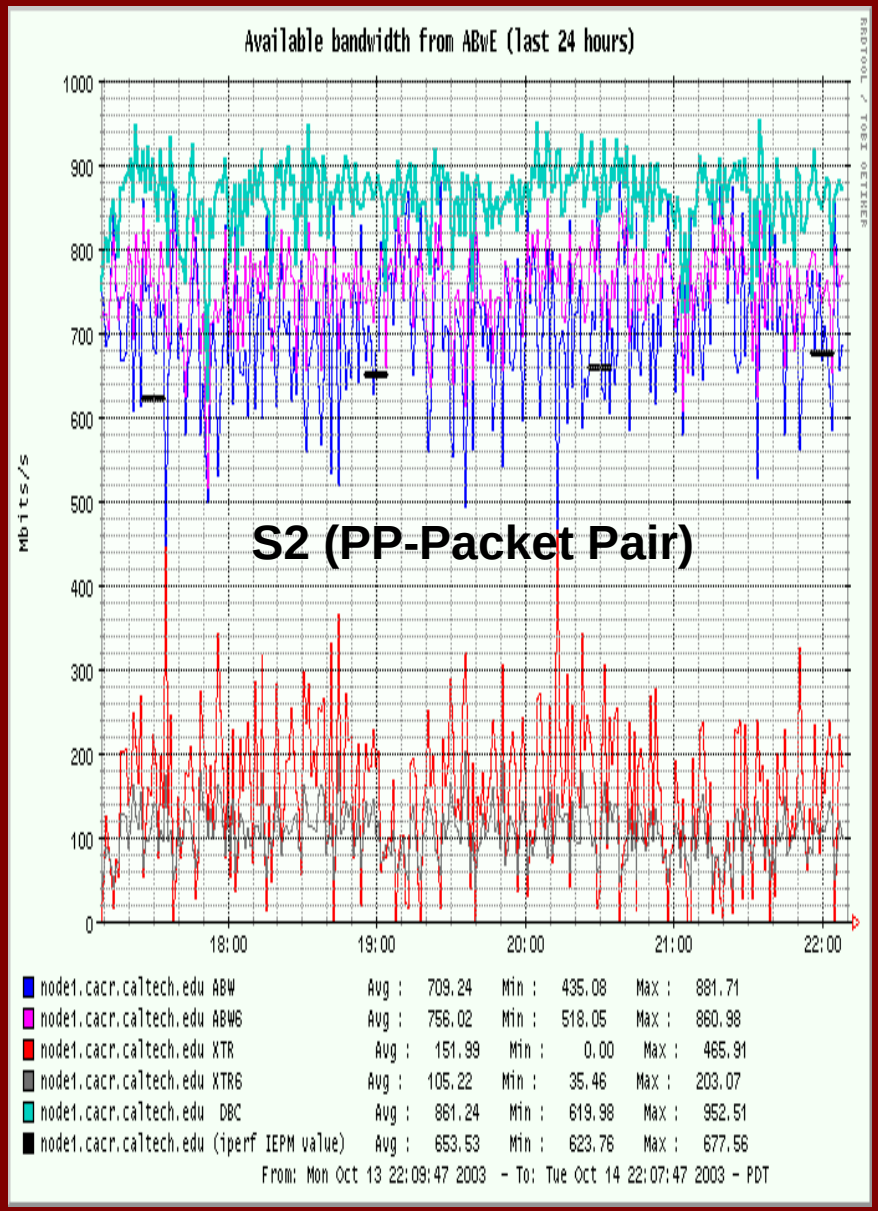


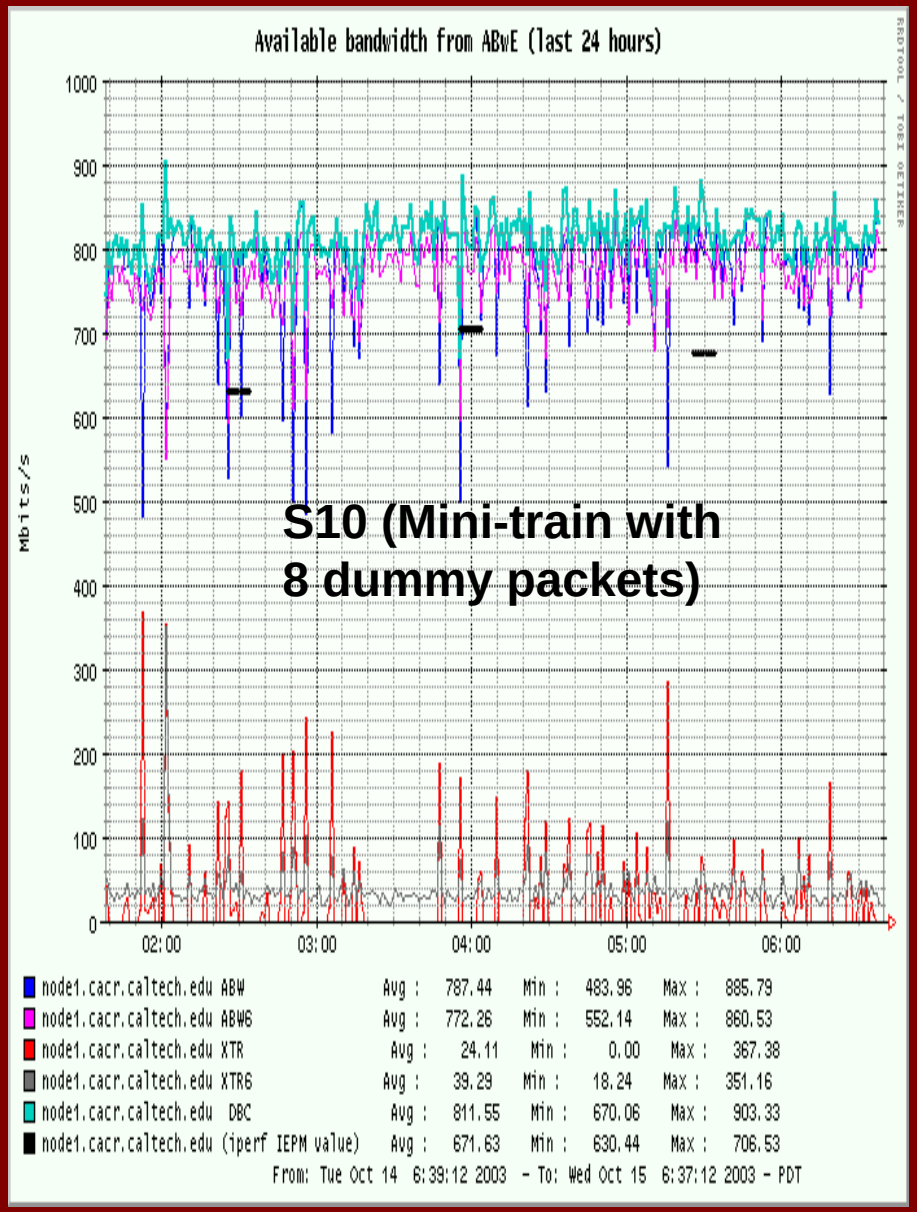
Solution LP – Long packets (9k) (creates micro-bottlenecks)

↑ Relevant packets ↑ cause a dispersion

Solution nP – n dummy Packets (mini-train)

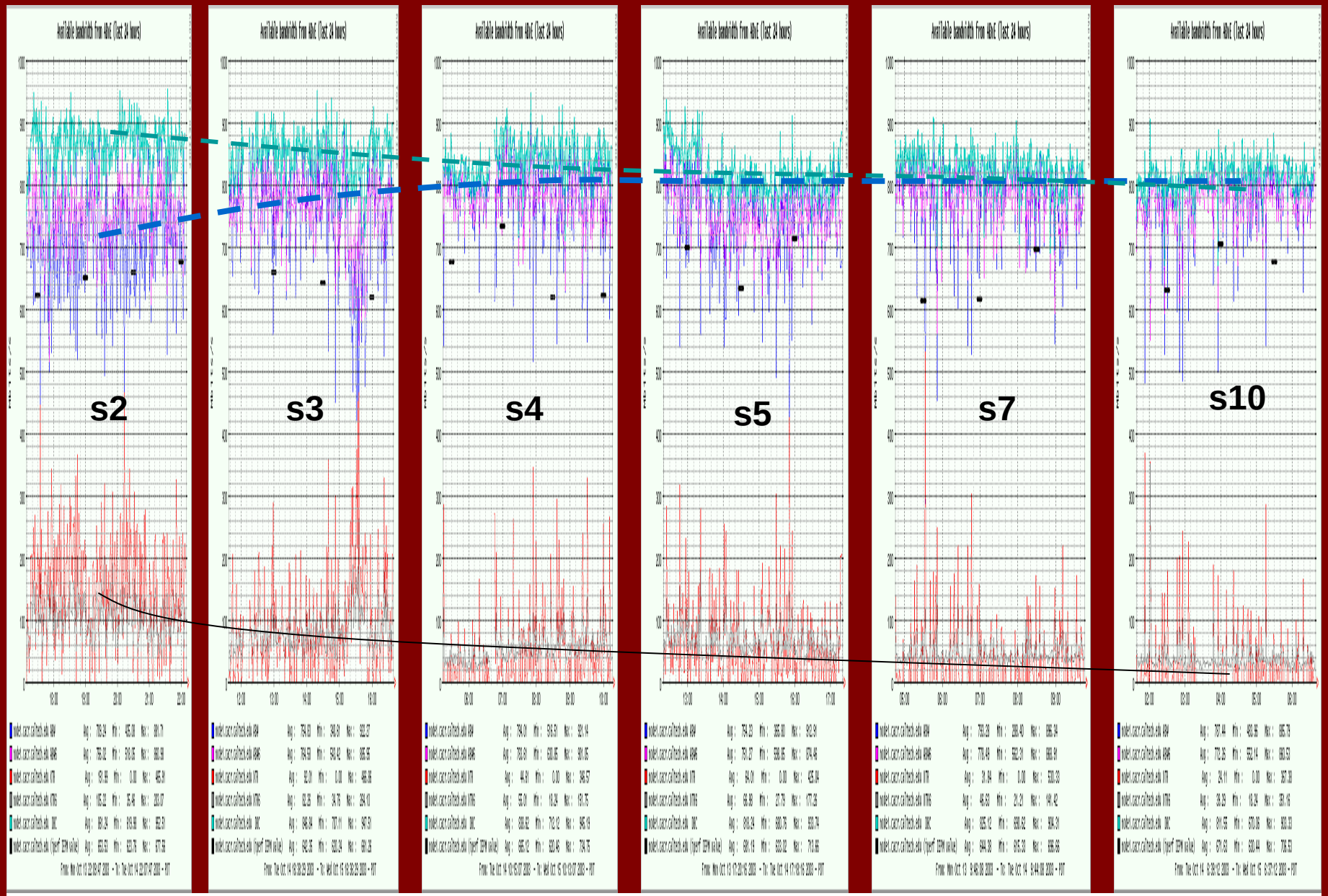
Measurement time
0.5 s to 2.5 s





PP versus TRAIN: ABW and DBC *merge* in TRAIN samples

(SLAC-CALTECH path)

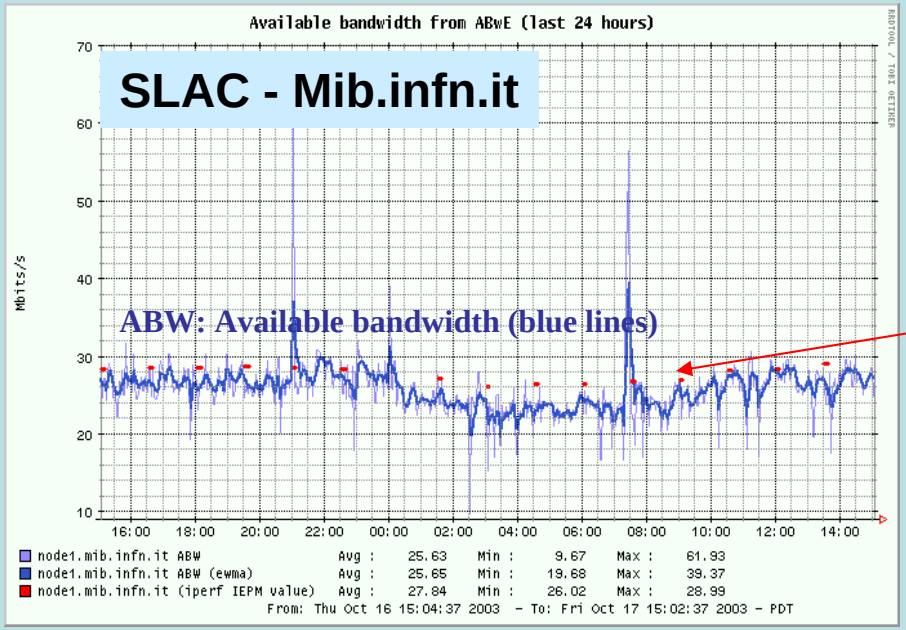


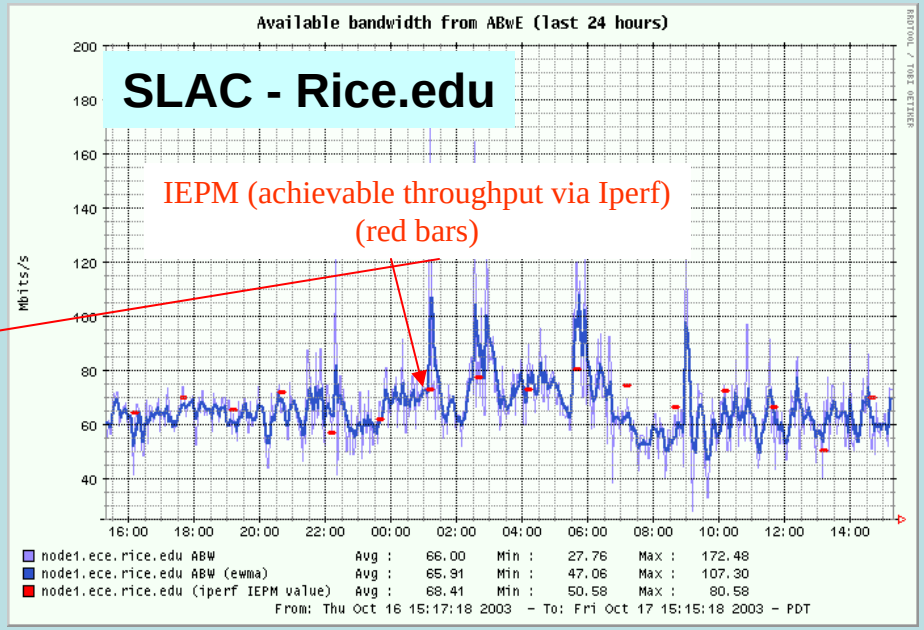
PP versus TRAIN: ABW and DBC *merge* in TRAIN samples

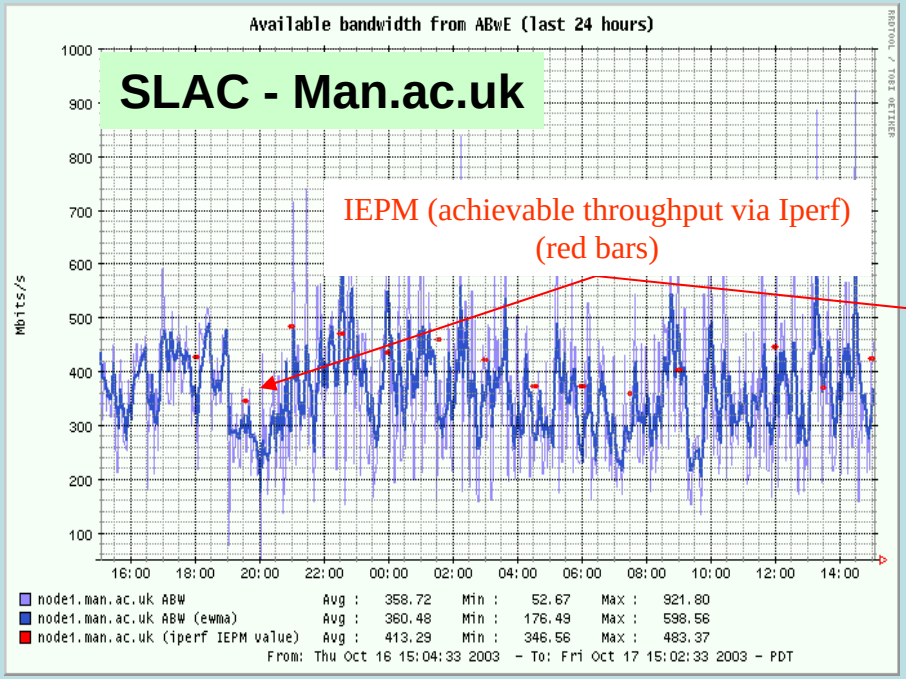
(SLAC-CALTECH path)

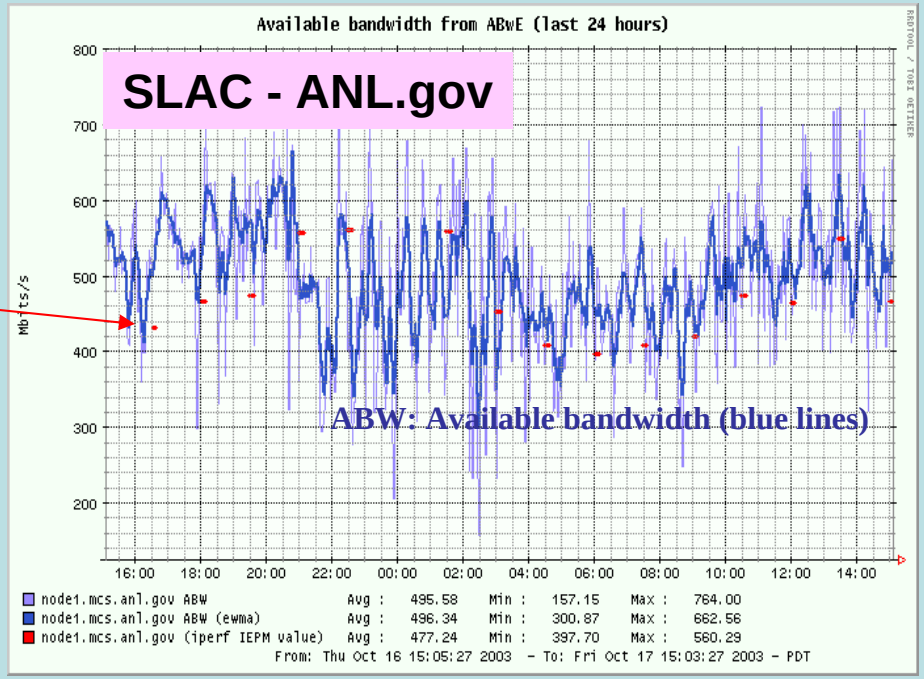
Compare long term Bandwidth statistics on real paths

ESNET, Abilene, Europe

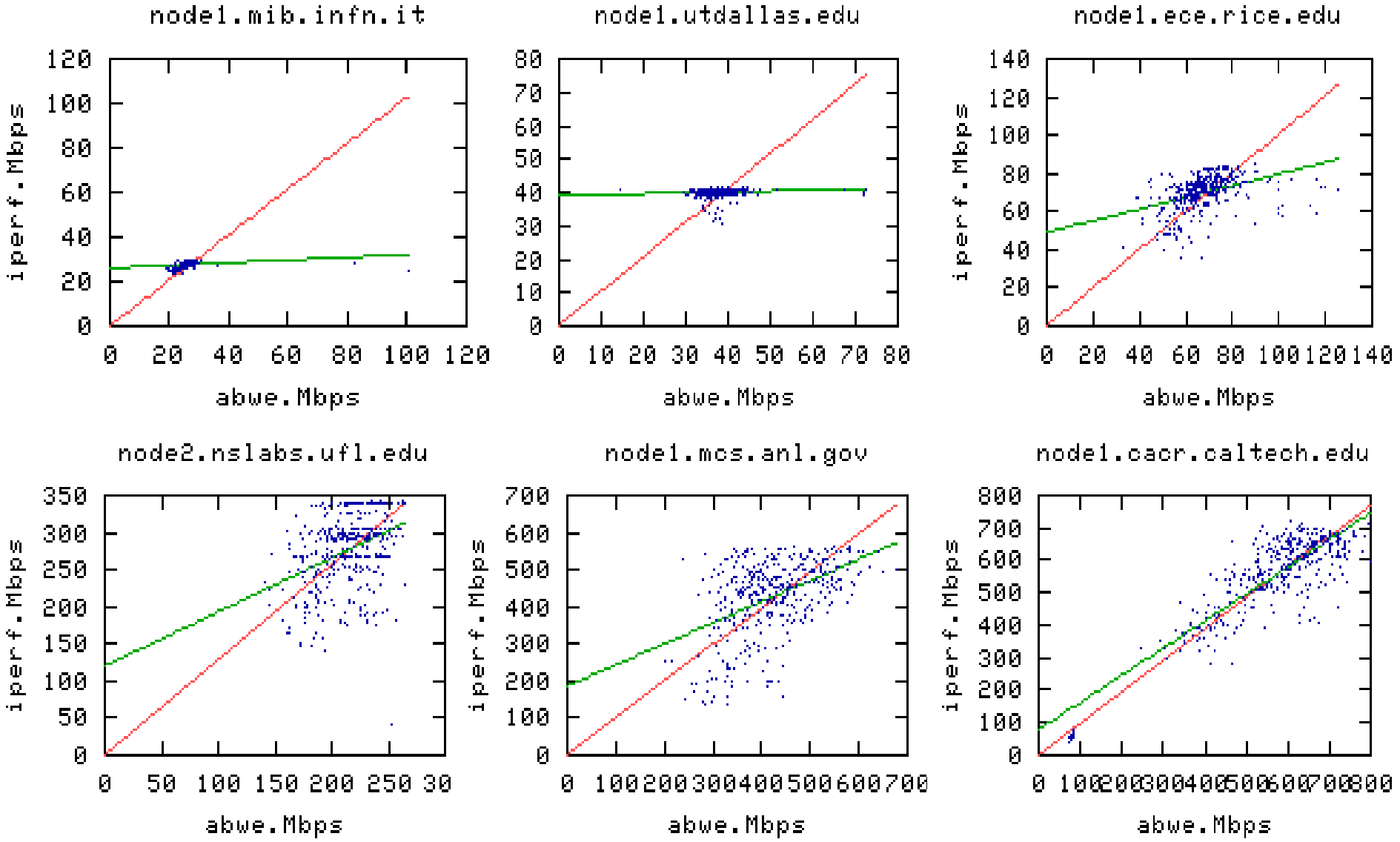








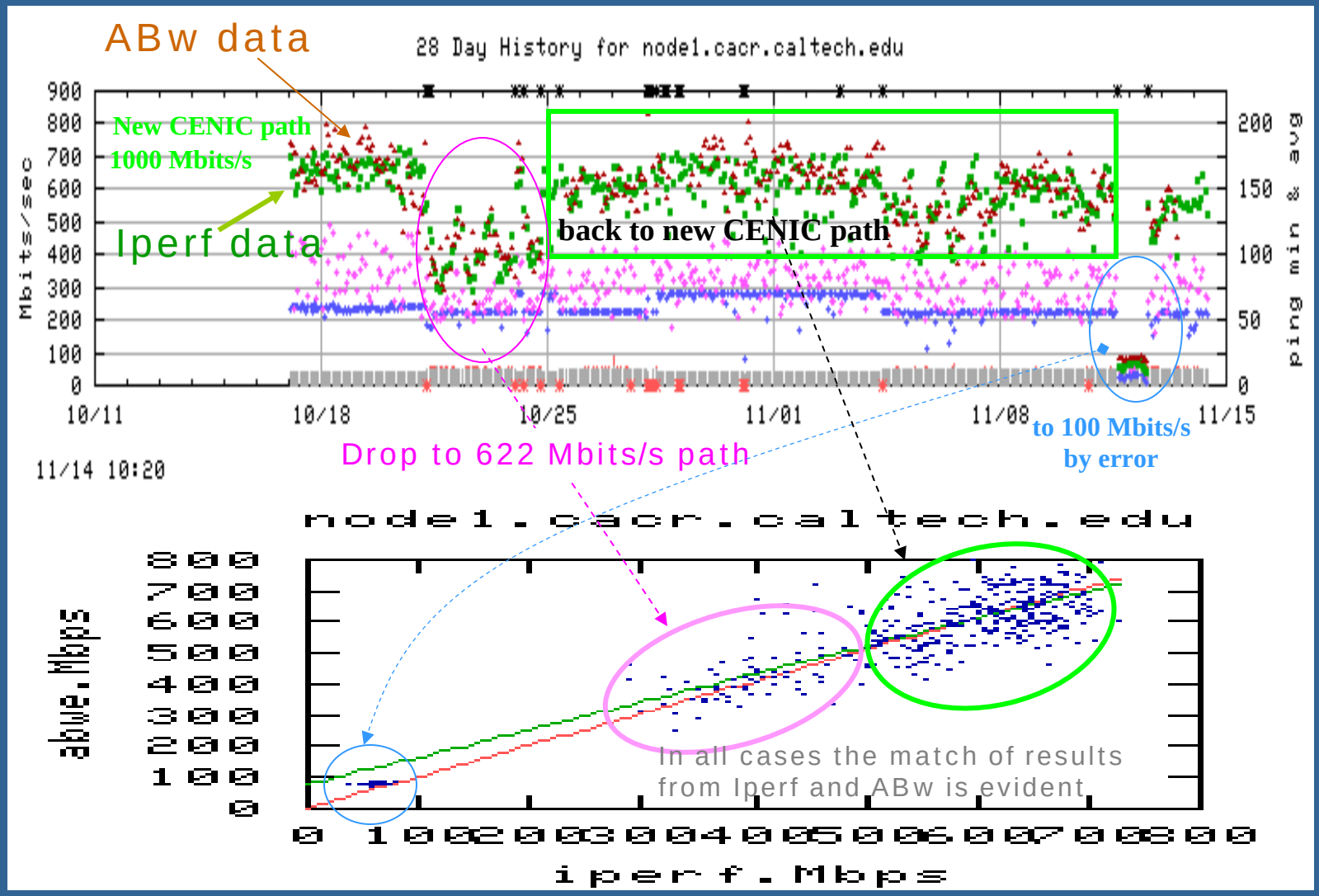
IEPM-Iperf vers. ABW (24 hours match)



Scatter plot graphs

Achievable throughput via Iperf versus ABw on different paths (range 20–800 Mbits/s)

(28 days history)



28 days bandwidth history

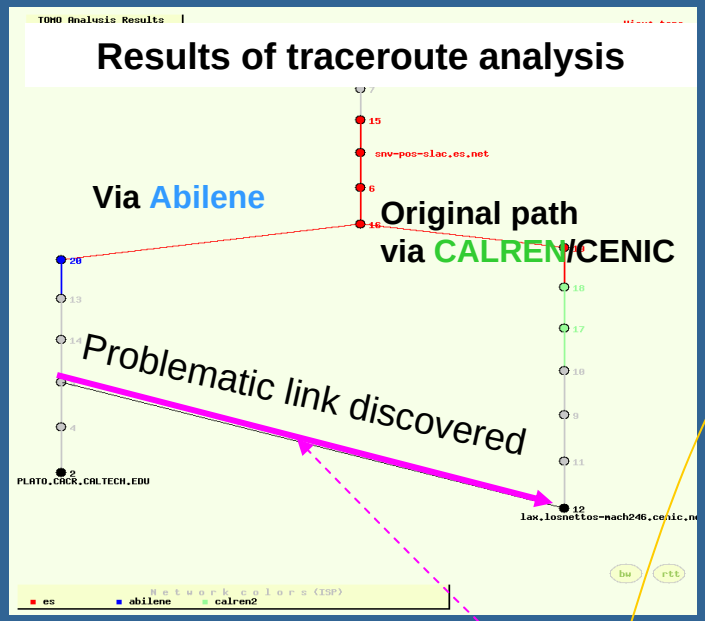
During this time we can see several different situations caused by different routing from SLAC to CALTECH

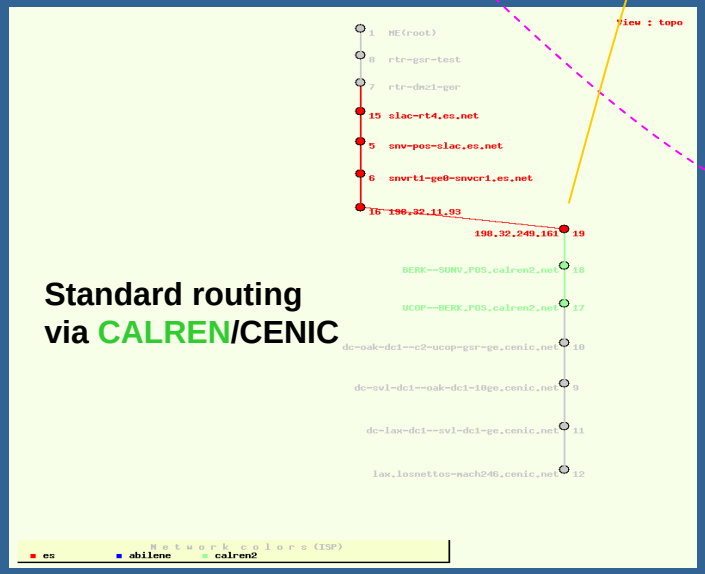
What we can detect with continues bandwidth monitoring

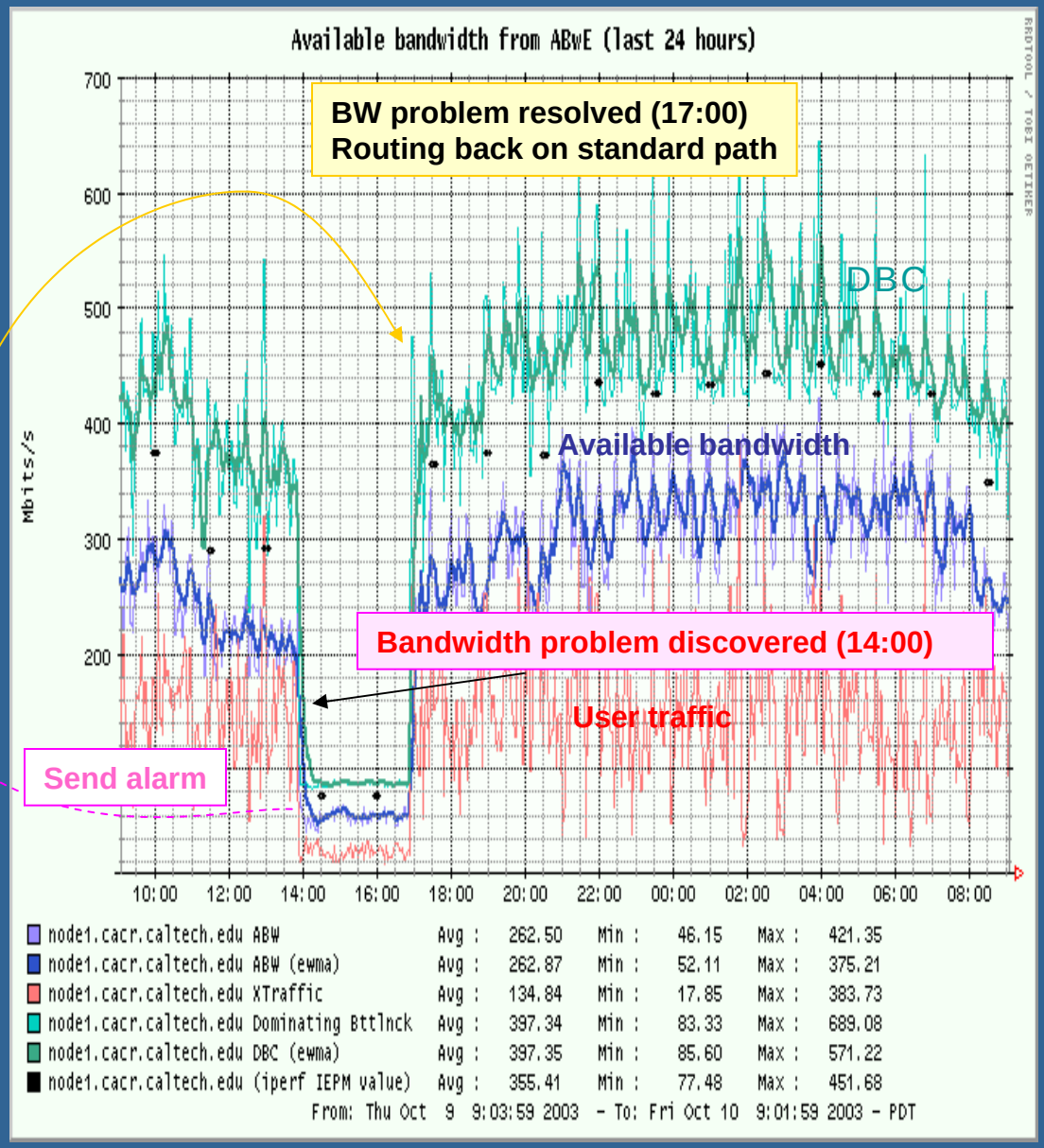
- Immediate bandwidth on the path
- Automatic routing changes when line is broken (move to backup lines)
- Unexpected Network changes (Routing changes between networks, etc.)
- Line updates (155 -> 1Giga, etc.)
- Extreme heavy load

ABw as Troubleshooting tool

(Discovering Routing problems and initiate alarming)



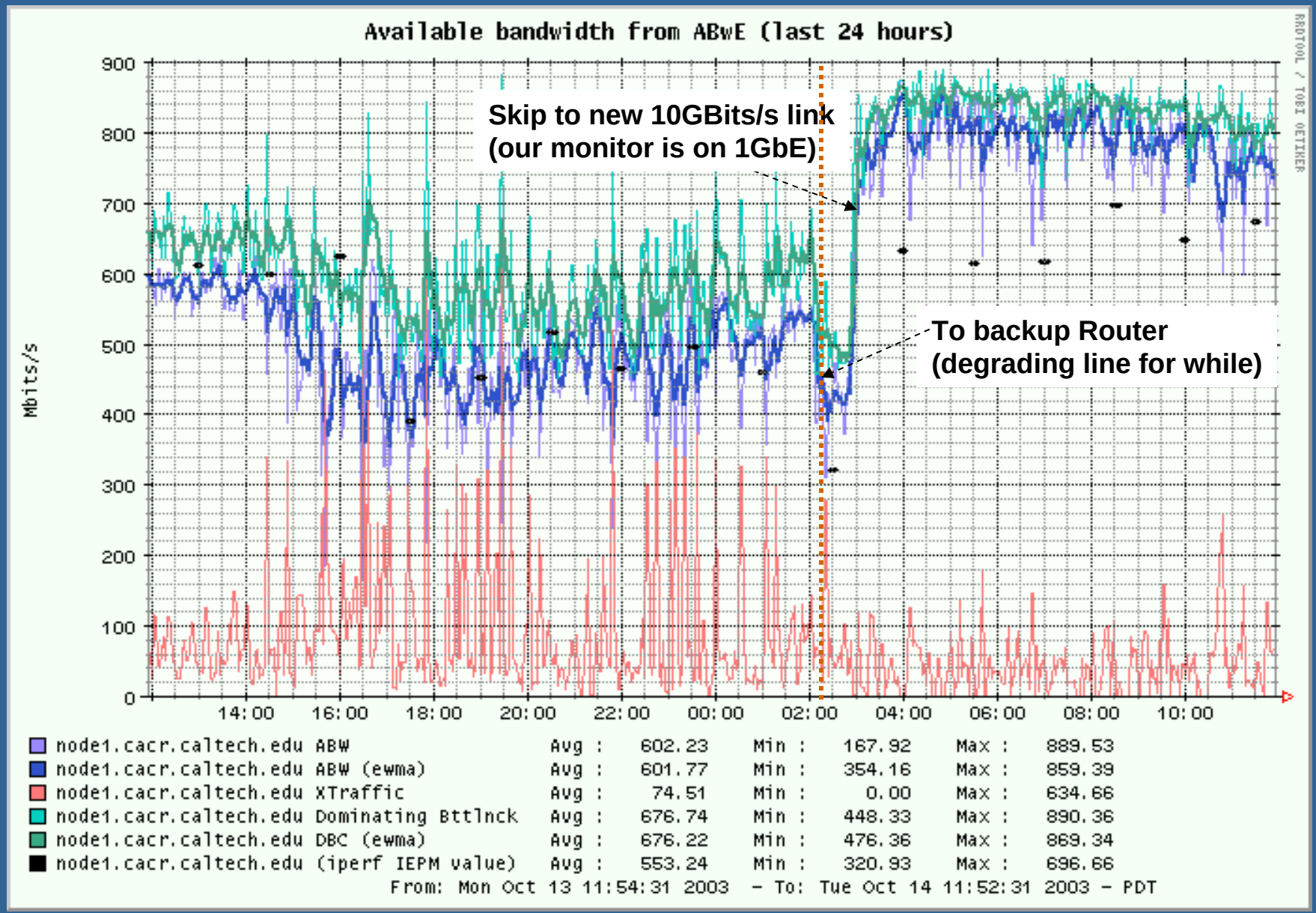




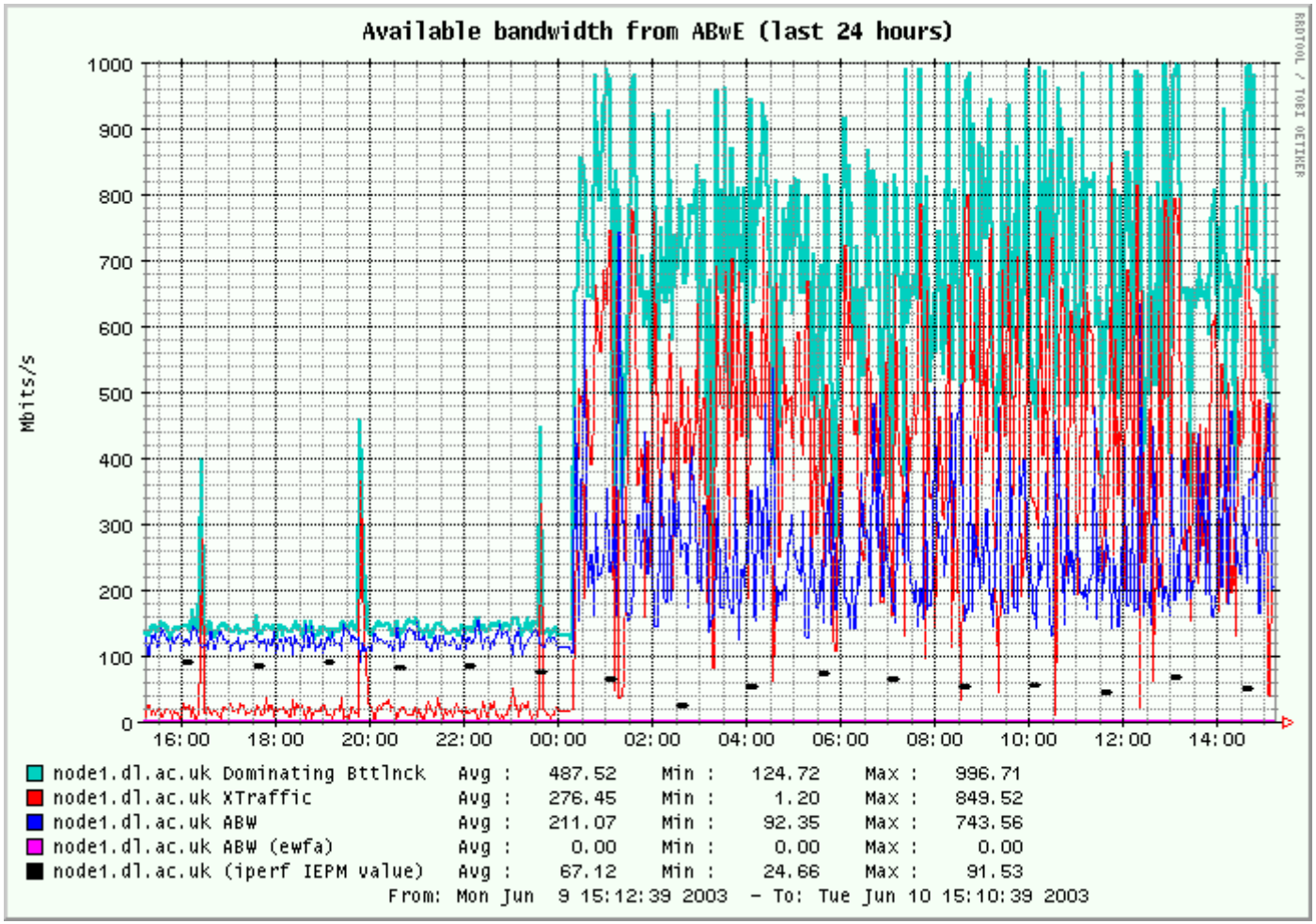
(Example from SLAC – CENIC path)

SLAC – CENIC path upgrade from 1 to 10 Gigabit

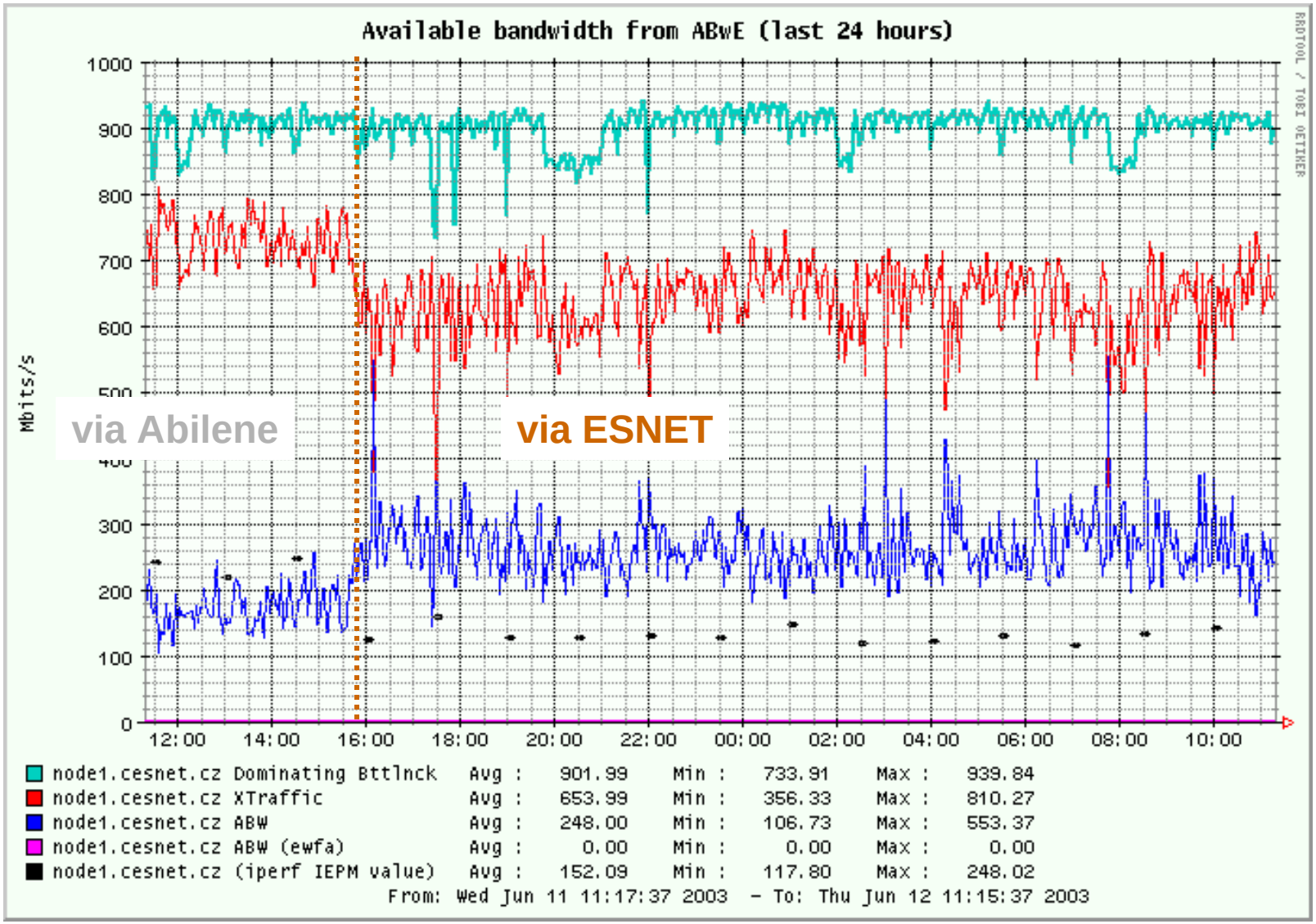
(Current monitoring machines allow monitor traffic in range 1 < 1000 Mbits only)

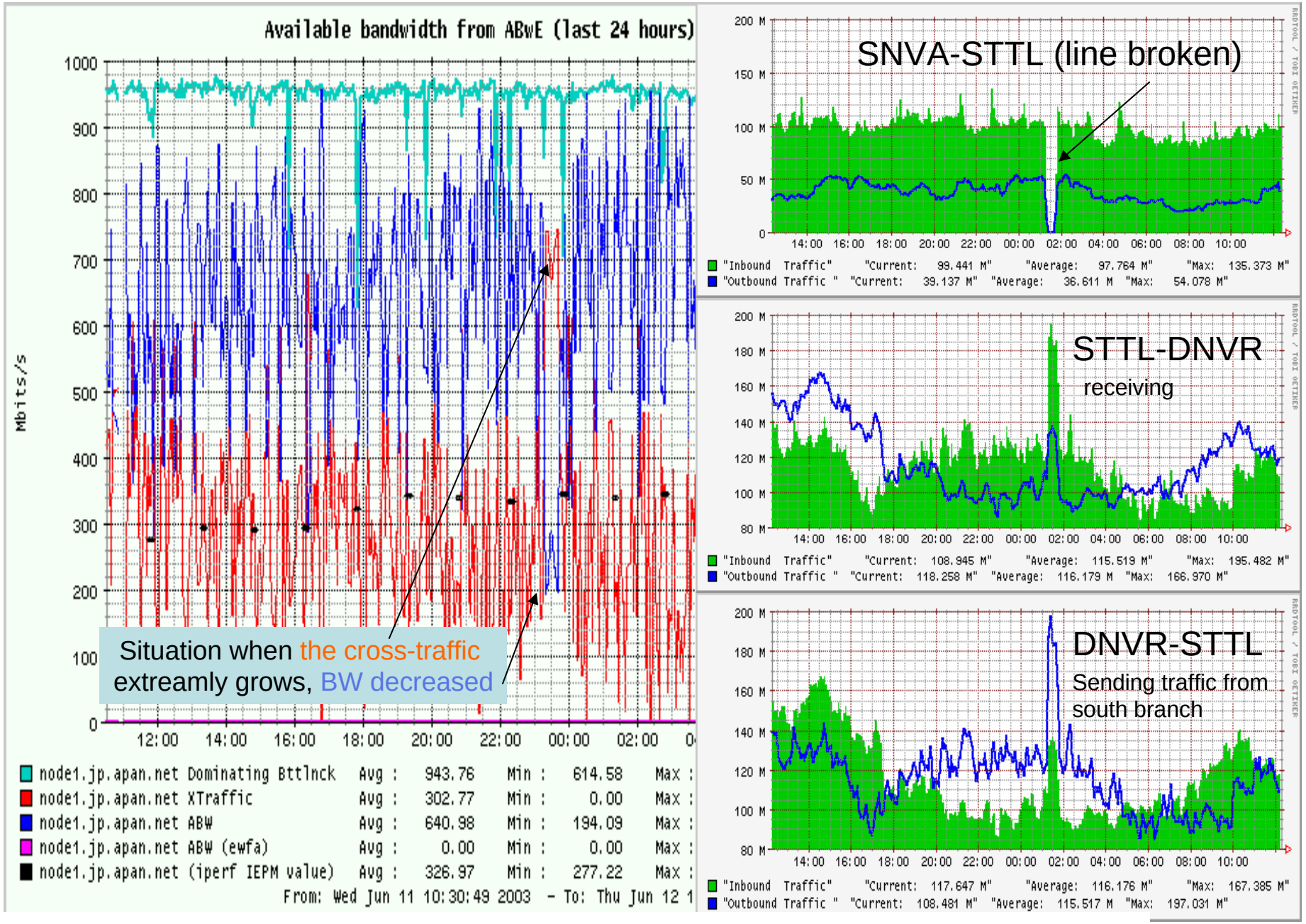


Upgrade 155Mbits/s line to 1000Mbits/s at dl.uk



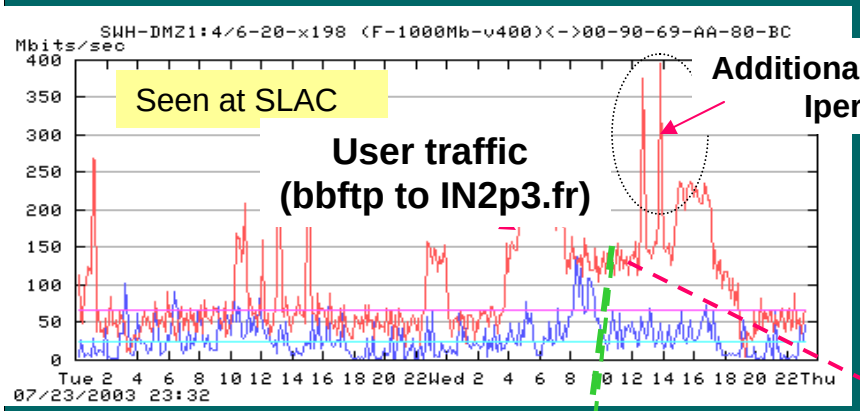
SLAC changed routing to CESNET



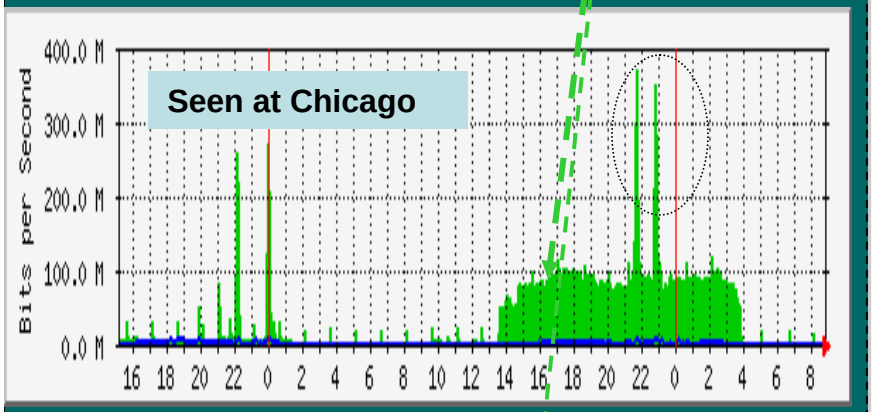


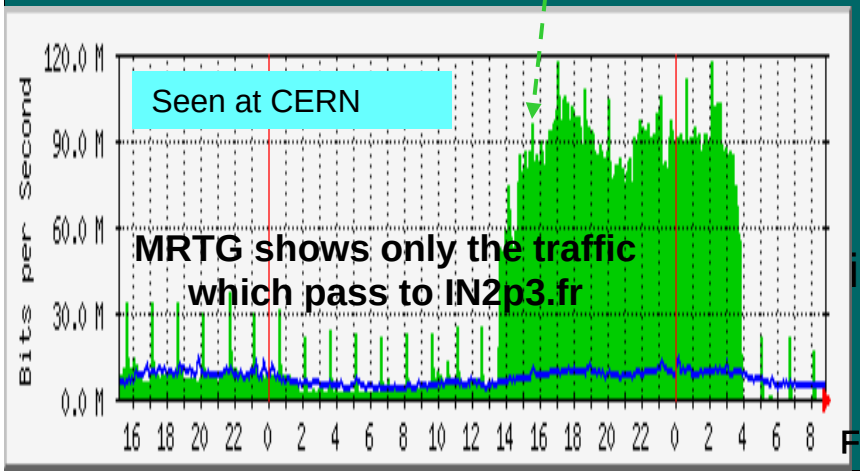
Abilene – automatic rerouting – June 11,2003

Typical SLAC traffic (long data transfer when physical experiment ends)



Additional traffic
Iperf





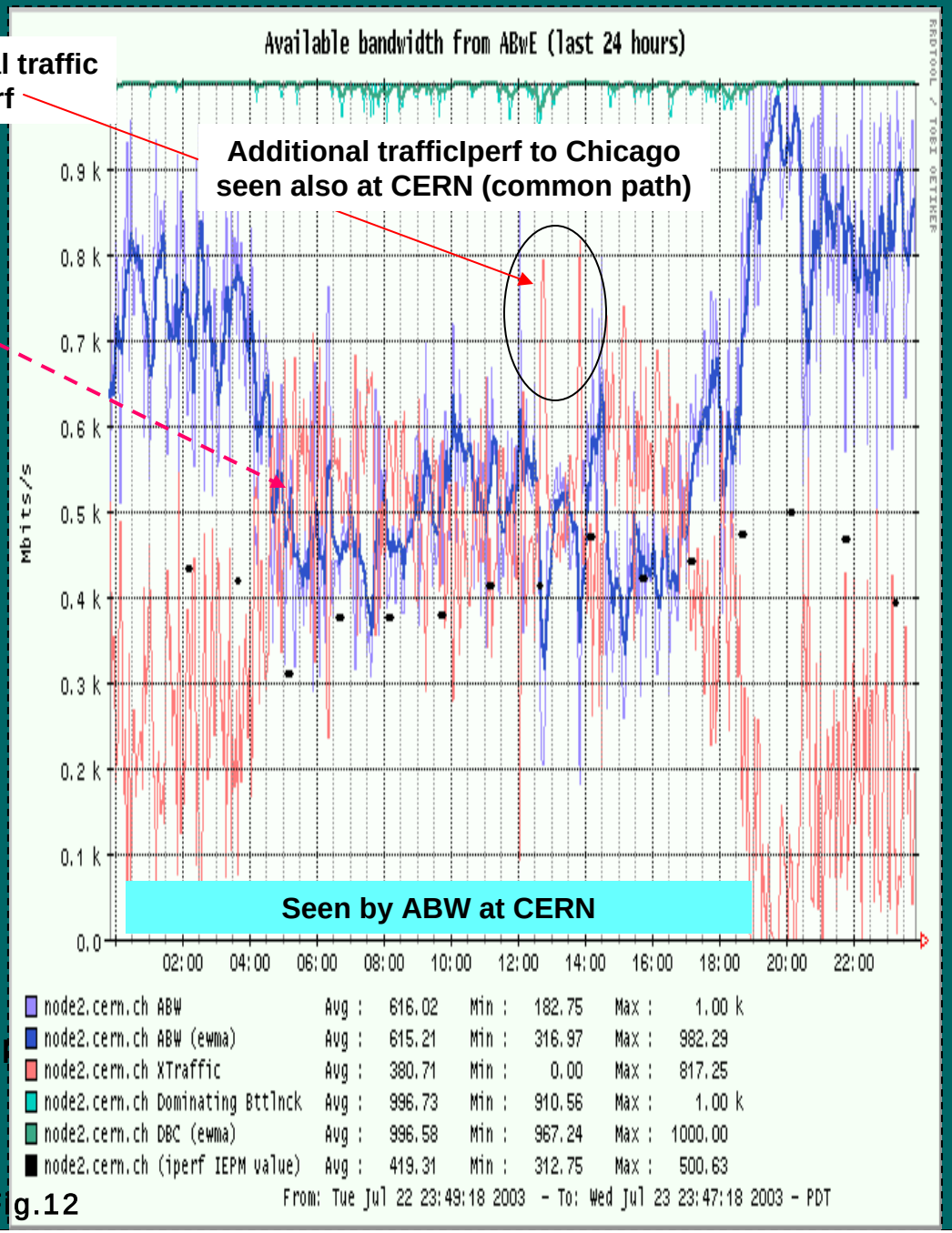


Fig.12

Abing new ABwE tool

- Interactive (reply < 1 second)
- Very low impact on the network traffic (40 packets to get value for destination)
- Simple and robust (responder can be installed on any machine on the network)
- Keyword function for protecting the client-server communication
- Measurements in both directions
- Same resolution as other similar methods

http://www-iepm.slac.stanford.edu/tools/abing

Thank you

References:

http://moat.nlanr.net/PAM2003/PAM2003papers/3781.pdf

http://www-iepm.slac.stanford.edu/tools/abing