NLANR, Internet2, and End-to-End performance

Scot Colburn

colburn@ucar.edu

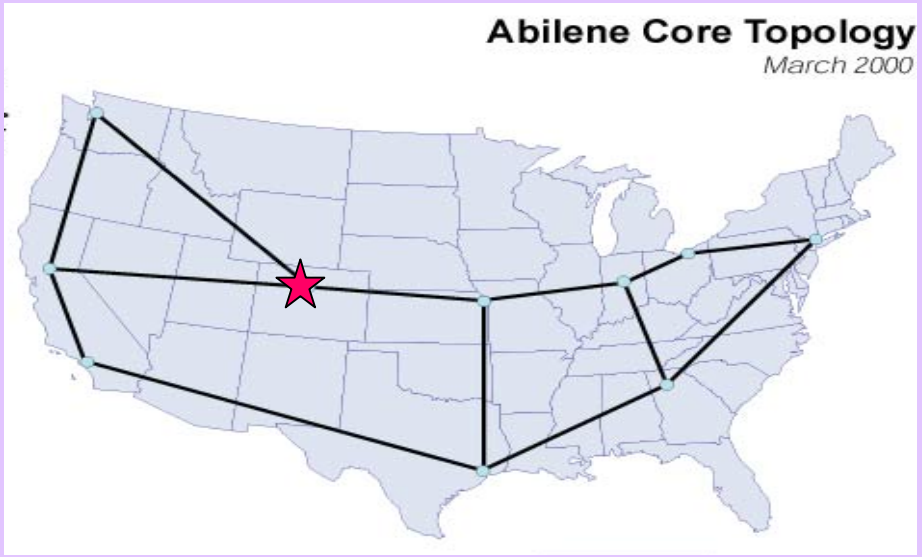
National Center for Atmospheric Research

May 21-22 2001

Florianópolis, Brasil

Who is NCAR?

- Operated by University Corporation for Atmospheric Research (UCAR)
- Center for Climate & Meteorology research
  - Global Climate Model
  - Supercomputing
- History of Internet leadership
  - One of first nodes on ARPANET
  - Operate Internet2 GigaPop



Why pursue End to End Performance?

- Give users high performance networking
- Enable interactive "Collaboratories" - "Collaborative laboratory"
- Enable "Access-grid" and "Earth-systems Grid"
- Move large scientific data-sets
- Provide good access to central resources

Who is solving the problem?

- **NLANR - National Laboratory for Applied Networking Research**
  - **http://www.nlanr.net**
- **Internet2 End-to-End Performance Initiative**
  - **http://www.internet2.edu/e2eperf/**
  - **Web100 to move data fast**

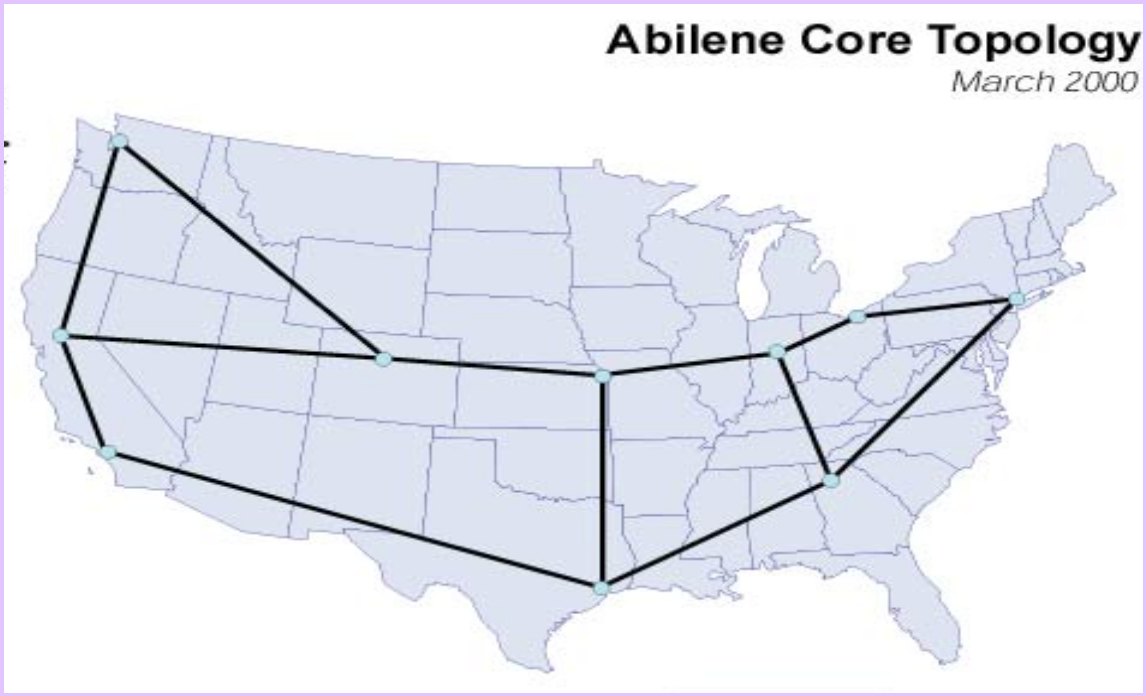
Who is NLANR?

- **NLANR - National Laboratory for Applied Network Research**
  - * **Application/User support (UIUC/NCSA)**
    - » **Autobuf, Iperf, nettest**
  - * **Engineering Services (CMU/PSC/NCAR)**
    - » **Web100 , TAAD (Traffic Analysis and Automatic Diagnosis), TCP performance tools**
  - * **Measurement and Analysis(UCSD/SDSC)**
    - » **Passive Monitoring and Analysis (PMA)**
    - » **Active Measurement Program (AMP)**
  - * **Find them at www.nlanr.net**

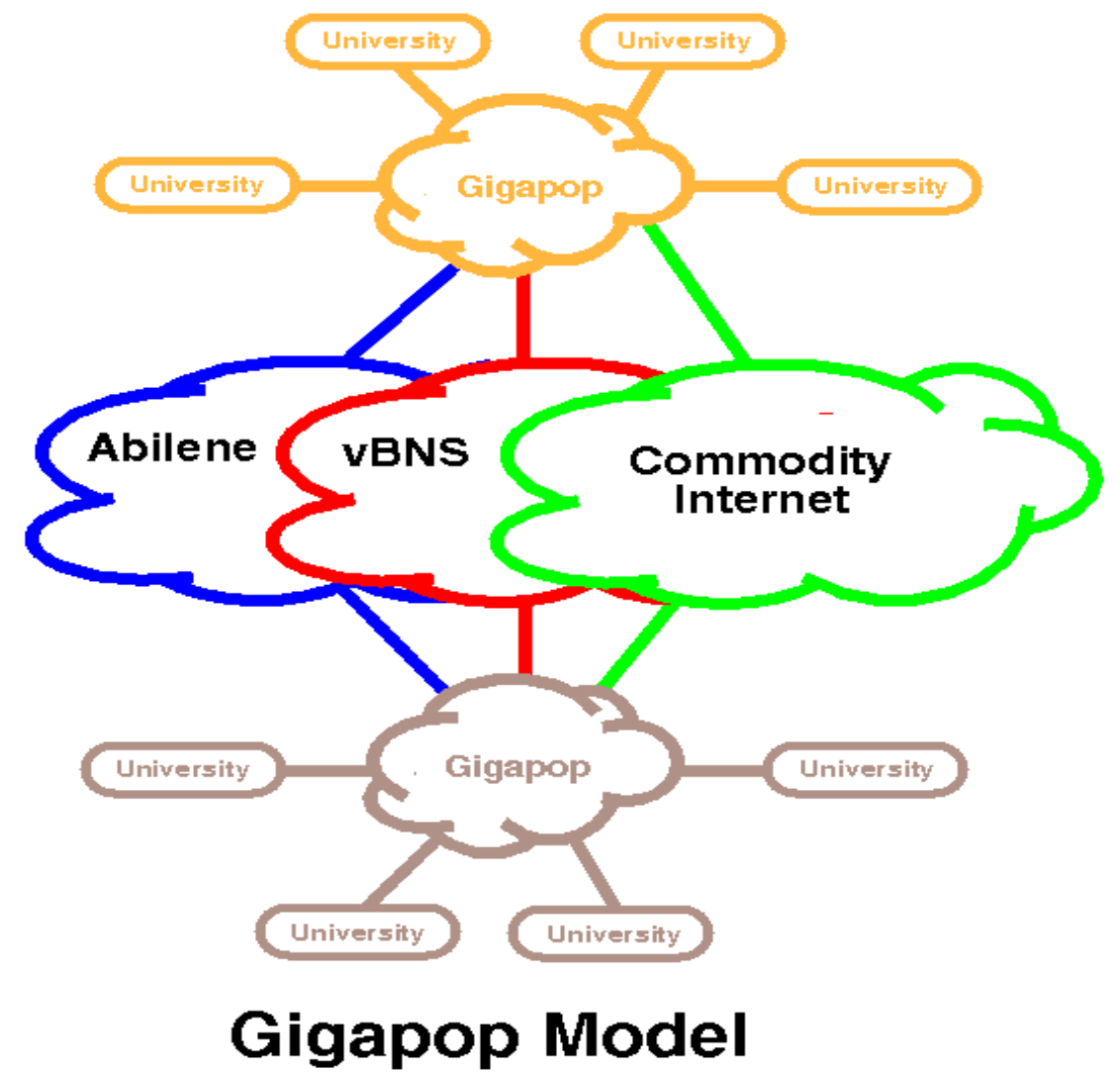
NLANR helps optimize fast networks

- vBNS - very high performance Backbone Network Service
  - built in 1995 by MCI-NSF partnership
  - now vBNS+ is a commercial network
- Abilene, the Internet2 backbone
  - Launched in 1999
  - 180 universities, OC-48 backbone
  - Built in partnership with Cisco and Nortel and Qwest

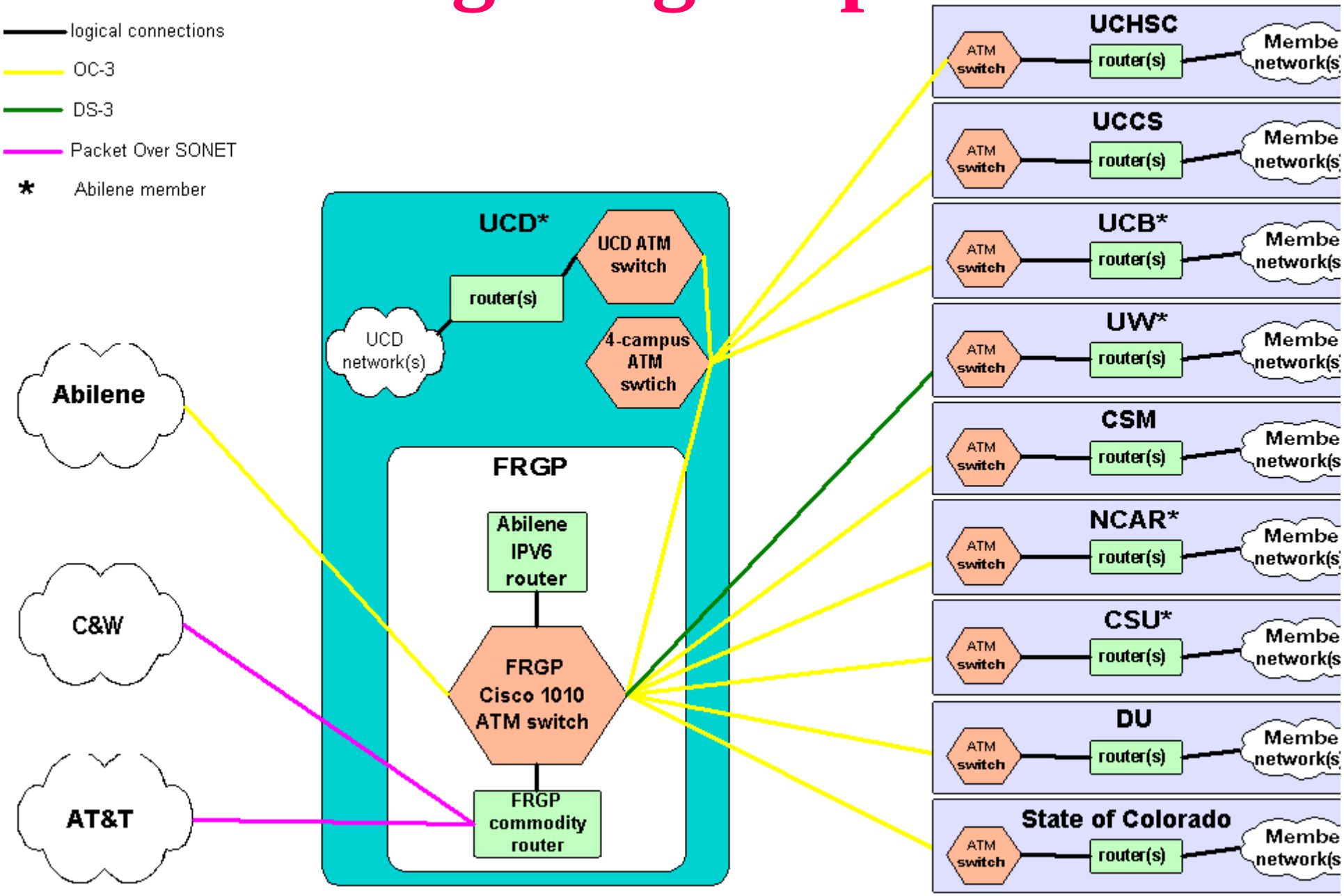
Abilene Network



Networks link Gigapops



Front Range GigaPop



Who is Internet2?

- Universities doing advanced networking research.
- Groups researching QoS, IPv6, Multicast, Measurement, Routing, Security, Topology.
- Network Operations Center (NOC) run by Indiana University.

Cool Abilene/I2 URLs

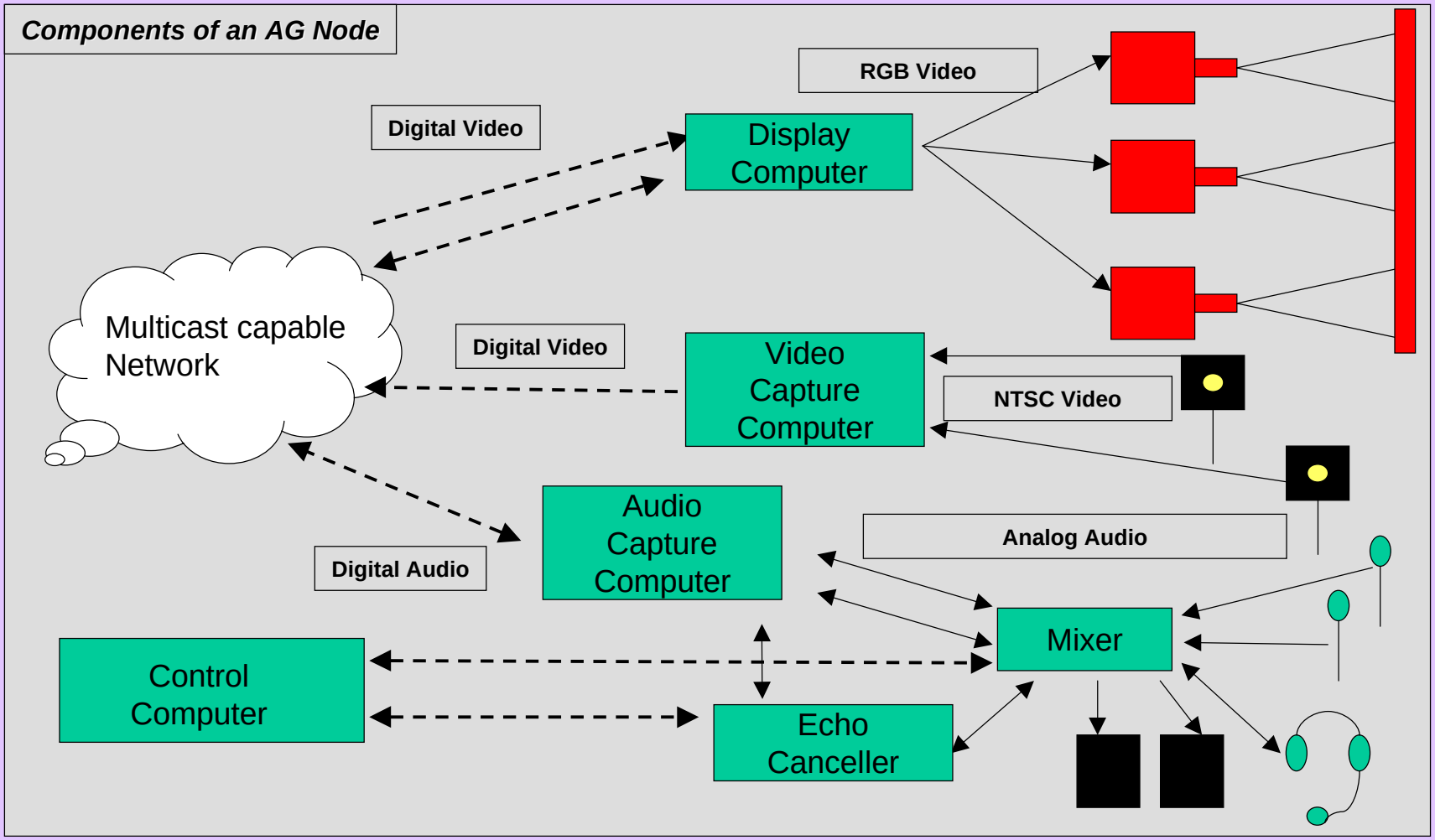
- **Abilene website**
  - **http://www.ucaid.edu/abilene/**
- **Internet2 website**
  - **http://www.internet2.edu/**
- **Abilene NOC at Indiana University**
  - **http://www.abilene.iu.edu/index.cgi**
- **“Live” Internet2 traffic map**
  - **http://hydra.uits.iu.edu/~abilene/traffic**

Fast Networks Enable Cool Applications : Access-Grid

- 

http://www-fp.mcs.anl.gov/fl/accessgrid/
- Group-to-group communications
- Meeting rooms with high-end audio and visual technology
- Collaborative access to computing facilities
- Developed by Argonne National Laboratories, University of Chicago

Access-Grid Node



Access-Grid installation at Argonne National Lab

Fast Networks enable Earth Systems Grid

- **Accelerated Climate Prediction Initiative (ACPI) needs easy access to huge data sets**
- **Earth Systems Grid (ESG) will provide flexible distributed data analysis and high-speed data transport between climate research centers.**
- **ESG is sponsored by Department of Energy (DOE)**

  http://www.scd.ucar.edu/css/esg/

Earth Systems Grid

Earth Systems Grid built with expertise and existing code from:

- **Distributed-Parallel Storage System (DPSS)**
- **Globus - computation grid**
- **Storage Access Coordination System (STACS)**
- **Program for Climate Model Diagnosis and Intercomparison (PCMDI)**

End-to-End performance problem remains

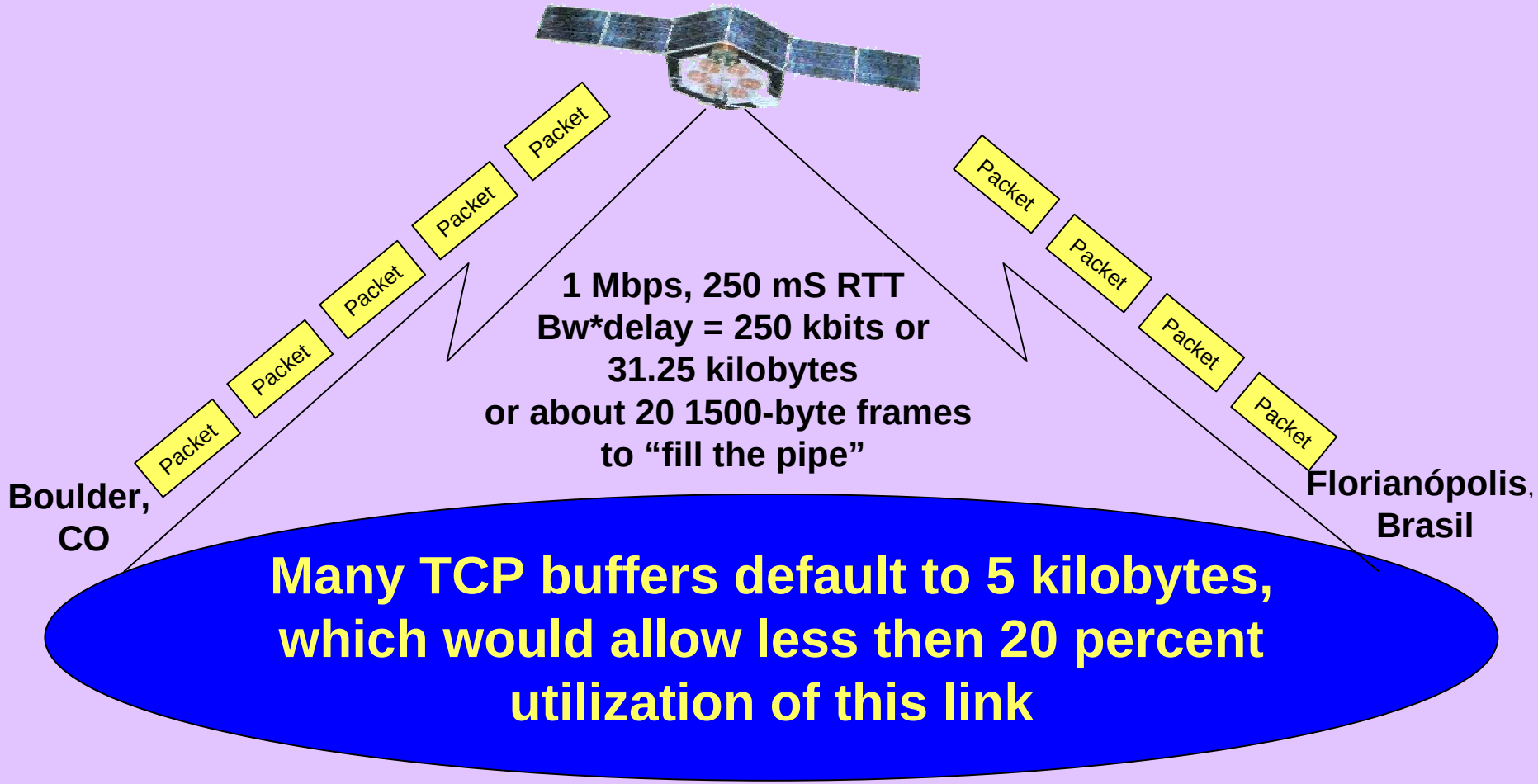
- Despite powerful networks, applications don’t get optimal throughput.
- Many scientists would be happy just to get good FTP performance.
- Problem often lies in the implementation of TCP (Transmission Control Protocol).
- TCP usually doesn’t provide adequate buffer space for high-performance transfers.

How to increase network performance

- Decrease packet loss
- Increase link bandwidth
- Increase Maximum Segment Size
- Decrease Round-Trip-Time

But a fast network is not enough...

Bandwidth-Delay Problem



The Wizard Gap

- **The "Wizard Gap" - the performance difference between well-tuned and non-tuned TCP sessions - has increased from 3:1 to 300:1 in the last ten years.**
- **Today, a well tuned connection may carry 1 Gbps, while a non-tuned connection will get 3 Mbps.**

from Matt Mathis at NLANR/I2 Joint Techs conference,
http://www.psc.edu/~mathis/papers/JTechs200105/index.html

Bandwith-delay solutions

- Manually set TCP buffer size in application
  - tedious
- Manually set TCP buffer size in OS
  - wasteful of memory
- Auto-tune in application
  - application specific

Can’t we automate this?

Web100 moves data fast

- Web100 is a Cisco sponsored, NSF funded partnership between NCAR, NCSA, and PSC.
- Seeks ubiquitous deployment of fast TCP code.
- Instruments the TCP stack.
- Exposes TCP stack values.
- Extends and integrates TCP improvements, primarily auto-tuning of TCP transmit and receive buffer sizes.

The Web100 Solution

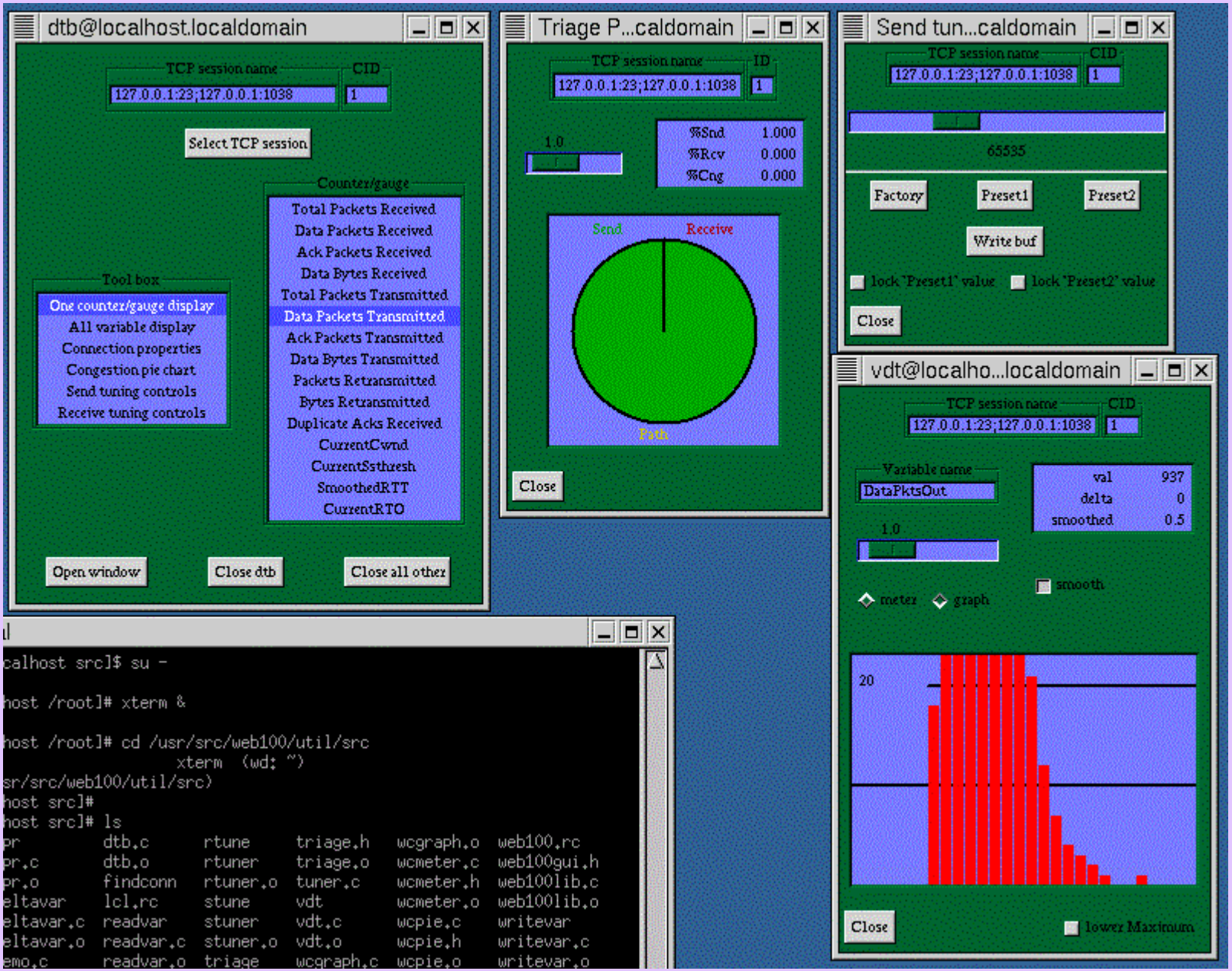
- Implement per-session TCP MIB in kernel
- Similar to UNIX *netstat* information, but more variables and more useful information
- Write-variables will allow user-level TCP-session performance tuning based on real-time congestion feedback from TCP session
- Also allows multitude of user-level display and diagnostic tools regarding TCP behavior during real-time

Web100 Implementation

- ~1,200 diff lines against Linux 2.2.14
- API is through /proc virtual filesystem and/or kernel hooks
- About 25 variables readable
- Counters updated continuously in kernel; /proc updates each time accessed
- One instance of data structure for each TCP session in /proc
- *curses* and gtk demo/example interfaces

Web100 Sample Demo Output

```
128.182.61.238.22 <-> 128.182.61.156.1022                    ESTABLISHED
-------------------------+-------------------------+-------------------------
PktsIn                1974 | PktsOut               1951 | Enabled:
DataPktsIn             972 | DataPktsOut           1002 |   SACK       N
AckPktsIn             1975 | AckPktsOut             949 |   ECN        N
DataBytesIn          19823 | DataBytesOut         74651 |   Timestamps N
DupAcksIn                0 | PktsRetran               0 |
                           | BytesRetran              0 |
-------------------------+-------------------------+-------------------------
loss episodes        0     | cwnd               1453792 | winscale rcvd        0
timeouts             0     | max cwnd           1453792 | rwin rcvd       986816
TO after FR          0     | ssthresh                 0 | max rwin rcvd   986880
                           | min ssthresh             0 | winscale sent        0
                           | max ssthresh             0 | rwin sent        32120
                           |                            | max rwin sent    32120
---------------------+----+-----------------+-------+-------+------------------
rto (ms)           20 | rtt (ms)           1 | mss      1448 | Rate
min rto (ms)       20 | min rtt (ms)       0 | min mss  1448 |  Out (kbps)    0.1
max rto (ms)       20 | max rtt (ms)       1 | max mss  1448 |  In  (kbps)    0.0
---------------------+---------------------+--------------+------------------
Overall rate-controlling effects (only valid if we are the sender):
 aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
 Receiver:(S)topped,(A)pp,(B)ufsize / Path:(C)ongestion / Sender:(b)ufsize,(a)pp
```


dtb@localhost.localdomain
TCP session name
CID
127.0.0.1:23;127.0.0.1:1038
1
Select TCP session
Counter/gauge
Total Packets Received
Data Packets Received
Ack Packets Received
Data Bytes Received
Total Packets Transmitted
Data Packets Transmitted
Ack Packets Transmitted
Data Bytes Transmitted
Packets Retransmitted
Bytes Retransmitted
Duplicate Acks Received
CurrentCwnd
CurrentSsthresh
SmoothedRTT
CurrentRTO
Tool box
One counter/gauge display
All variable display
Connection properties
Congestion pie chart
Send tuning controls
Receive tuning controls
Open window
Close dtb
Close all other
Triage P...caldomain
TCP session name
ID
127.0.0.1:23;127.0.0.1:1038
1
1.0
%Snd 1.000
%Rcv 0.000
%Cng 0.000
Send
Receive
Path
Close
Send tun...caldomain
TCP session name
CID
127.0.0.1:23;127.0.0.1:1038
1
65535
Factory
Preset1
Preset2
Write buf
lock 'Preset1' value
lock 'Preset2' value
Close
vdt@localho...localdomain
TCP session name
CID
127.0.0.1:23;127.0.0.1:1038
1
Variable name
DataPktsOut
val 937
delta 0
smoothed 0.5
1.0
smooth
meter
graph
20
Close
lower Maximum
calhost src]$ su -
host /root]# xterm &
host /root]# cd /usr/src/web100/util/src
xterm (wd: ~)
sr/src/web100/util/src)
host src]#
host src]# ls
pr dtb.c rtune triage.h wcgraph.o web100.rc
pr.c dtb.o rtuner triage.o wcmeter.c web100gui.h
pr.o findconn rtuner.o tuner.c wcmeter.h web100lib.c
eltavar lcl.rc stune vdt wcmeter.o web100lib.o
eltavar.c readvar stuner vdt.c wcpie.c writevar
eltavar.o readvar.c stuner.o vdt.o wcpie.h writevar.c
emo.c readvar.o triage wcgraph.c wcpie.o writevar.o

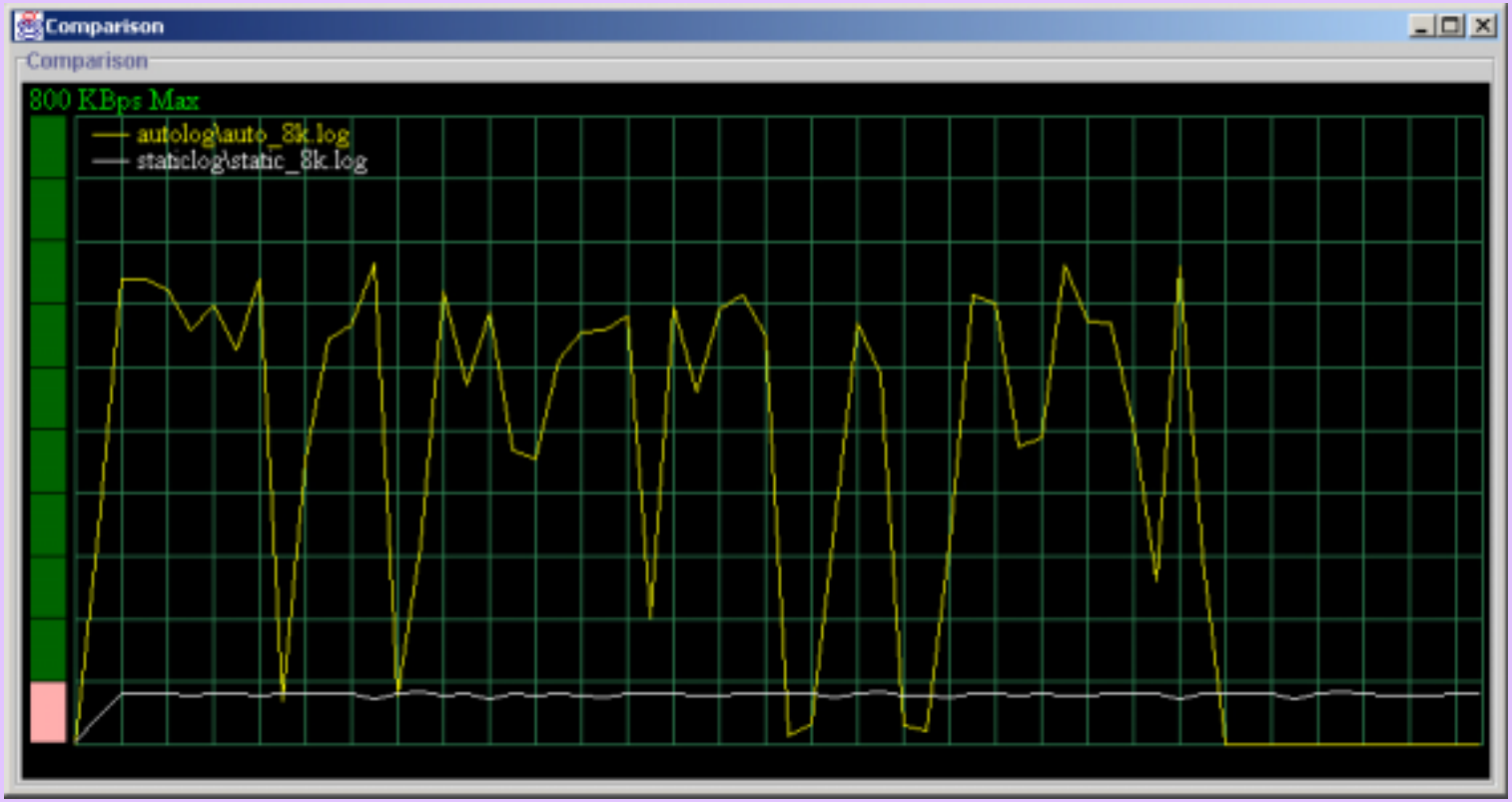
Web100 still under construction

- No public release yet - current version Alpha0.2. Will be open source eventually.
- 16 Alpha testers.
- Expect iterative improvement from user feedback.
- No Autotuning yet, but manual viewing/setting of variables possible.
- Web100 URL: http://www.web100.org

Other NLANR TCP tuning tools

- Available from NLANR DAST - Distributed Applications Support Team
- Autobuf autotuning FTP server and client
  - tests link with ICMP before file transfer
  - modified NcFTP client, WuFTP server
- Iperf Internet Performance tester
  - Version 1.2 recommends TCP window sizes
  - TCP and “raw” UDP transmission tests

AutoNcFTP



- 583% performance enhancement
- http://dast.nlanr.net/Features/Autobuf/

Questions? Perguntas?

colburn@ucar.edu