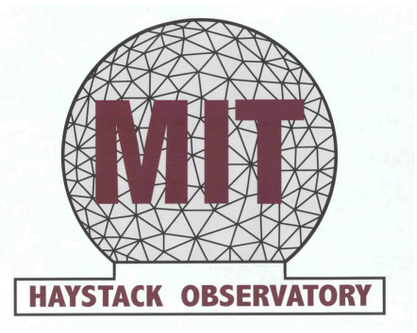
e-VLBI Overview

David Lapsley
dlapsley@haystack.mit.edu



Outline

- Introduction
- Types of networks
- Basic Transmission Protocols
- Current Global Connectivities
- Current Issues
  - ‘last mile’ connectivity
- e-VLBI Development
- VSI-E Initiative

Introduction

- What is e-VLBI?
- e-VLBI architecture
- Examples of e-VLBI

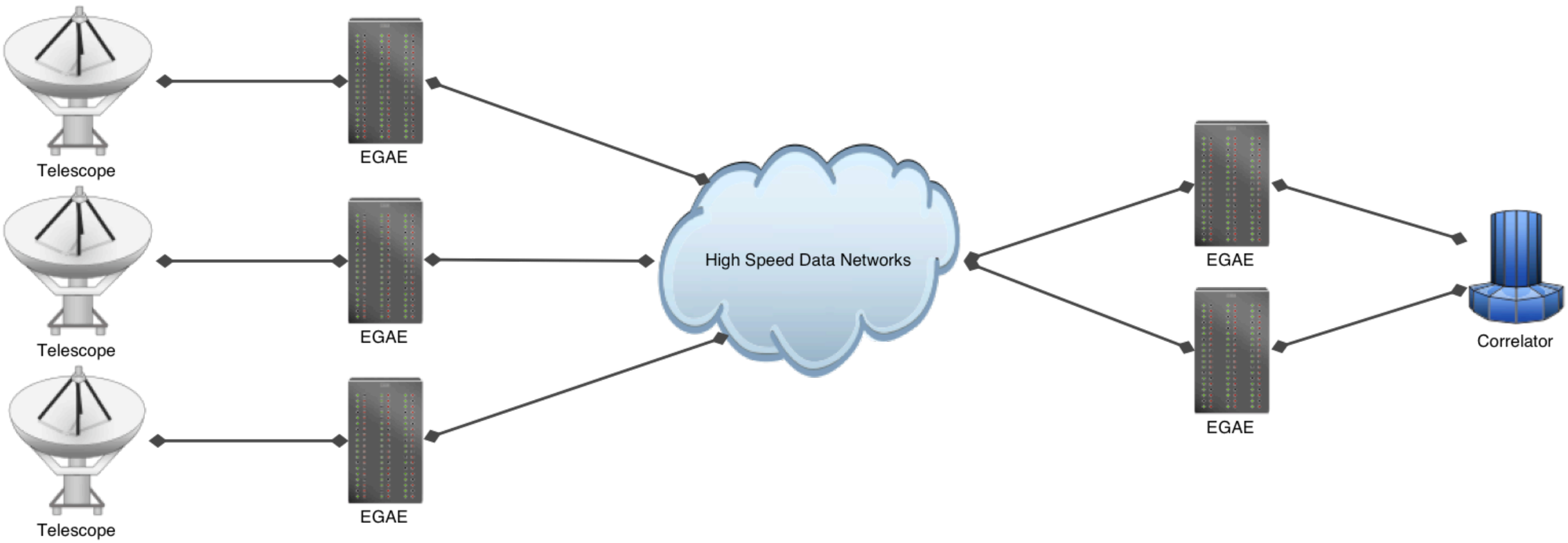
e-VLBI

- VLBI
  - Data is recorded onto magnetic media (e.g. tape or hard disk)
  - Data shipped to central site
  - Data correlated - result published 4d - 15 weeks later
- Use the network instead of storage media
  - Transmit data in real-time or near-real-time from instrument (telescope) to processing center
  - Many advantages...

Advantages of e-VLBI

- Higher sensitivity:
  - increase in bandwidth means more data bits
- Faster turnaround of results
- Lower costs:
  - no media pool, therefore no transport costs
- Real-time diagnostics:
  - enables real-time reconfiguration
- Capture of transient phenomena

e-VLBI Architecture



1. Data Acquisition

2. Encapsulation
Rate limiting
Marking
(Re-)Transmission
Mode selection

3. Delay
Loss
Bottlenecks
Other users

4. Data extraction
Buffering
Synchronization
QoS feedback
Mode selection

5. correlation

Kashima to Westford Experiment

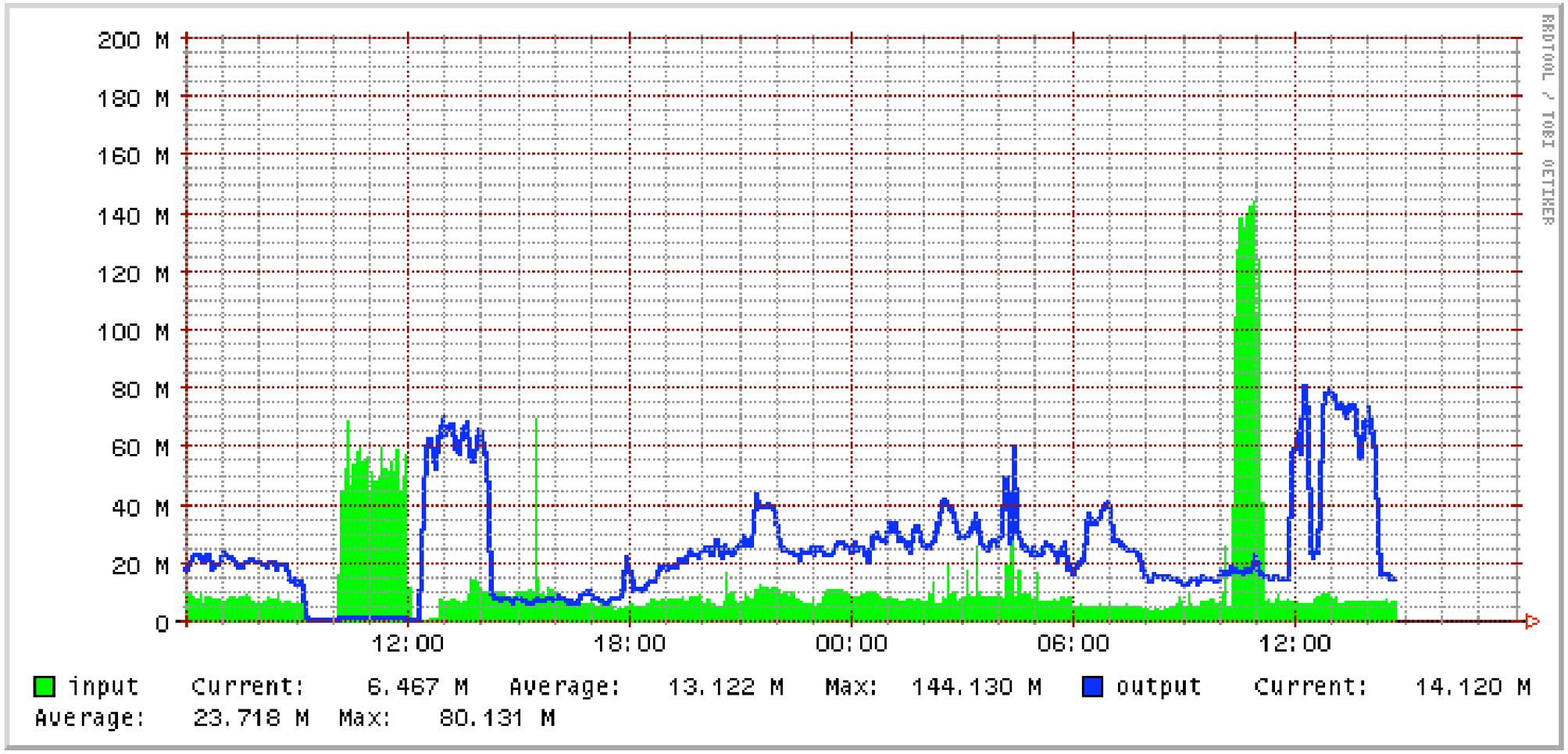
- "UT1" intensive
- UT1 estimate within 24 hours
- 40GB of data per station collected over 2 hours at 2 stations
- Total volume of data transferred:
  - Kashima to Westford: 41.54GB
  - Westford to Kashima: 41.54GB
- Average transfer rates:
  - Kashima to Westford: 107 Mbps
  - Westford to Kashima: 44.6 Mbps

Results

- UT1 estimate within 24 hours

Event	Time	Elapsed Time (hh:mm:ss)
First Scan		
Start transfer from Ka - Wf	Fri Jun 27 11:06:01 EDT 2003	00:06:05
Complete transfer from Ka - Wf	Fri Jun 27 11:12:06 EDT 2003	
Entire Dataset		
Start transfer from Ka - Wf	Fri Jun 27 11:20:04 EDT 2003	00:50:49
Complete transfer from Ka - Wf	Fri Jun 27 12:11:03 EDT 2003	
Start transfer from Wf - Ka	Fri Jun 27 13:16:24 EDT 2003	02:04:02
Complete transfer from Wf - Ka	Fri Jun 27 15:20:26 EDT 2003	
Processing		
Detected Fringes from first scan	Fri Jun 27 11:53:00 EDT 2003	00:53:00
Completion of correlation (Wf)	Sat Jun 28 01:19:00 EDT 2003	14:59:00
Estimated UT1-TAI (Ka)	Sat Jun 28 08:59:00 EDT 2003	21:59:00

Results



Graph showing traffic from NYCM-SINET
Courtesy Masaki Hirabaru of Communications Research Laboratory, Japan
Retrieved on 6/27/2003 from:
http://winger.uits.iu.edu/snapp/show-graph.cgi?title=nycm-sinet&rrdname=nycm-sinet.rrd

Acknowledgements

- The members of the experiment team included:
  - John Ball (Haystack), Kevin Dudevoir, Haystack, Dave Gordon,(NASA/GSFC), Masaki Hirabaru(CRL),Tetsuro Kondo(CRL),Yasuhiro Koyama(CRL),David Lapsley(Haystack),Hiro Osaki(CRL), Mike Poirier(Westford), Mike Titus(Haystack), Hisao Uose(NTT Laboratories), Alan Whitney(Haystack).
- Thanks also to:
  - Internet2, Super-SINET, Galaxy Network team (CRL, NTT, NAO, and ISAS)

Areas for Improvement

- NFS Tuning
- NFS elimination
  - Direct transfer to Mark 5's at correlator site
- High speed Transport Protocols
  - HSTCP, FAST, STCP, others
- Reduce number of network bottleneck links
  - identify configuration issues, investigate alternate routes

Types of Networks

- NSF defines three classes of Research & Education networks beyond the commodity Internet:
  - Production Networks
    - high-performance, always available and dependable (e.g. ESnet, DREN, NREN, Abilene). 24x7 reliability.
  - Experimental Networks
    - high-performance trials of cutting-edge networks, based on advanced application needs unsupported by existing production networks' services. Provide delivered experimental services on a persistent basis, encourage experimentation.
  - Research Networks
    - smaller-scale network prototypes. Enable basic scientific and engineering network research and testing. Not persistent, don't support applications.

- Networks can also be classified according to the technology used:
  - Circuit switched
  - Packet switched networks
  - Optically switched networks
    - Wavelengths are switched
    - At the ingress/egress, wavelengths are converted to/from other protocols (e.g. SONET, ATM, Ethernet, etc.)
  - Layer 2 networks
    - ethernet, atm, ppp
  - Layer 3 networks (IP)
    - packets are routed
- Each network type has its own characteristics

Circuit v. Packet Switching

- Circuit switching
  - establishes end to end circuit (or connection) with dedicated resources (bandwidth and buffer) prior to data transmission (e.g. telephone network)
  - highly reliable and predictable quality of service, but not suitable for all applications
- Packet Switching
  - data transported in small “packets” of information
  - no connection setup required prior to data transmission (e.g. Internet)
  - best effort service, higher data rates, statistical multiplexing makes better use of network resources

Circuit v. Packet Switching

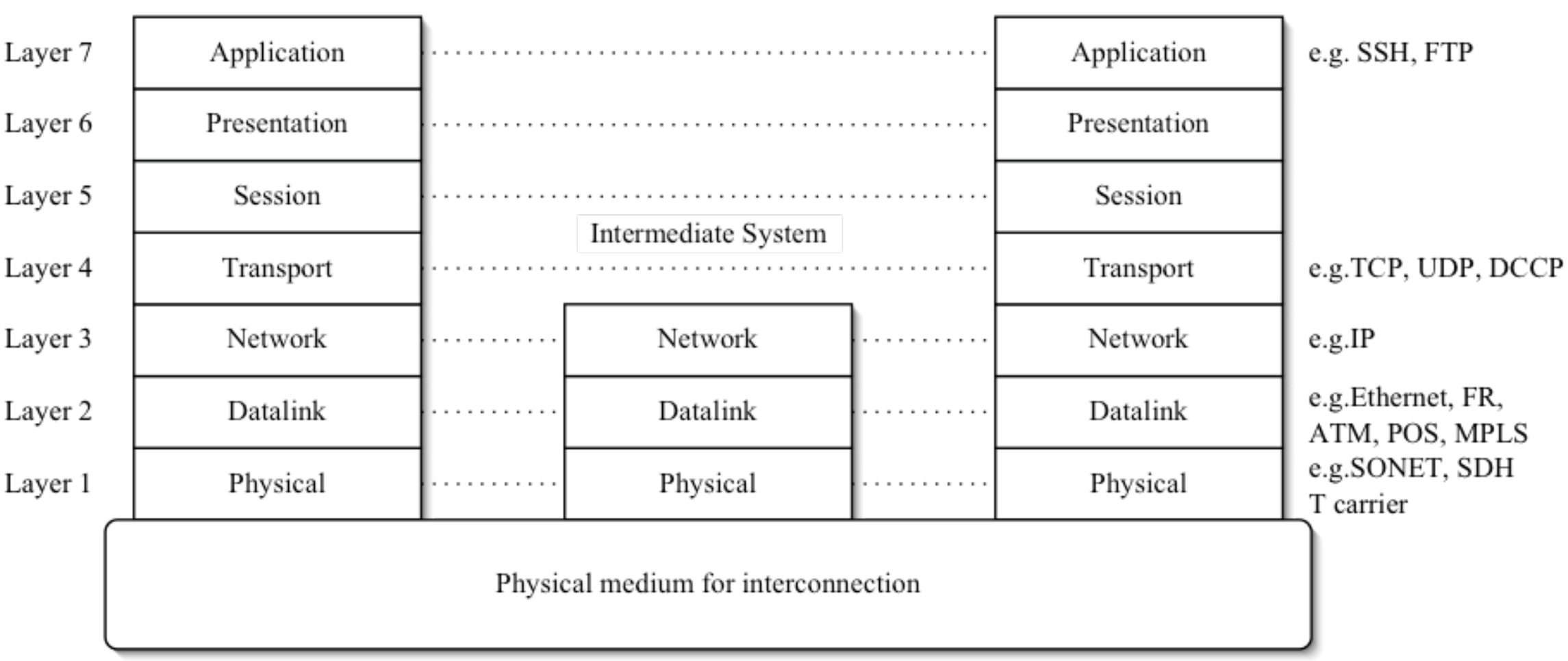
- Applications:
  - In the past, Circuit switched networks supported real-time services such as voice
  - In the past, Packet switched networks supported non-real-time services such as data
  - In recent years, evolution towards Hybrid Networks
    - Integration of circuit and packet switching (e.g. DSL, VoIP)
    - Evolution towards multi-service networks
      - e.g. Internet Protocol
      - e.g. Multi-Protocol Label Switching (MPLS)
      - e.g. Asynchronous Transfer Mode (ATM)

Networking Trends

- Optical is cheaper
- Gigabit ethernet
- 10 Gigabit ethernet
- Ethernet in the Wide Area Network
- Commodity layer 2 Gigabit ethernet switches
  - Much cheaper than other alternatives

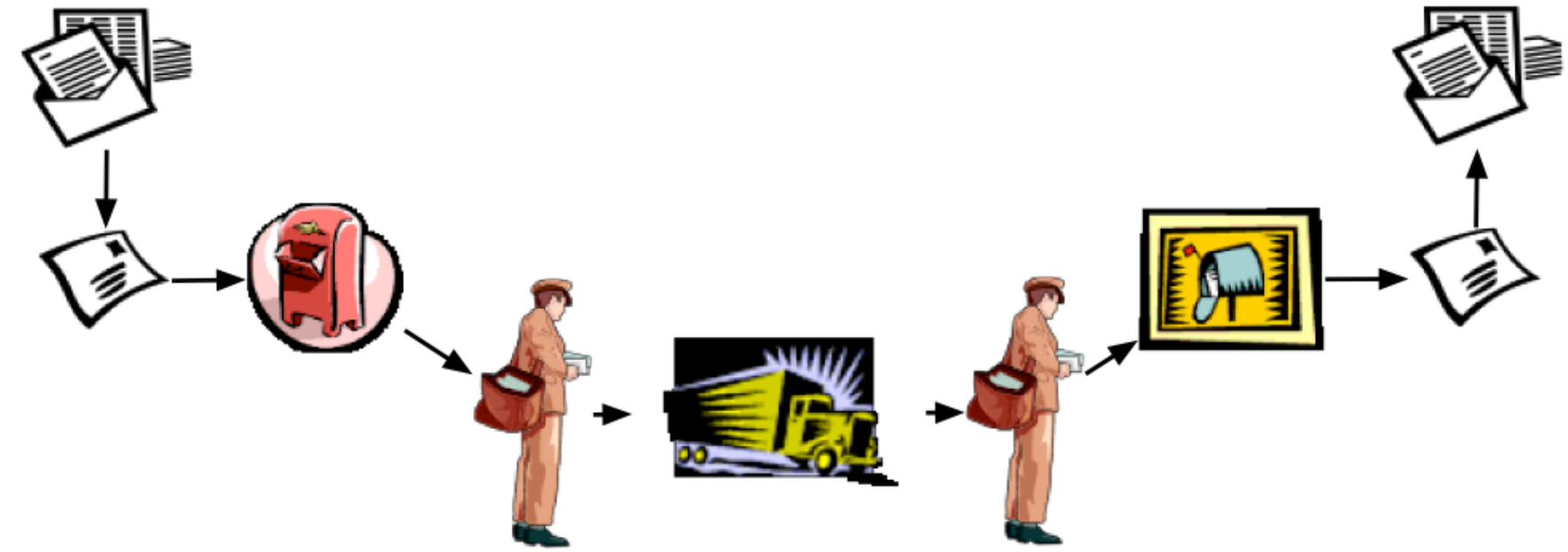
Basic Transmission Protocols

- Data transported over networks using layered protocol stack (layering enables decomposition of complex problem, abstraction and re-use)
  - Open Systems Interconnect (OSI) reference model developed by International Standards Organization (ISO)



Layered Transport

- Analogy: mailing a letter



Layer 3: IP Layer

- Packet Switching
- Simple
- Unreliable
  - higher layers add reliability
- Designed to operate over heterogeneous networks
- Provides addressing and encapsulation

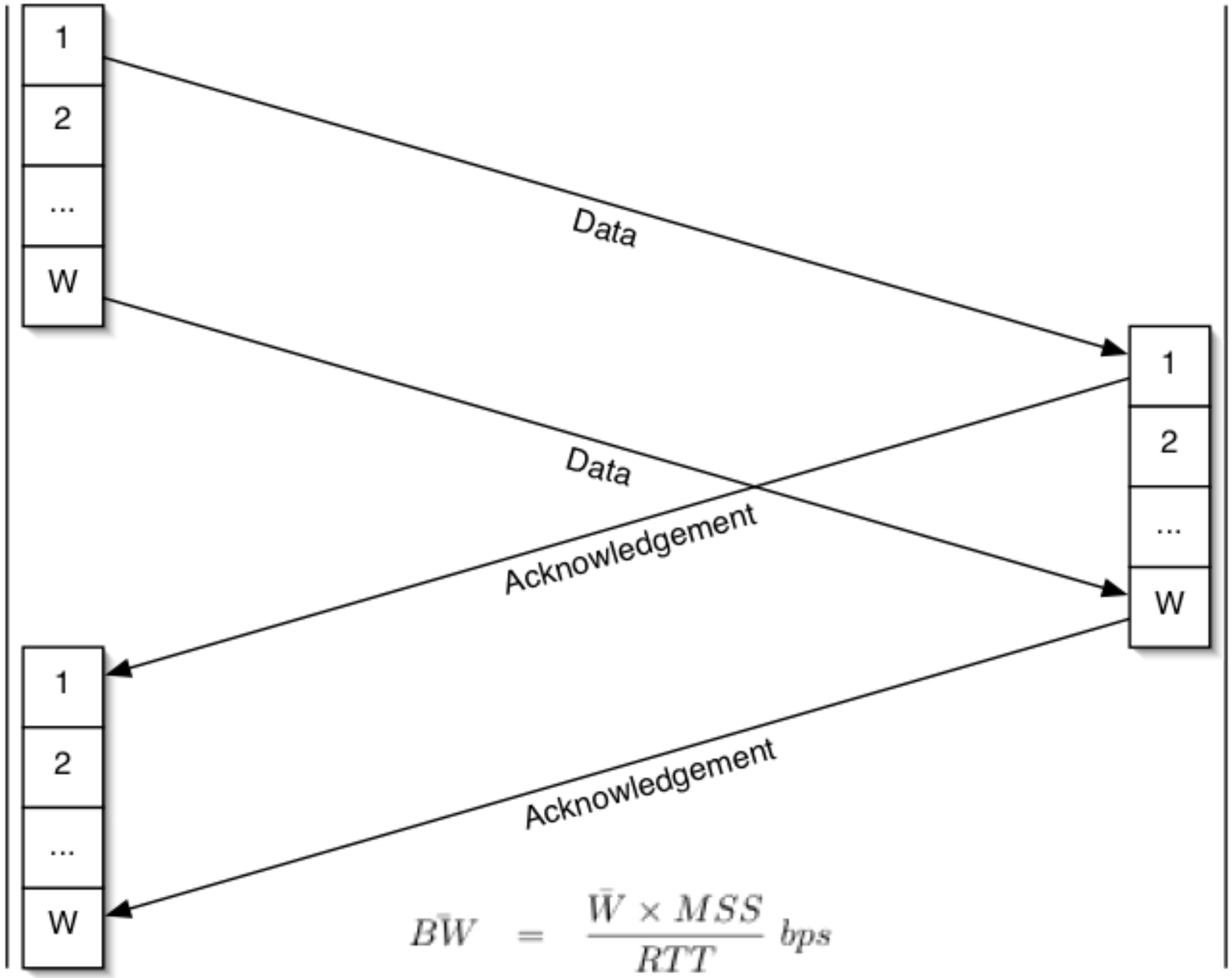
Layer 4: Transport Layer

- Transport layer is responsible for providing additional services on top of IP.
  - IP transfers packets from host to host, transport layer transfers packets from host/port to host/port
- Two main transport layer protocols:
  - Transmission Control Protocol(TCP)
    - reliable, end-to-end delivery, congestion avoidance and control
  - User Datagram Protocol (UDP)
    - lightweight, unreliable, end-to-end delivery
  - Datagram Congestion Control Protocol (DCCP)
    - new, minimal general purpose transport-layer protocol

TCP

- Used by many common applications: FTP, Telnet, SSH, SMTP
- Implements windowed congestion avoidance and control
  - allows connections to make use of network bandwidth, while ensuring that network doesn't go into congestion collapse
- Foundation on which TCP is built was laid in 1988
  - In response to Internet Congestion collapse in October 1986
  - Van Jacobson proposed TCP flow control in 1988
- Since then, many enhancements have been made (basic scheme remains the same)

Windowed Flow Control



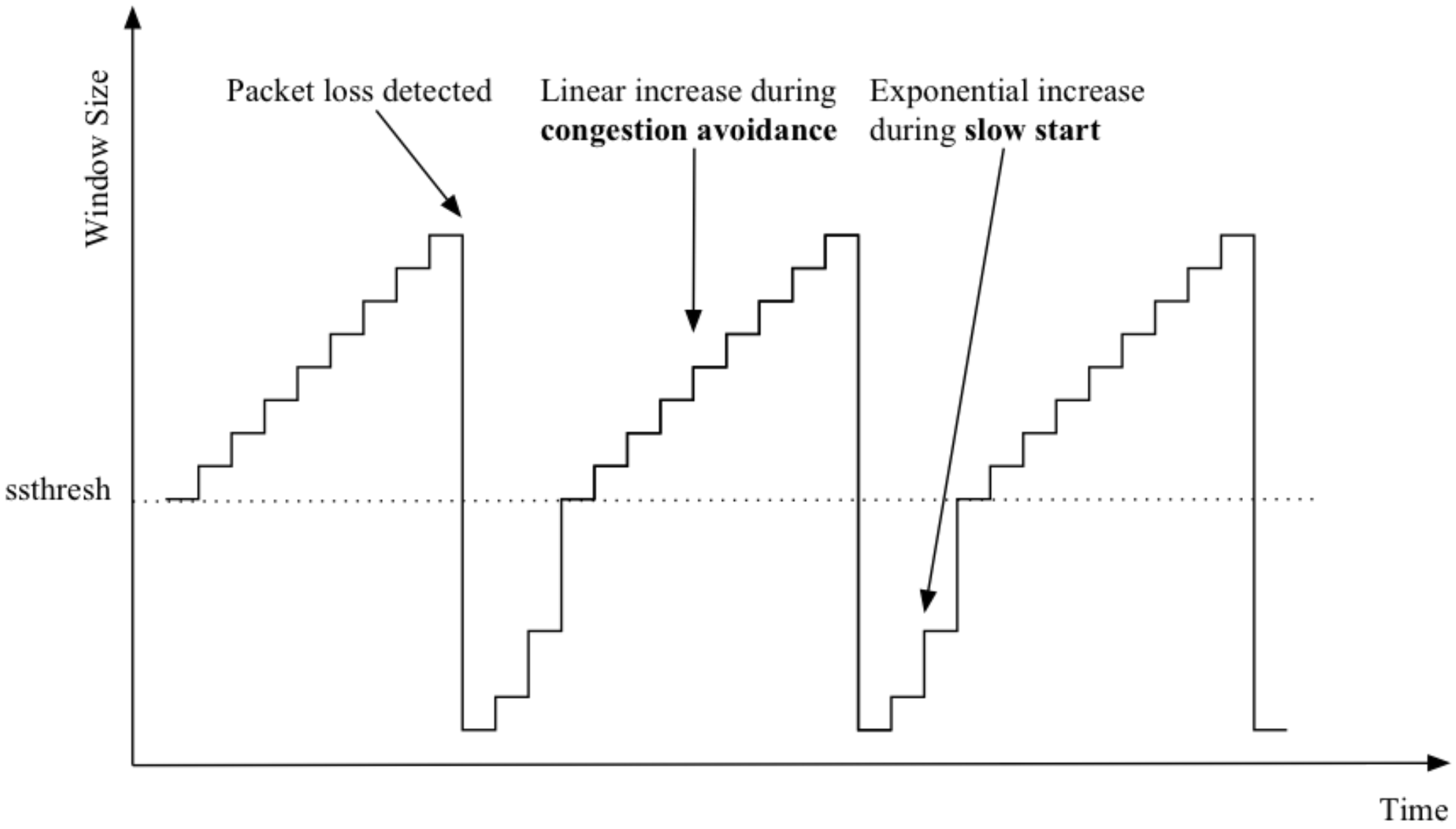
$$\bar{BW} = \frac{\bar{W} \times MSS}{RTT} \; bps$$

Sources may only transmit at most W packets per round trip time . W varies with time as the level of congestion in the network varies.

Average bandwidth (BW) is a function of average window size (W), Maximum Segment Size (MSS) and Round Trip Time (RTT).

Data packets are acknowledged by receiver. Missing acknowledgements cause the sender to retransmit.

Basic TCP Congestion Avoidance and Control



1. Source starts with a window size of 1 and increases exponentially(slow start) to "ssthresh" and then linearly (congestion avoidance) after that

2. When source detects a packet loss it sets "ssthresh" to half the window size and reduces window size to 1.

TCP Variants

- "Traditional"
  - TCP Tahoe (Jacobson 1988)
  - TCP Reno (Jacobson 1990)
  - TCP Vegas (Brakmo and Peterson 1994)
- High Speed
  - Fast AQM Scalable TCP (Jin, Wei, Low 2003)
  - High Speed TCP (Floyd 2003)
  - Scalable TCP (Kelly 2002)

TCP over Big, Fat Pipes

TCP average throughput and window size given by the Mathis equation:

$$B\bar{W} = \frac{MSS}{RTT}\sqrt{\frac{1.5}{p}}\ bps \quad (1)$$

$$\bar{W} = \sqrt{\frac{1.5}{p}}\ packets \quad (2)$$

where, $B\bar{W}$ is the average bandwidth, MSS is the Maximum Segment Size, RTT is the Round Trip Time, $\bar{W}$ is the average window size and p is packet loss.

For $B\bar{W} = 10Gbps$, $MSS = 1500B$, $RTT = 100ms$:

$$p = 2.16 \times 10^{-10} \quad (3)$$

$$\bar{W} = 83,333\ pps \quad (4)$$

This is equivalent to 1 packet loss per 5,556 s (1.54 hours)! Not realistic in shared networks!

TCP over Big, Fat Pipes

- Solutions (ascending order of preference):
  - Use UDP
    - huge negative impact on other users, not reliable, but provides access to maximum amount of bandwidth
  - Use rate-based flow control
    - must be designed to be “TCP Friendly”
  - Use low loss (dedicated) links(e.g. latest Internet land speed record)
  - Use multiple parallel TCP streams(e.g. BBFTP)
  - Modify TCP stack to allow it to open its window faster while still maintaining some degree of fairness with regular TCP sessions (e.g. FAST, HSTCP, STCP, etc.)

TCP over Big, Fat Pipes

- Tools
  - TCP Friendly Rate-based Flow Control
    - SABUL: http://www.dataspaceweb.net/sabul.htm
    - TSUNAMI: http://www.indiana.edu/~anml/anmlresearch.html
  - Multiple-Parallel TCP Sessions
    - http://doc.in2p3.fr/bbftp/
- Modified TCP Stacks
  - http://www.icir.org/floyd/hstcp.html
  - http://netlab.caltech.edu/FAST/index.html
  - http://www-lce.eng.cam.ac.uk/~ctk21/scalable/

TCP over Big, Fat Pipes

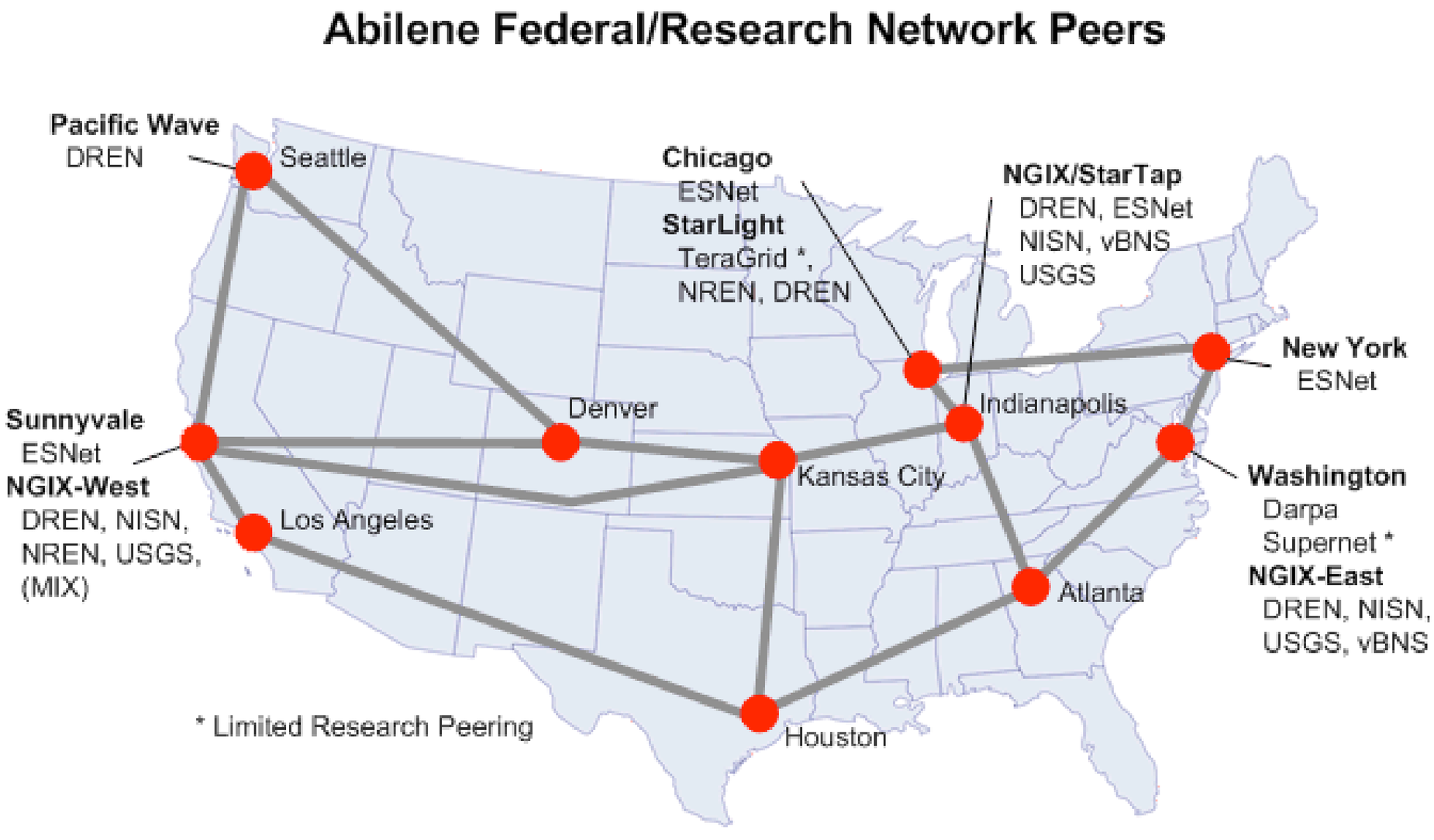
- Example of low loss links:
  - http://www-iepm.slac.stanford.edu/lsr2/
- Tuning TCP
  - http://www.psc.edu/networking/perf_tune.html
- Diagnostic Tools
  - http://dast.nlanr.net/Projects/Iperf/
  - http://www.employees.org/~bmah/Software/pchar/
  - ftp://ftp.ee.lbl.gov/tcpdump.tar.Z
  - http://www.tcptrace.org/
- Summary available at:
  - web.haystack.mit.edu/staff/dlapsley/

Current Global Connectivities

Global Connectivities

- VLBI locations spread across the world
- Transporting data at high speeds between sites involves working with a collection of research and education networks around the world
- These networks have backbone bandwidths ranging from 100's of megabits per second to 10 Gbps

Abilene Domestic Connectivity

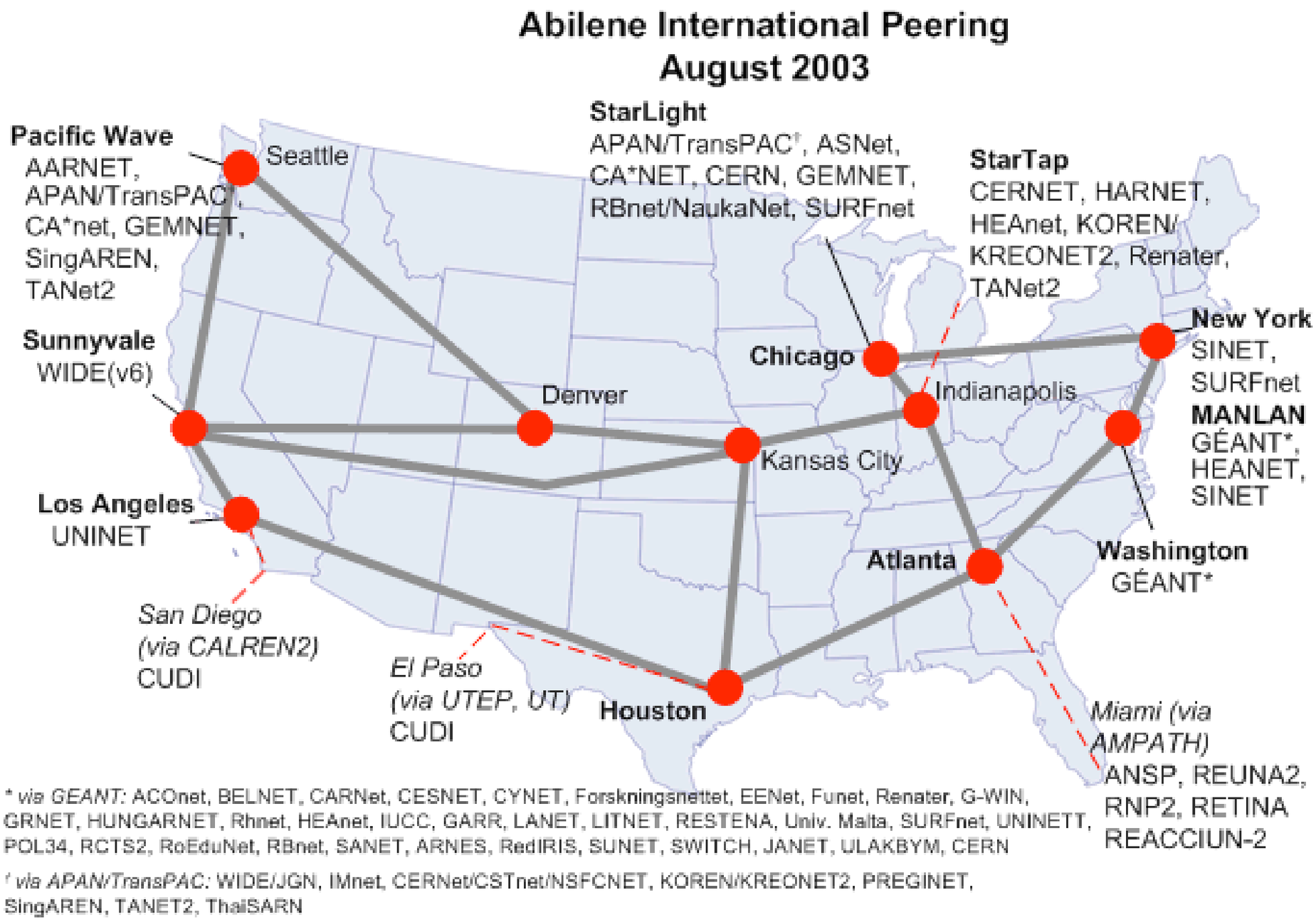


9/12/03

Courtesy Internet2. Available at:

http://abilene.internet2.edu/peernetworks/domestic.html

Abilene International Peering

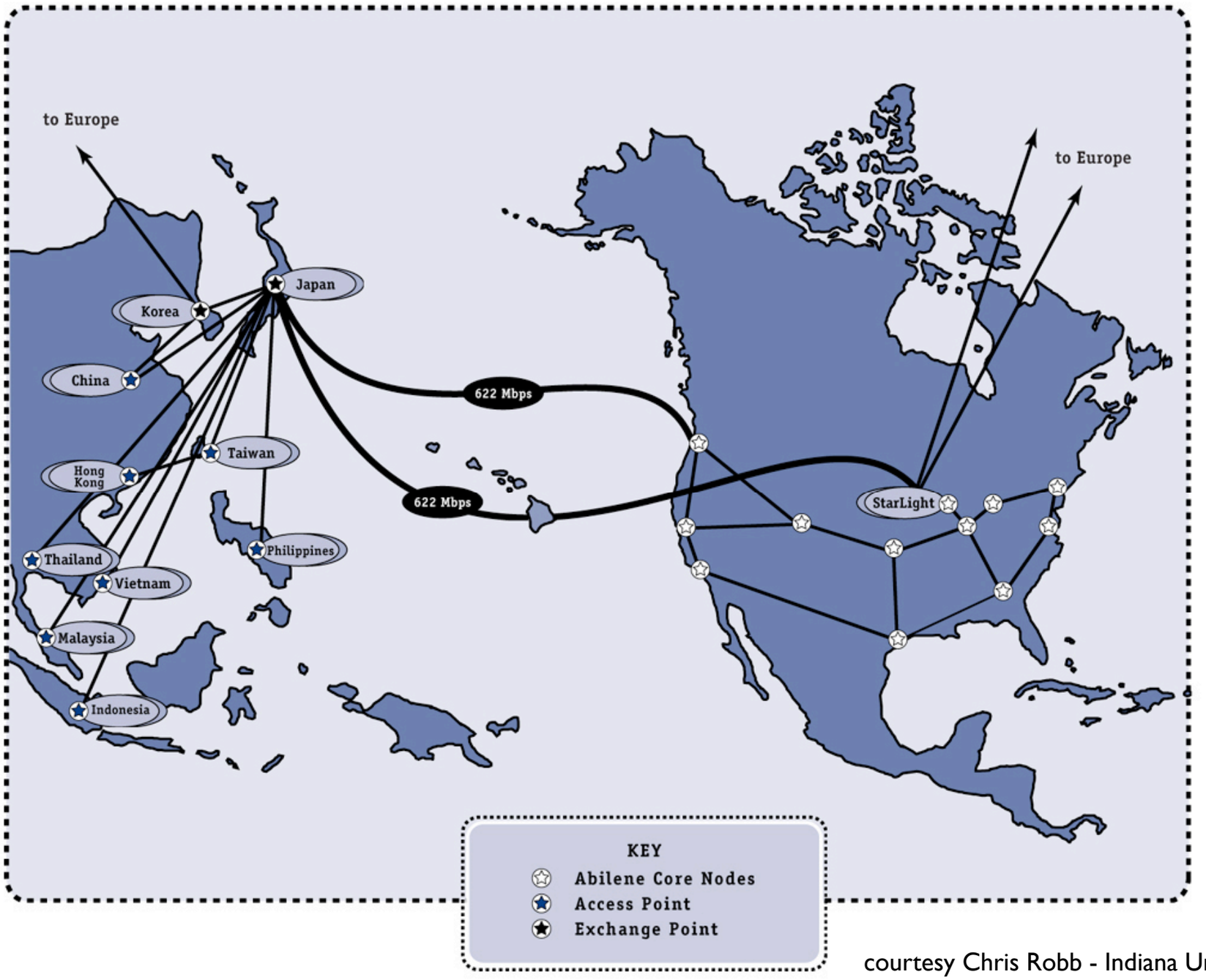


9/12/03

Courtesy Internet2. Available at:

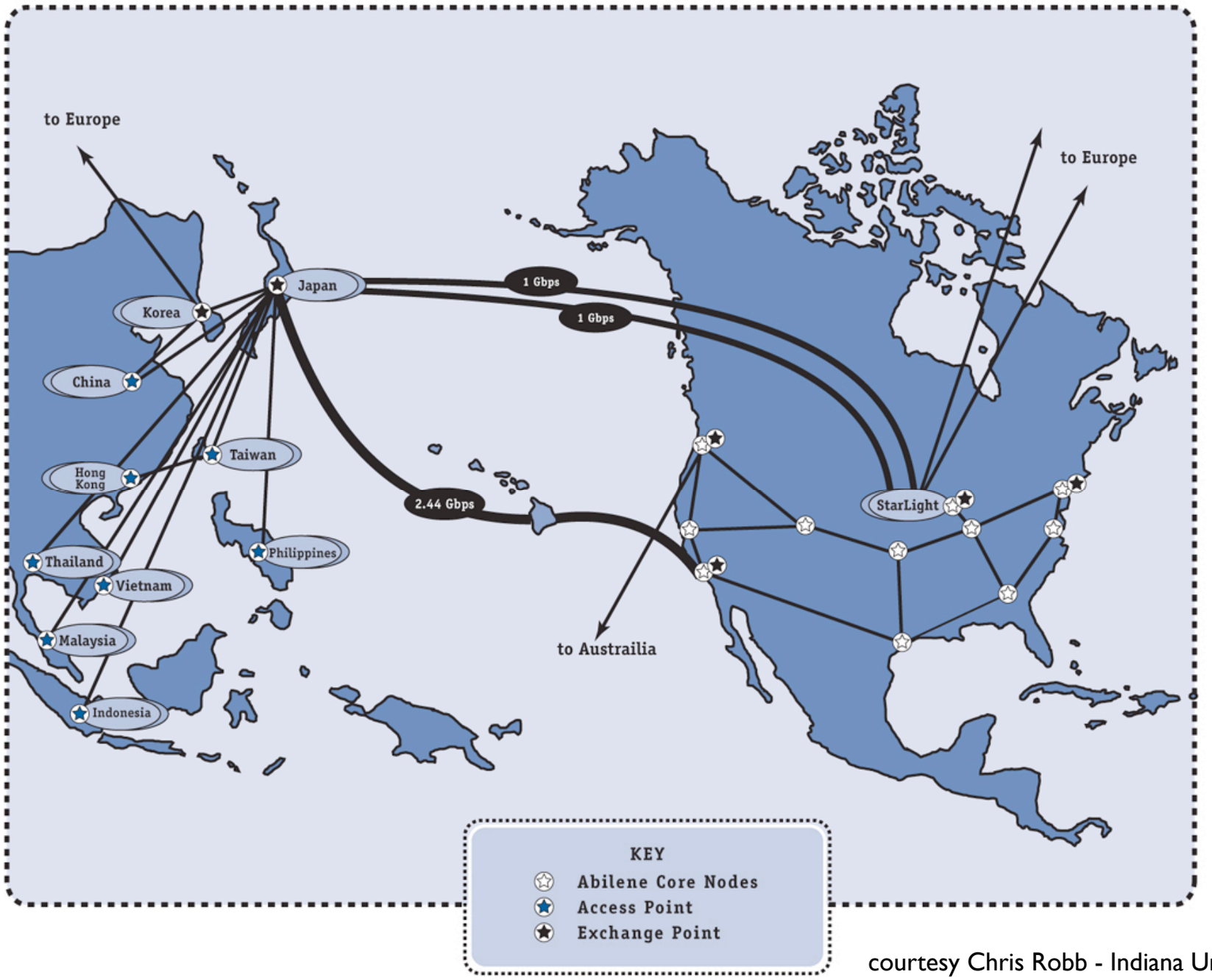
http://abilene.internet2.edu/peernetworks/domestic.html

APAN



courtesy Chris Robb - Indiana University

APAN Planned Upgrades



courtesy Chris Robb - Indiana University

Other International Networks

- Of interest to e-VLBI
  - SINET (Japan)
  - GEANT (Europe)
  - SURFnet (The Netherlands)
  - G-Win (Germany)
  - REUNA (Chile)
  - AARNET (Australia)
  - JANET (United Kingdom)

Other International Networks

Courtesy Internet2. Available at:

http://abilene.internet2.edu/peernetworks/peer-by-region.html

Americas	Asia-Pacific	Europe-Middle East
Argentina (RETINA) Brazil (RNP2/ANSP) Canada (CA*net) Chile (REUNA) Mexico (CUDI) United States (Abilene, vBNS) Venezula (REACCIUN-2)	Australia (AARNET) China (CERNET, CSTNET, NSFCNET) Hong Kong (HARNET) Japan (SINET, WIDE, IMNET, JGN) Korea (KOREN, KREONET2) Singapore (SingAREN) Philippines (PREGINET) Taiwan (TANET2) Thailand (UNINET, ThaiSARN)	Austria (ACOnet) Belgium (BELnet) Croatia (CARnet) Czech Rep. (CESnet) Cyprus (Cynet) Denmark (UNI-C) Estonia (ESnet) Europe (GEANT) Finland (FUnet) France (RENATER) Germany (G-Win) Greece (GRnet) Hungary (HUNGARnet) Iceland (ISnet) Ireland (HEAnet) Israel (IUCC) Italy (GARR) Latvia (LATNET) Lithuania (LITNET) Luxembourg (RESTENA) Netherlands (SURFnet) Norway (UNINETT) Poland (PCSS) Portugal (FCCN) Romania (RNC) Russia (RIPN) Slovakia (SANET) Slovenia (ARNES) Spain (RedIris) Sweden (SUNET) Switzerland (SWITCH) United Kingdom (JANET) *CERN

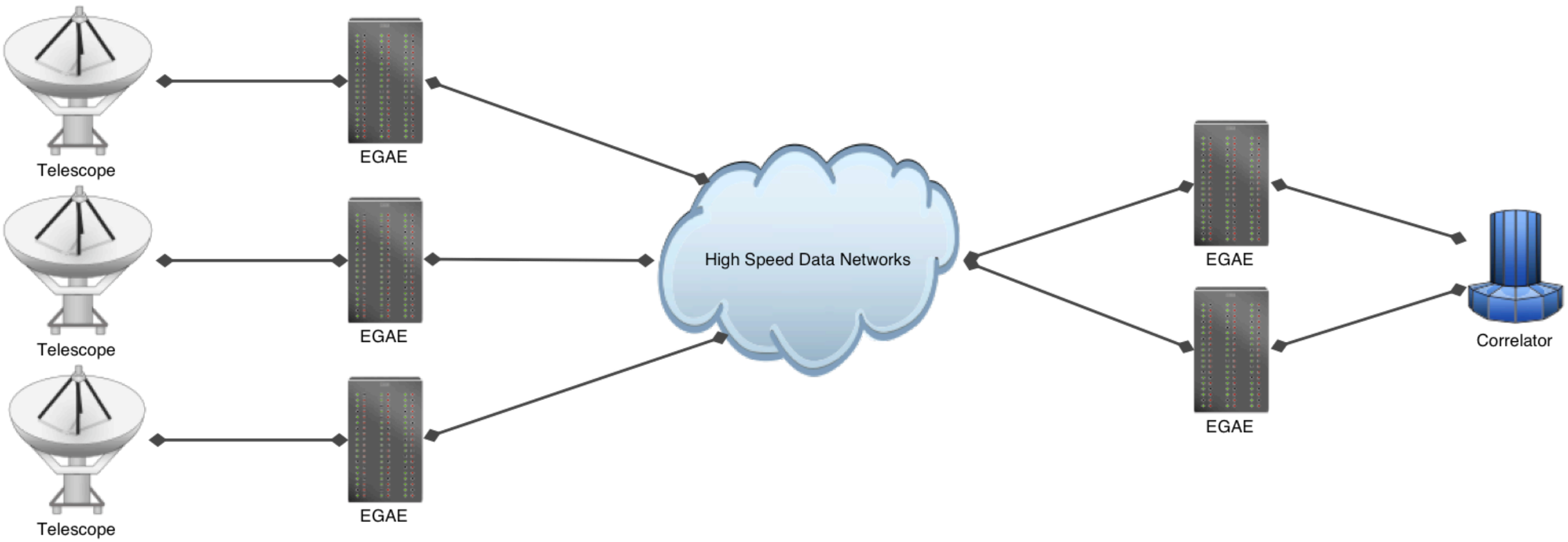
Current Issues

- ‘Last mile’ connectivity
- Network bottlenecks well below advertised rates
- Performance of transport protocols
  - untuned TCP stacks
  - fundamental limits of regular TCP
- Throughput limitations of COTS hardware
  - Disk-I/O - Network

e-VLBI Development at Haystack

- Experiment Guided Adaptive Endpoint:
  - Interfaces VLBI hardware to IP networks and transmits VLBI data
    - Uses low priority “scavenged bandwidth”
      - Abilene “less-than-best-effort” service
      - Statistical multiplexing on Research/Commerical networks
    - Adapts transmission rates to suit network congestion
      - Development of VLBI Transport Protocol
    - Allows characteristics of adaptive behaviour to be determined by high level experimental profile
      - VEX for Astronomical applications
      - XML based profile for generic scientific applications

Architecture



1. Data Acquisition

2. Encapsulation
Rate limiting
Marking
(Re-)Transmission
Mode selection

3. Delay
Loss
Bottlenecks
Other users

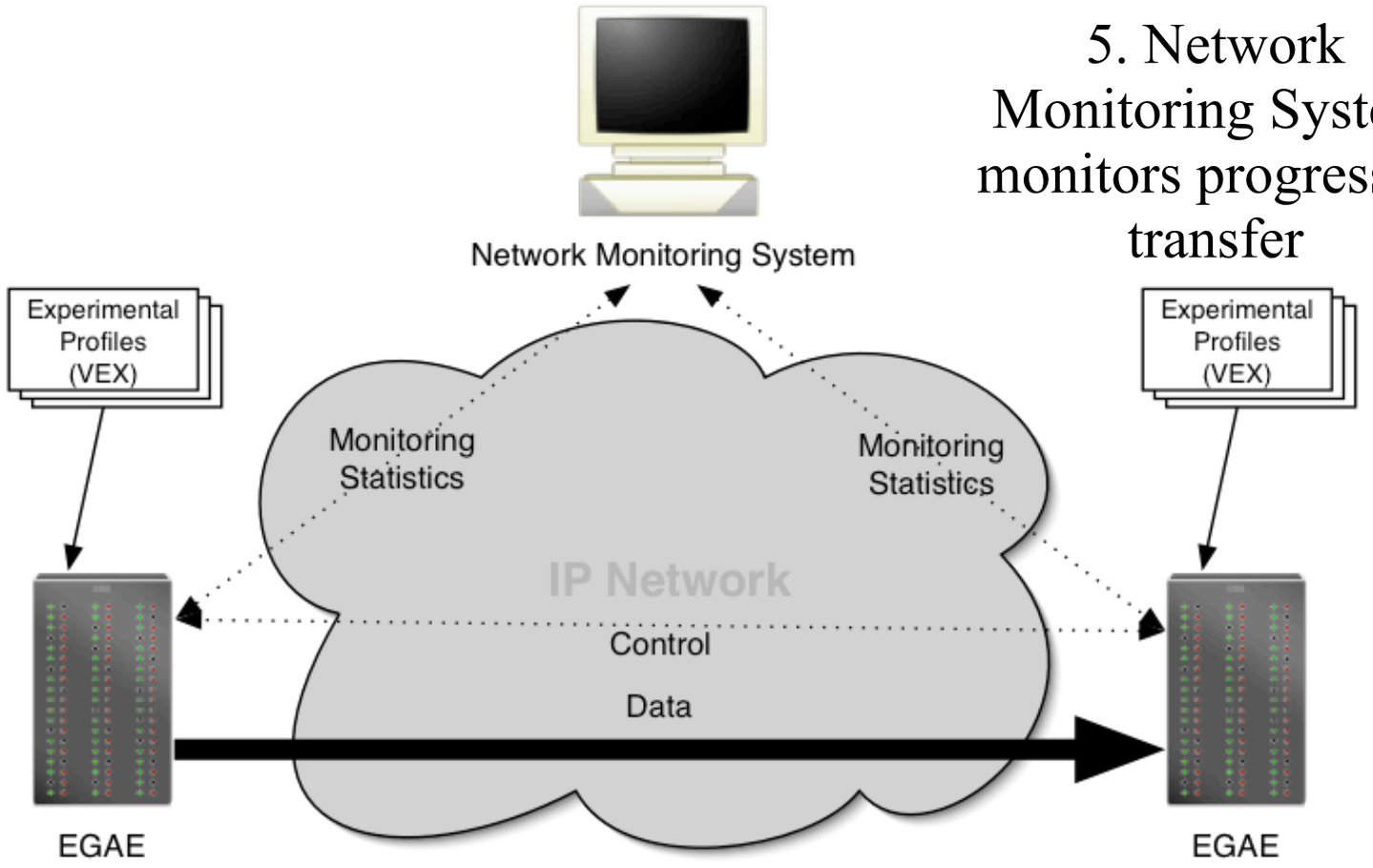
4. Data extraction
Buffering
Synchronization
QoS feedback
Mode selection

5. correlation

e-VLBI with EGAE

1. Astonomical + EGAE Profile downloaded to Stations (Telescope sites) and EGAEs

2. Station personnel oversee transfer

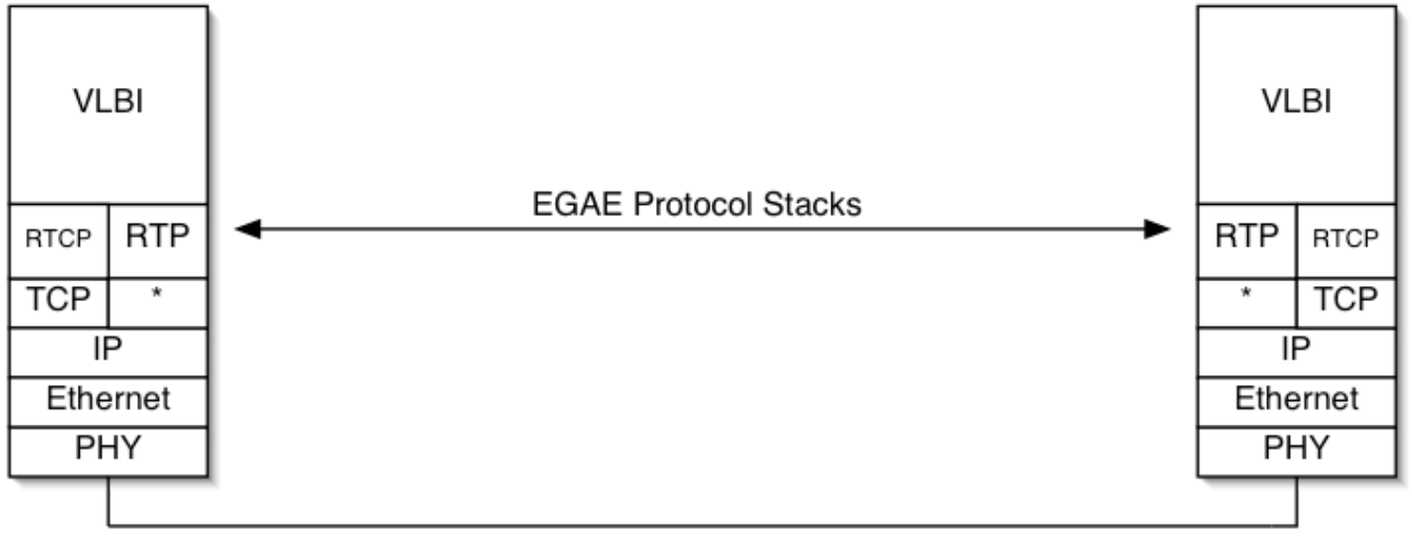


5. Network Monitoring System monitors progress of transfer

7. Successful data correlation!

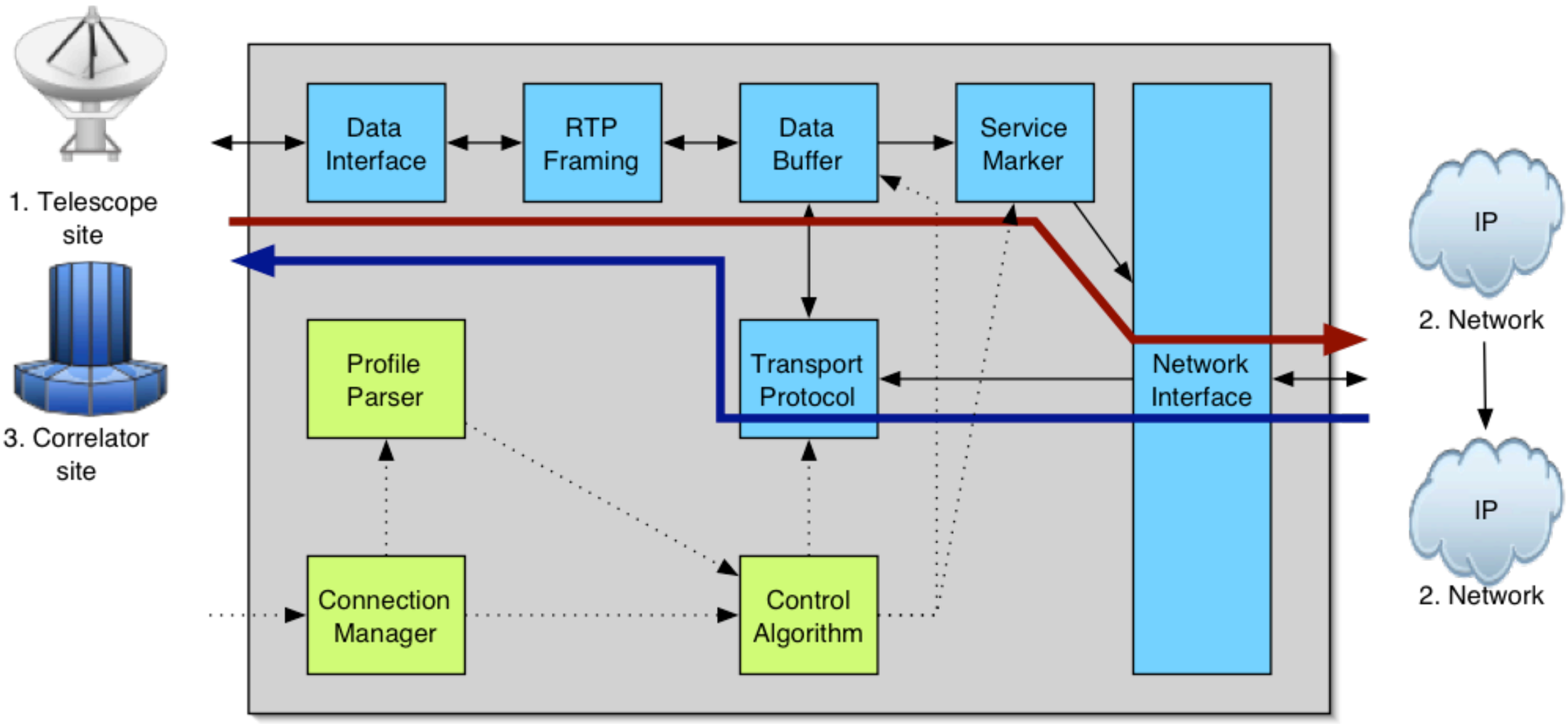
6. Real-time monitoring of a single data channel to verify setup

3. Transfer of VLBI data using RTP (RTCP for control channel and QoS feedback)



4. Data unpacked and transmitted to correlator or disc

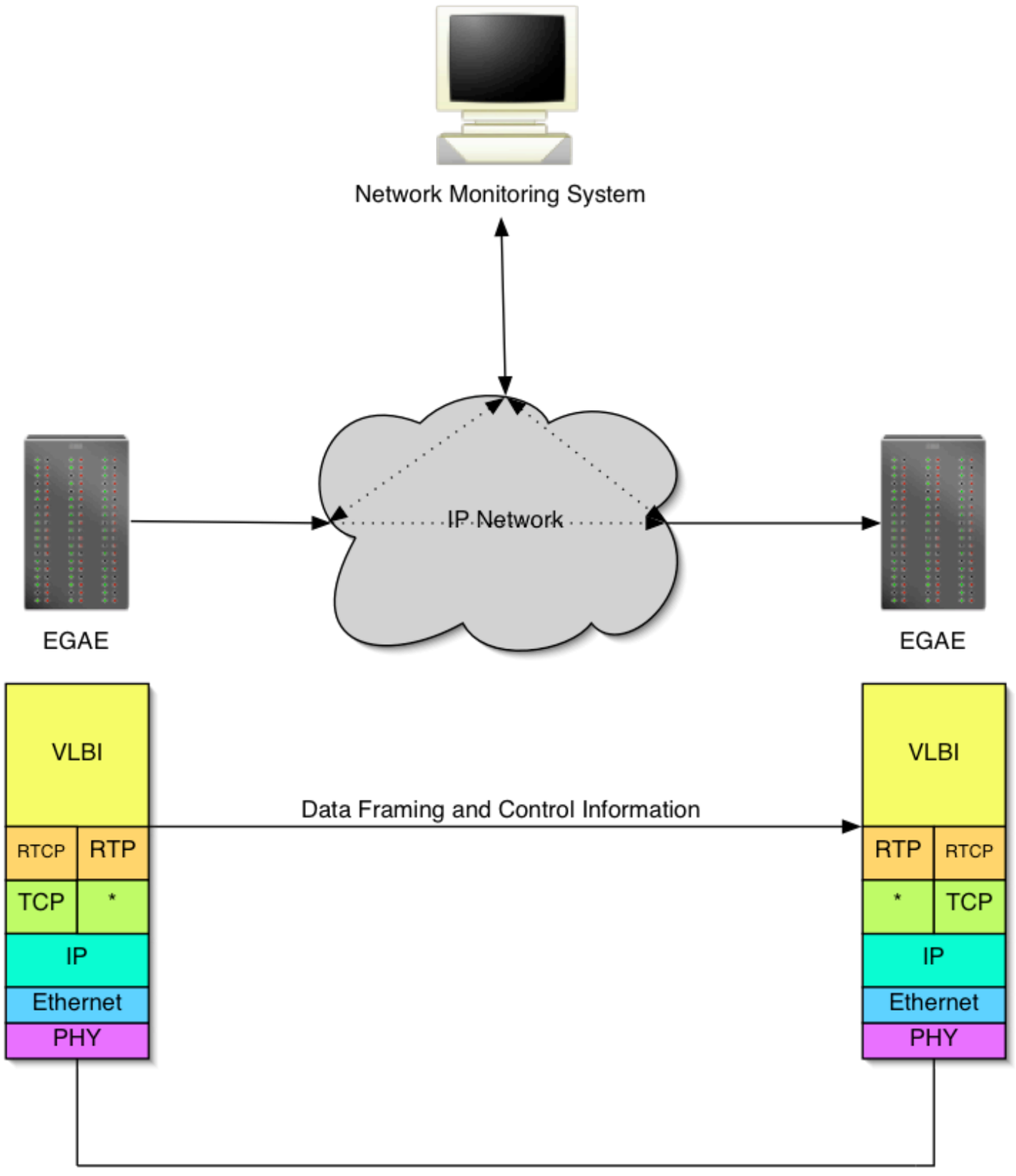
Experimental Guided Adaptive Endpoint Architecture



Red line: Telescope to Network

Blue line: Network to Correlator

Monitoring Architecture



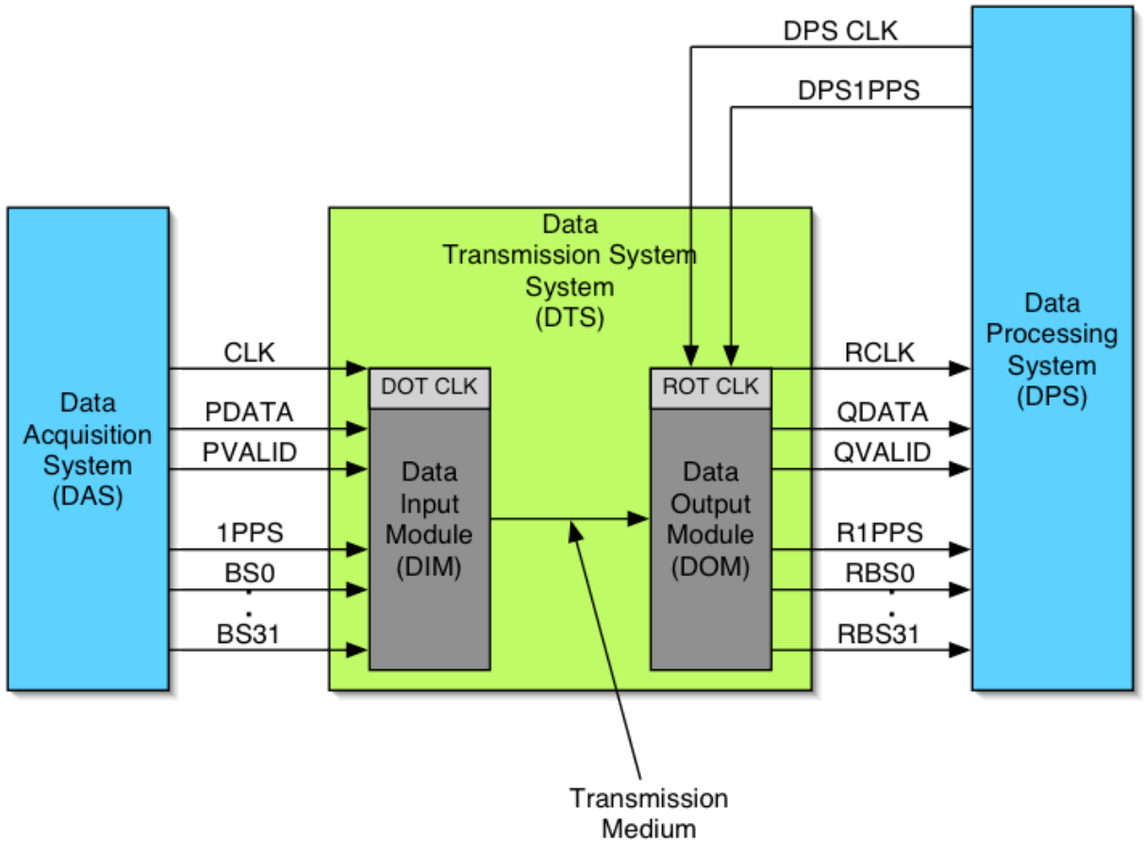
VSI-E

- VLBI Standard Interface - Electronic
- Follows in the footsteps of VSI-Hardware and VSI-Software
- International standard for electronic transport of VSI data
- facilitates inter-working of e-VLBI equipment around the world

RTP and VSI-E

- e-VLBI Workshop Dwingeloo 2003, decided to adopt RTP for transport of VSI-E data:
  - RTP has wealth of implementation and operational experience
  - RTP seen as internet-friendly by the network community:
    - attention to efficiency, attention to resource constraints, attention to scaling issues
- Draft RTP Profile developed by John Wroclawski from MIT LCS

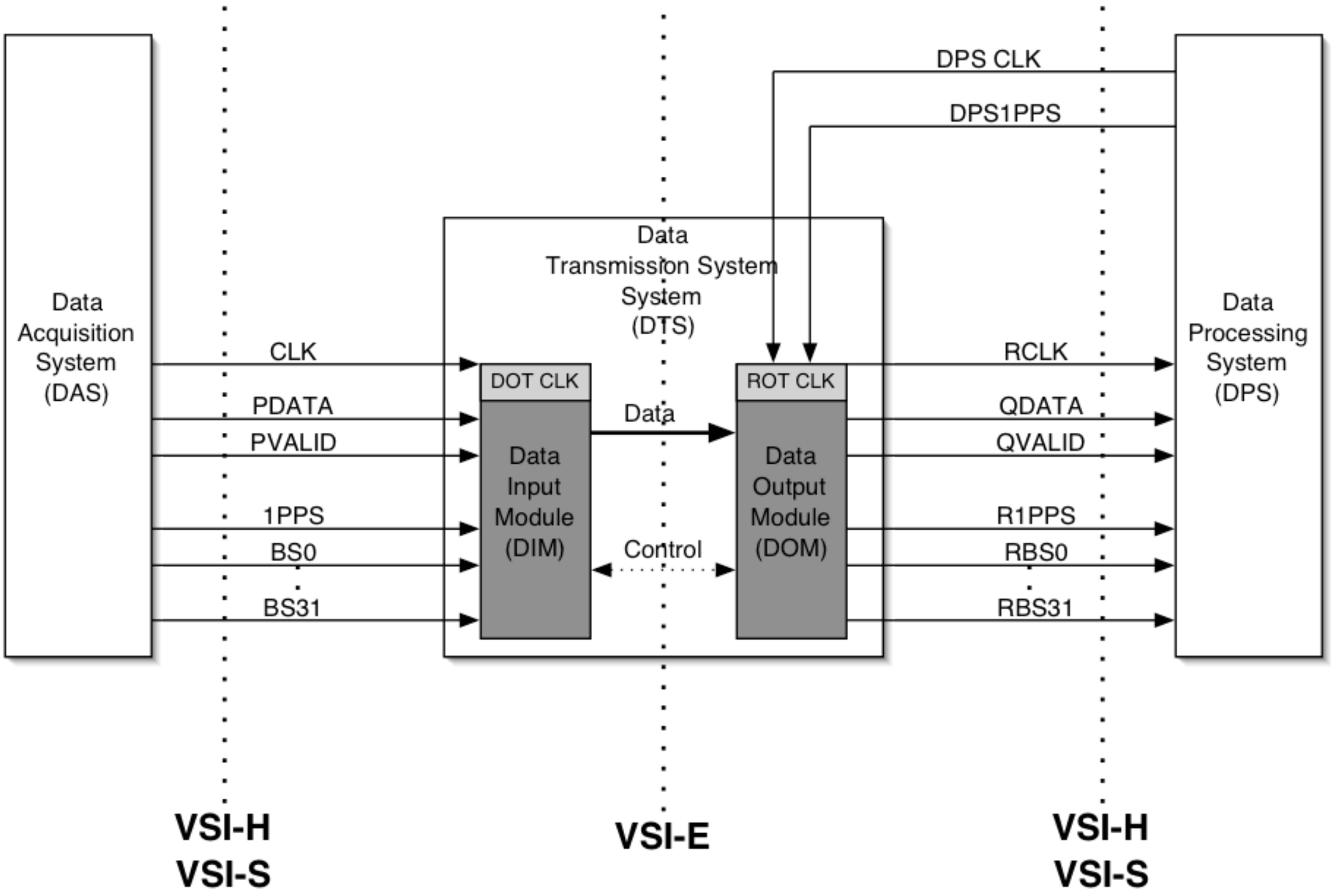
VSI-E Model



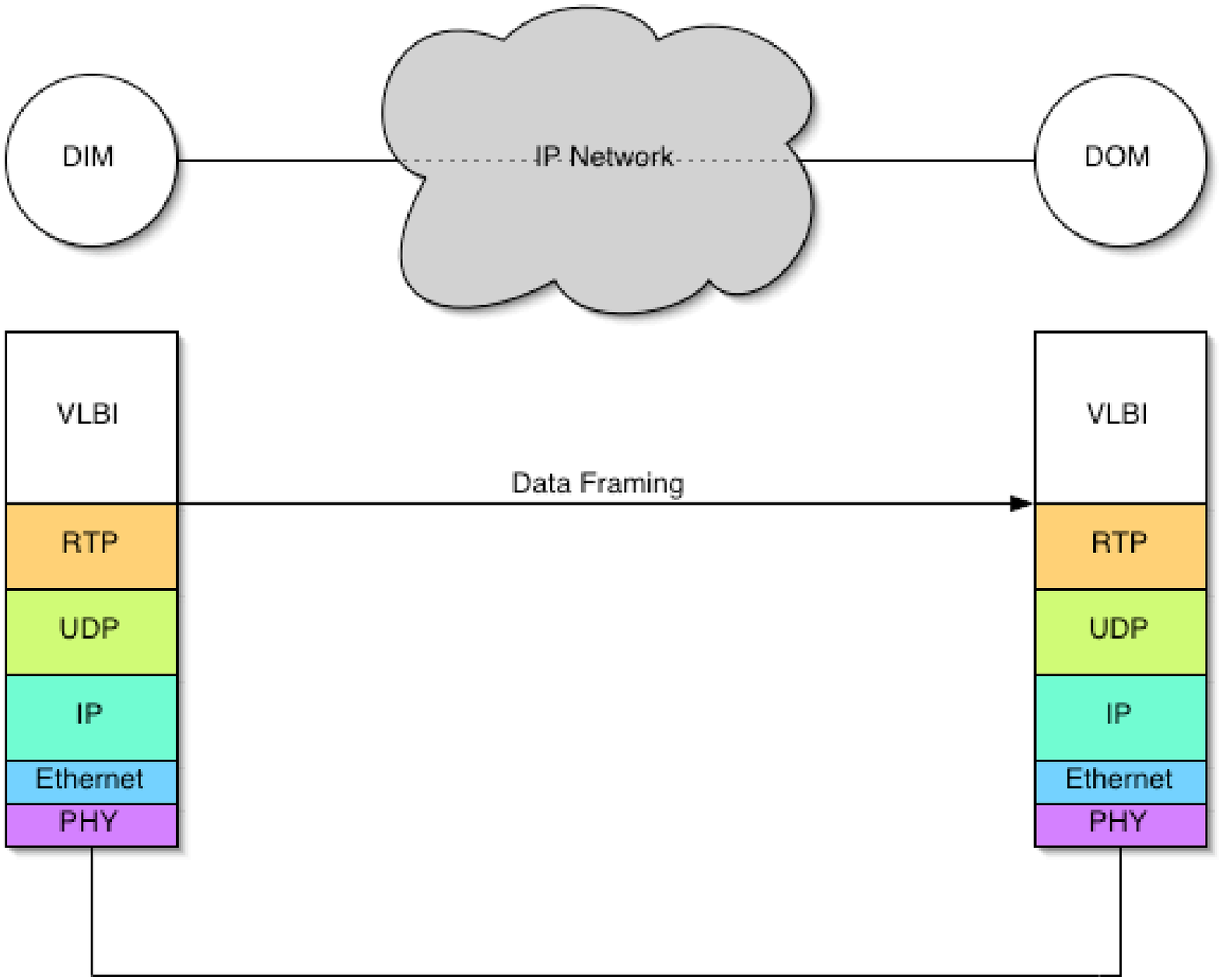
Note

CLK = A clock accompanying the bit streams.Provides a reference frequencey for the DIM
PVALID = a signal that specifies the "validity" of the bit streams
PDATA = a standard 8-bit ASCII asynchronous serial data stream
1PPS = A 1 pulse per second tick which defines corresponding parallel data bits
BS0..BS31 = 32 parallel bit-streams, all sampled by the DIM at the same rate
DOT CLK = Data Observe Time Clock: master clock within the DIM used to time tag samples.
ROT CLK = Requested Observe Time Clock: maintains the reference time to which the re-constructed data are to be synchronized
RCLK = clock accompanying the reconstructed bit streams
QDATA = a standard 8-bit ASCII serial data stream
QVALID = 1-bit global signal indicating that the reconstructed data are judged by the DOM to be correct
R1PPS = reconstructed 1PPS accompanying the bit streams
RBS0..RBS31 = reconstructed bit streams. Accurate reproductions of the active sampled bit-streams transferred from the DIM
DPSCLK = a clock from the DPS which acts as a frequency reference for the DOM
DPS1PPS = a 1-pps tick used to set an internal DOM clock called the Requested Observe Time clock to an integer-second epoch

VSI-E


Data
Acquisition
System
(DAS)
CLK
PDATA
PVALID
1PPS
BS0
BS31
Data
Transmission System
System
(DTS)
DOT CLK
Data
Input
Module
(DIM)
Data
Control
ROT CLK
Data
Output
Module
(DOM)
DPS CLK
DPS1PPS
RCLK
QDATA
QVALID
R1PPS
RBS0
RBS31
Data
Processing
System
(DPS)
VSI-H
VSI-S
VSI-E
VSI-H
VSI-S

e-VLBI Transport over RTP



References

- RFC791. Internet Protocol. J. Postel.
  - http://www.ietf.org/rfc/rfc0791.txt?number=791)
- RFC 3550: RTP: A Transport Protocol for Real-Time Applications
  - http://www.ietf.org/rfc/rfc3550.txt?number=3550
- RFC 3551: RTP Profile for Audio and Video Conferences with Minimal Control
  - http://www.ietf.org/rfc/rfc3551txt?number=3551
- RFC768: User Datagram Protocol
  - http://www.ietf.org/rfc/rfc0768.txt?number=768

References

- RFC793: Transmission Control Protocol DARPA Internet Program Protocol Specification
  - http://www.ietf.org/rfc/rfc0793.txt?number=793
- draft-ietf-dccp-spec-04.txt: Datagram Congestion Control Protocol
  - http://www.ietf.org/internet-drafts/draft-ietf-dccp-spec-04.txt
- http://www.internet2.edu
- http://www.dante.net/geant
- http://www.apan.net/home/index1.htm
- http://web.haystack.mit.edu/staff/dlapsley/