Network Middleware for High-Performance Networking

Martin Swany
Introduction and Motivation

- Networks are increasingly critical for science and education
- Data Movement is a key problem
- Network speeds can increase dramatically but users’ throughput increases much more slowly
  - Source: DOE
The Phoebus project aims to help bridge the performance gap by bringing revolutionary networks to users.

Phoebus is another name for the mythical Apollo in his role as the “sun god”.

Phoebus is based on the concept of a “session” that enables multiple adaptation points in the network to be composed.

Phoebus provides a gateway for legacy applications to use advanced networks.
End-to-End Session
Session Layer

- A *session* is the end-to-end composition of *segment-specific* transports and signaling
  - More responsive control loop via reduction of signaling latency
  - Adapt to local conditions with greater specificity
  - Buffering in the network means retransmissions need not come from the source
Session Layer Benefits

- A session layer provides explicit control over adaptation points in the network.
- Transport protocol
  - Rate-based to congestion based
  - Shorter feedback loops
- Traffic engineering
  - Map between provider-specific DiffServ Code Points / VLANs
- Authorization and Authentication
  - Rich expression of policy via e.g. the Security Assertion Markup Language (SAML)
Phoebus Signaling

- Phoebus speaks to the control plane to provision network resources
  - Can allocate circuits from the OSCARS IDC
    - Which underlies ION
  - Also, direct communication with DRAGON

- Once the connection is established to the Phoebus node, traffic can begin to flow
  - Could be sent over an existing link if unable to provision

- Phoebus can finish the connection over the commodity network if the allocation times out
The eXtensible Session Protocol (XSP) can be used to manage a multi-layer connection.

Session Layer Protocol

- PSP
- TCP
- Layer2
- TCP
- PSP
Deployment Plans

Internet2 Network - IP Network

- Phoebus nodes at router POPs
  - Internet2 Phoebus Gateway

- Map showing various cities and nodes connected by lines.
  - Key locations include Seattle, WA, Portland, OR, Sunnyvale, CA, Los Angeles, CA, etc.
  - Cities and nodes labeled with their respective addresses.

- Diagram highlighting the Internet2 Network's infrastructure with switching nodes and router nodes.

- Map labeled with "Internet2 Phoebus Gateway."
Phoebus Authentication

- Password
  - SQLite/MySQL/File backends
- Trusted Host/Subnet
- GSI
  - Globus-based
- Anonymous
  - The session has no identifying information
- Accepted authentication handler can be set on a per host/per subnet basis
Implementation - Library

- The client library provides compatibility with current socket applications
  - AF_LSL
- On Linux, LD_PRELOAD is used for function override
  - `socket()`, `bind()`, `connect()`, `setsockopt()`...
  - Allows Un*x binaries to use the system without recompilation
- Prototype working on MacOS X
Implementation - Intercept

- Intercept the TCP connection with IP Tables (on Linux)
- Redirect to local forwarding process
- Establish connection with appropriate service nodes or end node
  - Based on policy
- Transparent to end hosts
Phoebus XIO Driver

- Provides a modular Phoebus transport driver for use with the Globus Toolkit
- Based on the TCP XIO driver
- Simplifies use of Phoebus Gateways
  - Eliminates need for shim library
Phoebus and GridFTP

- `globus-gridftp-server` loads the Phoebus XIO driver when requested
- `globus-url-copy` extended to support Phoebus-based transfers
  - with `-ph` flag or explicitly with `–dcstack`
- Support for advanced features
  - 3rd party transfers
  - Parallel streams
Windows Support

- SOCKS proxy support in development
  - Working in the lab
- Java COGKit GridFTP
  - globus-url-copy
- Firefox Plugin jTopaz
  - available, uses Java implementation
50ms Latency, .001% loss
50ms Latency, .01% loss
Acknowledgements

 UD Students
  ✷ Ezra Kissel
    ✷ Matt Rein, Jason Zurawski, Omer Arap

 Internet2:
  ✷ Aaron Brown, Guy Almes (now at Texas A&M), Eric Boyd, Rick Summerhill, John Vollbrecht, Matt Zekauskas, Jeff Boote

 US Department of Energy Office of Science, Mathematical, Information and Computational Sciences (MICS) Program
  ✷ Early Career Principal Investigator program
End

🌟 Thank you for your attention
🌟 Questions?
TCP Overview

- TCP provides reliable transmission of byte streams over best-effort packet networks
  - Sequence number to identify stream position inside segments
  - Segments are buffered until acknowledged
  - Congestion (sender) and flow control (receiver) “windows”
  - Everyone obeys the same rules to promote stability, fairness, and friendliness
- Congestion-control loop uses ACKs to clock segment transmission
  - Round Trip Time (RTT) critical to responsiveness
- Conservative congestion windows
  - Start with window $O(1)$ and grow exponentially then linearly
  - Additive increase, multiplicative decrease (AIMD) congestion window based on loss inference
  - “Sawtooth” steady-state
  - Problems with high bandwidth delay product networks

\[ BW = \frac{mss \cdot C}{rtt \sqrt{p}} \]
The End to End Arguments

Why aren’t techniques like this already in use?

Recall the “End-to-End Arguments”

- E2E Integrity
  - Network elements can’t be trusted
  - Duplication of function is inefficient
- Fate sharing
  - State in the network related to a user
- Scalability

Network transparency
- Network opacity

The original assumptions regarding network scalability and complexity may not hold true any longer