SCTP

How can SCTP and High-Speed networking go together?

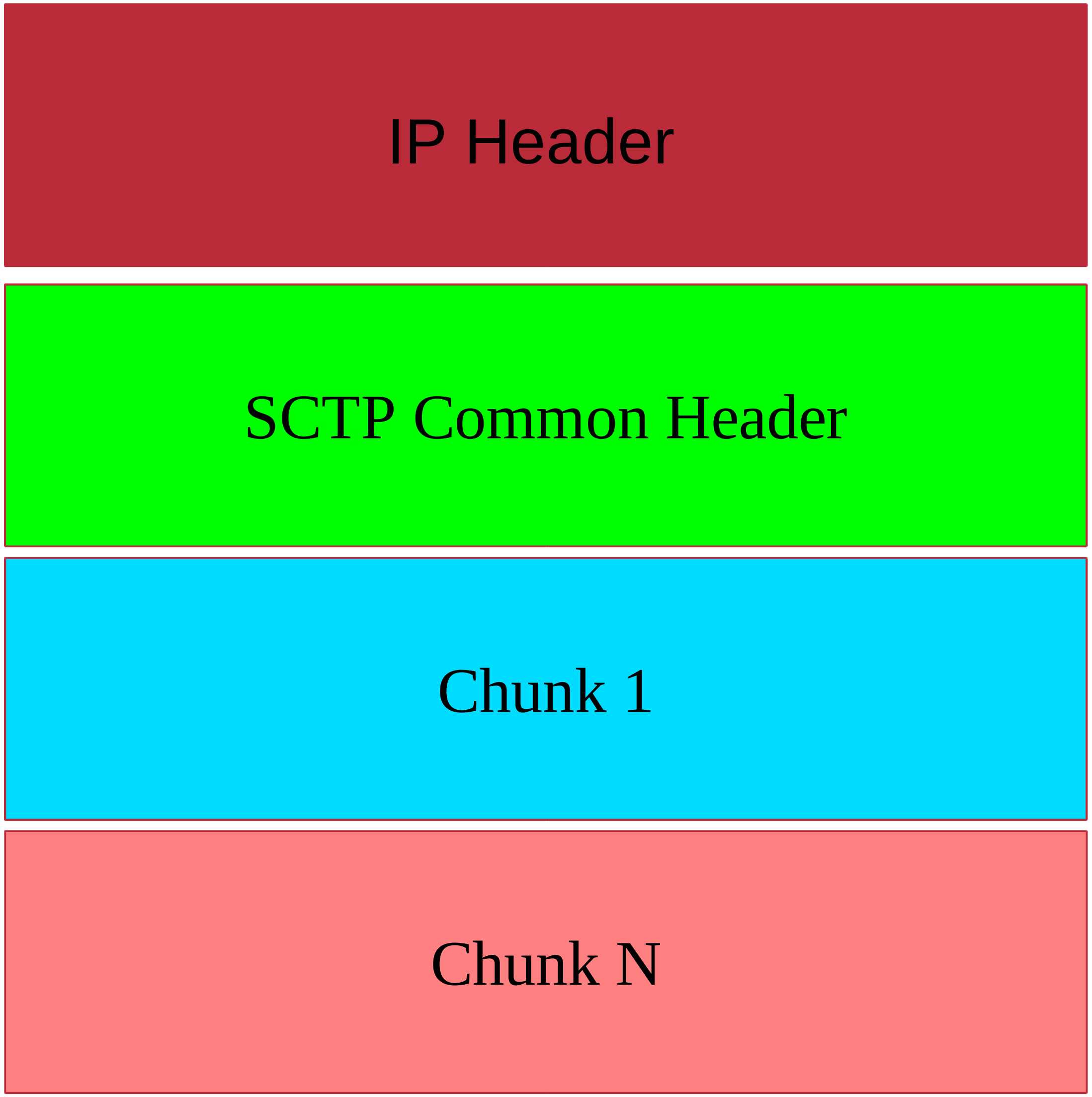
Talk Objectives: What I plan to cover

- A feature synopsis of SCTP, the what you get
- A brief survey on SCTP, how it works in a nut-shell
- A view into why one might want to use SCTP for high speed networking, the why
- And an interesting look at a feature currently rolling out that uses SCTP

Features of SCTP

- Reliable data transfer w/SACK
- Congestion control and avoidance (TCP friendly)
- Message boundary preservation
- PMTU discovery and message fragmentation
- Message bundling
- Multi-homing support
- Multi-stream support
- Unordered data delivery option
- Security cookie against connection flood attack (SYN flood)
- Built-in heartbeat (reachability check)
- Extensibility

SCTP Packet With IP Header



IP Header

SCTP Common Header

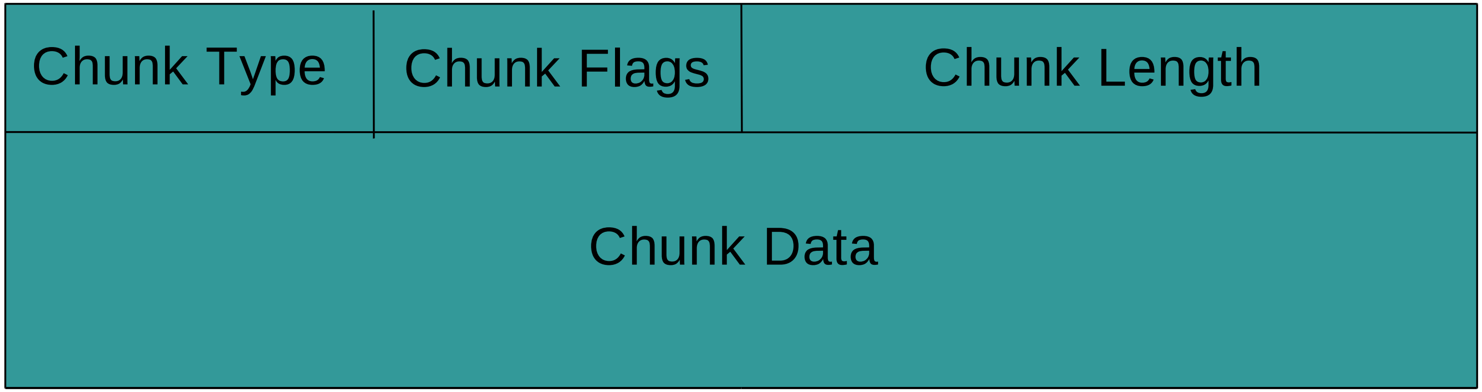
Chunk 1

Chunk N

SCTP Common Header

<table>
<tr><td>Source Port</td><td>Destination Port</td></tr>
<tr><td colspan="2">Verification Tag</td></tr>
<tr><td colspan="2">CRC-32c Checksum</td></tr>
</table>

SCTP Chunks



Chunk Type

Chunk Flags

Chunk Length

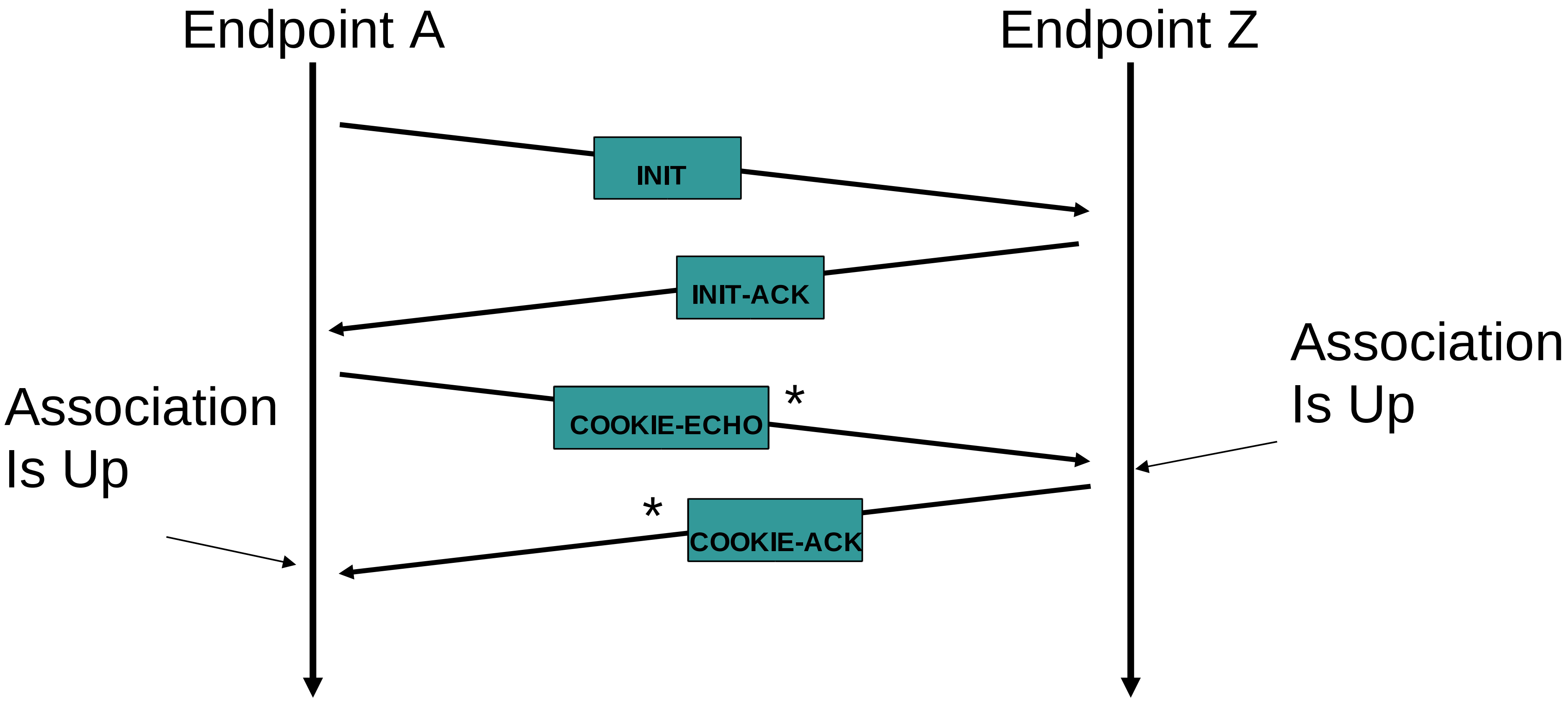
Chunk Data

SCTP Chunk Header Fields

- Chunk Type: 8-bit value indicating the type of chunk
- Chunk Flags: 8-bit flags, defined on per chunk type basis
- Chunk Length: 16-bit length in bytes, including the chunk type, chunk flags, and chunk length fields.

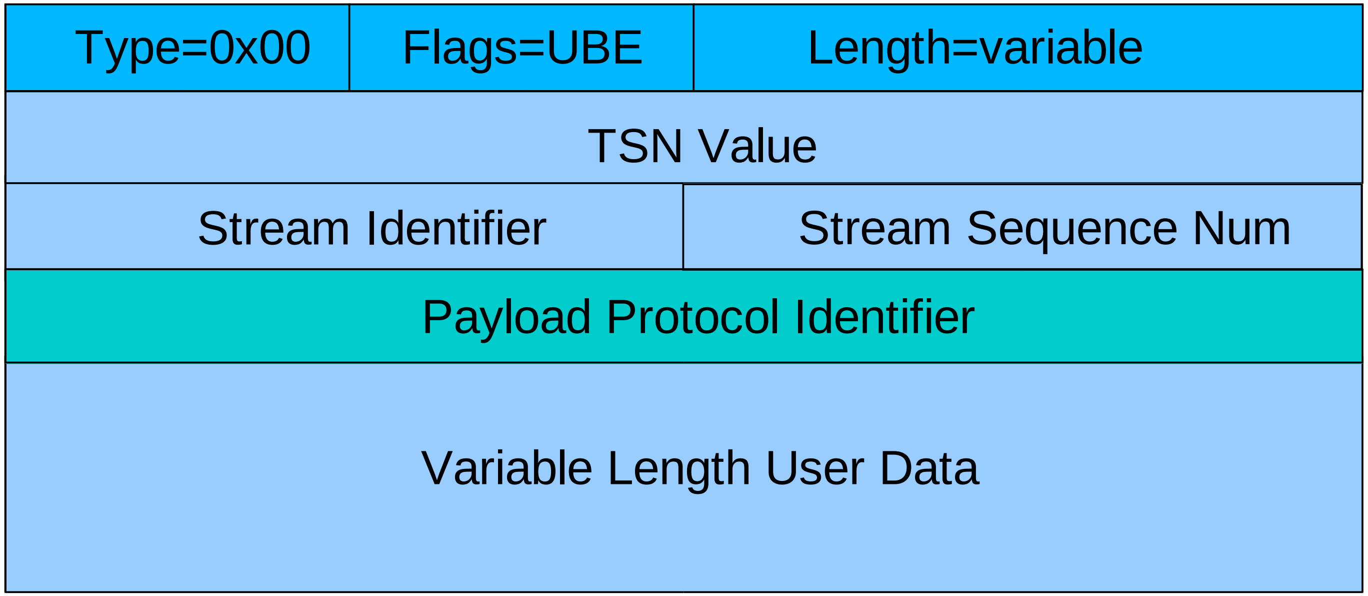
  Note that chunks are padded to 32-bit boundaries within an SCTP packet. Any padding bytes (0x00) used are NOT included in the chunk length

Setting Up an Association



* -- User data can be attached

DATA Chunk



Type=0x00	Flags=UBE	Length=variable
TSN Value		
Stream Identifier	Stream Sequence Num	
Payload Protocol Identifier		
Variable Length User Data		

- **Flag Bits ‘UBE’ are used to indicate:**

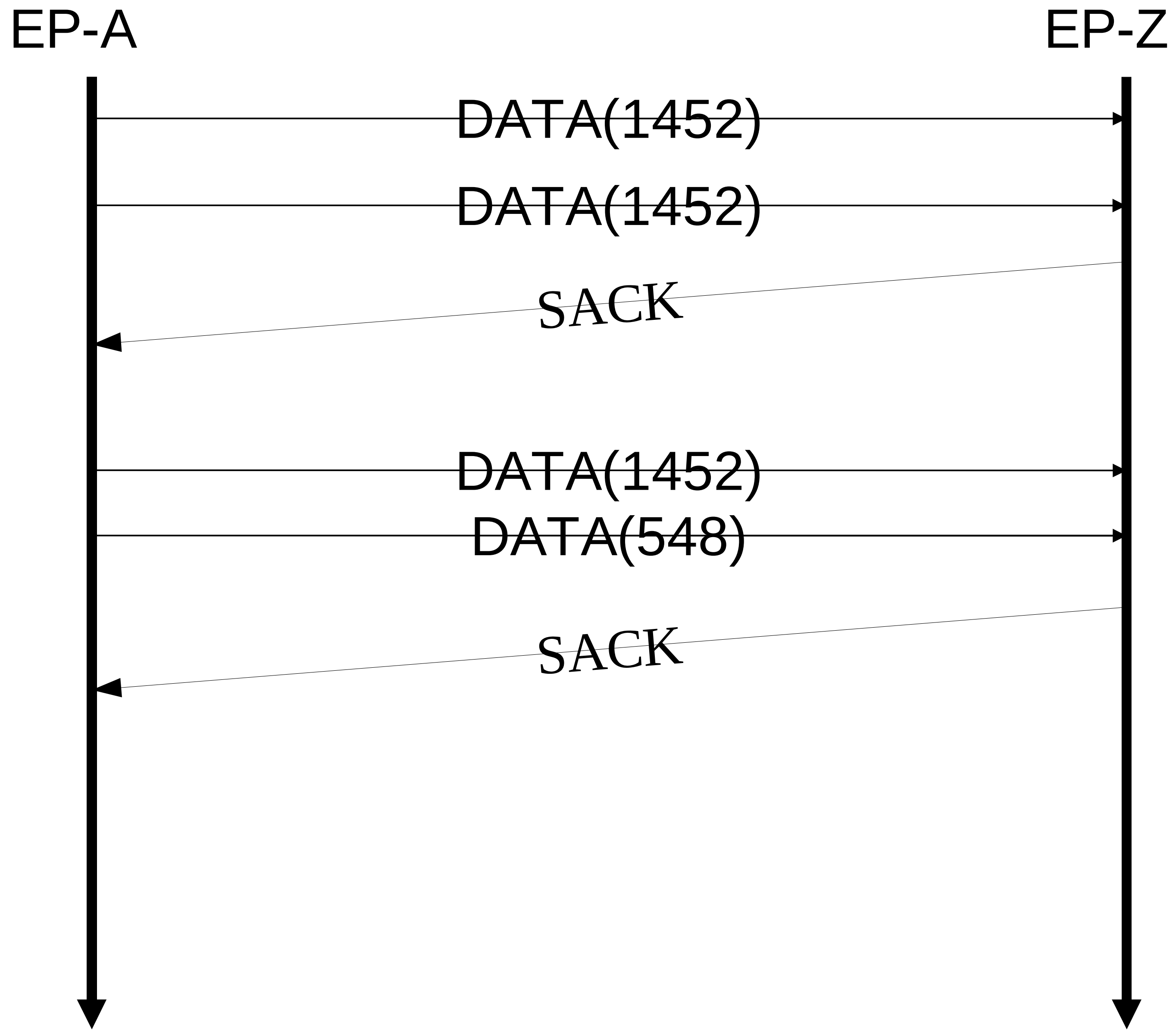
  U – Unordered Data

  B – Beginning of Fragmented Message

  E – End of Fragmented Message

- **A user message that fits in one chunk would have both the B and E bits set**

Data Transfer Example

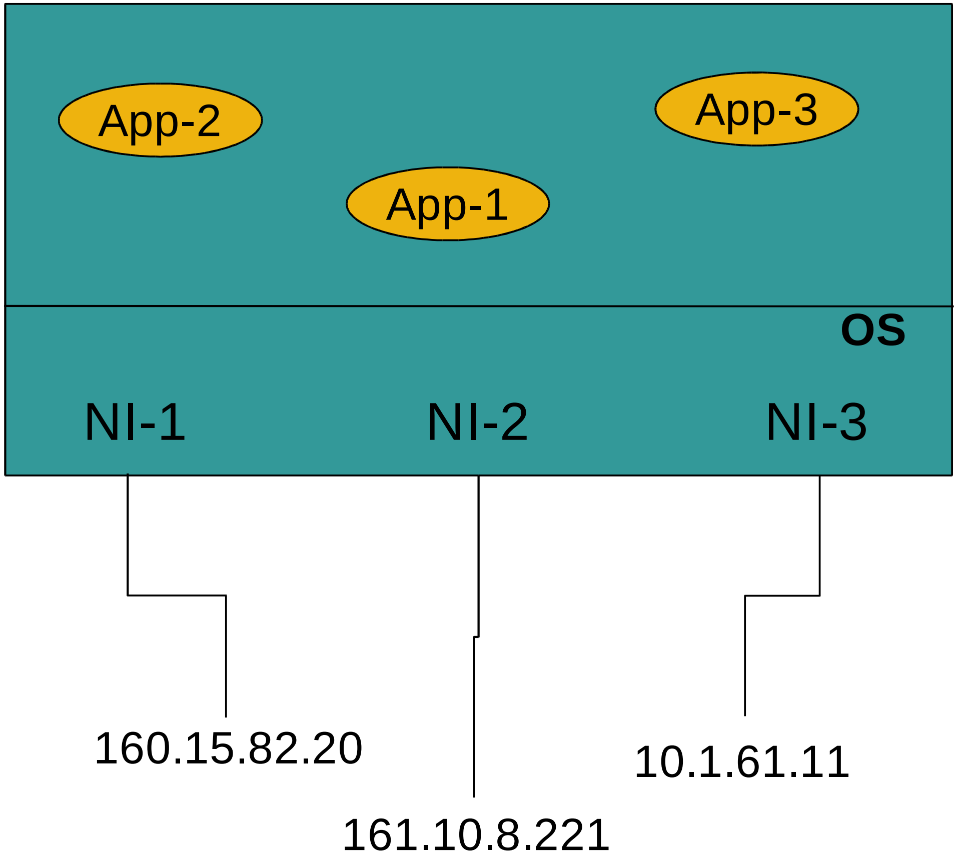


SCTP Streams

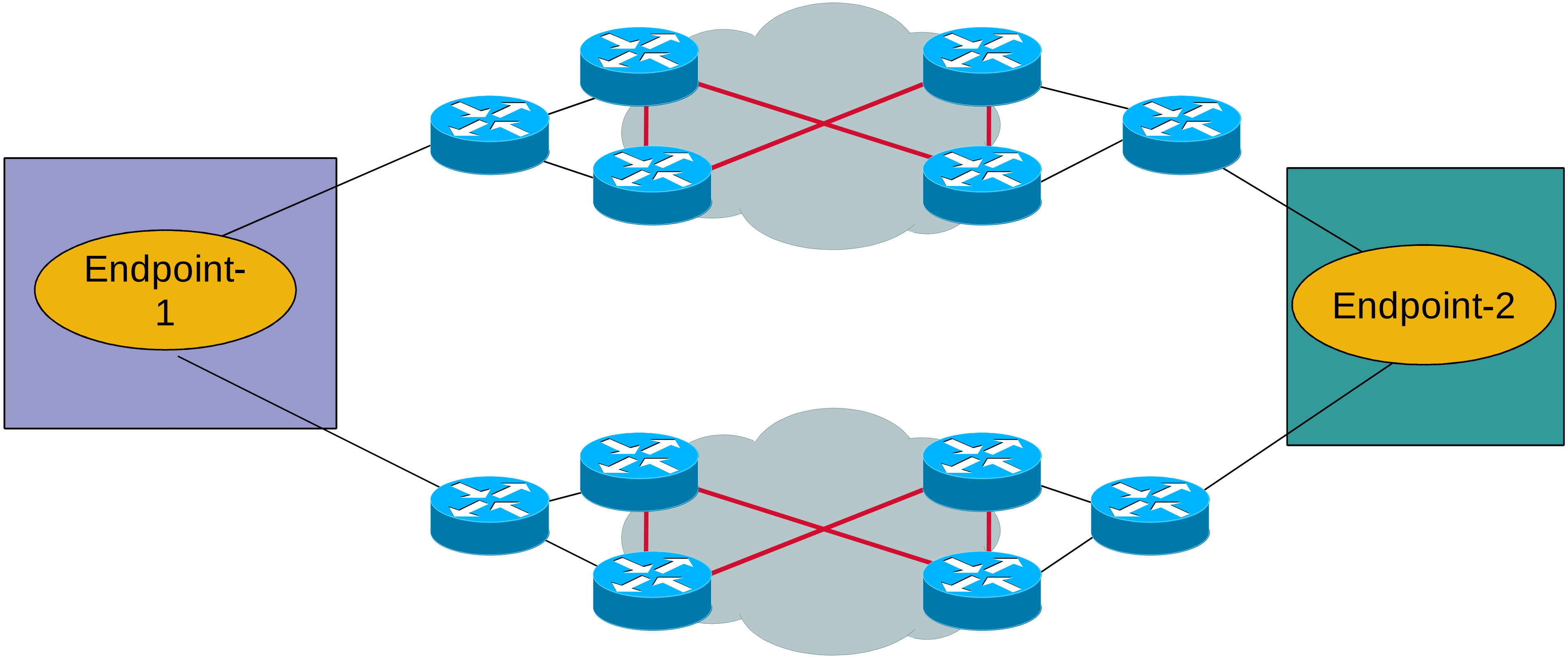
- **Streams separate ordering from retransmission: streams are used for message ordering while the Transport Sequence Number (TSN) is used for retransmission.**
- **Each Data Chunk holds a Stream Identifier and a Stream Sequence Number in addition to the TSN.**
- **Messages sent in different streams are delivered without respect to one another.**
- **This allows an application to avoid head of line blocking if data loss occurs.**

IP Multi-homing

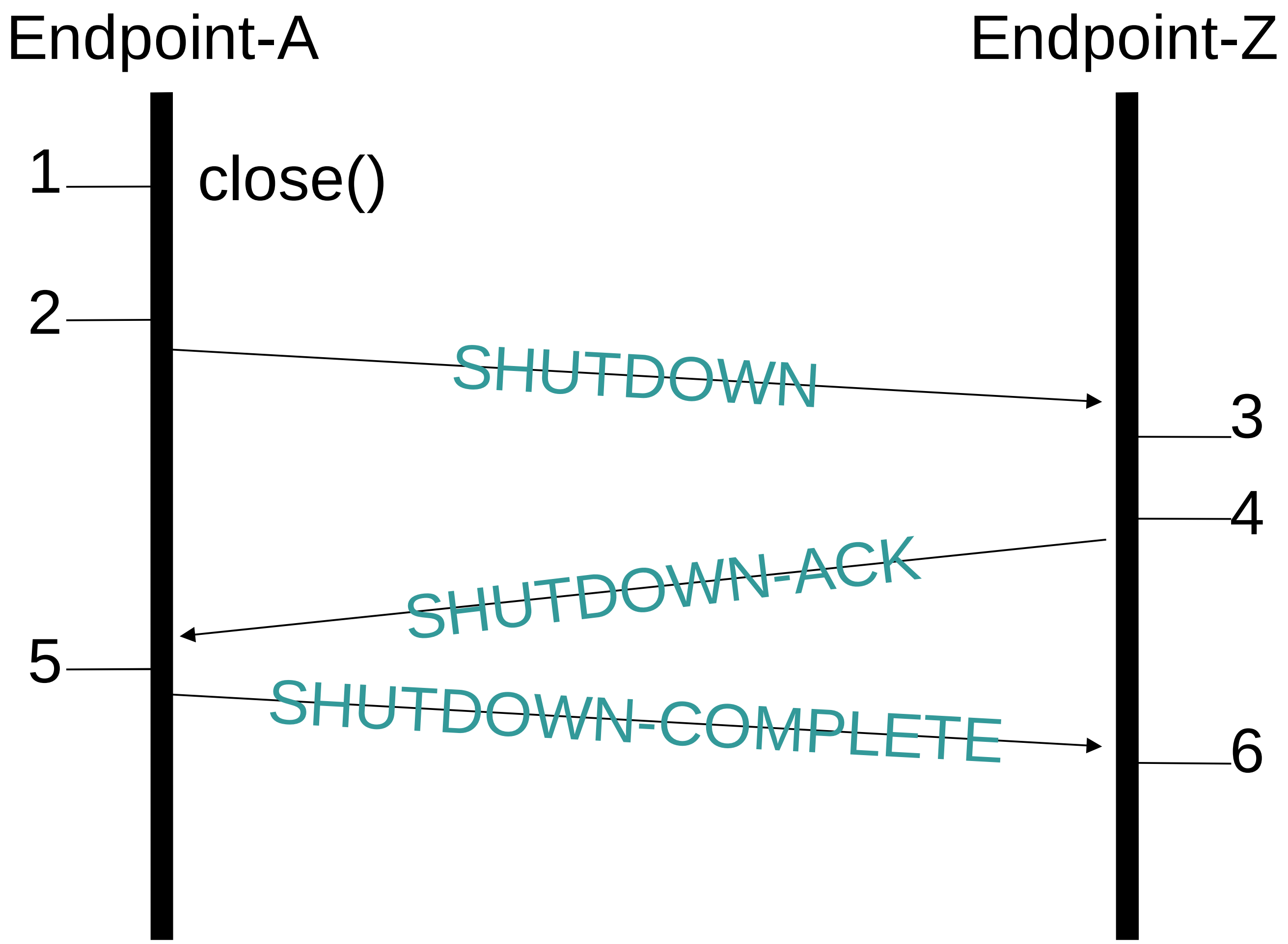
- **An SCTP association may encompass more than one IP address**



Maximum Path Diversity



The Shutdown Handshake



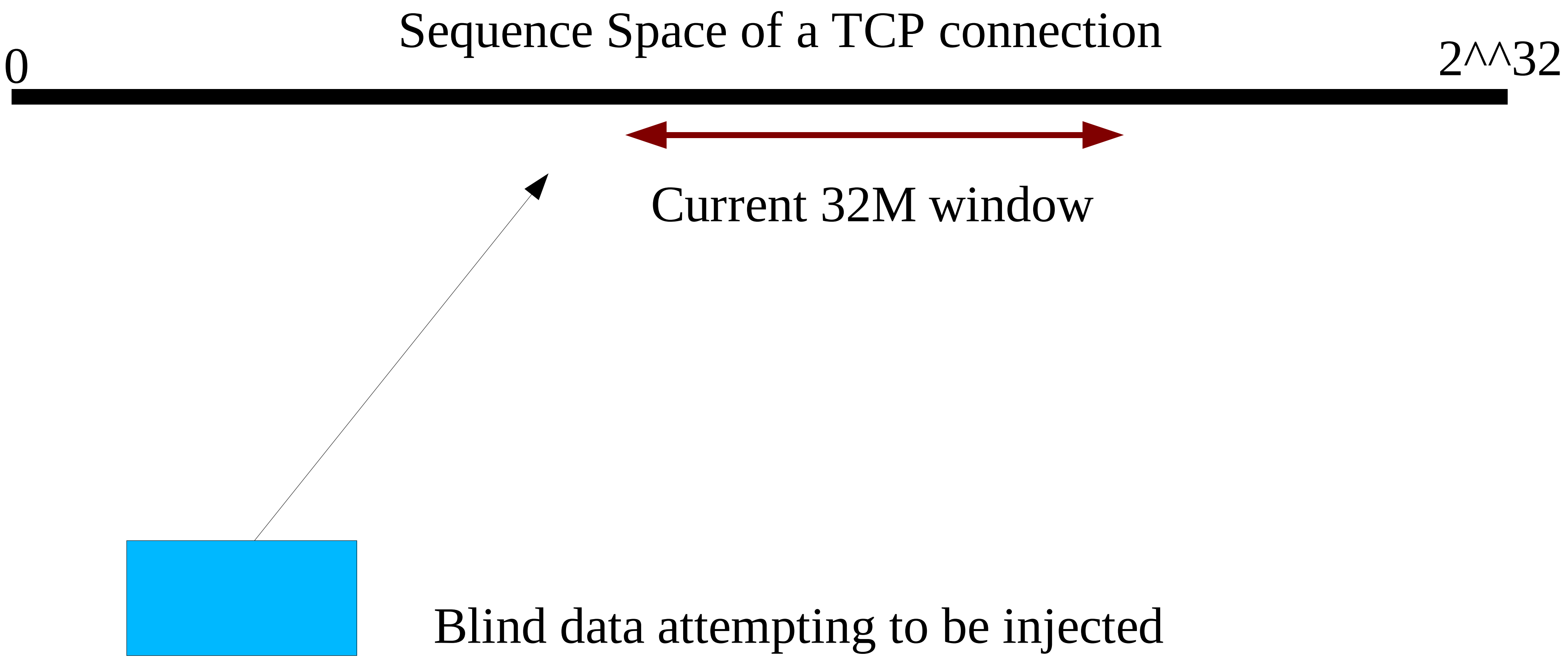
So what can and cannot SCTP do for HS networking?

- **SCTP uses the same congestion control mechanisms as TCP thus many of the same problems that are applicable to TCP and HS networking are resident in SCTP**
- **The same “fixes” we come up with for TCP can be applied to SCTP, for example some SCTP implementations already have implemented Sally Floyd's High-Speed TCP algorithm.**
- **But is there more that SCTP can offer than just cloning TCP fixes?**

More from SCTP

- SCTP provides message boundaries, for RDMA and zero copy networking solutions this is a boon.
- SCTP provides a stronger checksum (CRC32c), this both adds cost in terms of CPU but also strength in terms of data integrity. Offload engines will need to be developed for CRC32c if SCTP is to compete in the HS networking world (some of these are already being offered by Intel – 80314).
- SCTP provides a 32 bit window (rwnd) so no window scaling is needed.
- SCTP provides stronger security against blind attacks, consider the next slide:

A blind attack


Sequence Space of a TCP connection
0
2^^32
Current 32M window
Blind data attempting to be injected

Blind Attack

- A Blind attacker need only guess 128 sequence numbers to find an acceptable one which would be injected into a connection .
- Most TCP implementations will only throw out packets when the ACK value is ahead of the current sndnxt value.
- This means that an attacker need only guess 2 ACK values for every packet.
- Thus, a blind attacker only needs to generate 256 packets to penetrate a HS-TCP connection.
- Note this of course assumes that the port numbers can be guessed (which is usually quite easy).

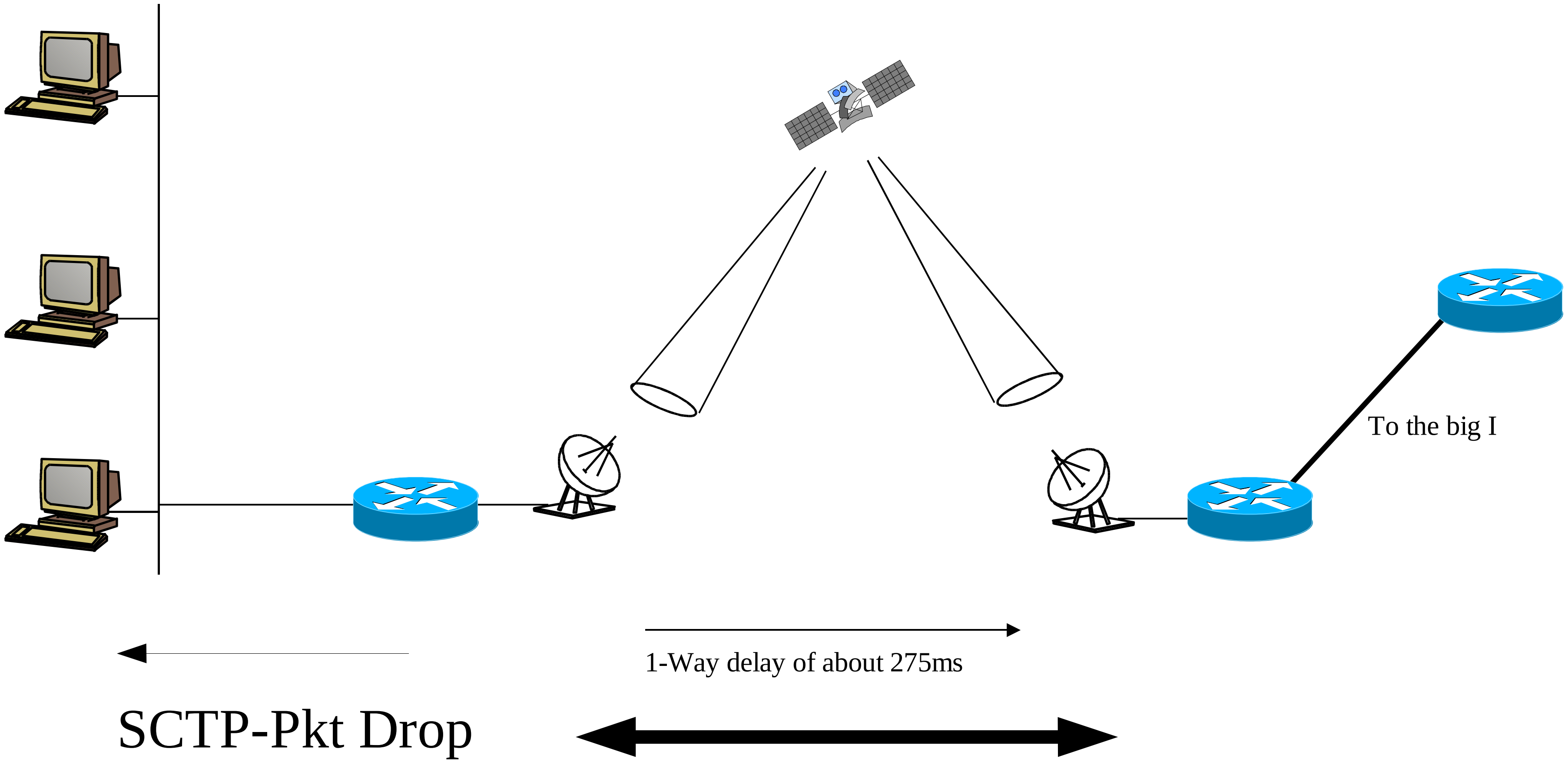
So how does SCTP deal with this?

- An attacker needs to first of all guess the V-Tag, a 32 bit random value, to penetrate an association.
- This yields a 1 in 4 billion chance of a blind data packet being injected, or the need to send 4 billion packets.
- Add to this that an SCTP TSN (the unit for reliability in SCTP) does not correspond to the window, but to chunks. This means an attacker has even more randomness to guess (about another 16 bits of random numbers using our 32Meg window and a 1500 byte MTU size).
- This gives SCTP a considerable advantage in defense against blind attacks over TCP.

Extensibility

- SCTP provides for extensibility via the chunk and parameter types.
- This extensibility is designed similar to IPv6 so that basic SCTP implementations can still properly inter-operate with one that has an extension.
- This extensibility means that if a HS networking idea needed to be incorporated into SCTP it could be facilitated with ease.
- An example of this extensibility can be found in some of the satellite work we are doing at Cisco.

Errors vs Congestion



To the big I

1-Way delay of about 275ms

SCTP-Pkt Drop

Cisco RBSCP

SCTP Packet Drop

- Cisco's RBSCP protocol distinguish packet drops due to congestion vs due to link errors.
- When an SCTP packet is dropped due to error, a report is sent to the sender so that retransmissions can occur without a cwnd adjustment.
- Normal SACK's traveling over the satellite link from the remote end can also get stamped with a “bandwidth report” which allows an SCTP sender to make faster cwnd adjustments both up and down.

Summary

- SCTP offers some unique features that COULD be used by the HS networking community.
- Some work needs to be done in CRC offload so that SCTP's stronger data integrity will not slow it down when working in HS networks.
- SCTP's extensibility can provide a pathway to the future that will allow improvements for HS networking!

Reference Materials

- [SCTP reference book] Stream Control Transmission Protocol (SCTP): A Reference Guide, R. Stewart and Q. Xie, Addison-Wesley, 2002, ISBN 0-201-72186-4
- RFC 2960: Stream Control Transmission Protocol, October 2000
- RFC 3309: SCTP Checksum Change, September 2002
- [I-G] draft-ietf-tsvwg-sctpimpguide-10: SCTP Implementer's Guide

SCTP Programming References

- [sockets API] draft-ietf-tsvwg-sctpsocket-07: Sockets API Extensions for SCTP
- UNIX Network Programming, Volume 1, Third Edition, Stevens-Fenner-Rudoff, Addison-Wesley, 2004, ISBN 0-13-141155-1

SCTP Extensions Drafts

- [PR-SCTP] draft-ietf-tsvwg-prsctp-03: SCTP Partial Reliability Extension (soon to be RFC)
- [Add-IP] draft-ietf-tsvwg-addip-sctp-08: SCTP Dynamic Address Reconfiguration
- [Pkt-Drop] draft-stewart-sctp-pktdrprep-00: SCTP Packet Drop Reporting
- [Auth] draft-tuexen-sctp-auth-chunk-00: Authenticated Chunks for SCTP

Online References

- http://www.sctp.org

  Also reachable with HTTP over SCTP!

- http://www.ietf.org/html.charters/tsvwg-charter.html

  All current work on SCTP is done in the IETF TSVWG

- sctp-impl on mailer.cisco.com

Closing Questions

- **Questions?**