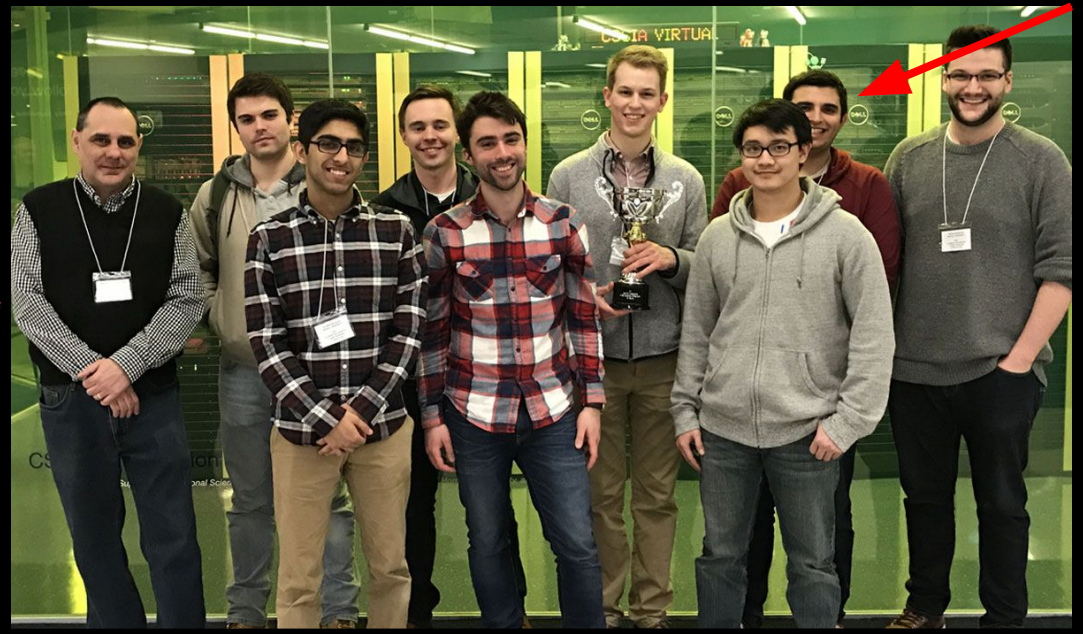
BGP Fundamentals

Aaron Atac - Subspace

NANOG™

About Me

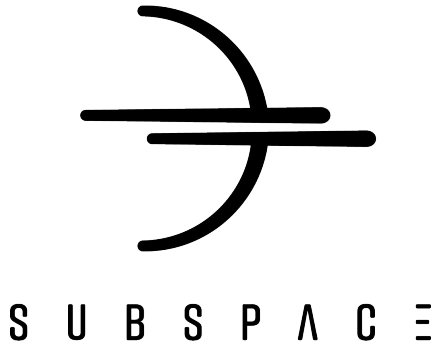


About Me


NANOG 67
Chicago, IL
June 13-15, 2016


NANOG 68
Dallas, TX
October 17-19, 2016

Where I've been and where I am now



ORD

July 2018

LAX

Now

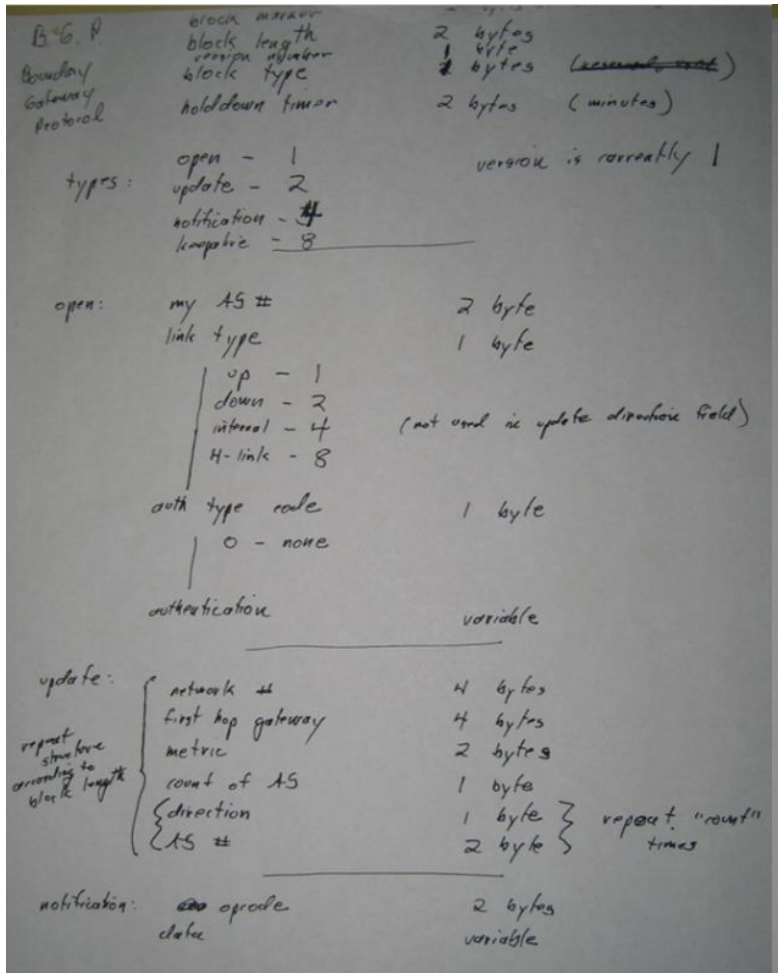
BGP (Border Gateway Protocol)

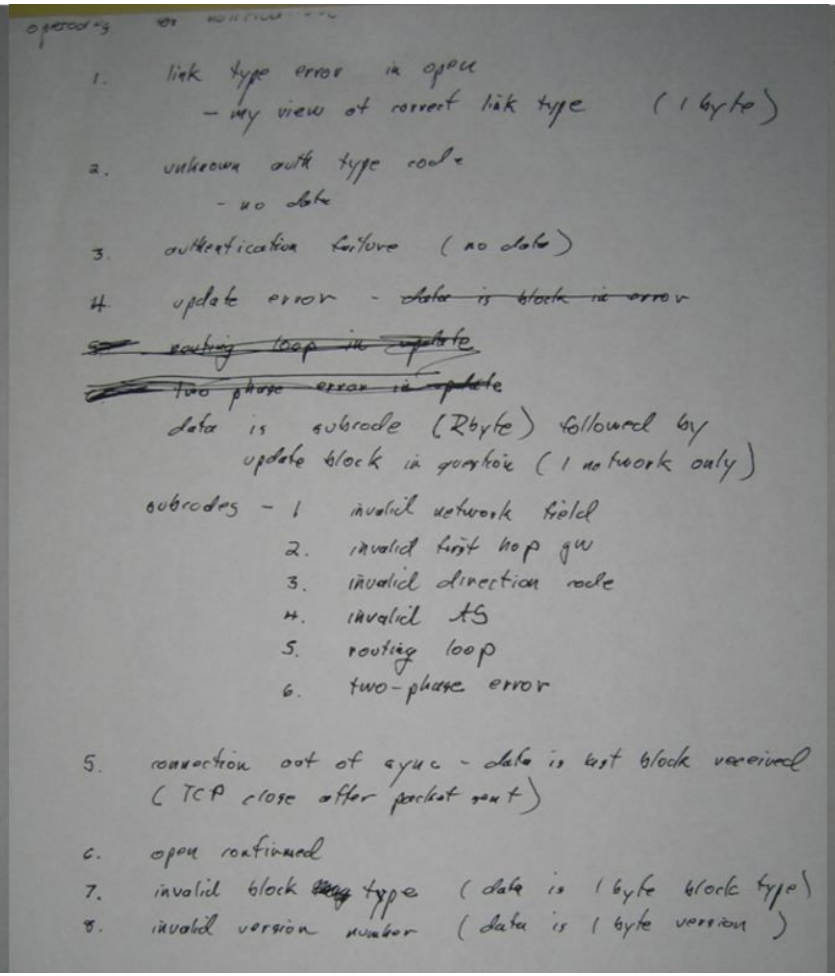
Originally sketched over lunch by Kirk Lougheed and Yakov Rekhter, with the help of Len Bosack at IETF 12, January 1989.

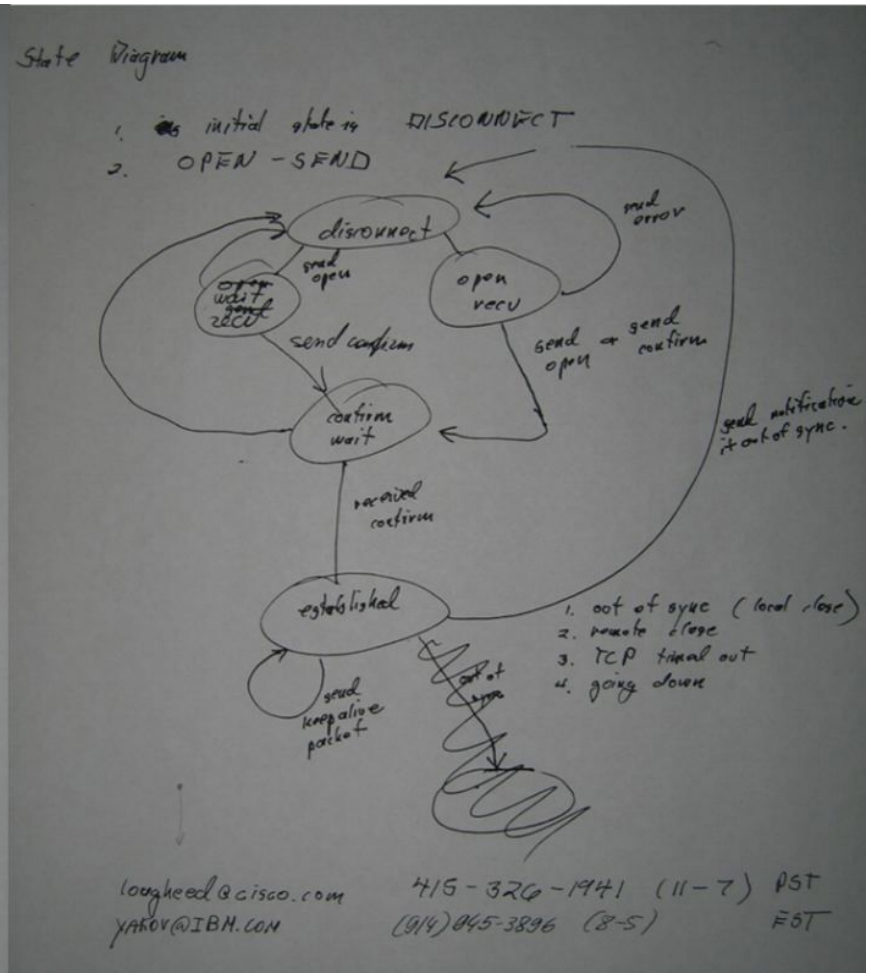
32 years folks.

- AKA TNP (Three Napkin Protocol) - Design by pragmatism
- Original napkins apparently got too much ketchup on them, they then wrote these three pieces of paper:

BGP (Border Gateway Protocol)

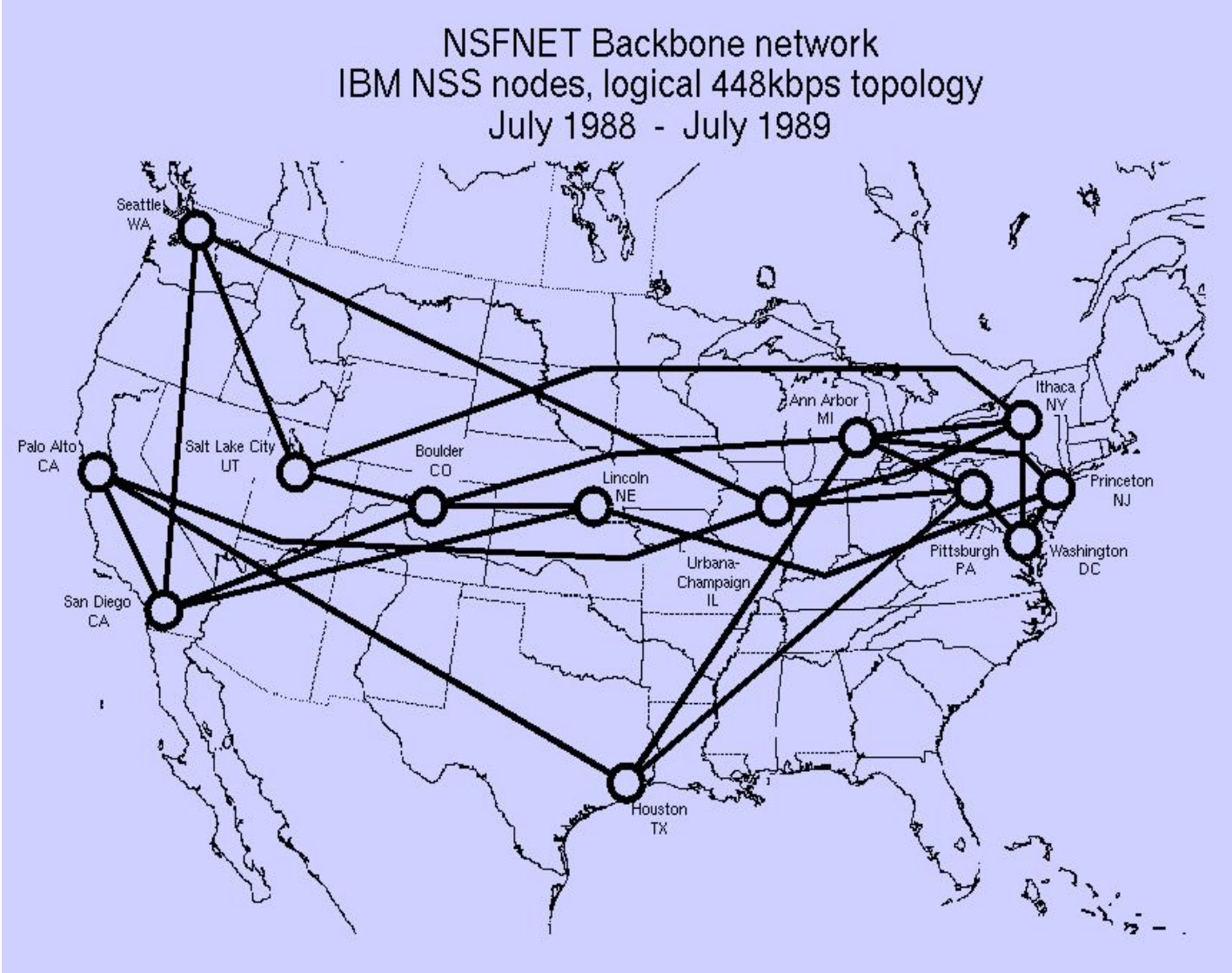






The network at that time

- 13 Nodes
- 170 Networks
- By July 1988 all were interconnected at 1.5 Mbps (T-1)
- IBM RT PCs connected by token ring IEEE 802.5



Claffy, Kimberly C.; Braun, Hans-Werner; Polyzos, George C. (August 1994). "Tracking long-term growth of the NSFNET". *Communications of the ACM*.

NANOG™

What and why?

- EGP and Policy Based Routing in the New NSFNET Backbone RFC1092 - February 1989
  - EGP as reachability protocol
  - “It should be noted that the use of EGP is only viewed as an interim measure until better inter autonomous system protocols are defined and widely deployed for gateways used by regional networks.”
  - “The EGP model assumes an engineered spanning tree topology, however, the NSFNET (due to the presence of backdoor routes) does not fit into this model.”

What and why?

- The NSFNET Routing Architecture RFC1093 - February 1989
  - “In the longer run the hope is to replace the EGP interface with a new inter Autonomous System protocol. Such a new protocol should also allow to move the filtering of network numbers or Autonomous Network number groups to the regional gateways in order for the regional gateways to decide as to what routing information they wish to receive.”

Fundamentals of BGP-1

- A Border Gateway Protocol (BGP) RFC1105 - June 1989
- Incremental updates instead of periodic
- Use TCP/179 unicast as a reliable transport
- Have Autonomous Systems (AS) make up AS_PATHs
    - Used to provide information and prevent loops

Fundamentals of BGP-2

- A Border Gateway Protocol (BGP) RFC1163 - June 1990
- Path Attributes
  - Mandatory vs optional, transitive vs non-transitive attributes
  - ORIGIN, AS_PATH, and NEXT_HOP are the only mandatory attributes.
- “The notion of Up/Down/Horizontal relations present in RFC1105 has been removed from the protocol.”
- Marker field

Fundamentals of BGP-3

- A Border Gateway Protocol 3 (BGP-3) RFC1267 - October 1991
- Optimize and simplify the exchange of information about previously reachable routes
- Connection Collision Detection

Fundamentals of BGP-4

- A Border Gateway Protocol 4 (BGP-4) RFC1771 - March 1995
- Support CIDR!
  - Encode reachability information as variable length prefixes instead of fixed length based on classful networks
- LOCAL_PREF attribute

The Basics of Paths

- The mandatory attributes that have to be propagated
  - ORIGIN: IGP or EGP or Unknown
  - AS_PATH: Shortest path wins
  - NEXT_HOP
- Others
  - LOCAL_PREF: Highest wins
  - MULTI_EXIT_DISC (MED): Lowest wins
  - Communities: Value that you can filter on.
    - Some are “well-known” meaning they have been defined and reserved.
    - Examples: NO_ADVERTISE, NO_EXPORT, NO_PEER, BLACKHOLE, GRACEFUL_SHUTDOWN.

What were some of the original problems?

- Scalability / iBGP full-mesh requirement
  - Fix Autonomous System Confederations for BGP RFC1965 - June 1996
    - Partition AS into sub-ASes
      - Only mesh within sub-AS
  - BGP Route Reflection - An alternative to full mesh IBGP RFC1966 - June 1996
    - Hub-and-spoke

What were some of the original enhancements?

- Multiprotocol Extensions for BGP-4 RFC2283 - February 1998
  - BGP-4 was only carrying routing information for IPv4
- Protection of BGP Sessions via the TCP MD5 Signature Option RFC2385 - August 1998
  - Marker field introduced in BGP-2
    - Limited value, under-specified, under-implemented

What were some of the original enhancements?

- Capabilities Advertisement with BGP-4 RFC2842 - May 2000
  - Version number wasn't going to be efficient.
  - Why we still are on BGP-4 in February of 2021
  - BGP speaker telling the other side its supported features
- Route Refresh Capability for BGP-4 RFC2918 - September 2000
- Carrying Label Information in BGP-4 RFC3107 - May 2001

Newer Enhancements

- Graceful Restart Mechanism for BGP RFC4724 - January 2007
- BGP Support for Four-octet AS Number Space RFC4893 - May 2007
- The TCP Authentication Option RFC5925 - June 2010
- An Infrastructure to Support Secure Internet Routing RFC6480 - February 2012
- A Profile for Route Origin Authorizations (ROAs) RFC6482 - February 2012
- BGP Prefix Origin Validation RFC6811 - January 2013

Newer Enhancements

- BGP Monitoring Protocol (BMP) RFC7854 - June 2016
- Advertisement of Multiple Paths in BGP RFC7911 - July 2016
- BGP Administrative Shutdown Communication RFC8203 - July 2017
- Default External BGP (EBGP) Route Propagation Behavior without Policies RFC8212 - July 2017
- Graceful BGP Session Shutdown RFC8326 - March 2018
- Support for Adj-RIB-Out in the BGP Monitoring Protocol (BMP) RFC8671 - November 2019
- Extended BGP Administrative Shutdown Communication RFC9003 - January 2021

Filtering and Validity of Paths

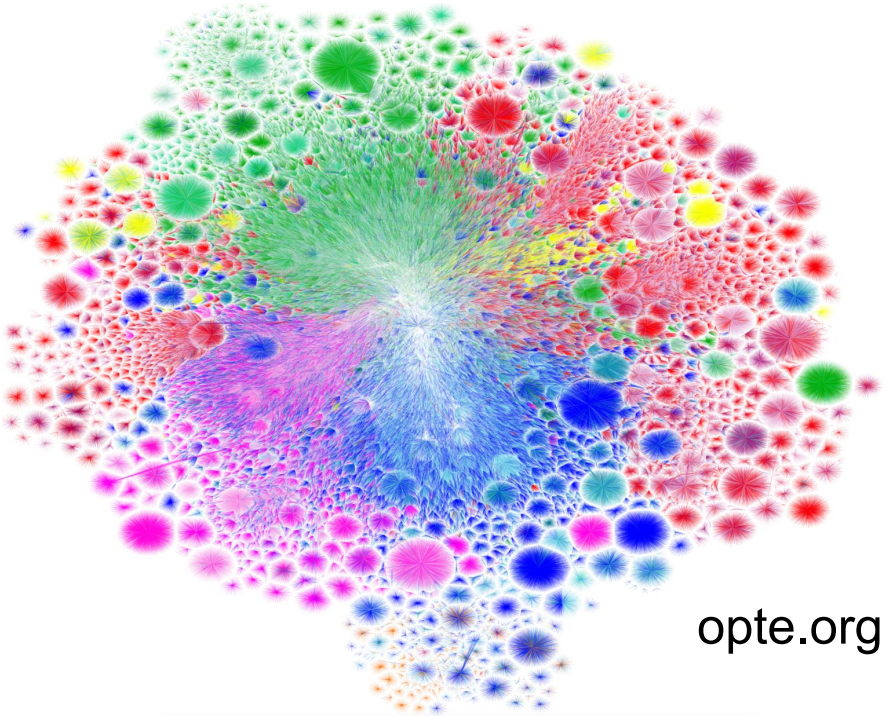
- Received Prefixes
  - You don’t have to accept everything you receive.
  - You can modify what you receive.
    - Example: Change the ORIGIN of a path matching a certain community.
  - Different networks will give you different views, and these views may not be all that they seem.
    - Views may contain short paths, but they may be geographically distant.
    - Views may contain effects of others misconfigurations.
    - Views may contain lossy paths.
    - Views may contain hijacked and/or leaked paths.

Filtering and Validity of Paths

- What to do with these received prefixes and views of all these different paths?
  - Filter malicious and illegitimate (aka bogon) routes
  - Apply business logic and traffic engineering desired
  - Don’t leak them (unintentionally announcing prefixes to a different network that will cause undesirable traffic shifts)
- IRR/route/route6/AS-SET, PeeringDB, bgpq4, and prefix lists
- BGP Prefix Origin Validation RFC6811 - January 2013
- NLNOG BGP Filtering Guide

The Land of ISPs

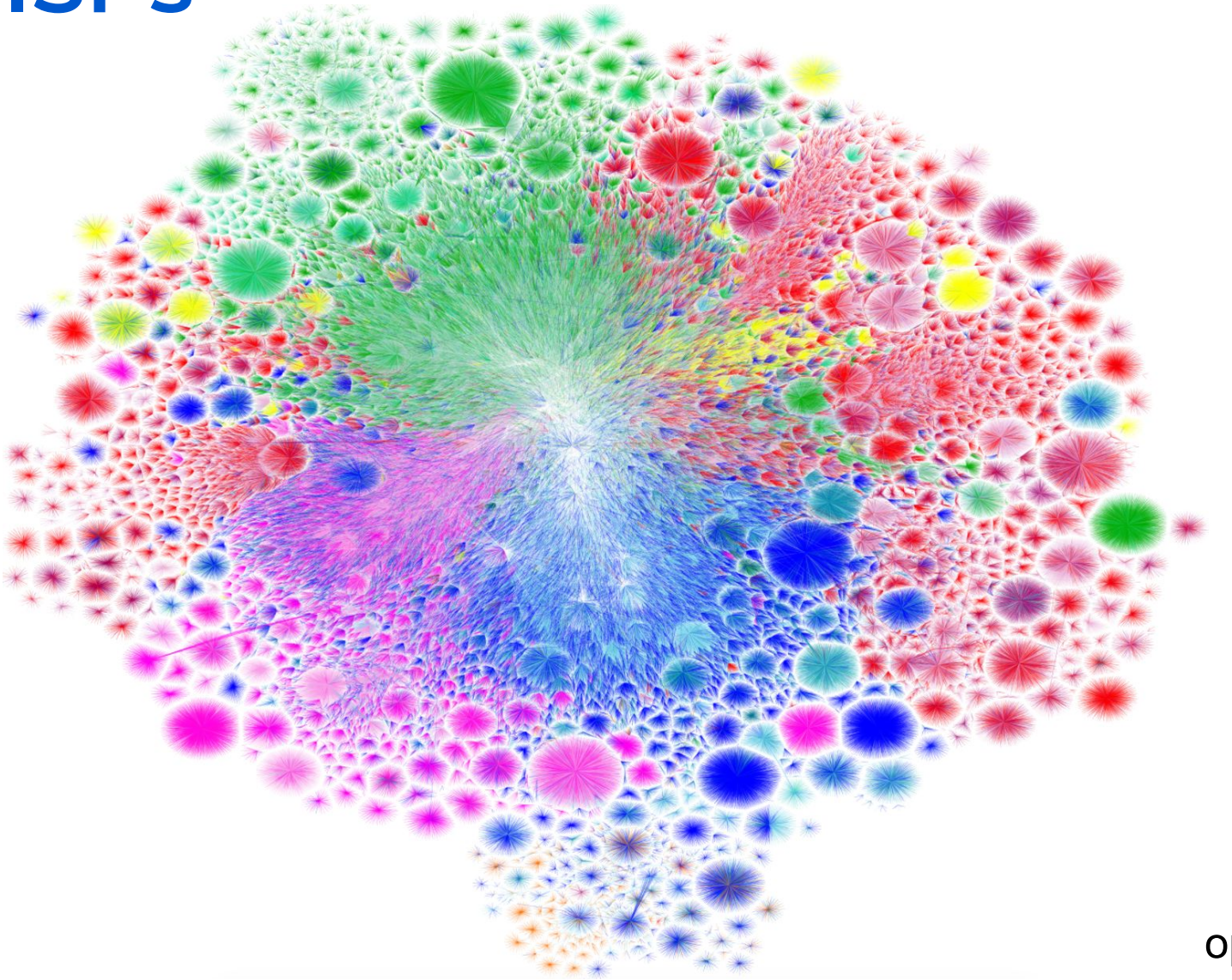
- IX aka Internet Exchange
  - Internet Exchange BGP Route Server RFC7947
- PNI/SFI:
  - Private Network Interconnect
  - Settlement Free Interconnect
- IP Transit aka Transit aka DIA aka Direct Internet Access
- The more you peer directly with networks the less you will need to use your transit connection.



opte.org

Asia Pacific
Europe/Middle East/Central Asia/Africa
North America
Latin American and Caribbean
RFC1918 IP Addresses
Unknown

The Land of ISPs



Asia Pacific
Europe/Middle East/Central Asia/Africa
North America
Latin American and Caribbean
RFC1918 IP Addresses
Unknown

Routing Policy

- Announcing and receiving prefixes sounds so simple…yet it’s not when you look under the hood at what the decision-making humans do to influence paths.
- Traffic engineering applies both to how you announce and how you receive.

External Peering and Traffic Engineering

- Most specific always wins
- ORIGIN
- MED
- Communities
- AS_PATH prepending
  - BGP assumes shortest AS_PATH is best, not always the case
  - AS_PATH length doesn't always compare with geographic distance
  - Congestion
  - Loss
  - Return path
- Maybe I don't actually want to peer with you…at least at this location

Volatility of the DFZ

- Routes flap
  - Route Flap Damping exists, but is often too aggressive and does more harm than good. Vendor default settings are generally too reactive to small offenders.
  - Even if you're bringing down infrastructure it doesn't mean you must withdraw an aggregate. Announce aggregates externally, route to more specific internally.
  - If you really really must, use the RIPE-580 guidelines, not the vendor defaults!
- Routes leak
  - This is unfortunately common and can subject eyeballs to degraded experiences.
- Routes get hijacked
  - Unfortunately common as well, fat-fingering happens, other times it's intentional, serial independent actors or govts/nation-states.

Volatility of the DFZ

- DFZ = Default Free Zone aka BGP Global Routing Table
- AS6447 RouteViews Wed Feb 24 02:20:00 2021 AET
  - IPv4: 903983
  - IPv6: 110993
- Desperately Seeking Default - Geoff Huston - APNIC
  - Poor orchestration and misguided security principles of some organizations and their routing policy can as well lead to the bgp global routing table looking a lot different depending on the feed you're looking at.

Public Information Gathering

- There's some pretty cool public collectors, feeds, and probes out there:
  - University of Oregon Route Views Project
  - RIPE RIS Live, RIPEstat, RIPE Atlas (shout out Global Traceroute)
  - CAIDA BGPStream, ASRank
  - NLNOG RING
  - The Oracle Internet Intelligence Routing 3D Visualization

Q & A

NANOG™

Thank You

NANOG™