Crusoe

Traceroute
A brief overview

Ruairí Carroll

Agenda

Why am I making this presentation?

- Long time unhealthy obsession with traceroute
- Spent 2024 trying to build a startup around traceroute and monitoring
- Going to try impart some of what I learned over the years here today
- It’s been a few years since RAS last gave a talk about traceroute

A quick trip down memory lane

- Created by Van Jacobson in circa 1989
  - v1.0 Tue Feb 28 23:50:05 PDT 1989
- Relied on the very recently created raw sockets patch which enabled ping
- Showed up first in 4.3-BSD-Reno

A quick trip down memory lane

BSD Man Page

↗

Implemented by Van Jacobson from a suggestion by Steve Deering. Debugged by a cast of thousands with particularly cogent suggestions or fixes from C. Philip Wood, Tim Seaver, and Ken Adelman.

Original Version - Overview

- Probes are UDP based
- Used TTL field in IP Packet to enumerate each host
- “Seq” (index of probes sent) used to increment Dst Port
- RTT calculated as time received - time sent, in milliseconds

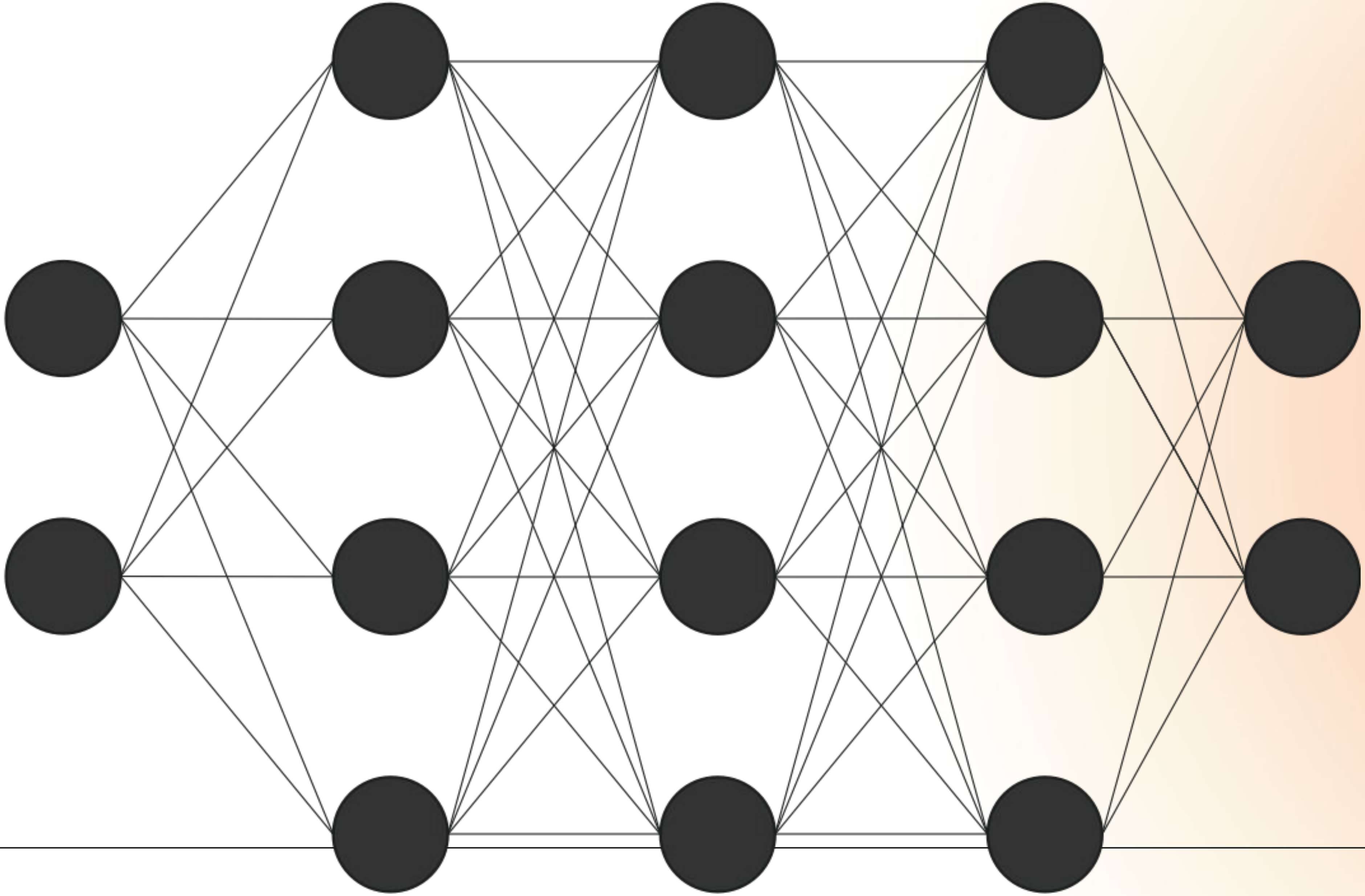
Original Version - Overview

The very first traceroute (never released) used ICMP ECHO_REQUEST datagrams as probe packets. During the first night of testing it was discovered that more than half the router vendors of the time would not return an ICMP TIME_EXCEEDED for an ECHO_REQUEST. traceroute was then changed to use UDP probe packets.

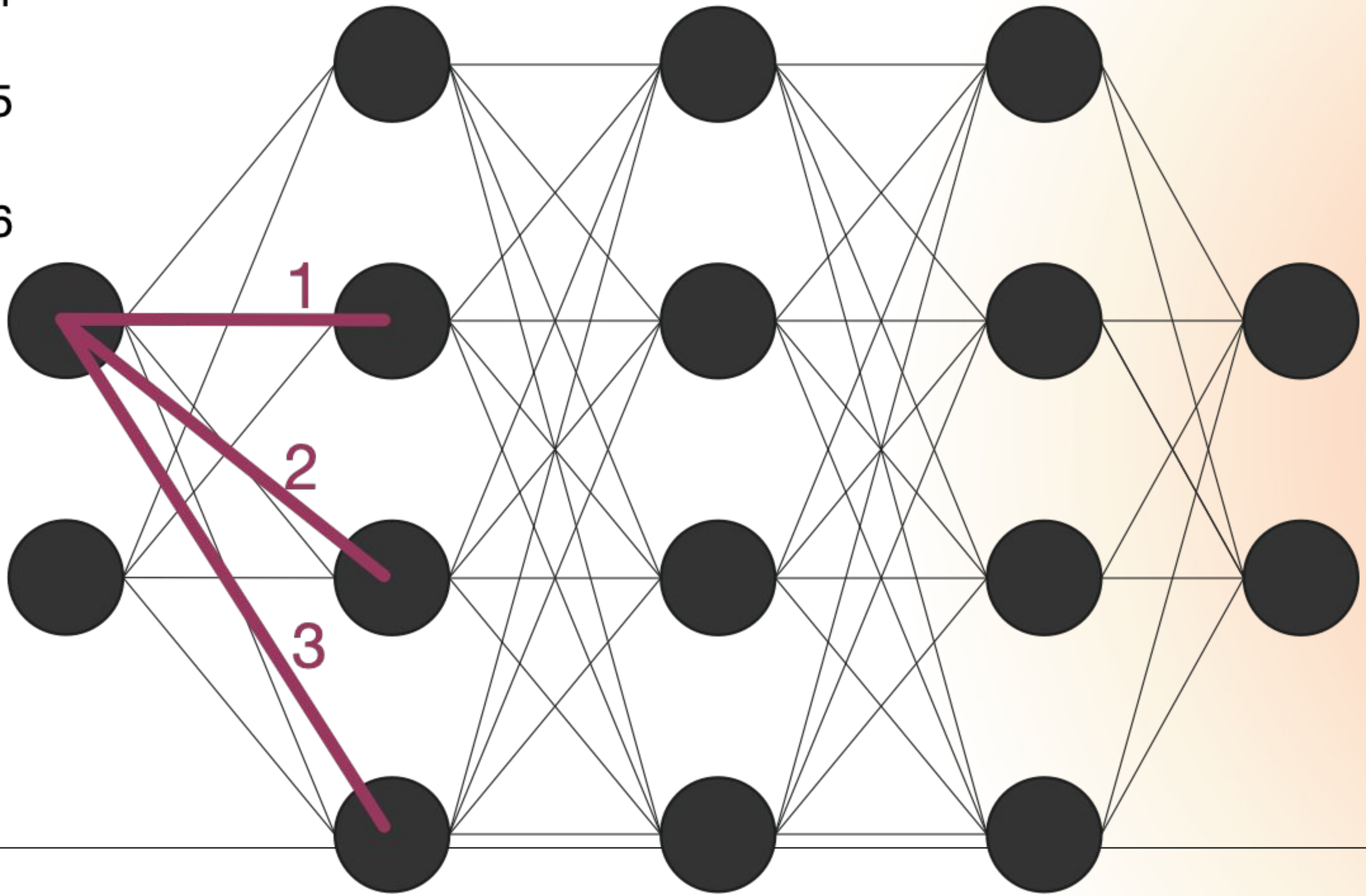
OpenBSD man page for traceroute

Original Version - Overview

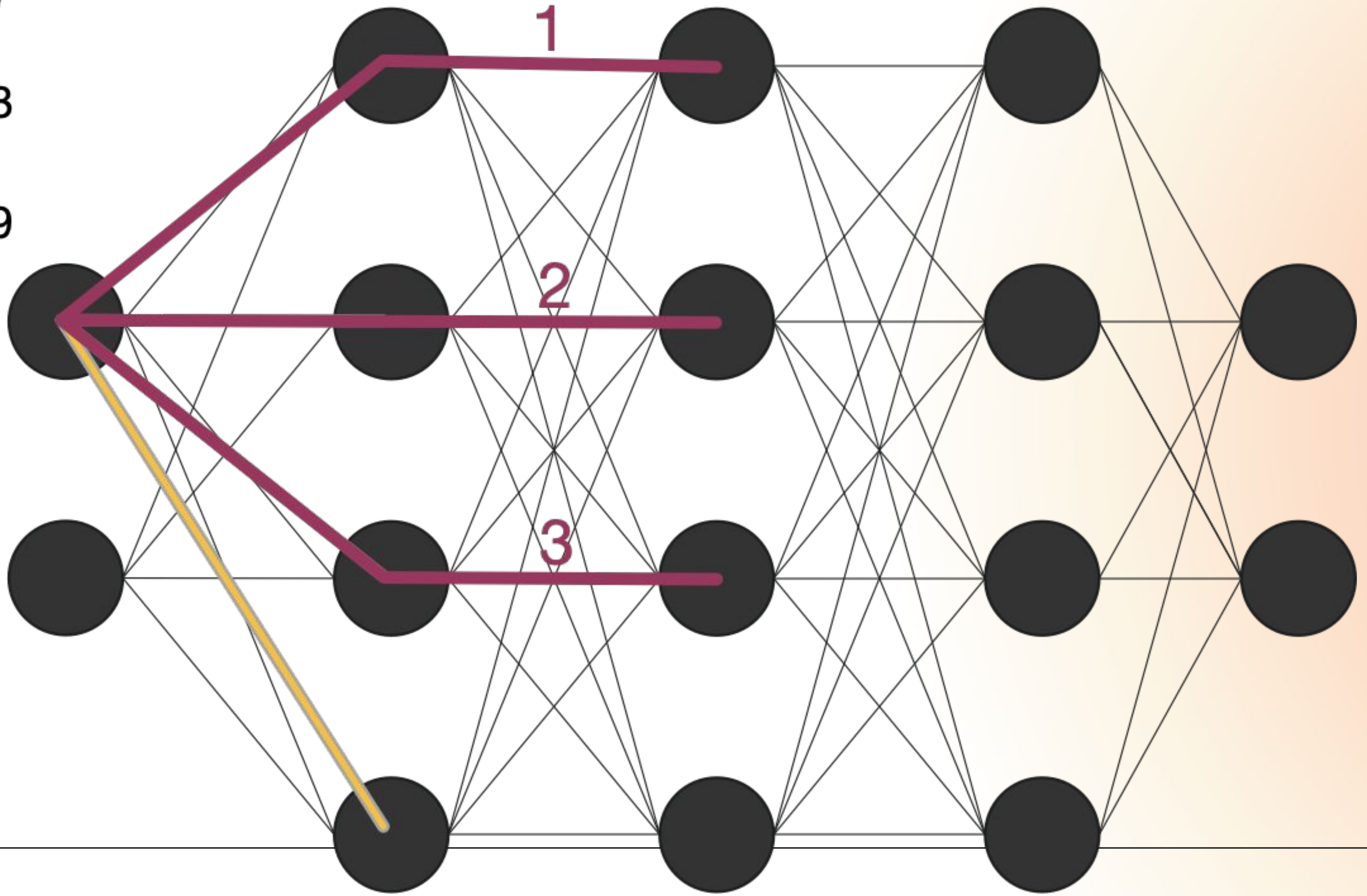
```
[yak 72]% traceroute allspice.lcs.mit.edu.
traceroute to allspice.lcs.mit.edu (18.26.0.115), 30 hops max
1  helios.ee.lbl.gov (128.3.112.1)  0 ms  0 ms  0 ms
2  lilac-dmc.Berkeley.EDU (128.32.216.1)  19 ms  19 ms  19 ms
3  lilac-dmc.Berkeley.EDU (128.32.216.1)  39 ms  19 ms  19 ms
4  ccngw-ner-cc.Berkeley.EDU (128.32.136.23)  19 ms  39 ms  39 ms
5  ccn-nerif22.Berkeley.EDU (128.32.168.22)  20 ms  39 ms  39 ms
6  128.32.197.4 (128.32.197.4)  59 ms  119 ms  39 ms
7  131.119.2.5 (131.119.2.5)  59 ms  59 ms  39 ms
8  129.140.70.13 (129.140.70.13)  80 ms  79 ms  99 ms
9  129.140.71.6 (129.140.71.6)  139 ms  139 ms  159 ms
10  129.140.81.7 (129.140.81.7)  199 ms  180 ms  300 ms
11  129.140.72.17 (129.140.72.17)  300 ms  239 ms  239 ms
12  * * *
13  128.121.54.72 (128.121.54.72)  259 ms  499 ms  279 ms
14  * * *
15  * * *
16  * * *
17  * * *
18  ALLSPICE.LCS.MIT.EDU (18.26.0.115)  339 ms  279 ms  279 ms
```



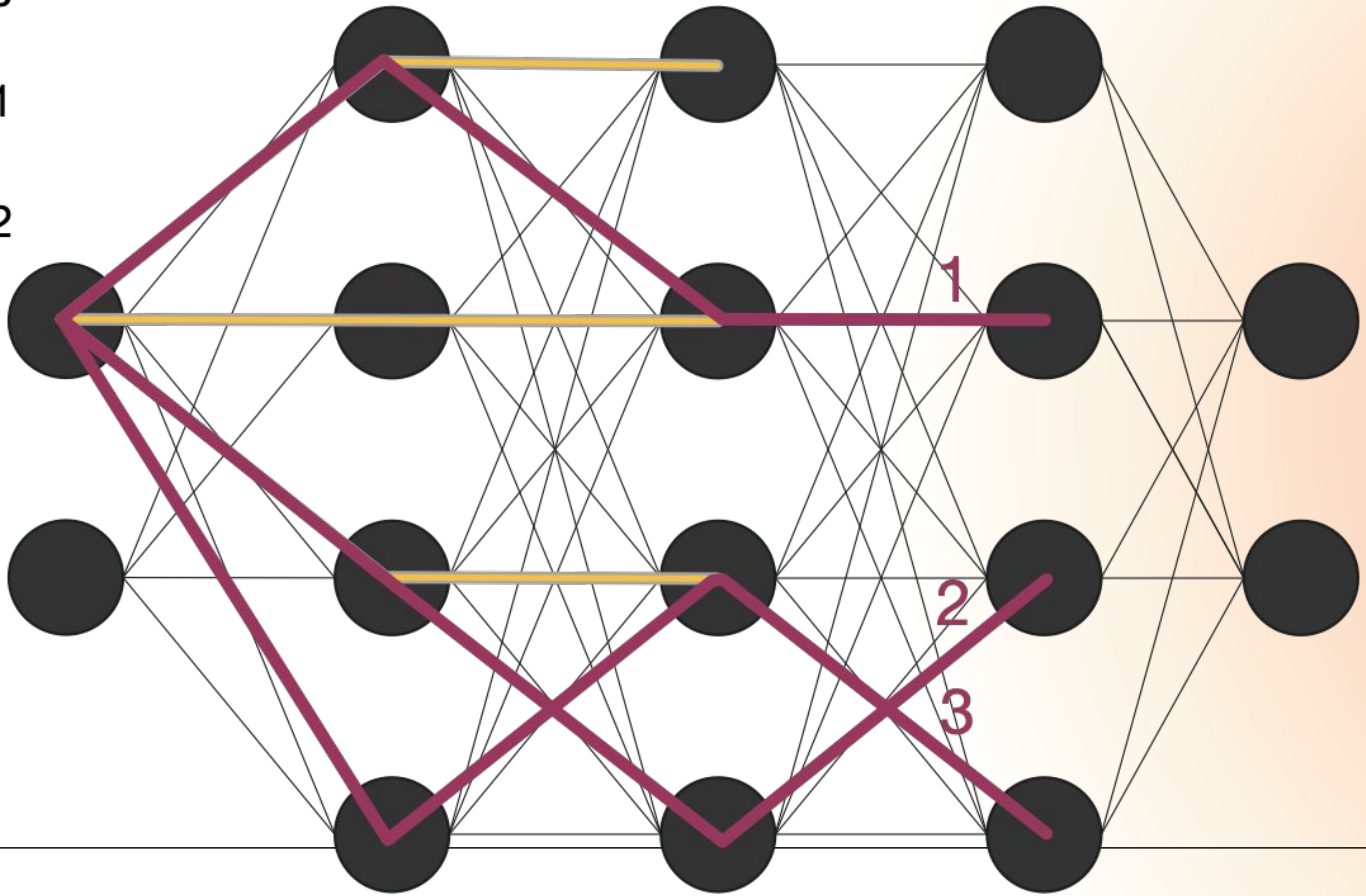
1: TTL 1
dport 33434
2: TTL 1
dport 33435
3: TTL 1
dport 33436



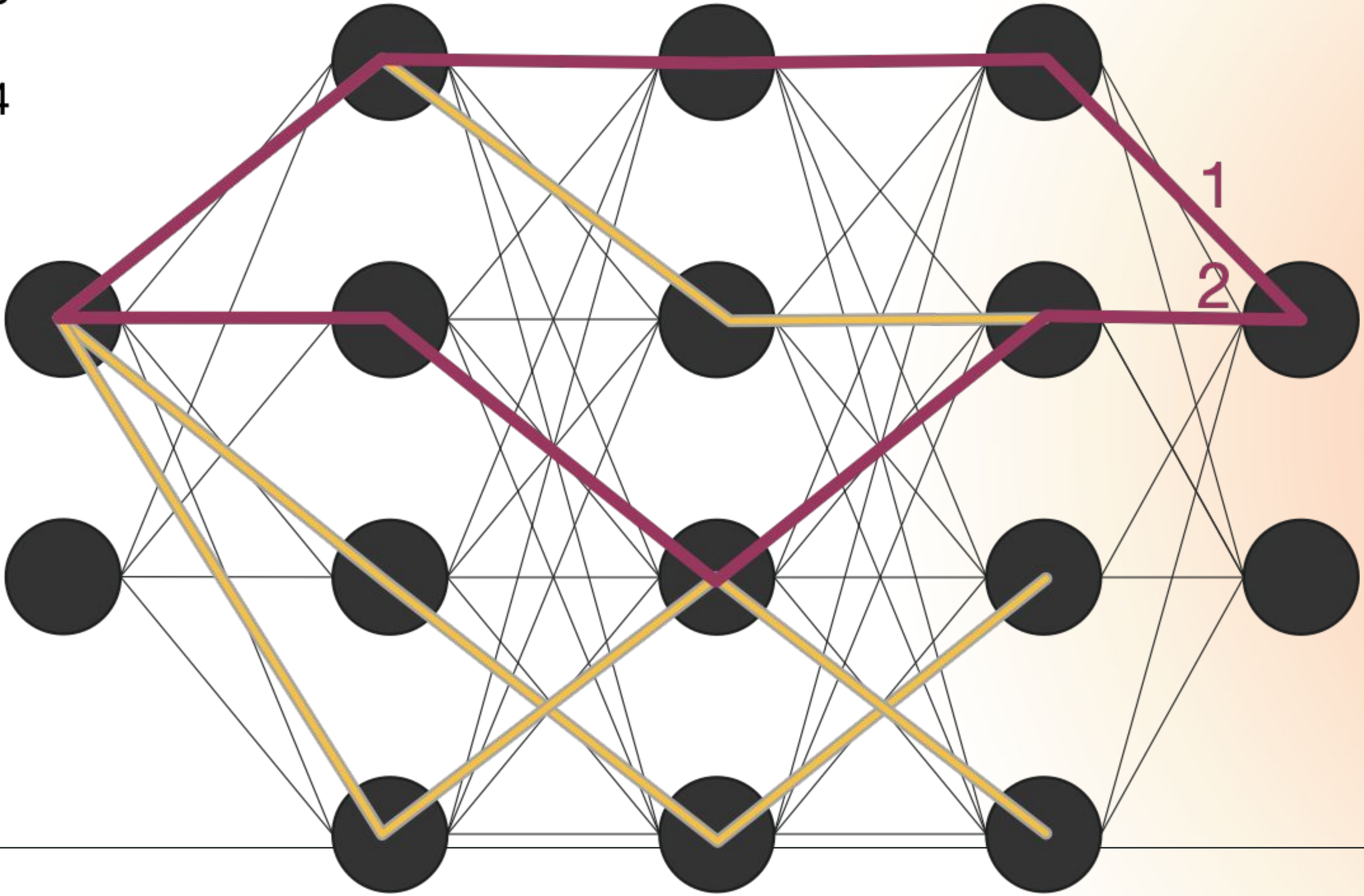
1: TTL 2
dport 33437
2: TTL 2
dport 33438
3: TTL 2
dport 33439



1: TTL 3
dport 33440
2: TTL 3
dport 33441
3: TTL 3
dport 33442



1: TTL 4
dport 33443
2: TTL 4
dport 33444



A note on hashing

- ICMP vs UDP vs TCP
- Src/Dst IP, Prefix

A note on hashing

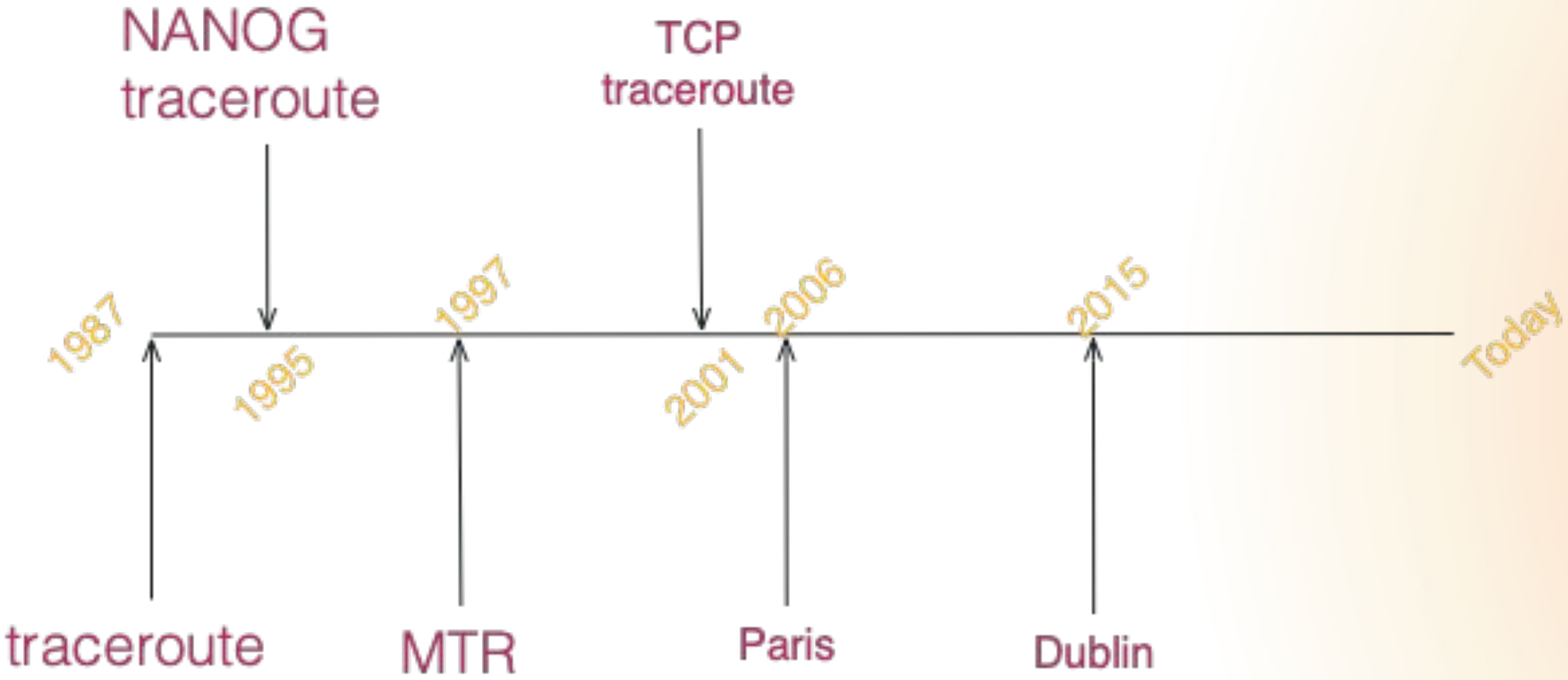
ICMP

```
1.|-- 46.23.89.65
2.|-- 46.23.91.2
3.|-- 46.244.4.217
4.|-- 80.249.211.140
5.|-- 141.101.65.103
6.|-- 1.1.1.1
```

UDP

```
1.|-- 46.23.89.65
2.|-- 46.23.91.2
      46.23.91.3
3.|-- 46.244.4.217
      46.244.6.161
4.|-- 80.249.211.140
5.|-- 141.101.65.107
      141.101.65.93
6.|-- ???
```

Timeline



NANOG
traceroute
TCP
traceroute
1987
1995
1997
2001
2006
2015
Today
traceroute
MTR
Paris
Dublin

NANOG traceroute

- Based on “original” traceroute
- Does AS lookup, microsecond time resolution, Path MTU discovery, Parallel Probing
- Does not change the algorithm from original

MTR

- Combines ping with traceroute
- Defaults to ICMP ECHO_REQUEST probes
- ECHO_REQUEST used to gauge packet loss
- UDP/TCP mode increments src port by 1 per TTL

TCP traceroute

- Defaults to TCP!
- Useful for environments where firewalls are setup to filter ICMP
- Does not vary entropy

Paris traceroute

- Attempt to identify more “complex” topologies that include diamonds and circles
- Varies fields used in multi-path calculations to generate entropy
- Does not vary entropy

Dublin traceroute

- Built to handle NAT and ECMP environments
- Builds on the work from Paris
- Has libraries for re-use to build on

OpenBSD - Recent updates

- Does all the hop probing in parallel
- Does async DNS lookups

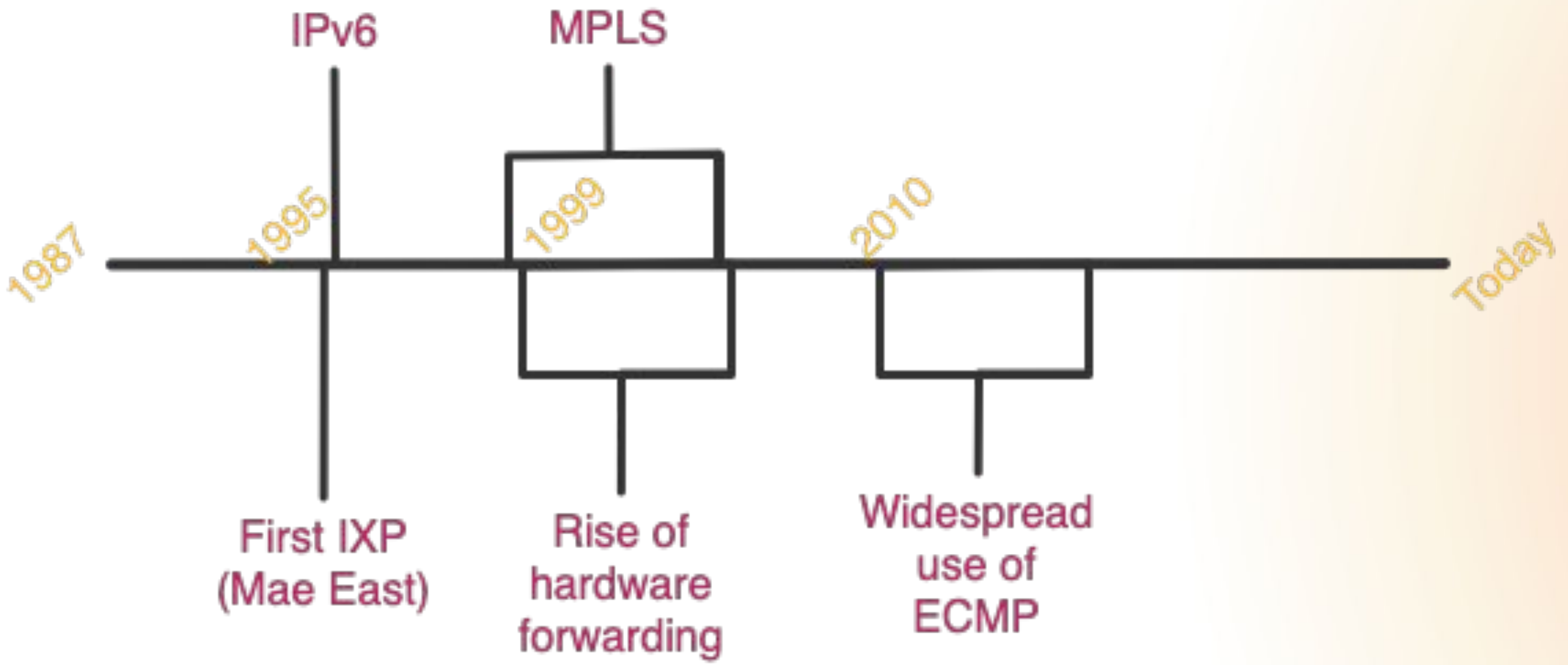
Networks Then

- Relatively flat campus networks
- Hierarchical internet backbone - NSNET had T1-T3 backbone in 1991 [1] [2]
- Software forwarding
- Multi-path, Tunnels and other “fun” technologies for traffic not really present
  GRE shows up in 1994

Networks Now

- ASIC forwarding the norm - even with a rise of DPDK/high performance linux routing
- Overlay heavy; MPLS, VxLAN, GRE very common
- IPv6
- IXPs, peering and a more connected internet
- Move to Layer 3 “inside the firewall”
  Load balancing/ECMP/UCMP used frequently

Networks Now



Effects on traceroute

IXPs, peering and a more connected internet

- Peering and move to asymmetric internet
- Forward and reverse paths can diverge

Effects on traceroute

IPv6

- IPv6 traceroute
- Flowlabel vs L4 hashing

Effects on traceroute

Overlays

- MPLS tunnel
- VxLAN/GRE
- GRE

Effects on traceroute

ASIC

- Hardware forwarding
- Weak control planes; Latency, Rate Limiting
- Ingress interface
  RFC 792
  The address of the gateway or host that composes the ICMP message. Unless otherwise noted, this can be any of a gateway's addresses.

Effects on traceroute

Move to Layer 3 “inside the firewall”

- Rise of ECMP in the data center
  - Per flow
  - Per ip
  - Per prefix
- Different hashing strategies
  - Src/Dst IP/Prefix
  - Layer3+Layer4
- BGP as-path multipath-relax at edge
- Per-packet load balancing

Summary

- Jacobson’s traceroute is one of the most widely used network measurement tools
- Loose relationship between changes in the network to changes in traceroute
- Reading traceroute can be like reading tea leaves

Summary

- Windows `tracert` is the most historically correct version of traceroute
- Asymmetry is part of the internet, yet hard for traceroute to reason about
- Measuring reverse path is very difficult today

Questions?

We’re hiring!

- Reykjanesbæ, Iceland
  - Network Data Centre Technician, Senior
  - Manager, Data Center Engineering
  - Hardware Data Center Operations Technician

Crusoe

Thank you!