THE INTERNET IS BROKEN

Why Public Internet Routing Sucks

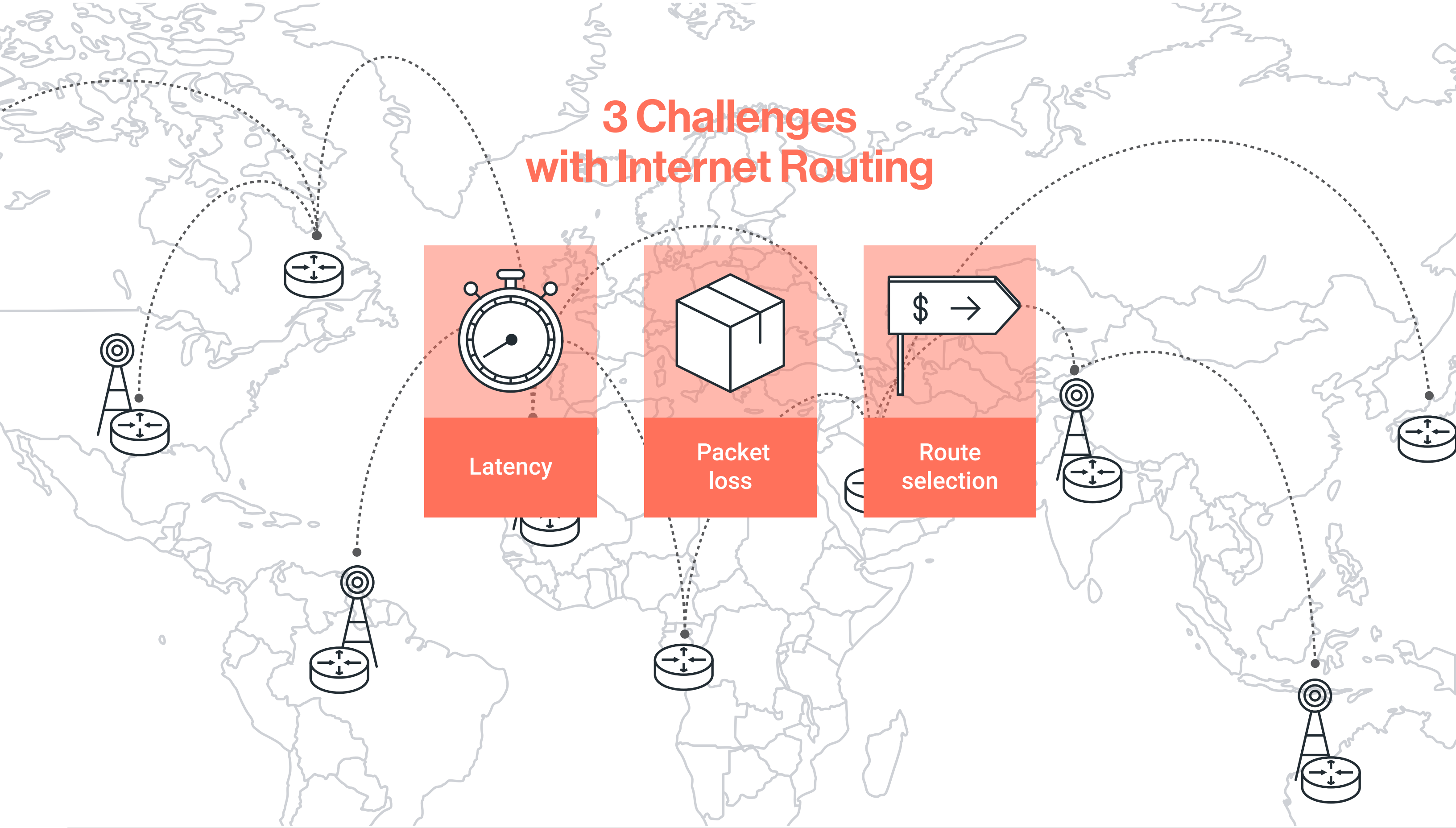
The Internet is Broken

Anyone with hands on experience in setting up long haul VPNs over the Internet knows it is not a pleasant exercise. Even if you factor out the complexity of appliances and the need to work with old relics like IPSec, managing latency, packet loss and high availability remains a huge problem on the Internet. Service providers also know this (and make billions on MPLS).

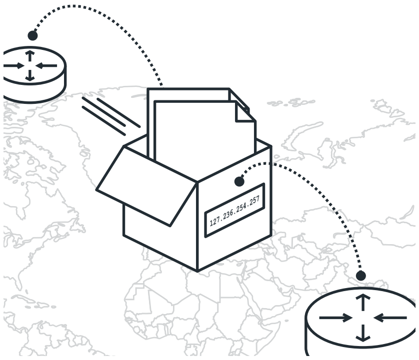
The bad news is that it is not getting any better. It doesn't matter that available capacity has increased dramatically. The problem is in the way providers are interconnected and with how global routes are (mis)managed. It lies at the core of how the Internet was built, its protocols, and how service providers implemented IP routing. The same architecture that allowed the Internet to cost-effectively scale to billions of devices also set its limits.

Addressing these challenges requires a deep restructuring in the Internet fabric of and core routing - and should form the foundation for possible solutions. There isn't going to be a shiny new router that would magically solve it all.

The problem is in the way providers are interconnected and with how global routes are mismanaged. It lies at the core of how the Internet was built, its protocols, and how service providers implemented their routing layer.



IP Routing's Historical Baggage: Simplistic Data Plane



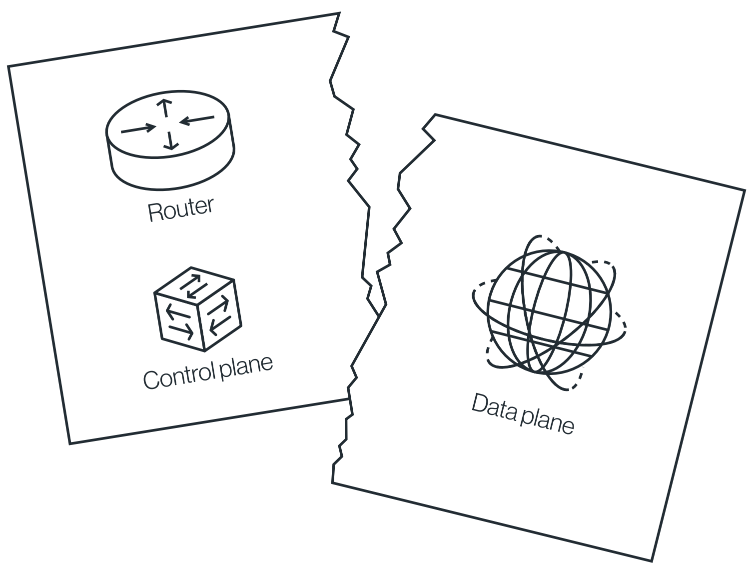
Whether the traffic is voice, video, HTTP, or email, the Internet is made of IP packets. Billions of them travel to destinations carry information. If they are lost along the way, it is the responsibility of higher level protocols, such as TCP, to recover them. Packets hop from router to router, only "aware" of their next hop, and their ultimate destination. Routers are the ones making the decisions about the packets. When a router receives a packet, it performs a calculation according to its routing table - identifying the best next hop for the packet.

From the early days of the Internet, routers were shaped by technical constraints. There was a shortage of processing power available to move packets along their path the "data plane". Access speeds and available memory were limited, so routers had to rely on custom hardware that performed minimal processing per packet and had no state management. Communicating with this restricted data plane had to be extremely simple and infrequent. Routing decisions were moved out to a separate process, the "control plane," that instructed the data plane on the next hop.

This separation of control and data planes allowed architects to build massively scalable routers, handling millions of packets per second. However, even as processing power increased on the data plane, it wasn't really used. The paradigm was, and still is, that the control plane makes all the decisions, the data plane executes the routing table, and apart from routing table updates, they hardly communicate. Getting "feedback" from the data plane was simply out of the question.

A modern router does not have any idea how long it actually took a packet to reach its next hop, or whether it reached it at all. Are neighbors congested? Maybe and maybe not. The router doesn't even know if it is congested itself. And to the extent it does have information to share, it will not be communicated back to the control plane, where routing decisions are actually made.

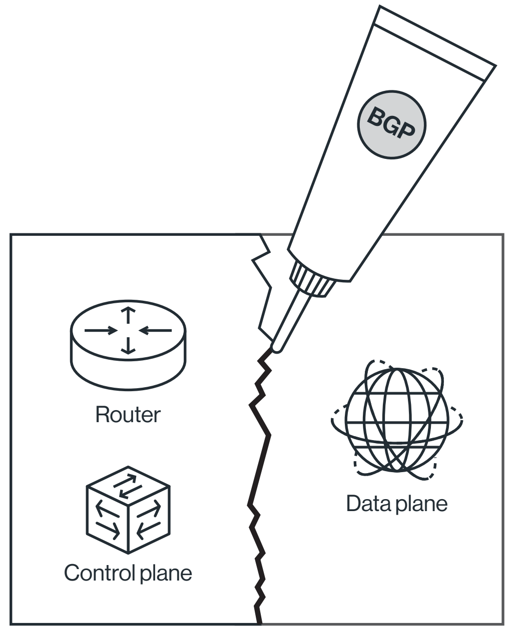
Ironically, this limited exchange between the control plane and the data plane was taken to the extreme in OpenFlow and Software-defined Networking (SDN): the separation of control plane and data plane into two different machines.This might be a good solution for cutting costs in the data center, but to improve global routing it makes more sense to substantially increase information sharing between the control plane and the data plane.



BGP - The Protocol Behind Internet Routing

BGP is the routing protocol that glues the Internet together.

BGP was able to scale relatively unchanged, since being drafted on a napkin in 1989, focusing on propagating reachability information. In very simple terms, its task is to communicate the knowledge of where an IP address (or a whole IP subnet) originates from, and what is the best router through which it could be reached.



BGP involves routers connecting with their peers, and exchanging information about which IP subnets they originate, and also "gossip" about IP subnets they learned about from other peers. As these "rumors" propagate between the peers and across the globe, they are appended with the accumulated rumour path from the originator (this is called the **AS-Path**). As more routers are added to the path, the "distance" grows.

Here is an example of what a router knows about a specific subnet **45.62.176.0**, shown using Hurricane Electric's excellent looking glass service. It learned about this subnet from multiple peers, and selected the shortest **AS-Path**. This subnet originates from **AS 13150** the rumour having reached the router across **AS 5580**. Now the router can update its routing table accordingly.

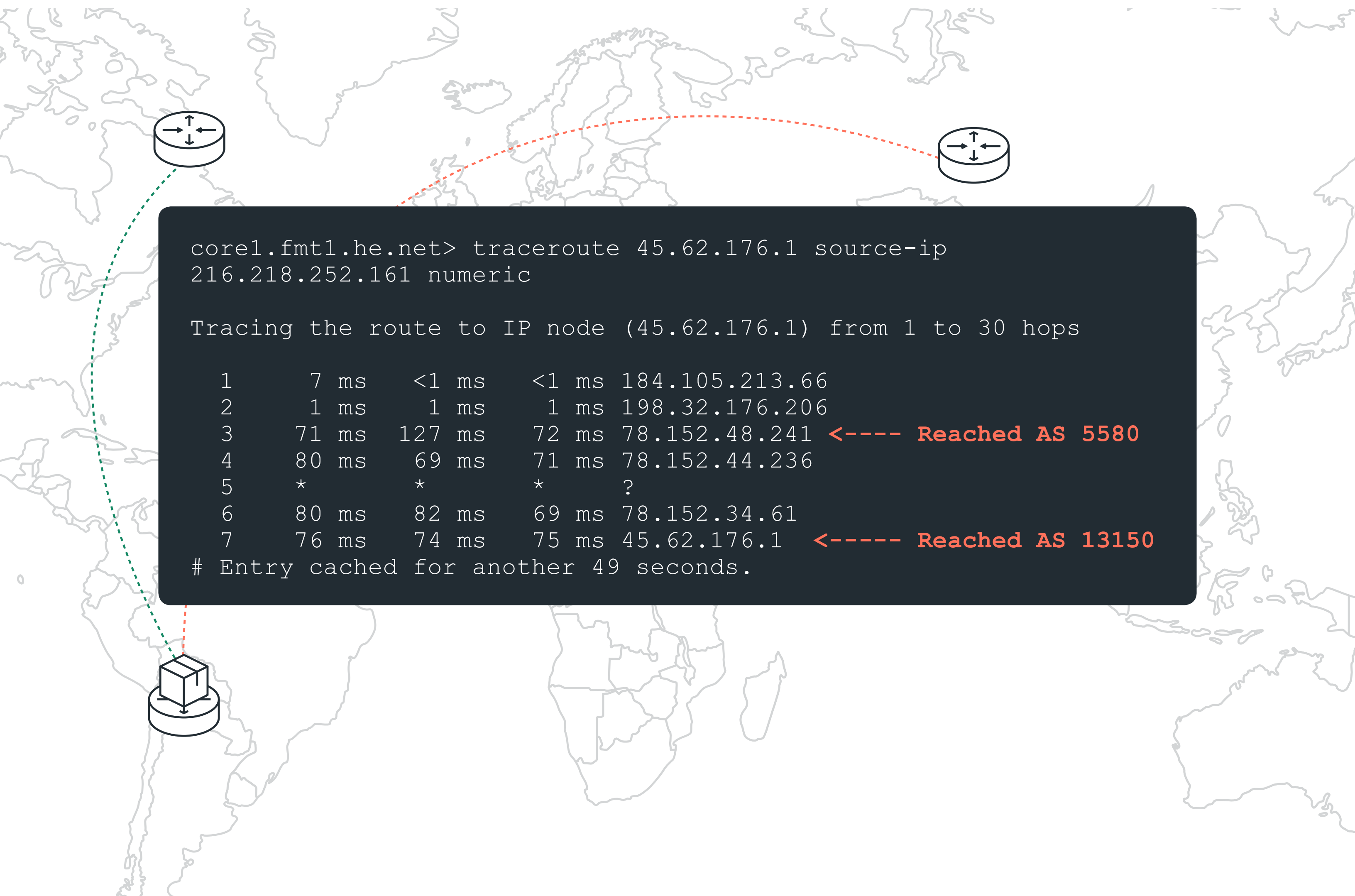
```
core1.fmt1.he.net> show ip bgp routes detail 45.62.176.0
  Number of BGP Routes matching display condition : 11
       S:SUPPRESSED F:FILTERED s:STALE
1       Prefix: 45.62.176.0/24,  Status: BI,  Age: 1d14h45m31s <--- The
Freshest Rumor
         NEXT_HOP: 198.32.176.206, Metric: 15, Learned from Peer:
216.218.252.165 (6939)
          LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
         AS_PATH: 5580 13150 <---- This traces the rumour about this subnet
2       Prefix: 45.62.176.0/24,  Status: I,  Age: 7d5h25m3s
         NEXT_HOP: 206.72.210.212, Metric: 95, Learned from Peer:
216.218.252.178 (6939)
          LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
         AS_PATH: 5580 13150
            COMMUNITIES: 6939:1111
[snip]
```

BGP is a very elegant protocol, and we can see why it was able to scale with the Internet: it requires very little coordination across network elements. Assuming the routers running the protocols are the ones that are actually routing traffic, BGP has a built-in resiliency. When a router fails, so will the routes it propagated, and other routers will be selected.

BGP has a straightforward way of assessing distance: it uses the AS-Path, so if it got the route first-hand it is assumed to be closest. Rumored routes are considered further away as the hearsay "distance" increases. The general assumption is that the router that reported the closest rumor is also the best choice to send packets.

If we want to see how traffic destined for this IP range is actually routed, we can use Traceroute. Note that in this case, there was a correlation between the AS-Path, and the path the actual packets traveled. As we will show later, it doesn't necessarily work that way.

BGP has a straightforward way of assessing distance: it uses the AS-Path, so if it got the route first-hand it is assumed to be closest.



```
core1.fmt1.he.net> traceroute 45.62.176.1 source-ip
216.218.252.161 numeric

Tracing the route to IP node (45.62.176.1) from 1 to 30 hops

  1      7 ms   <1 ms   <1 ms 184.105.213.66
  2      1 ms    1 ms    1 ms 198.32.176.206
  3     71 ms  127 ms   72 ms 78.152.48.241 <---- Reached AS 5580
  4     80 ms   69 ms   71 ms 78.152.44.236
  5     *       *       *     ?
  6     80 ms   82 ms   69 ms 78.152.34.61
  7     76 ms   74 ms   75 ms 45.62.176.1  <----- Reached AS 13150
# Entry cached for another 49 seconds.
```

BGP doesn't know if a specific path has 0% or 20% packet loss. Also, using the AS-Path as a method to select smallest latency is pretty limited. It's like calculating the shortest path between two points on the map by counting traffic lights, instead of miles, along the way.

Here is a real life example of what can happen. A straightforward route between Hurricane Electric (HE), a tier 1 service provider, as seen from **Singapore**, to an **IP address in China**. It has a **path length of 1**.

```
core1.sin1.he.net> show ip bgp routes detail 115.239.XXX.XXX
     Number of BGP Routes matching display condition : 3
          S:SUPPRESSED F:FILTERED s:STALE
   1        Prefix: 115.224.0.0/12,  Status: BI,  Age: 26d4h57m34s
             NEXT_HOP: 202.97.32.90, Metric: 1855, Learned from Peer:
216.218.252.166 (6939)
              LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
             AS_PATH: 4134
   2        Prefix: 115.224.0.0/12,  Status: I,  Age: 2h45m1s
             NEXT_HOP: 202.97.32.97, Metric: 1954, Learned from Peer:
216.218.252.184 (6939)
              LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
             AS_PATH: 4134
                COMMUNITIES: 6939:1111
   3        Prefix: 115.224.0.0/12,  Status: I,  Age: 2h45m2s
             NEXT_HOP: 202.97.32.97, Metric: 1954, Learned from Peer:
216.218.252.164 (6939)
              LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
             AS_PATH: 4134
                COMMUNITIES: 6939:1111
          Last update to IP routing table: 26d4h57m34s, 1 path(s)
installed:

 # Entry cached for another 52 seconds.
```

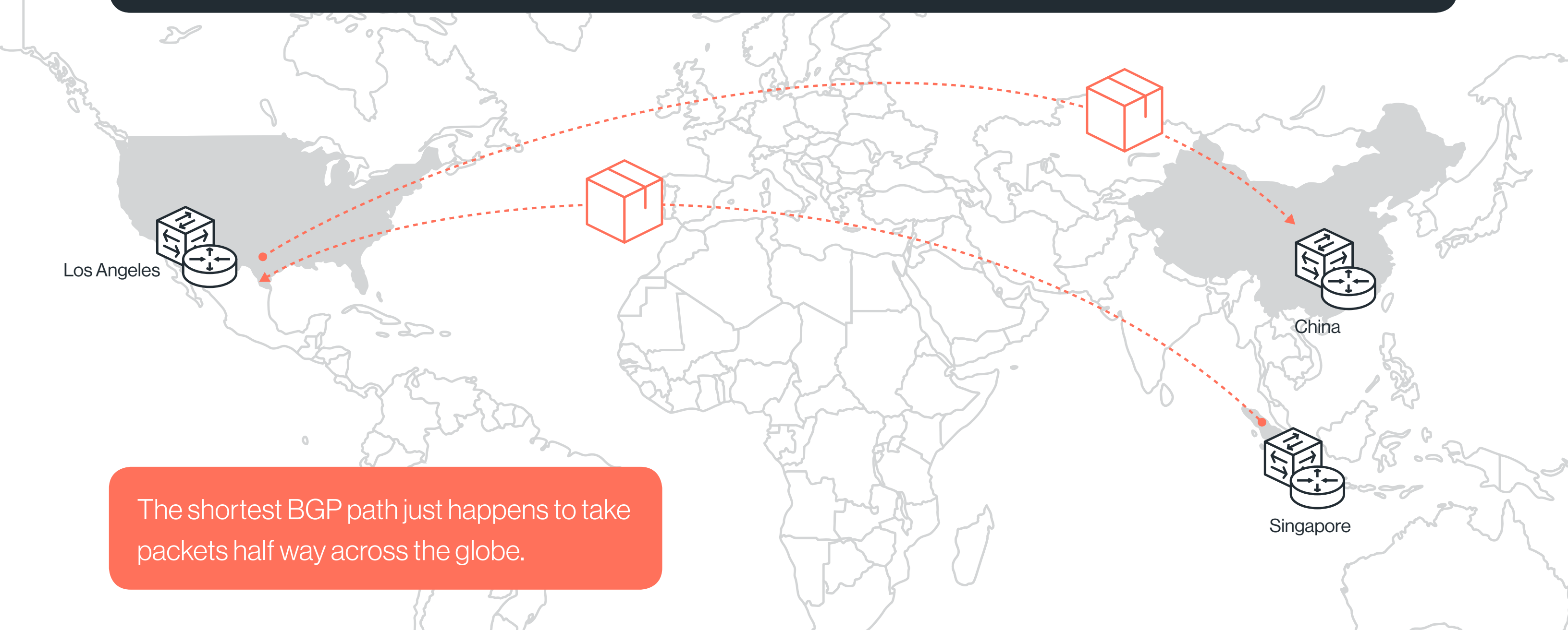
But if we trace the path the packets actually take from Singapore to China, the story is very different: **packets seems to make a "connection" in Los Angeles.**

This is a very long journey. Why did this packet travel all the way to the US West Coast to get from Singapore to China? **Simply because HE peers with China Telecom in Los Angeles.** So every packet from anywhere within the HE autonomous system will go through Los Angeles to reach China Telecom.

```
core1.sin1.he.net> traceroute 115.239.XXX.XXXX source-ip 27.50.XXX.XXX
numeric

 Tracing the route to IP node (115.239.248.245) from 1 to 30 hops

  1     69 ms    49 ms    57 ms 184.105.223.189
  2     94 ms   110 ms   102 ms 184.105.222.105
  3    211 ms   209 ms   215 ms 184.105.223.105
  4    202 ms   199 ms   199 ms 72.52.92.121 <------ a.k.a. 100ge2-1.
core1.lax1.he.net:
  5    195 ms   199 ms   201 ms 64.71.131.134
Wauza! We're in Los Angeles!
  6    400 ms   400 ms   397 ms 202.97.49.81
  7    627 ms   647 ms   652 ms 202.97.50.121
  8     *        *       662 ms 202.97.35.89
  9    549 ms   550 ms   550 ms 202.97.50.253
 10    530 ms   467 ms   430 ms 202.97.82.110
 11    546 ms   549 ms   556 ms 115.233.166.242
 12     *        *        *      ?
 13    589 ms   678 ms   673 ms 122.224.7.186
 14     *        *        *      ?
 15     *        *        *      ?
 16    658 ms   672 ms   670 ms 115.239.XXX.XXX
 # Entry cached for another 4 seconds.
```



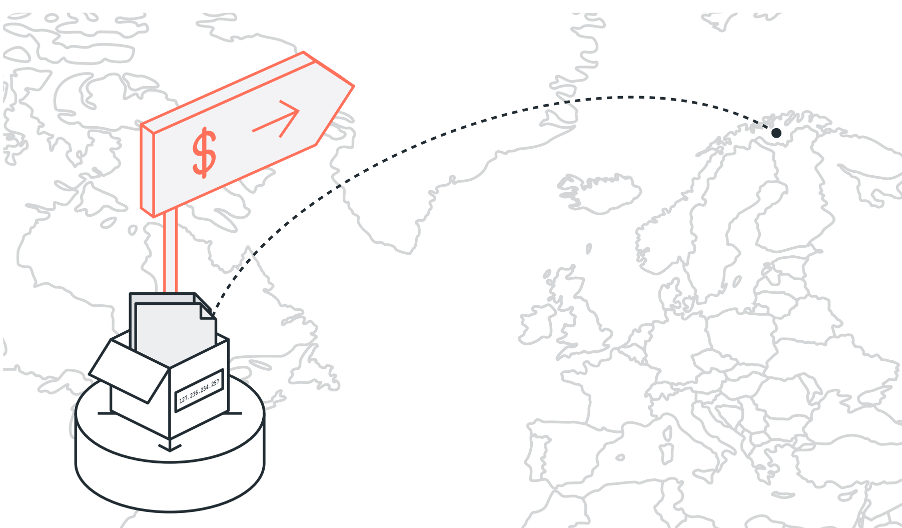
The shortest BGP path just happens to take packets half way across the globe.

BGP Abused:
BGP Meets the Commercial Internet

To work around BGP's algorithms, the protocol itself extends to include a host of manual controls to allow manipulation of the "next best hop" decisions.

Controls such as 3 is enough, local preference (prioritizing routes from specific peers), communities (allow peers to add custom attributes, which may then affect the decisions of other peers along the path), and AS-path prepending (manipulates the propagated AS-path) allow network engineers to tweak and improve problematic routes and to alleviate congestion issues.The relationship between BGP peers on the Internet is a reflection of commercial contracts of ISPs. Customers pay for Internet traffic. Smaller service providers pay larger providers, and most pay tier 1 providers. Any non-commercial relationship has to be mutually beneficial, or very limited.

BGP gives service providers the tools to implement these financial agreements:



- Service providers usually prefer routing traffic for "paying" connections.
- Service providers want to quickly get rid of "unpaid" packets, rather than carrying them across their backbone (so called "hot potato" routing).
- Sometimes, service providers will carry the packets over very long distances just to get the most financially beneficial path.

All this comes at the expense of best path selection.

Customers pay for Internet traffic. Smaller service providers pay larger providers, and most pay tier 1 providers.

The relationship between BGP peers on the Internet is a reflection of commercial contracts of ISPs.

Here is an example closer to home. Cato Networks publishes the range **185.114.121.0/24** via two service providers: **AS1299 (TeliaSonera)**, and **AS558 (Hibernia/Atrato)**. We can see this route from the two service providers. We can see that the path through **AS5580** is extremely long (artificially, with path prepending), but the short path is all the way down in #12. Why? **AS1299** is one of the largest tier 1 providers, sending traffic through it will typically be more expensive. So routes are tweaked using the "local preference" option of BGP, which overrides shortest path selection.

```
core1.mci3.he.net> show ip bgp routes detail 185.114.121.1
  Number of BGP Routes matching display condition : 12
       S:SUPPRESSED F:FILTERED s:STALE
1       Prefix: 185.114.121.0/24,  Status: BI,  Age: 1h28m49s
         NEXT_HOP: 206.223.119.45, Metric: 135, Learned from Peer:
216.218.252.168 (6939)
          LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
         AS_PATH: 5580 13150 13150 13150 13150
2       Prefix: 185.114.121.0/24,  Status: I,  Age: 1h28m49s
         NEXT_HOP: 206.108.34.245, Metric: 235, Learned from Peer:
216.218.252.147 (6939)
          LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
         AS_PATH: 5580 13150 13150 13150 13150
            COMMUNITIES: 6939:1111
3       Prefix: 185.114.121.0/24,  Status: I,  Age: 1h28m49s
         NEXT_HOP: 206.126.236.204, Metric: 290, Learned from Peer:
216.218.252.169 (6939)
          LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
         AS_PATH: 5580 13150 13150 13150 13150
            COMMUNITIES: 6939:1111
4       Prefix: 185.114.121.0/24,  Status: I,  Age: 1h28m49s
         NEXT_HOP: 206.126.115.25, Metric: 310, Learned from Peer:
216.218.252.148 (6939)
          LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
         AS_PATH: 5580 13150 13150 13150 13150
            COMMUNITIES: 6939:1111
[snip]

11      Prefix: 185.114.121.0/24,  Status: I,  Age: 1h28m49s
         NEXT_HOP: 206.72.210.212, Metric: 545, Learned from Peer:
216.218.252.178 (6939)
          LOCAL_PREF: 100,  MED: 1,  ORIGIN: igp,  Weight: 0
         AS_PATH: 5580 13150 13150 13150 13150
            COMMUNITIES: 6939:1111
12      Prefix: 185.114.121.0/24,  Status: E,  Age: 1h10m28s
         NEXT_HOP: 213.248.73.209, Metric: 0, Learned from Peer:
213.248.73.209 (1299)
          LOCAL_PREF: 70,  MED: 48,  ORIGIN: igp,  Weight: 0
         AS_PATH: 1299 13150
            COMMUNITIES: 6939:2000
       Last update to IP routing table: 1h28m51s, 1 path(s) installed:
# Entry cached for another 56 seconds.
```

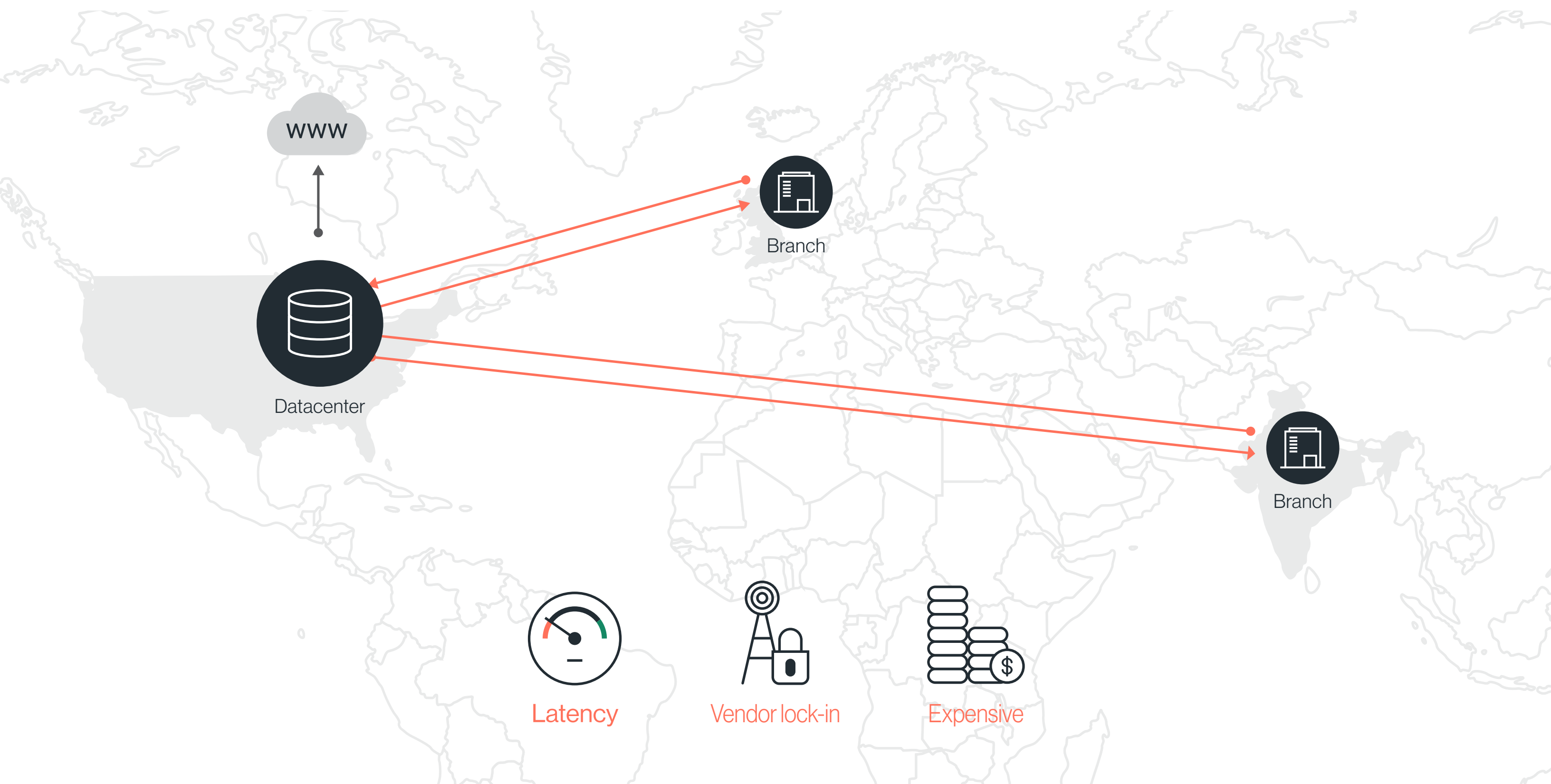
The MPLS Racket

To address these problems, service providers came up with an alternative, offering private network services. These networks were built on their own backbones, which internally used MPLS as the routing protocol. MPLS is in many ways the opposite of BGP. Instead of an open architecture, MPLS uses policy based end-to-end routing. A packet's path through the network is predetermined, which makes it suitable only for private networks. This is why an MPLS is sold by a single provider, even if the provider patched together multiple networks behind the scenes to reach customer premises.

MPLS is a control plane protocol. It has many of the same limitations as BGP: routing is decided by policy, not real traffic conditions, such as latency or packet loss. (However, it does have some limited capabilities in capacity reservation and congestion avoidance). Thus, providers needed to be very careful about bandwidth management to maintain their SLA.

The combination of single vendor lock-in and the need for careful planning and overprovisioning to maintain SLA, made these private network services premium, expensive products. As the rest of the Internet, with its open architecture, became increasingly competitive and cost-efficient, MPLS came under huge pressure. As a backbone implementation, MPLS is not likely to ever become affordable. Most traffic is, and will remain on the Internet.

The combination of single vendor lock-in and the need for careful planning and overprovisioning to maintain SLA, made these private networks a premium and expensive product.



A Way Forward

The Internet just works. Not flawlessly, not optimally, but packets generally reach their destination. The basic structure of the Internet has not changed much over the past few decades, and has proven itself probably beyond the wildest expectations of its designers. However, it has also cemented key limitations:

The Data Plane is Clueless

Routers, which form the data plane, are built for traffic load, and are therefore stateless, and have no notion of individual packet or traffic flows.

Control Plane Intelligence is Limited

Because the control plane and the data plane are not communicating, the routing decisions are not aware of packet loss, latency, congestion, or actual best routes.

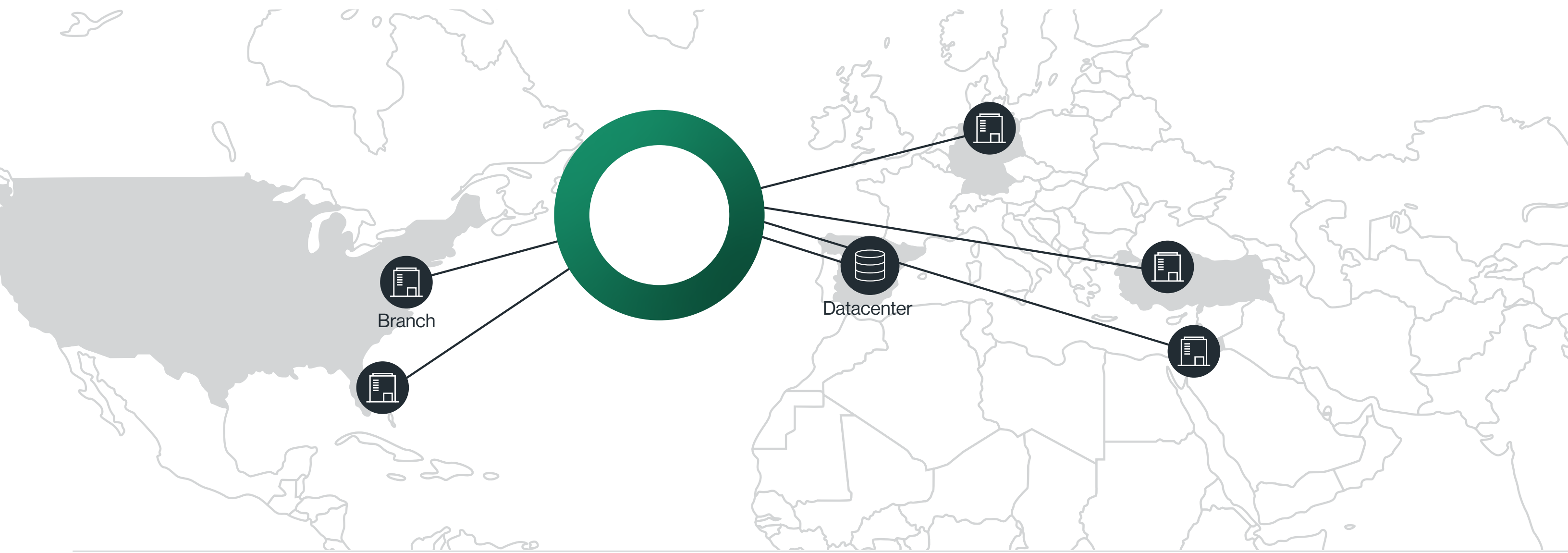
Shortest Path Selection is Abused

Service providers' commercial relationships often works against the end user interest when selecting the best path.

These problems can be addressed by:

- Converging the data plane and control plane routing statefully
- Dynamically selecting the best path

This is the cloud networking platform built by Cato Networks. Unlike legacy MPLS solutions, Cato overcomes the limitations of public internet routing with an affordable global connectivity solution.



About Cato Networks

Cato, the cloud-native carrier, provides the only secure managed SD-WAN service built with the global reach, self-service, and agility of the cloud. Cato replaces MPLS and multiple networking and security point solutions with a converged WAN transformation platform built for the digital business. Using Cato, customers easily migrate from MPLS to SD-WAN, improve global connectivity to on-premises and cloud applications, enable secure branch internet access everywhere, and securely and optimally integrate cloud datacenters and mobile users into the network.

For more information:

www.CatoNetworks.com

@CatoNetworks