A pourquoi story, also dubbed an "origin story", is also used in mythology, referring to narratives of how a world began, how creatures and plants came into existence, and why certain things in the cosmos have certain yet distinct qualities

-- Wikipedia

L2VPN: the Pourquoi Story
(yeah, we're talking BGP)

NLNOG Day 2023

Kireeti Kompella
Chief Engineer, AWAN BU
Juniper Networks

Filesystem Guy (1990-1996)

- ACSC, 1990-1994; worked on UniTree
  - https://ntrs.nasa.gov/api/citations/19950017700/downloads/19950017700.pdf
- SGI, 1994-1995; worked on xLV, xFS
  - https://irix7.com/techpubs/007-2825-002.pdf
- NetApp, 1995-1996; worked on WAFL
  - https://community.netapp.com/fukiw75442/attachments/fukiw75442/data-ontap-discussions/2334/1/WAFL.pdf

Kernel Guy/Microkernel Guy (1997)

- Juniper Networks, 1997; worked on device drivers (ifd/ifl/ifa/iff)
- Had to learn PPP, Cisco HDLC from scratch: header, layer 2 rewrite, hellos/keepalives; how the layer 3 proto is indicated
- Same for Frame Relay (DLCIs) and ATM (VPIs, VCIs), but now with sub-interfaces (ifls)
- Moved on to route tables and nexthops
- What I *really* wanted was to work on routing protocols

Important later

Traffic Engineering (1998)

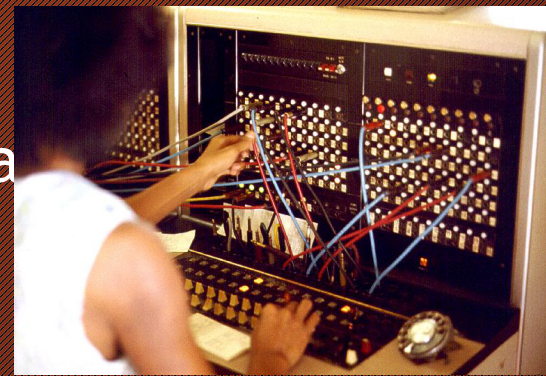
- Traffic Engineering: big requirement from UUnet
- Mike O'Dell: "tell the router what you want [i.e., source + dest + TE constraints] and let the router connect the dots"
- TE requirements (RFC 2702, Awduche et al, Sep 1999)
- ISIS TE extensions (RFC 3784, Smit & Li, Jun 2004)
- RSVP-TE signaling (RFC 3209, Awduche et al, Dec 2001)
- CSPF (not standardized; variant of Dijkstra's SPF)

TED

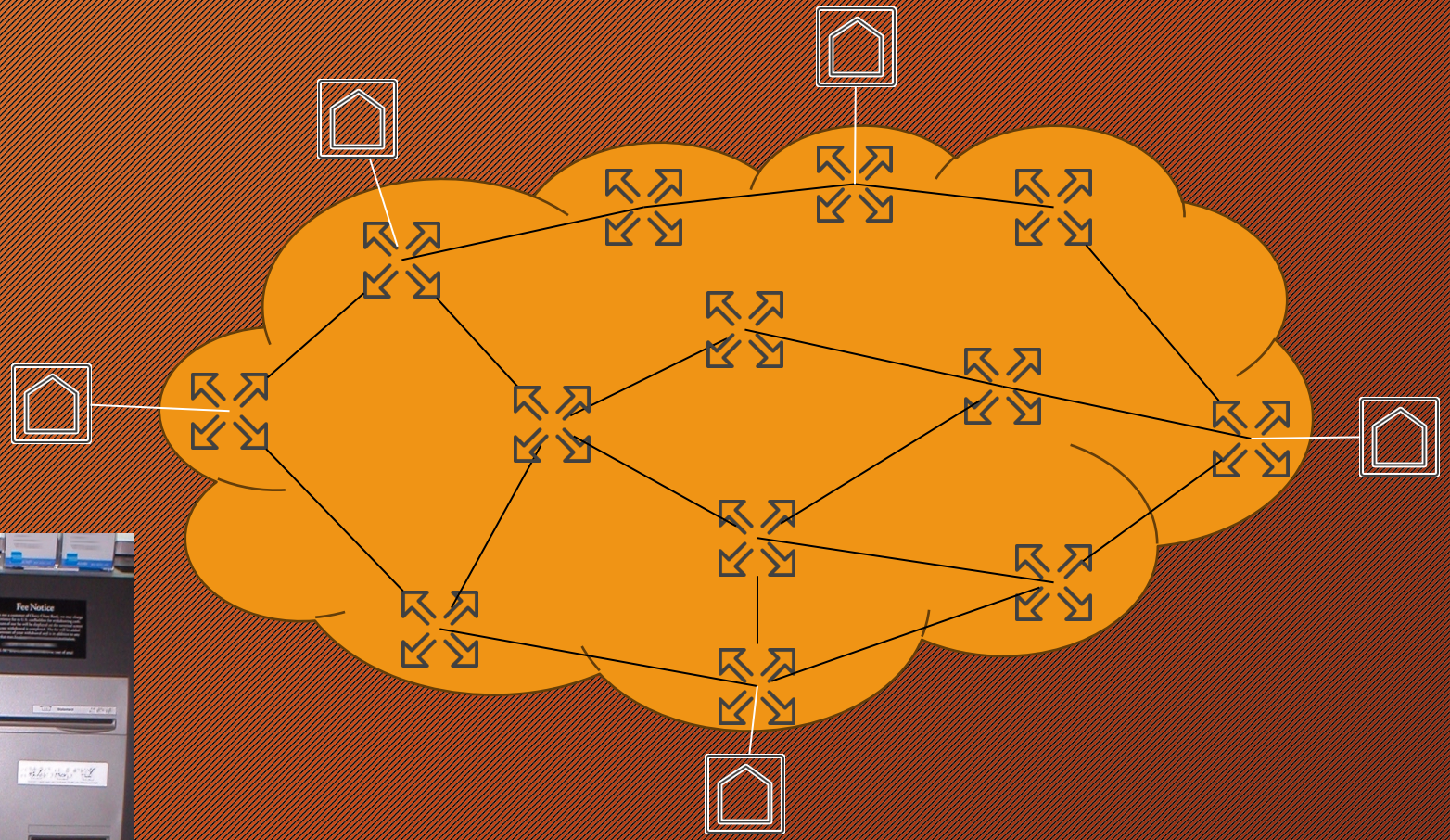
- My first exposure to rpd code

Circuit cross-connect (CCC) (1999)

- MPLS = multi-protocol “above” and “below”
  - Below: run over any layer 2 encap
  - Above: carry any type of traffic (not just IP)
- Can you replace an ATM network (say) with MPLS?
  - Tail circuits would have to remain ATM; “core” would change to MPLS
- Generalize: Can you carry any Layer 2 frame over MPLS?
- CCC was designed to connect a pair of {PPP, Cisco HDLC, Frame Relay, ATM or Ethernet} ifls across a pair of RSVP-TE tunnels
- CCC could also directly cross-connect a pair of ifls
- Translational Cross-Connect (TCC) was a related technology tha connected IP traffic between unlike Layer 2 ifls



ATM Network (“before”)



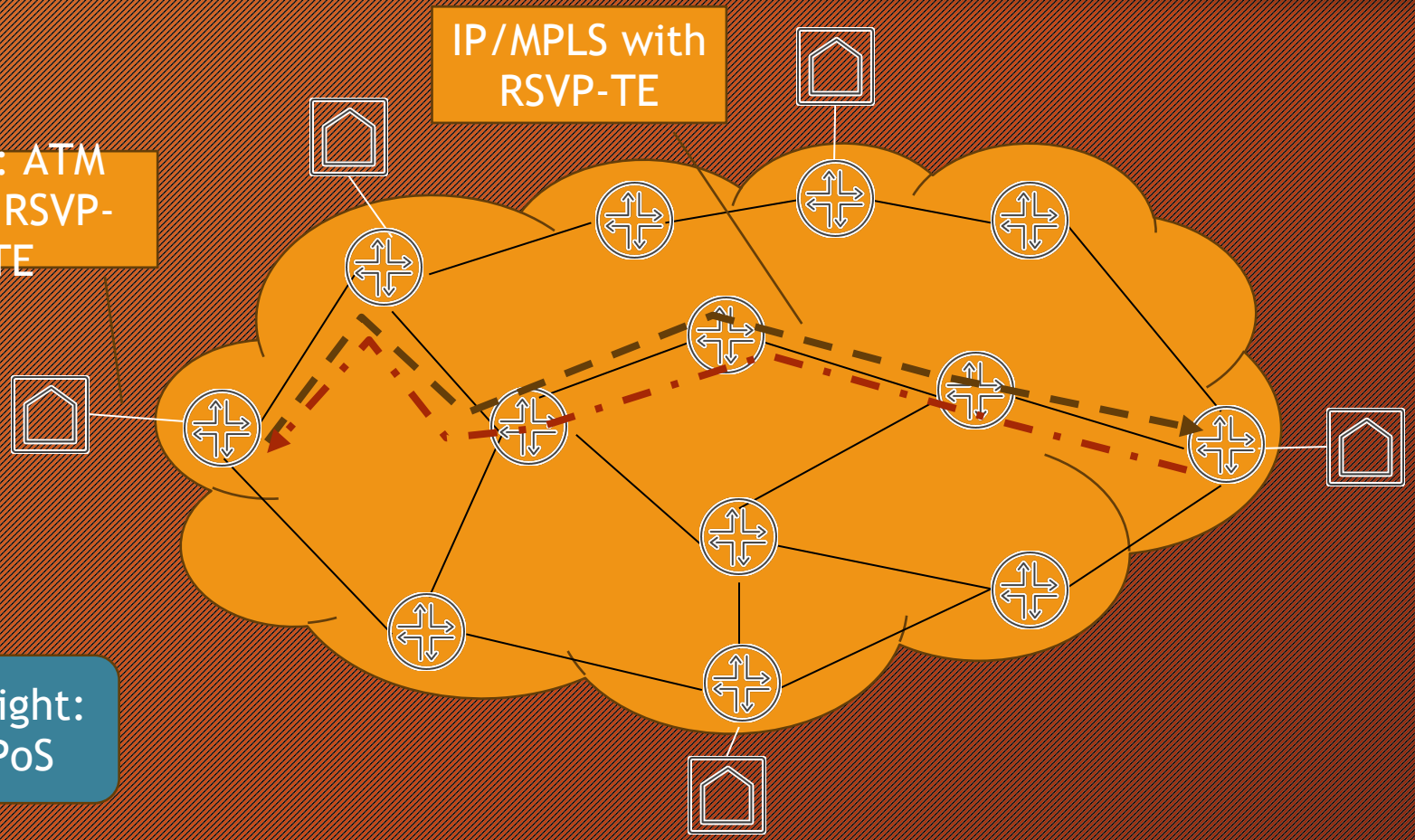
Theory: SP builds an ATM network to carry “multi-protocol” traffic + QoS

Gives ATM handoffs to customers for whatever traffic they may have (AAL-5, voice, etc.)

Does Traffic Engineering

However, IP and ATM were unhappy bedfellows

ATM over MPLS (“after”)



Different fight: ATM vs. PoS

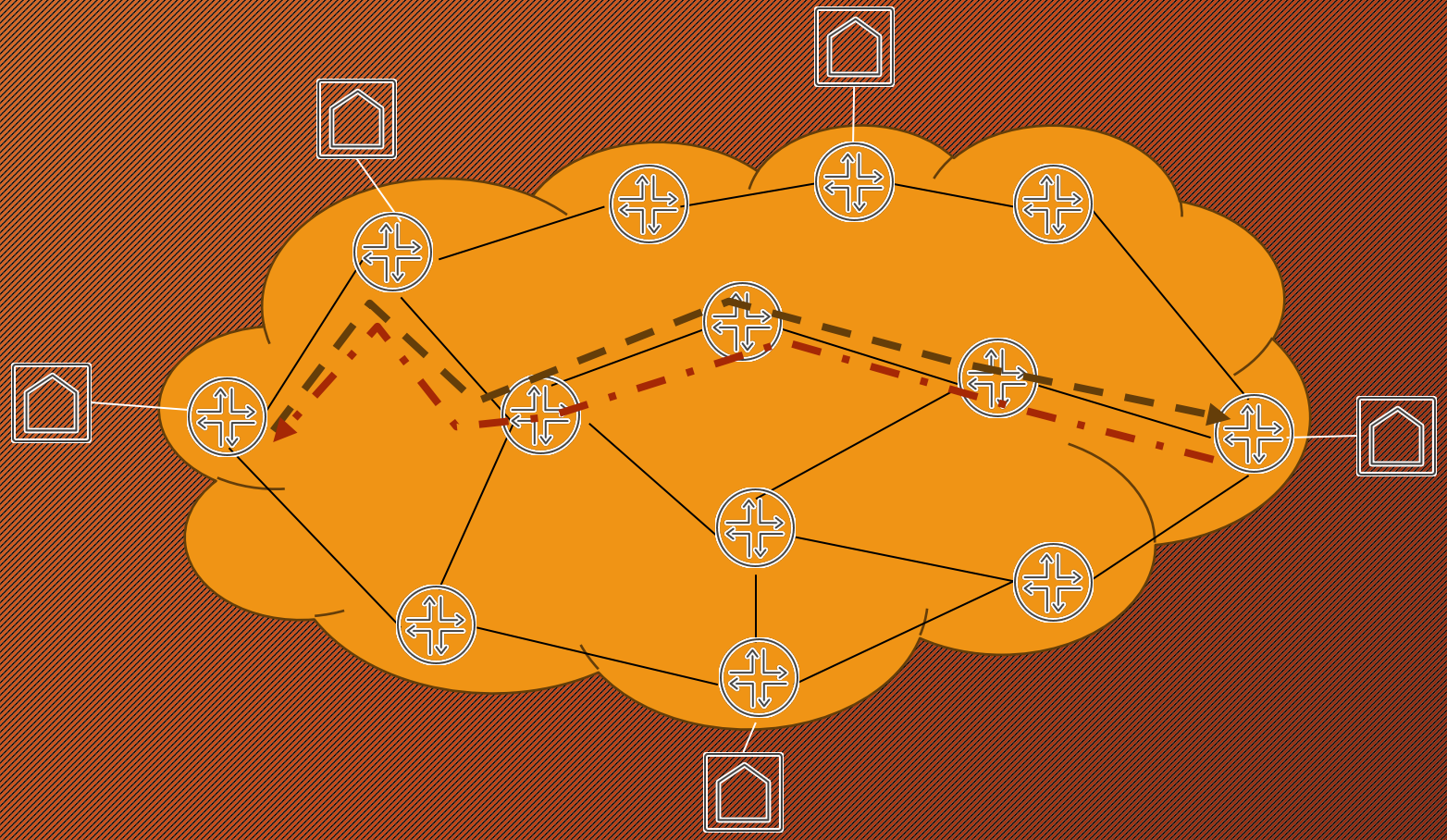
Steps

1. Replace ATM switches with Layer 3 (IP/MPLS) routers
2. Switch ATM encaps to MPLS/IP over PPP (“packet over sonet”)
3. Run IP and MPLS over PPP in core network
4. Carry ATM over MPLS for those customers who insist on ATM

CCC: top-to-bottom project

- From cli to rpd to dcd to kernel to microkernel
- Had to write B-chip microcode as well
  - Bi (parsing incoming packets) and Bo (packet rewrite)
  - Interestingly, the ABCD chips were called the “Martini chipset”
- Done over three months of working 90% from home (pre-Covid ☺) over a 128kbps DSL line (!)
  - Mostly from 8pm to 5am
  - With an ear open for my sleeping three-month-old daughter

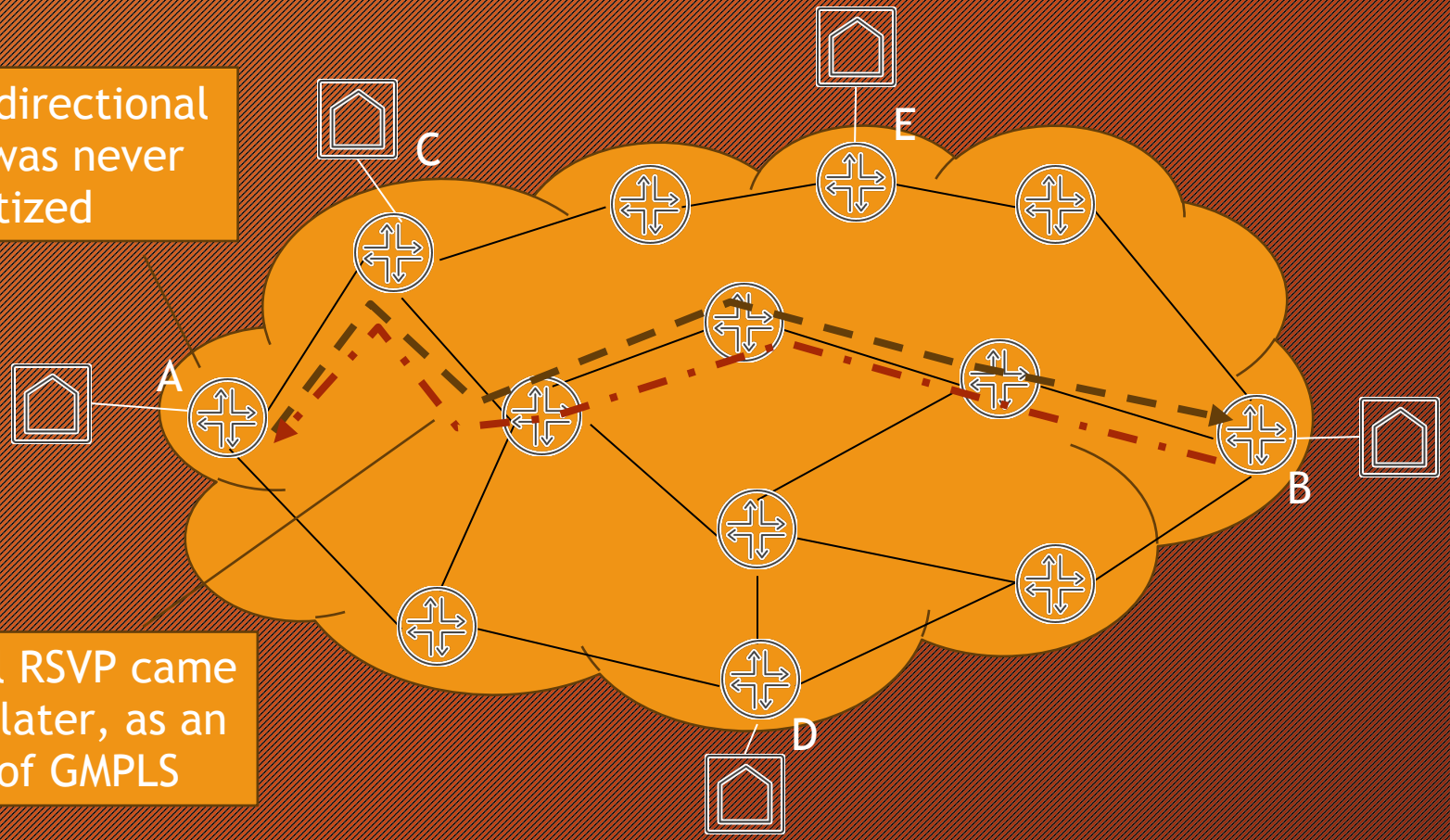
ATM over MPLS with CCC: issues



1. Must (manually) create *n*(n-1)* CCC connections for *n* endpoints
   - Prone to errors
2. No indication that all connections belong to one “VPN”
   - Debugging hell
3. Any topology can be created, but one must do so manually

Solution 1: “single-sided” provisioning

CCC had a bidirectional mode that was never productized



Bidirectional RSVP came many years later, as an outcome of GMPLS

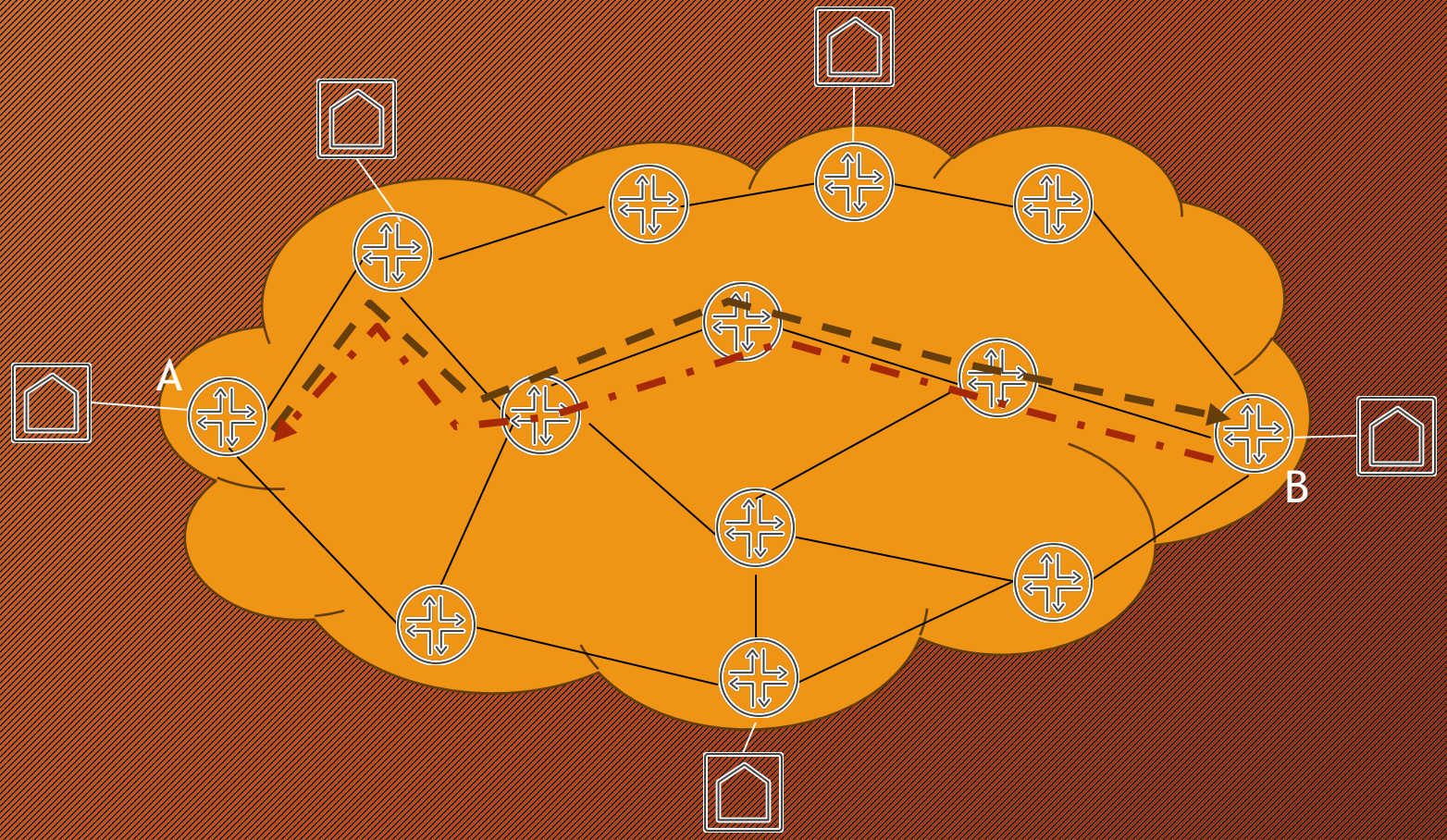
Provision one end of CCC connection for both directions: at router A, provision both the A->B and the B->A connections

Asymmetry is annoying:

- provision A for AB/BA
- provision A for AC/CA
- … or provision C for AC/CA?
- Similarly for D, E, …

Also, factor of 2: meh!

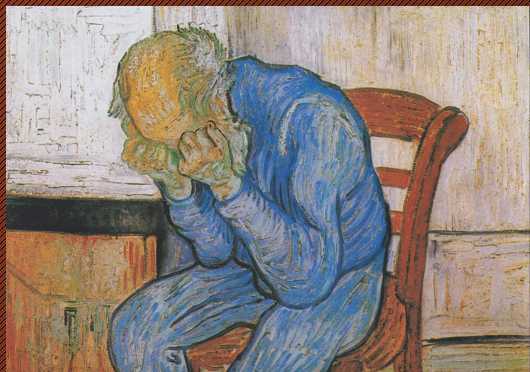
Solution 2: Use RFC 2547 technology (BGP)



- Provision each endpoint independently (“merely local intervention”)
- A single Layer 2 VPN (all endpoints are related)
- Provisioning complexity: *n* rather than *n*^2
- Managing becomes easier
- Adding a new site is easy
- *Unified* approach to VPNs

Problems to solve

- Need *n-1* labels in each BGP advertisement
  - Solved by using label blocks (base plus range)
- Need hub-and-spoke/dual-hub-and-spoke topologies
  - Solved by using Route Targets (same as for RFC 2547)
- Need to standardize?
  - Nah, too much opposition! (RFC 6624 is *Informational*)
- Need Transparent LAN Service ("LAN Emulation over MPLS")
  - Adapt technology (RFC 4761 – yes, this one is Standards Track)



Objections from Nay-sayers

- “You cain’t use BGP for Layer 2 information”
  - Actually, no Layer 2 information was carried in BGP L2VPN, just config info
  - Now, of course, MAC addresses are carried in BGP, and that’s just fine!
  - Also, flow information (Layer 4+) is also carried in BGP (flow-spec)
- You can’t do per connection QoS
  - You can if you want to badly enough – but mostly, people didn’t want it
- BGP is too complicated for most people
  - Bogus, bogeyman objection to scare people (but it worked often!)
- Rinse, repeat

Not a perfect fit (Layer 2 vs. IP)

- IP: destination-based forwarding
  - IP forwarding doesn't care what the source is
- Layer 2: circuit-oriented
  - Source is implicitly or explicitly part of forwarding
  - Also, per-circuit QoS ☺
  - Especially needed for MAC address learning (VPLS, EVPN)
- This is primarily a data plane problem
  - But it needs some help from control plane
  - Hence label blocks

This was never brought up as an objection to using BGP or IP VPN technology!

Technology Choices

- Use LDP for peer-to-peer signaling; tell each PE out-of-band (e.g., using an NMS) who the peer PEs are
- Use LDP for peer-to-peer signaling; use BGP for auto-discovery
  - Overly complex solution; rarely seen in practice
- Use BGP for both auto-discovery and for signaling (as in RFC 2547)
  - Can make effective use of Route Reflectors, Route Target filtering, ORFs
  - Inter-AS VPNs (“option B” and “option C”) work very much like RFC 2547

Adjacent Applications

- Initially, used BGP auto-discovery and signaling for L2VPN (aka VPWS, or point-to-point PWs)
- Used pretty much the same technology for Virtual Private LAN Service (aka Transparent LAN Service)
  - Control plane is very similar; data plane changed to include MAC learning
- BGP auto-discovery and signaling is used for EVPN (aka mac-vrf)
  - In EVPN, BGP carries, in addition to auto-discovery and labels, MAC and IP addresses
  - So much for “BGP MUST NOT be used to carry Layer 2 information” ☺

Vindication? Nah, no such thing

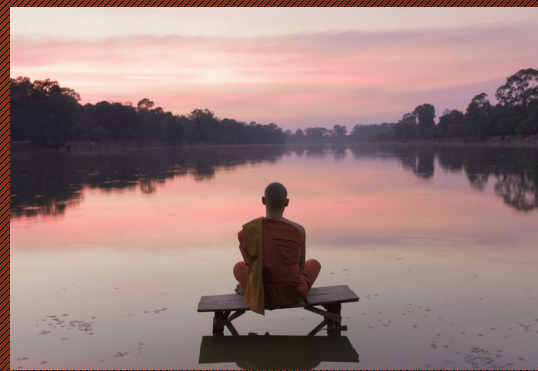
- Vigorous discussions and divergent opinions are the lifeblood of technical progress
  - But dogmatic adherence to one's own ideas, maybe not so much
  - … especially for commercial, not technical, reasons
- The industry wasted time, effort, … and will do so again
- At a conference, a speaker peddling "Rosen-style" mVPN made a statement to the effect "BGP was not designed to carry multicast VPN routes … too much scale and churn" …
  - … and got shredded by Yakov Rekhter
  - … and did the speaker learn? Sadly, no. Next conference, same message

What's Next?

- The chapter on Layer 2 VPNs is coming to a close …
  - The only interesting Layer 2 technology now is Ethernet …
  - … so the focus is on EVPN (BGP-based, of course!)
  - But there is a clear move to IP all around
- VPNs themselves are morphing to "SD-WAN" – do-it-yourself
- MPLS is changing too
- So much more is being done with BGP (flow-spec, link-state, CT)
- But what's "in the ground" will be there for quite a long time

Lessons Learned

- Find the right advisors
- Do your homework
- Trust your gut
- Be stubborn (but listen)!
- “The more strident their shouting, the more encouraged you should be”
- When done, move on ☺
  - Easy to say in retrospect!



The Patent I'm Proudest Of

US7136374B1
(applied 2001-05; issued 2006-11)

Abstract
A layer 2 transport network, and components thereof, supporting virtual network functionality among customer edge devices. Virtual private network configuration can be accomplished with merely local intervention by preprovisioning extra channel (or circuit) identifiers at each customer edge device and by advertising label base and range information corresponding to a list of channel (or circuit) identifiers.