

Mythology and Folklore of Network Protocol Design

Radia Perlman
Sun Microsystems Laboratories

Messages

- Dispel myths and “religion”
 - *“It’s not what you don’t know that’ll get you. It’s what you do know that ain’t true”*
Mark Twain
- Learn from mistakes
- Learn from cool ideas
- Be provocative. Start lively discussion

Messages for students

- Don't believe everything you hear
- Don't believe things you cannot understand
- There is a lot of randomness in whether papers get accepted
- Standards creation is more political than technical

General stuff to consider in designs

- Simplicity
- Scalability
- Manageability
- Robustness
- Ease of adding new features

First a bit of background in order
to make the examples clear

What are protocol layers?

- Just a way of thinking about the problem
- ISO defined 7 layers
- TCP/ IP suite claims it's only 4 layers, but has at least 6 of the ISO layers
 - Perhaps leaves out session layer, but “BEEP” was in fashion for awhile
- A lot of the layers get subdivided into others

Bridges, Routers, and Switches! Oh my!

- This discussion sheds light on how/ why things work today
- Need the background for some other examples

Why this whole layer 2/3 thing?

- Myth: bridges/ switches simpler devices, designed before routers
- OSI Layers
 - 1: physical

Why this whole layer 2/3 thing?

- Myth: bridges/ switches simpler devices, designed before routers
- OSI Layers
 - 1: physical
 - 2: data link (nbr-nbr)

Why this whole layer 2/3 thing?

- Myth: bridges/ switches simpler devices, designed before routers
- OSI Layers
 - 1: physical
 - 2: data link (nbr-nbr)
 - 3: network (create entire path)

Why this whole layer 2/3 thing?

- Myth: bridges/ switches simpler devices, designed before routers
- OSI Layers
 - 1: physical
 - 2: data link (nbr-nbr)
 - 3: network (create entire path)
 - 4 end-to-end

Why this whole layer 2/3 thing?

- Myth: bridges/ switches simpler devices, designed before routers
- OSI Layers
 - 1: physical
 - 2: data link (nbr-nbr)
 - 3: network (create entire path)
 - 4 end-to-end
 - 5 and above: boring

Definitions

- Repeater: layer 1 relay
- Bridge: layer 2 relay
- Router: layer 3 relay

Definitions

- Repeater: layer 1 relay
- Bridge: layer 2 relay
- Router: layer 3 relay
- OK: What is layer 2 vs layer 3?

Definitions

- Repeater: layer 1 relay
- Bridge: layer 2 relay
- Router: layer 3 relay
- OK: What is layer 2 vs layer 3?
 - My definition: layer 3 forwards, layer 2 does not

Definitions

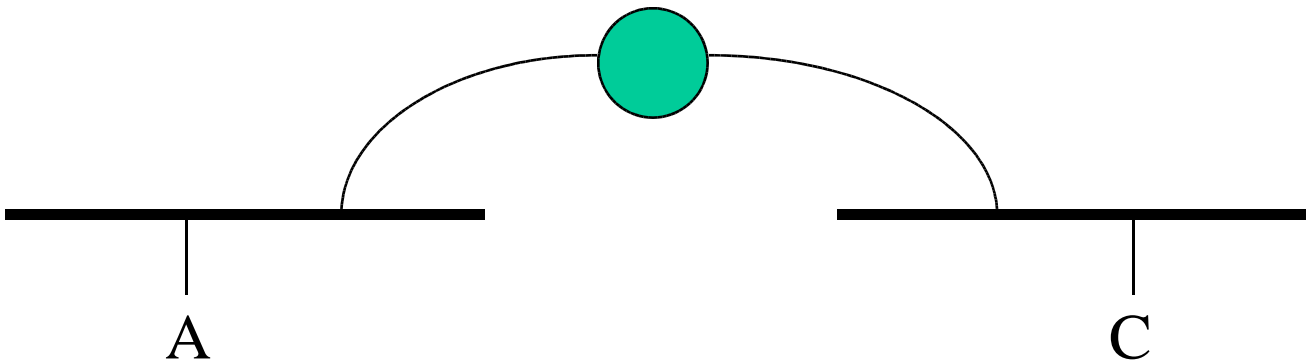
- Repeater: layer 1 relay
- Bridge: layer 2 relay
- Router: layer 3 relay
- OK: What is layer 2 vs layer 3?
 - True definition of a layer n protocol:
Anything designed by a committee whose charter is to design a layer n protocol

Layer 3 (DECnet, IP)

- Put source, destination, hop count on packet
- At the time DECnet was more prevalent, but it's logically equivalent to IP
- Then along came “the EtherNET”
 - rethink routing algorithm a bit, but it's a link!
- The world got confused. Built on layer 2
- I tried to argue: “*But you might want to talk from one Ethernet to another!*”
- “*Which will win? Ethernet or DECnet?*”

Problem Statement

Need something that will sit between two Ethernets, and let a station on one Ethernet talk to another



Basic idea

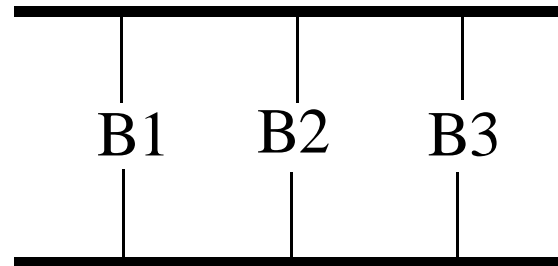
- Listen promiscuously
- Learn location of source address based on source address in packet and port from which packet received
- Forward based on learned location of destination

What's different between this and a repeater?

- no collisions
- with learning, can use more aggregate bandwidth than on any one link
- no artifacts of LAN technology (# of stations in ring, distance of CSMA/CD)

But loops are a disaster

- No hop count
- Exponential proliferation



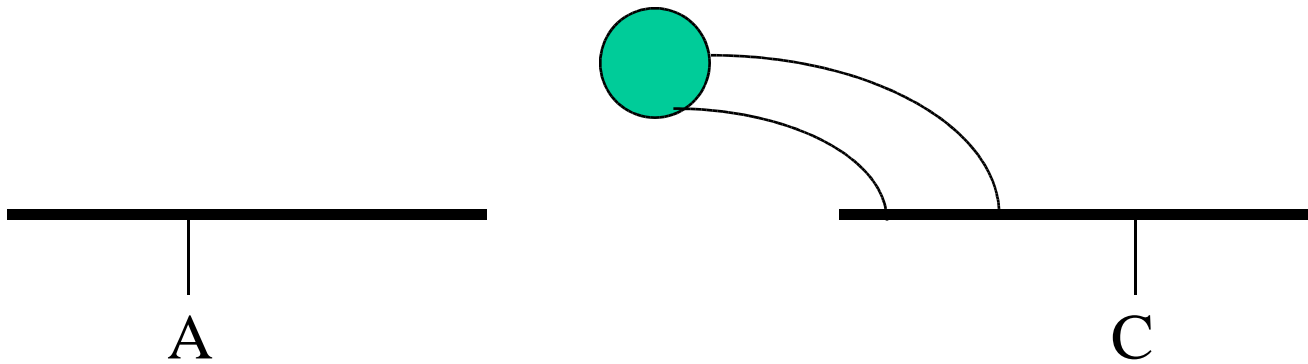
Thus the Spanning Tree Algorithm

*I think that I shall never see
A graph more lovely than a tree.
A tree whose crucial property
Is loop-free connectivity.
A tree which must be sure to span
So packets can reach every LAN.
First the Root must be selected
By ID it is elected.
Least cost paths from Root are traced
In the tree these paths are placed.
A mesh is made by folks like me.
Then bridges find a spanning tree.*

Bother with spanning tree?

- Maybe just tell customers “don’t do loops”
- First bridge sold...

First Bridge Sold



Myth

- Ethernet continues to be a successful technology

So what is Ethernet?

- CSMA/CD, right? Not any more, really...
- source, destination (and no hop count)
- limited distance, scalability (not any more, really)

Switches

- Ethernet used to be bus
- Easier to wire, more robust if star (one huge multiport repeater with pt-to-pt links)
- If store and forward rather than repeater, and with learning, more aggregate bandwidth
- Can cascade devices...do spanning tree
- We're reinvented the bridge!

Simple things people get wrong

- They get obsessed with esoteric stuff like “provable properties” of cryptographic algorithms, but miss basic system issues

Example: What is a version number?

Version Numbers

- What's the difference between a new protocol, and a new version of an existing protocol?
- For instance, why was CLNP a “new protocol”, and IPv6 a “new version of IP”?

Version Numbers

- What's the difference between a new protocol, and a new version of an existing protocol?
- For instance, why was CLNP a “new protocol”, and IPv6 a “new version of IP”?
 - Its name?
 - Who defines it?

Definition that makes sense to me

- New protocol means different protocol discriminator at layer n-1
- New version means same protocol discriminator, and version number distinguishes

If you distinguish with version number

- Then if the packet format is incompatible, the old version node must not try to parse
- Don't increase version number unless packet is incompatible
- Specify packet must be dropped if version number is bigger

Is IPv6 a new version of IPv4?

- IPv4 spec says “set version to 4”
- But doesn't say to look at it
- IPv6 format incompatible with IPv4
- So if you send an IPv4 node an IPv6 packet, it will do who knows what...

Result

- IPv6 needs new protocol type
- So IPv6 is a new protocol, not a new version of IP
- IPv6 does have a version number field, but it could be version 1

We learned our lesson, right?

- So IPv6 spec must say “drop if version number > 6 ”

We learned our lesson, right?

- So IPv6 spec must say “drop if version number > 6 ”
- Nope...just says “set this field to 6”

Another example: SSL

- They completely changed the format from SSLv2 to v3
- Not only didn't say to drop if version number greater...

Another example: SSL

- They completely changed the format from SSLv2 to v3
- Not only didn't say to drop if version number greater...
- **But moved the version number field!!!**

Parameters

Parameters

- Minimize these:
 - someone has to document it
 - customer has to read documentation and understand it
- How to avoid
 - architectural constants if possible
 - automatically configure if possible

Settable Parameters

- Make sure they can't be set incompatibly across nodes, across layers, etc. (e.g., hello time and dead timer)
- Make sure they can be set at nodes one at a time and the net can stay running

Parameter tricks

- IS-IS
 - pairwise parameters reported in “hellos”
 - area-wide parameters reported in LSPs
- OSPF
 - copied most of IS-IS, but got this wrong.
Use field in hello to refuse to talk if not identical!
- Bridges
 - Use Root’s values, sent in spanning tree
msgs

What's with IPv6?

What's with IPv6?

- A connectionless network layer consists of:
 - Source address
 - Destination address
 - Hop count

What's with IPv6?

- A connectionless network layer consists of:
 - Source address
 - Destination address
 - Hop count
- What could take so long?

What's with IPv6

- In 1992, IAB said
 - Gee, IP addresses not big enough
 - Why don't we use CLNP?

What's with IPv6?

- In 1992, IAB said
 - Gee, IP addresses not big enough
 - Why don't we use CLNP?
- What's CLNP?
 - ISO's version of IP
 - 20 byte addresses
 - Came with mature routing protocols, autoconfiguration, implemented by all major vendors

What's with IPv6?

- But some vocal IETF people said
 - We can't replace IP with ISO!
- Result
 - We may have missed our window
 - Giving a large committee 13 years (and counting) of time, you can generate lots of pages of specs
 - CLNP would have been fine
 - And we'd have bigger addresses now (and a simpler protocol)

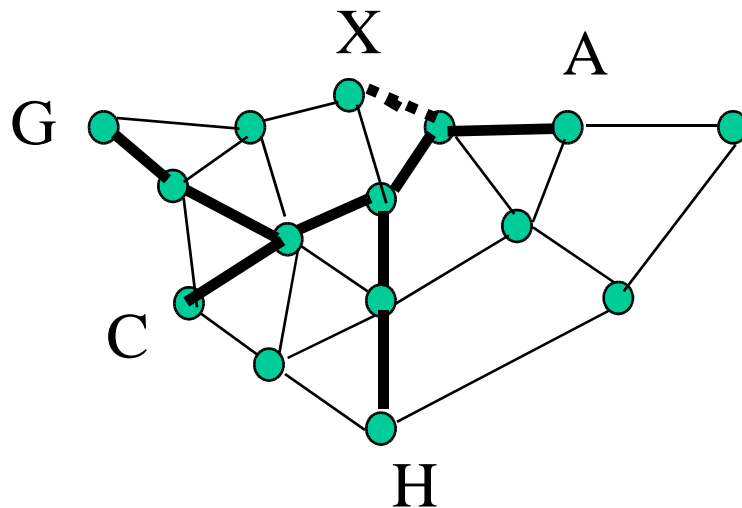
Result

- We may have missed our window
- Giving a large committee 13 years (and counting) of time, you can get lots of pages of specs
- CLNP would have been fine
- And we'd have bigger addresses now (and a simpler protocol than IPv6)
- Worse yet, we get IPv4 and NATs, and, if we migrate, very complex migration

What's with IP Multicast?

Multicast

- Ethernet: falls out of technology
- ATM: create VC. “Add member”



IP Multicast

- Idea: make it look “just like Ethernet”
 - globally unique multicast addresses
 - IP address 32 bits, top 4 bits=1110
 - anyone can request to listen. anyone can send without being a member
- So, start out with unchangeable “model”
 - signalling protocol to inform local rtr to send G

Problem: Can't be implemented

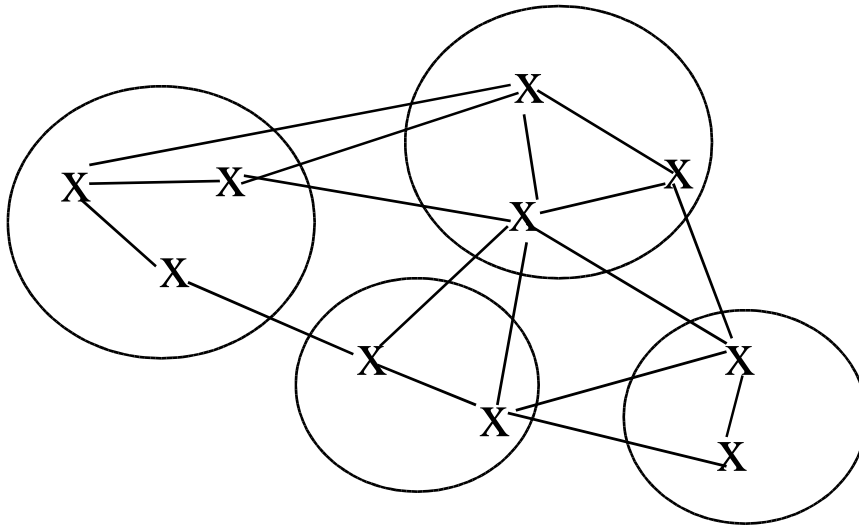
- various attempts:
 - flood and prune
 - send all data everywhere, in case someone in Albania wants to listen
 - if not interested, send “prune”
 - keep track of all (S,G) pairs nbr NOT interested in
 - MOSPF
 - routers keep track of all listeners for all groups

IP Multicast attempts

- Tree building like with ATM
 - send join towards Root
 - create tree
- Problems:
 - who is Root for G?
 - unscalable intradomain protocol to select a Root-candidate for G
 - how to administer addresses

IP Multicast

- So, came up with unscalable complex intradomain
- Then MSDP to piece domains together



How IP Multicast should look

- Two types
 - finding something (low bandwidth, can't set up tree). Just flood with RPF
 - conference call, etc. Find host H. Build tree to H. Have address of group be (H,G), where G only has to be unique to H

Self-Stabilization

- Bad things may happen
 - sick or malicious devices might corrupt databases or inject bad traffic
- Once bad device disconnected from the net, the network should return to normal operation
- How could it not?

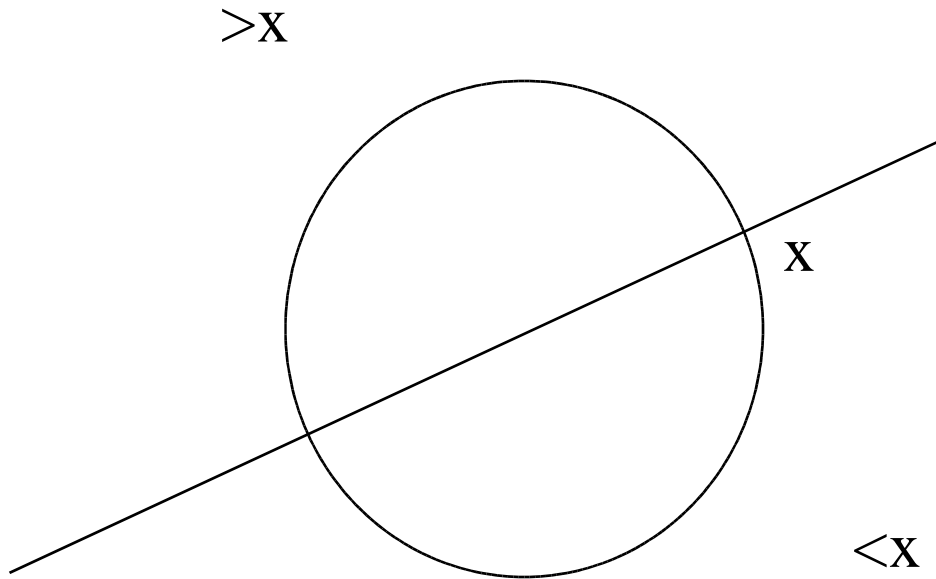
Link State Routing, ARPANET style

- Link state routing
 - figure out who your nbrs are
 - create LSP (who I am, who my nbrs are)
 - flood LSP, keep most recently generated LSP from each other router
 - use LSP database to calculate paths

How to flood

- Regular flooding is exponential
- But here, only flood each packet once (if newer than that in database)
- How to recognize packet is new?
- ARPANET
 - sequence number and age
 - sequence number circular
 - age increments after holding it for n seconds

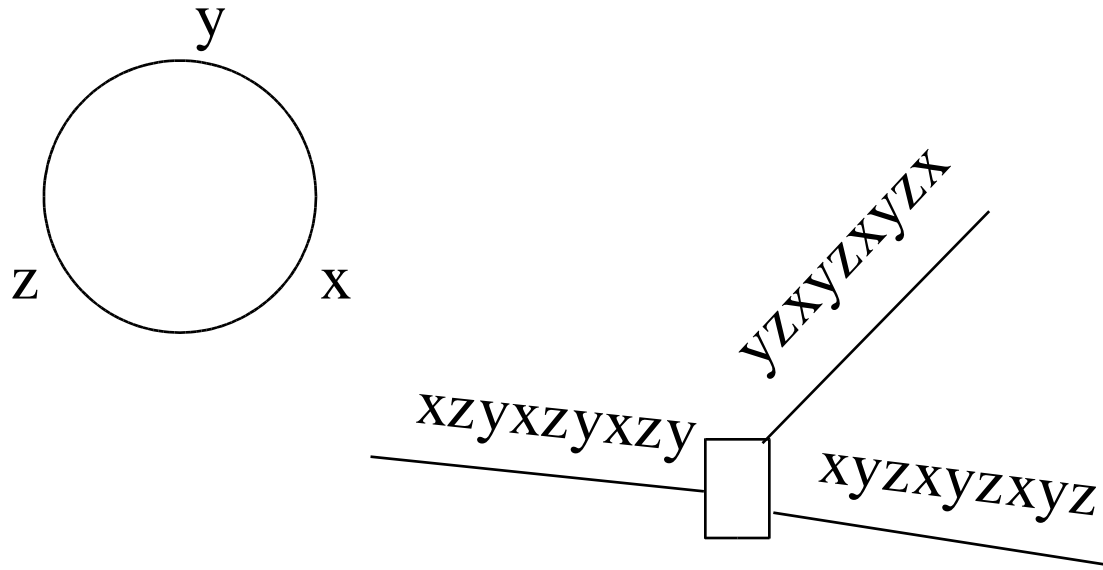
Arithmetic in circular space



ARPANET disaster

- symptom: net didn't work
- how do you diagnose and manage a network?
- Note: these guys were really really lucky!
- What had happened: Fred, a sick router, generate bad LSPs before dying, with sequence numbers x, y, z

ARPANET disaster



So how do you fix a broken net?

- Patched version of code that ignore LSPs from Fred
- One by one crashed systems (not easy!) and reloaded with patched code
- Only after all routers reloaded, can they be reloaded with correct version again

Robustness

- Can be designed to be self-stabilizing (paper from 1983)
- Paper claimed “but can’t expect the network to continue operating if faulty equipment is still connected”
- My thesis from 1988: routing with Byzantine robustness

Other things that I could rant about

- XML
- BGP

My complaint about how networking is taught

- It's taught like a trade school...all the details of the currently deployed stuff

How networking should be taught

- Cover conceptual problems
- And range of solutions
 - Interesting ideas, even if not currently deployed
 - Interesting ideas, even if never deployed
- Teach how to think critically
 - Don't believe everything in print
 - Don't assume what comes out of standards bodies is perfect

Mistakes standards bodies make

- “We should get extra credit because we didn’t look at any ideas done before”
- “We are too busy to answer basic questions, or explain anything”
- “If we revisit old decisions, we’ll lose 10 years worth of work”
 - If you find yourself in a hole, stop digging!

Lessons

- Always seems easy to start over with new thing. Always takes longer and comes out worse.
- Don't cast something in stone before there is a plausible way of realizing it
- Minimize configuration
- Don't just dive in and start doing stuff. Think about what problem you're solving before you try to come up with a solution.