Multicast Help Wanted: From Where and How Much?

Kevin Almeroth (almeroth@cs.ucsb.edu)
University of California—Santa Barbara

P2PM 2007

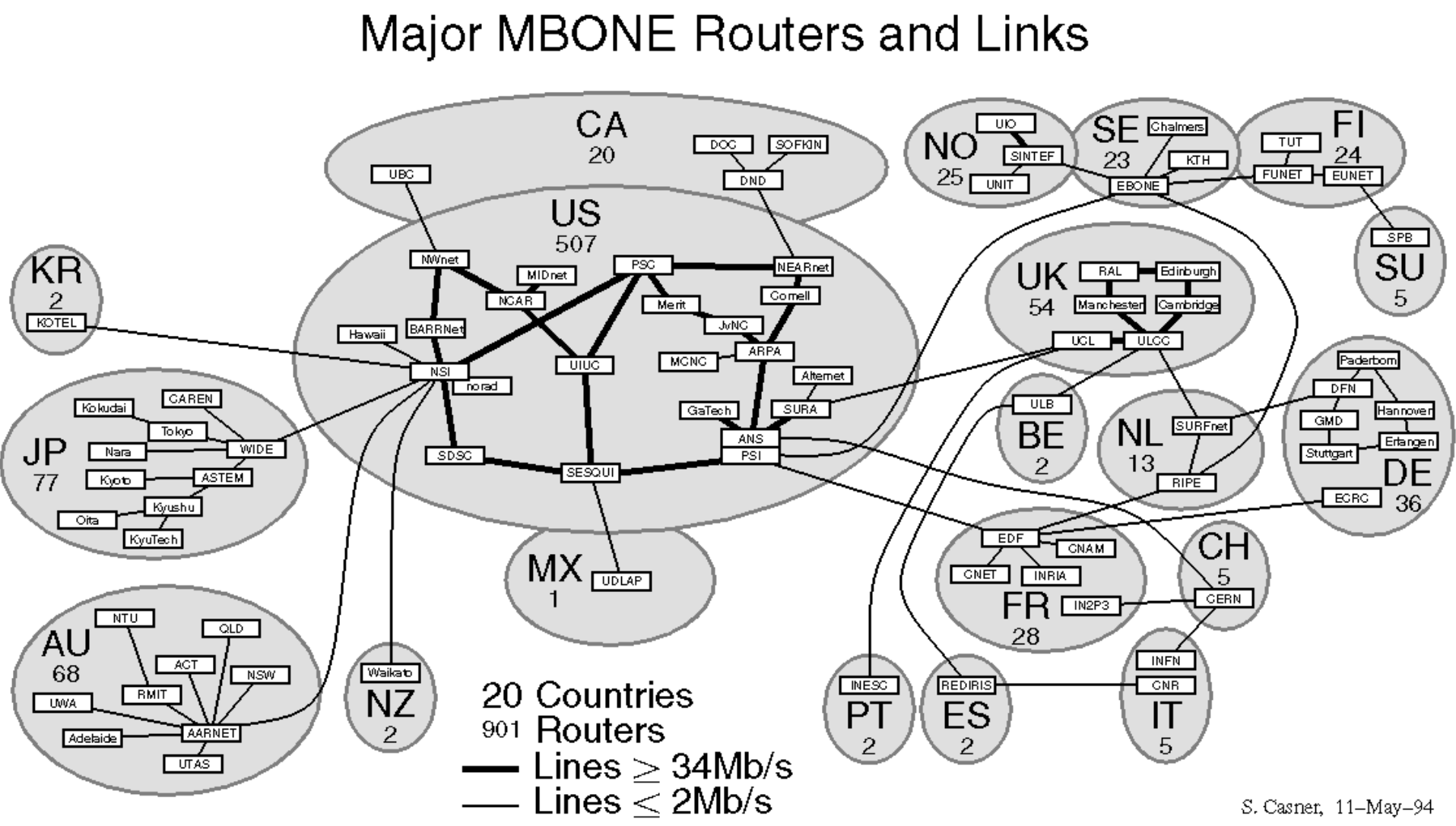
Terminology

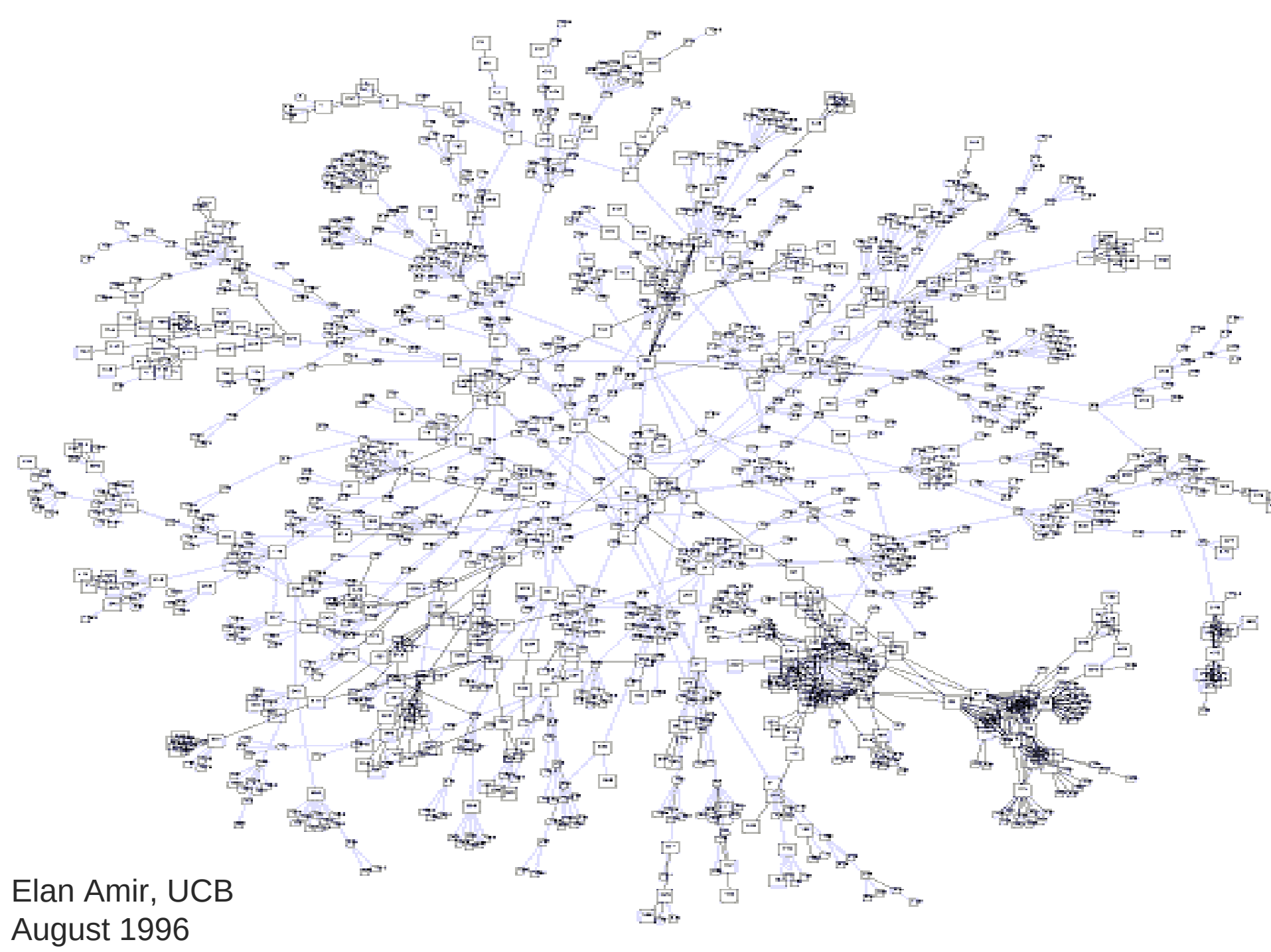
- Assume that everyone knows the terminology:
  - “native” multicast
  - ALM
    - Overlay (assumes proxies in the network)
    - End System (only at the edges/hosts)
- Assume everyone knows a bit about the native multicast protocol acronym soup:
  - DVMRP, PIM-DM, MOSPF
  - PIM-SM
  - IGMPv2, IGMPv3, MLDv2, MLDv3
  - MSDP
  - MBGP/BGP4+
  - ASM v. SSM

History

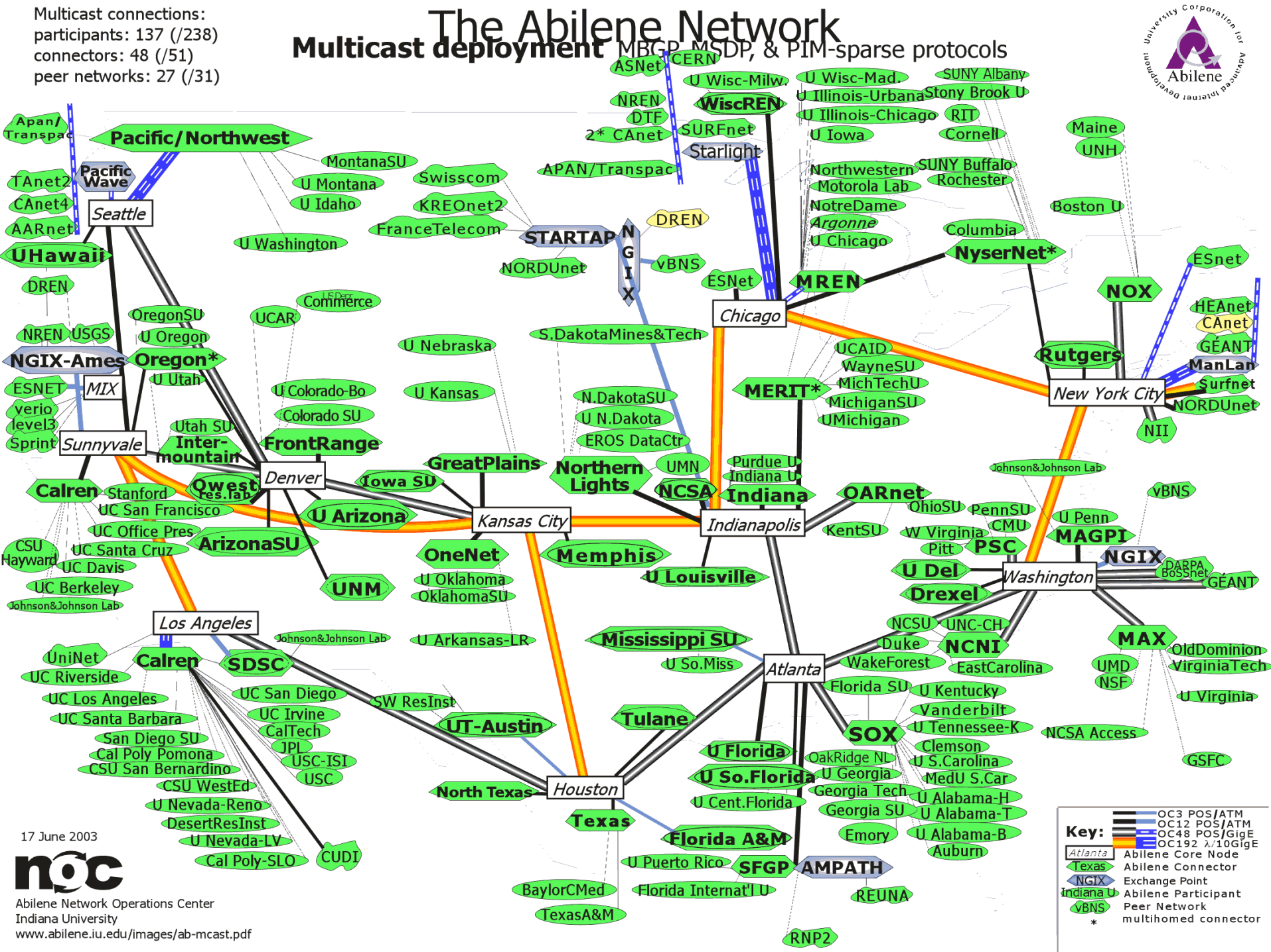
- It was all ALM (end system) in the beginning.
  - Started at the March 1992 IETF
  - The end-system software was called “mrouted”
  - Originally used DVMRP—a broadcast-and-prune protocol
  - Worked really well, but was small, and had no scale
- Eventually evolved into an overlay multicast environment
  - Some of the mrouted boxes were located in the core
- Eventually evolved into a hybrid environment
  - When “native” multicast support became available, the challenge became to connect the islands together.
- Eventually we got rid of the “MBone” and just had native

History


Major MBONE Routers and Links
CA
20
US
507
UBC
DOC
SOFKIN
DND
NWnet
MIDnet
PSC
NEARnet
NCAR
Cornell
Merit
JvNC
Hawaii
BARRNet
MCNC
ARPA
NSI
UIUC
norad
Alternet
GaTech
SURA
ANS
PSI
SDSC
SESQUI
KR
2
KOTEL
JP
77
Kokudai
CAREN
Tokyo
Nara
WIDE
Kyoto
ASTEM
Kyushu
Oita
KyuTech
AU
68
NTU
QLD
ACT
NSW
RMIT
UWA
Adelaide
AARNET
UTAS
NZ
2
Waikato
MX
1
UDLAP
NO
25
UIO
SINTEF
UNIT
SE
23
Chalmers
KTH
EBONE
FI
24
TUT
FUNET
EUNET
SPB
SU
5
UK
54
RAL
Edinburgh
Manchester
Cambridge
UCL
ULCC
BE
2
ULB
NL
13
SURFnet
RIPE
DE
36
Paderborn
DFN
Hannover
GMD
Erlangen
Stuttgart
ECRC
CH
5
CERN
FR
28
EDF
CNAM
CNET
INRIA
IN2P3
PT
2
INESC
ES
2
REDIRIS
IT
5
INFN
CNR
20 Countries
901 Routers
Lines ≥ 34Mb/s
Lines ≤ 2Mb/s
S. Casner, 11–May–94



Elan Amir, UCB
August 1996


Multicast connections:
participants: 137 (/238)
connectors: 48 (/51)
peer networks: 27 (/31)
The Abilene Network
Multicast deployment MBGP, MSDP, & PIM-sparse protocols
University Corporation for Advanced Internet Development
Abilene
Apan/ Transpac
Pacific/Northwest
Pacific Wave
TAnet2
CAnet4
AARnet
Seattle
UHawaii
DREN
MontanaSU
U Montana
U Idaho
U Washington
Swisscom
KREOnet2
FranceTelecom
STARTAP
NGIX
DREN
vBNS
NORDUnet
APAN/Transpac
ASNet
CERN
U Wisc-Milw.
NREN
DTF
2* CAnet
WiscREN
SURFnet
Starlight
U Wisc-Mad.
U Illinois-Urbana
U Illinois-Chicago
U Iowa
SUNY Albany
Stony Brook U
RIT
Cornell
Maine
UNH
Northwestern
Motorola Lab
NotreDame
Argonne
U Chicago
SUNY Buffalo
Rochester
Boston U
Columbia
NyserNet*
ESNet
MREN
Chicago
NOX
HEAnet
CAnet
GÉANT
ManLan
Surfnet
NORDUnet
NII
Rutgers
New York City
UCAID
WayneSU
MichTechU
MichiganSU
UMichigan
MERIT*
Johnson&Johnson Lab
vBNS
U Penn
MAGPI
NGIX
DARPA
BoSSnet
GÉANT
Washington
U Del
Drexel
PSC
W Virginia
Pitt
CMU
PennSU
OhioSU
KentSU
OARnet
Indianapolis
Indiana
Purdue U
Indiana U
UMN
NCSA
Northern Lights
GreatPlains
Iowa SU
Kansas City
Memphis
U Louisville
OneNet
U Oklahoma
OklahomaSU
U Arizona
ArizonaSU
UNM
Denver
FrontRange
Inter-mountain
Utah SU
Qwest
Res.lab
U Colorado-Bo
Colorado SU
U Kansas
U Nebraska
S.DakotaMines&Tech
N.DakotaSU
U N.Dakota
EROS DataCtr
Commerce
UCAR
OregonSU
U Oregon
Oregon*
U Utah
NGIX-Ames
MIX
NREN
USGS
ESNET
verio
level3
Sprint
Sunnyvale
Calren
Stanford
UC San Francisco
UC Office Pres
UC Santa Cruz
UC Davis
UC Berkeley
Johnson&Johnson Lab
CSU Hayward
Los Angeles
UniNet
Calren
SDSC
UC Riverside
UC Los Angeles
UC Santa Barbara
San Diego SU
Cal Poly Pomona
CSU San Bernardino
CSU WestEd
U Nevada-Reno
DesertResInst
U Nevada-LV
Cal Poly-SLO
CUDI
UC San Diego
UC Irvine
CalTech
JPL
USC-ISI
USC
SW ResInst
UT-Austin
Houston
North Texas
Texas
Johnson&Johnson Lab
U Arkansas-LR
Mississippi SU
U So.Miss
Tulane
U Florida
U So.Florida
U Cent.Florida
Florida A&M
U Puerto Rico
SFGP
AMPATH
Florida Internat'l U
BaylorCMed
TexasA&M
REUNA
RNP2
Atlanta
NCSU
UNC-CH
Duke
WakeForest
NCNI
EastCarolina
Florida SU
U Kentucky
Vanderbilt
U Tennessee-K
SOX
Clemson
U S.Carolina
MedU S.Car
U Alabama-H
U Alabama-T
U Alabama-B
Auburn
Emory
Georgia SU
Georgia Tech
OakRidge NL
U Georgia
MAX
OldDominion
VirginiaTech
U Virginia
UMD
NSF
NCSA Access
GSFC
Key:
OC3 POS/ATM
OC12 POS/ATM
OC48 POS/GigE
OC192 λ/10GigE
Atlanta Abilene Core Node
Texas Abilene Connector
NGIX Exchange Point
Indiana U Abilene Participant
vBNS Peer Network
* multihomed connector
17 June 2003
ngc
Abilene Network Operations Center
Indiana University
www.abilene.iu.edu/images/ab-mcast.pdf

It was doomed soon after the start.

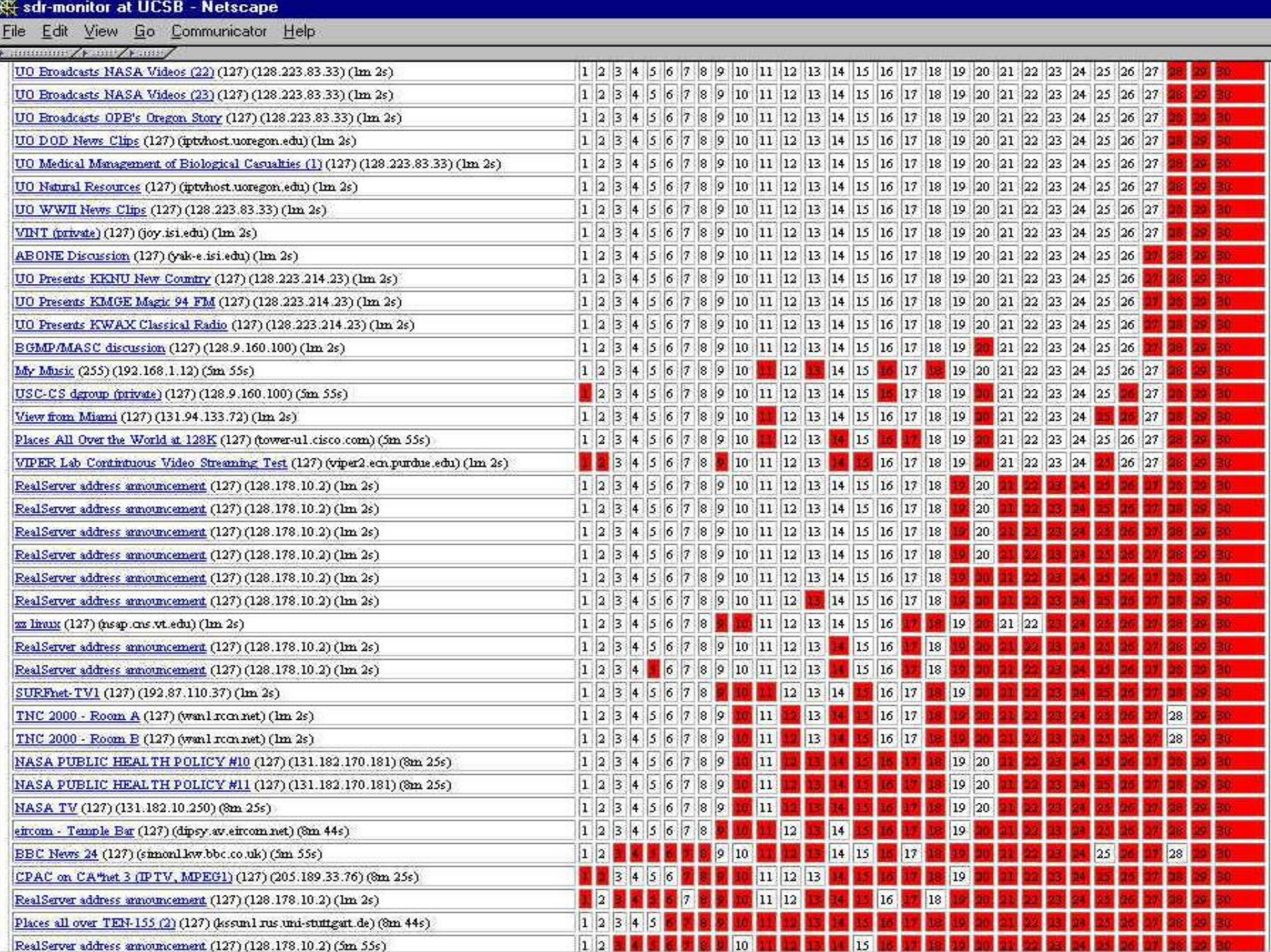
- Original architecture was based on Deering PhD dissertation which was for LAN-based multicast
  - We never got away from many of those assumptions
- The first step was a small one and it worked…
  - No scalability (broadcast and prune…), minimal requirements, but it worked!
- …but the second step was too big
  - Would only accept (nearly-)infinite scalability
  - "Small group multicast" was dismissed out-of-hand
    - See Ammar NOSSDAV 2003 keynote: http://www.cc.gatech.edu/fac/Mostafa.Ammar/nossdav-key.ppt

It was doomed soon after the start.

- The key application was streaming audio/video
  - Reliable data transfer didn’t enter into the picture until far too late
- And until very recently (surprisingly!), the economics of deployment and use were aggressively, proactively ignored
- In our defense, hindsight is 20/20

Original Problems

- Addressing
  - We only had “sd” then “sdr” to avoid address conflicts…and that was always broken
- Reliability and Congestion Control
  - “Not our problem”… wrong!
  - Solutions exist, but only recently are they compelling
- Security
  - “Not our problem”… right!! Oops, wrong! No, right!
- L2 address collision
  - That would have been an easy problem to fix. Duh!


sdr-monitor at UCSB - Netscape
File Edit View Go Communicator Help
UO Broadcasts NASA Videos (22) (127) (128.223.83.33) (1m 2s)
UO Broadcasts NASA Videos (23) (127) (128.223.83.33) (1m 2s)
UO Broadcasts OPE's Oregon Story (127) (128.223.83.33) (1m 2s)
UO DOD News Clips (127) (iptvhost.uoregon.edu) (1m 2s)
UO Medical Management of Biological Casualties (1) (127) (128.223.83.33) (1m 2s)
UO Natural Resources (127) (iptvhost.uoregon.edu) (1m 2s)
UO WWII News Clips (127) (128.223.83.33) (1m 2s)
VINT (private) (127) (joy.isi.edu) (1m 2s)
ABONE Discussion (127) (yak-e.isi.edu) (1m 2s)
UO Presents KKNU New Country (127) (128.223.214.23) (1m 2s)
UO Presents KMGE Magic 94 FM (127) (128.223.214.23) (1m 2s)
UO Presents KWAX Classical Radio (127) (128.223.214.23) (1m 2s)
BGMP/MASC discussion (127) (128.9.160.100) (1m 2s)
My Music (255) (192.168.1.12) (5m 55s)
USC-CS dgroup (private) (127) (128.9.160.100) (5m 55s)
View from Miami (127) (131.94.133.72) (1m 2s)
Places All Over the World at 128K (127) (tower-u1.cisco.com) (5m 55s)
VIPER Lab Continuous Video Streaming Test (127) (viper2.ecn.purdue.edu) (1m 2s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
zz linux (127) (nsap.cns.vt.edu) (1m 2s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
SURFnet-TV1 (127) (192.87.110.37) (1m 2s)
TNC 2000 - Room A (127) (wan1.rccn.net) (1m 2s)
TNC 2000 - Room B (127) (wan1.rccn.net) (1m 2s)
NASA PUBLIC HEALTH POLICY #10 (127) (131.182.170.181) (8m 25s)
NASA PUBLIC HEALTH POLICY #11 (127) (131.182.170.181) (8m 25s)
NASA TV (127) (131.182.10.250) (8m 25s)
eircom - Temple Bar (127) (dipsy.av.eircom.net) (8m 44s)
BBC News 24 (127) (simon1.kw.bbc.co.uk) (5m 55s)
CPAC on CA*net 3 (IP TV, MPEG1) (127) (205.189.33.76) (8m 25s)
RealServer address announcement (127) (128.178.10.2) (1m 2s)
Places all over TEN-155 (2) (127) (kssun1.rus.uni-stuttgart.de) (8m 44s)
RealServer address announcement (127) (128.178.10.2) (5m 55s)

Original Problems (cont)

- Routing
  - Lots of attempts, but all had fatal flaws: broadcast-and-prune and then network source discovery
  - Academia didn't help: many unrealistic assumptions
- Bridging between "islands"
  - Had performance impacts with mrouted and especially as the network grew…
  - Machines were slower so needing to send data all the way to application-space was problematic
- Deployment
  - Original deployment was driven by the "cool factor", but beyond that we had no plan and no real incentives

More Recent Problems

- Inter-domain and source discovery
  - Wow, we took a major wrong turn here!!
- Firewalls
  - Filtered all mcast traffic for a while, or rather, all UDP traffic and that means all multicast
  - Talked to vendors: “what is `m-u-l-t-i-c-a-s-t’?”
- Congestion control and reliability
  - I think Digital Fountain finally got this right… but the market never fell in love

More Recent Problems (cont)

- Deployment
  - Paid little attention to the issues ISPs care about
  - Paid little attention to application development
- Authentication/Authorization/Accounting (AAA)
  - Important to the ISPs
  - Important to service provides
  - …and the application developers need to be aware
- Monitoring/Troubleshooting/Management
  - The tools simply do not exist, or at least “shrink-wrapped tools with 800-number support lines”

The Biggie: Economics

- Users
  - don't care how they get content, they just want it
- ISPs
  - Multicast was a "service" they never got paid for
  - UUNet tried (UUcast) but the billing model was illogical: pay more when more users listening
- Content providers
  - L-O-V-E multicast because they pay less…
- Application developers
  - Good AAA requires implementing some non-scalable features, for example, tracking membership
  - The lesson of Starbust

The Biggie: Economics (cont)

- There are *some* benefits
  - But really, they are all second-order
- Access to more content… for less
  - Nobody chooses an ISP based on access to content (see recent AOL decision)
- ISPs could charge differently for multicast (but less than N*unicast)
  - Still hard to manage (see telephone company billing)
  - ISPs still lose money if they charge based on access
    - unless they are in an odd “sweet spot” on the curve

Why These Problems Happened

- The academic community was disconnected from reality
- Router vendors were clueless about long-term strategy
  - The goal became “product differentiation” (see PGM)
- The IETF was dominated by router vendors
  - Not on purpose, but ISPs couldn’t afford to care
- Not to keep bashing on the IETF, but…
  - The community chose to be very insular…

Current Problems

- State scalability and CPU processing
  - With large numbers of groups/members/sources, router resource consumption becomes an issue
  - And still that pesky problem of per-flow state
- Congestion control
  - Or because multicast is UDP and all UDP is blocked by firewalls
  - There are solutions, just depends whether apps will use them
- Security
  - Not data security but core protocol operation DoS security
- Monitoring/Troubleshooting/Management/AAA
  - Still important

Current Problems (cont)

- Architecture baseline?
  - Is it ASM? Or SSM? Or SGM? Or ???
  - What's my API?
- Deployment
  - The *one* fatal flaw is that for multicast to work, it has to be deployed everywhere
- Mobility
  - …and the problem wasn't hard enough already?!?

QED: Multicast failed

"Multicast could be the poster child for the irrelevance of the networking research community. Few other technologies (quality of service springs to mind) have generated so many research papers while yielding so little real-world deployment."

Bruce Davies, public review of ACM Sigcomm 2006 accepted paper, "Revisiting IP Multicast" by S. Ratnasamy, A. Ermolinskiy, S. Shenker
http://www.sigcomm.org/sigcomm2006/discussion/

Multicast is a success…

- …according to just about every metric except one
- Significant deployments:
  - Exchanges and securities trading companies
  - Enterprises and college campuses
    - Major companies use wide variety of apps
    - Campuses distribute CableTV using multicast (Northwestern)
  - Edge networks
    - Often called walled gardens
    - Examples: DSL and Cable TV (triple play)
  - Military networks
    - One statistic: “60% of our traffic is going to be multicast”
    - Need multicast support in ad hoc networks

The One Metric…

The ubiquitous deployment of a revenue generating native one-to-many and many-to-many infrastructure capable of securely and robustly supporting both reliable, TCP-friendly file transfer, all manner of streaming media (including seamless rate adaptation), and any style of audio/video conferencing (with minimal jitter and end-to-end delay)—all with only minimal additional router complexity, deployment effort, management needs, or cost.

In Fact…

- Multicast, as an academic-style research area, has been one of the more successful recent research areas
  - Original idea was generated in academia
  - Academic-based research has led to standardized and deployed protocols, industry/academia collaboration, companies, products, revenue, etc.
  - And these efforts continue…

But There Are Sad Truths

- The academic community became in-bred and allowed all manner of papers to be published.
  - We lost our discipline as a community
  - Spoiled multicast for a long time (maybe ever)

- Other areas in danger of the same result:
  - *QoS*: may save itself by broadly defining "QoS"
  - *Ad hoc networks*: saved itself based on military apps; evolving to "mesh networks"; but still spoiling as a research area

But There Are Sad Truths

- The community has become quite ossified
  - The IETF is not interested in adopting critical changes
    - For example: appropriate feedback for failed joins
    - In some cases, no good solutions exist
  - OS makers are slow to implement standards
    - For example: IGMPv3 and MLDv2 for SSM support
  - Application developers are hit multiple times
    - Which multicast model is being used and where?
    - Limited audience for most apps
    - Unclear what knowledge is needed and how to get it

Current Course Adjustments

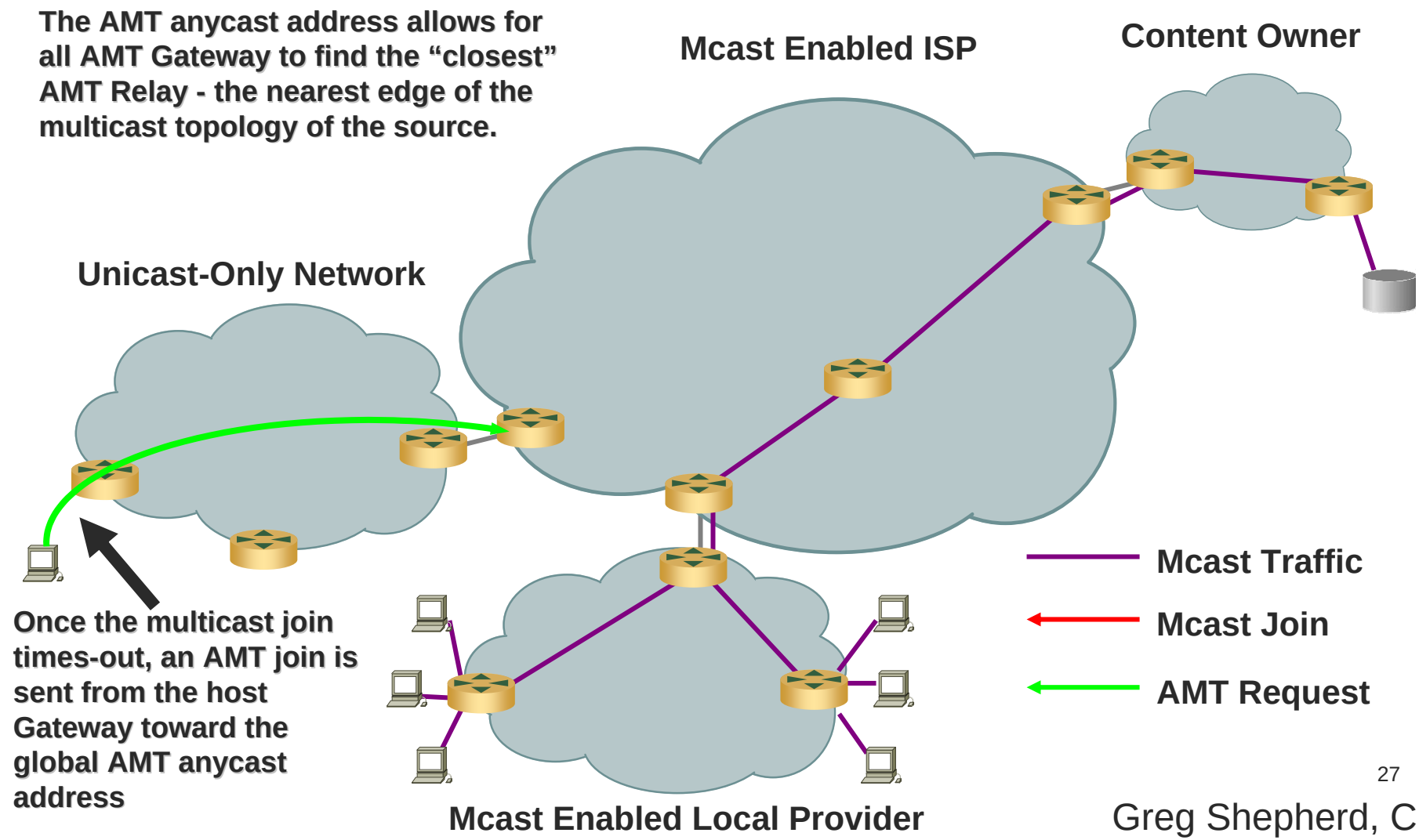
- IRTF SAM RG has a good mission
  - Need to invite MBONED community
- Continue work towards a hybrid solution
  - Solutions must be incrementally deployable
  - For example: AMT
- Continue focused work for specific applications
- Convince academic community to re-accept multicast
  - They still are in many cases (even Sigcomm did), but what they consider interesting are monolithic solutions
  - Need a place that accepts good, deployable solutions
  - Interest by the funding agencies would also help

Automatic Multicast Tunneling

- Automatic IP Multicast without explicit Tunnels
  - www.ietf.org/internet-drafts/draft-ietf-mboned-auto-multicast-*.txt
- Allows multicast content to reach unicast-only receivers
- Provide the benefits of multicast wherever multicast is deployed.
  - Hybrid solution
  - Multicast networks get the benefit of multicast
- Works seamlessly with existing applications
  - Requires only client-side shim (somewhere on client) and router support in some places

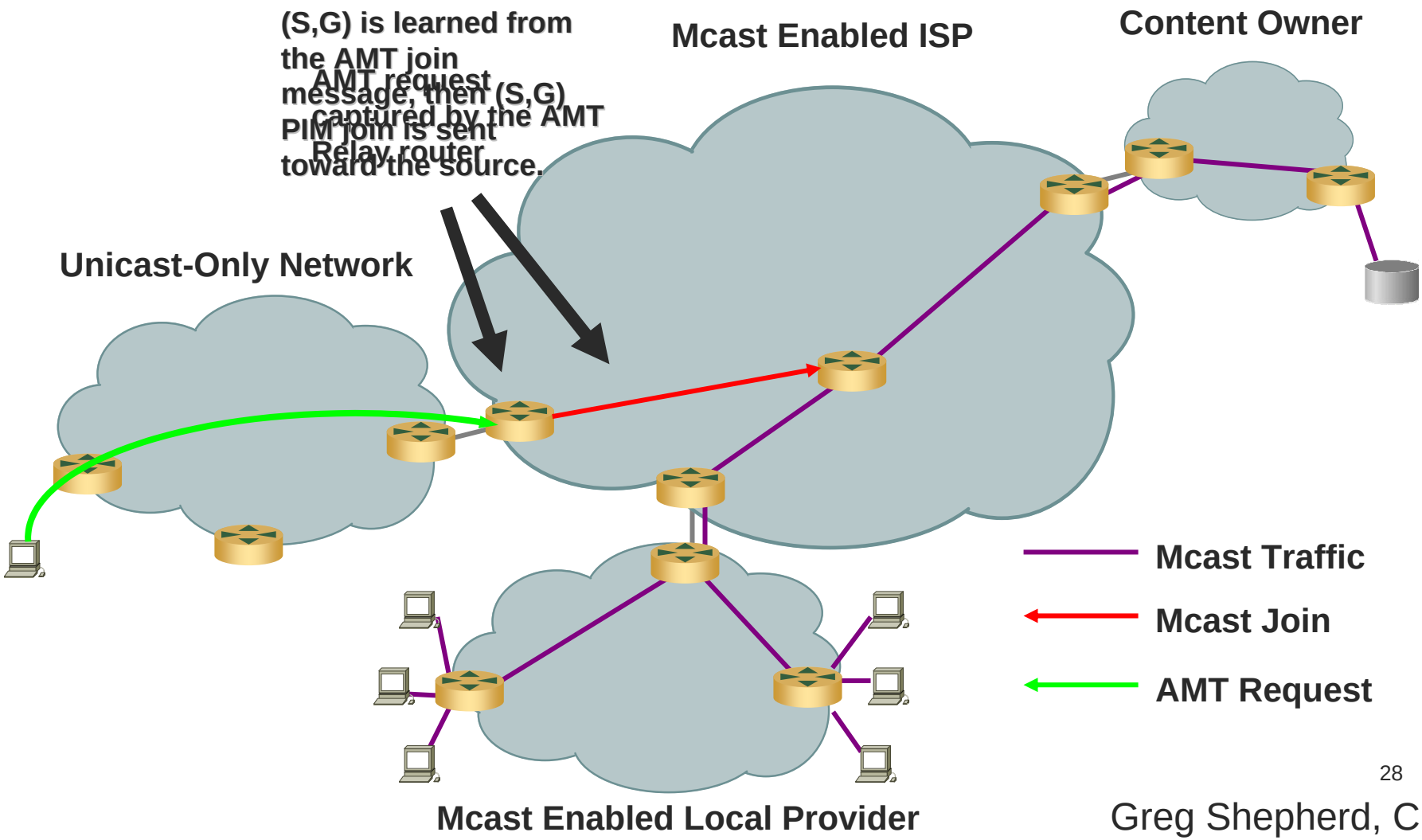
AMT

The AMT anycast address allows for all AMT Gateway to find the “closest” AMT Relay - the nearest edge of the multicast topology of the source.

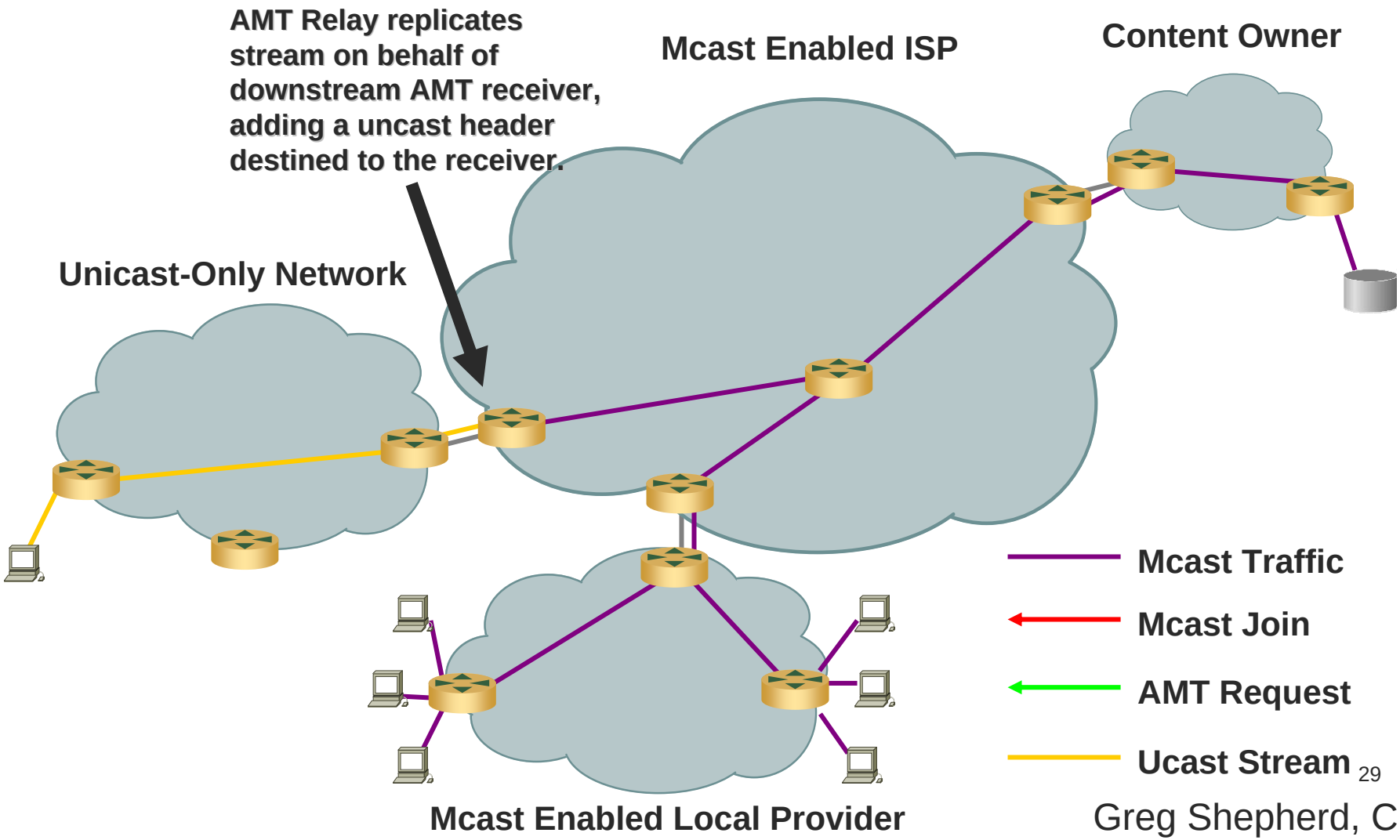


Once the multicast join times-out, an AMT join is sent from the host Gateway toward the global AMT anycast address

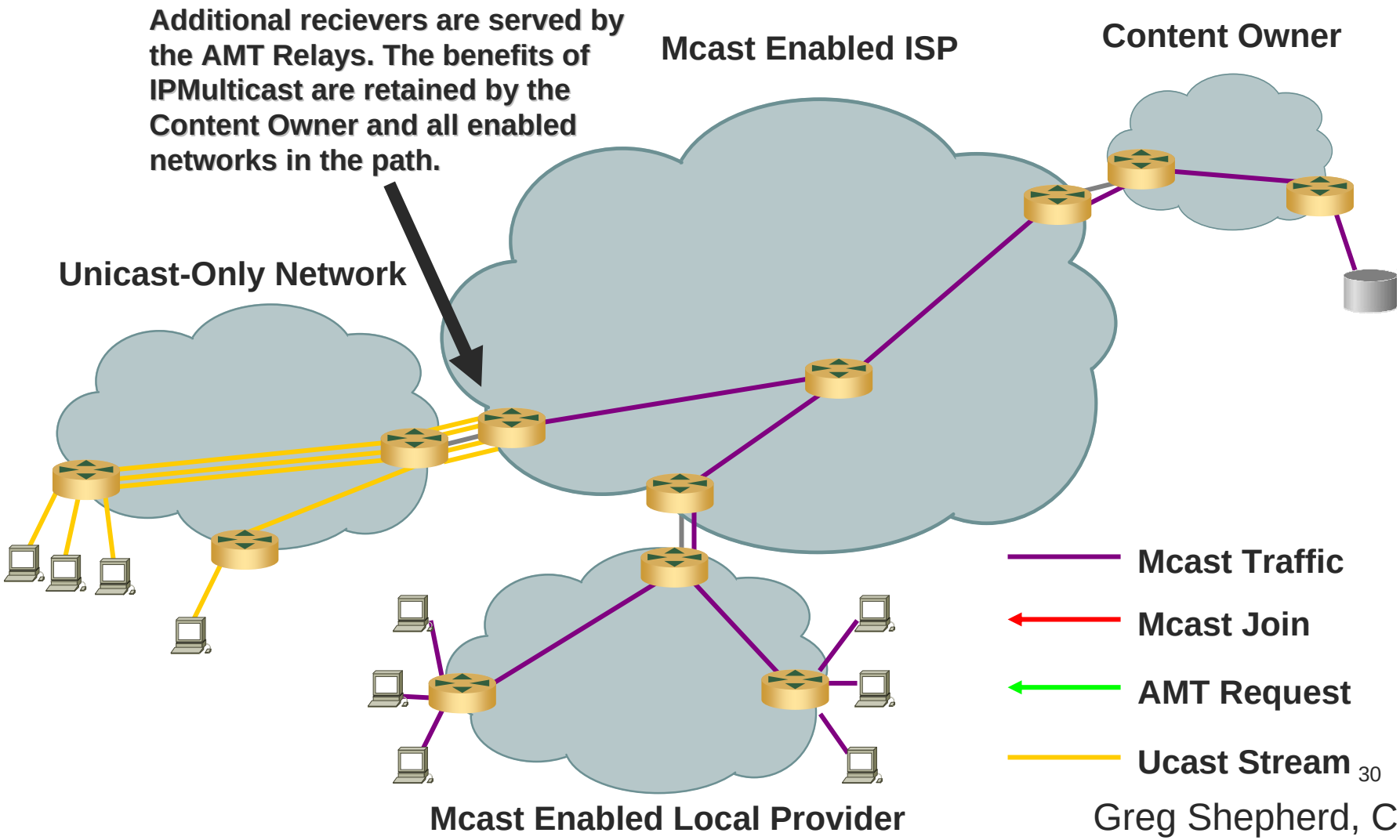
AMT


(S,G) is learned from the AMT join message, then (S,G) PIM join is sent toward the source.
AMT request captured by the AMT Relay router
Mcast Enabled ISP
Content Owner
Unicast-Only Network
Mcast Traffic
Mcast Join
AMT Request
Mcast Enabled Local Provider

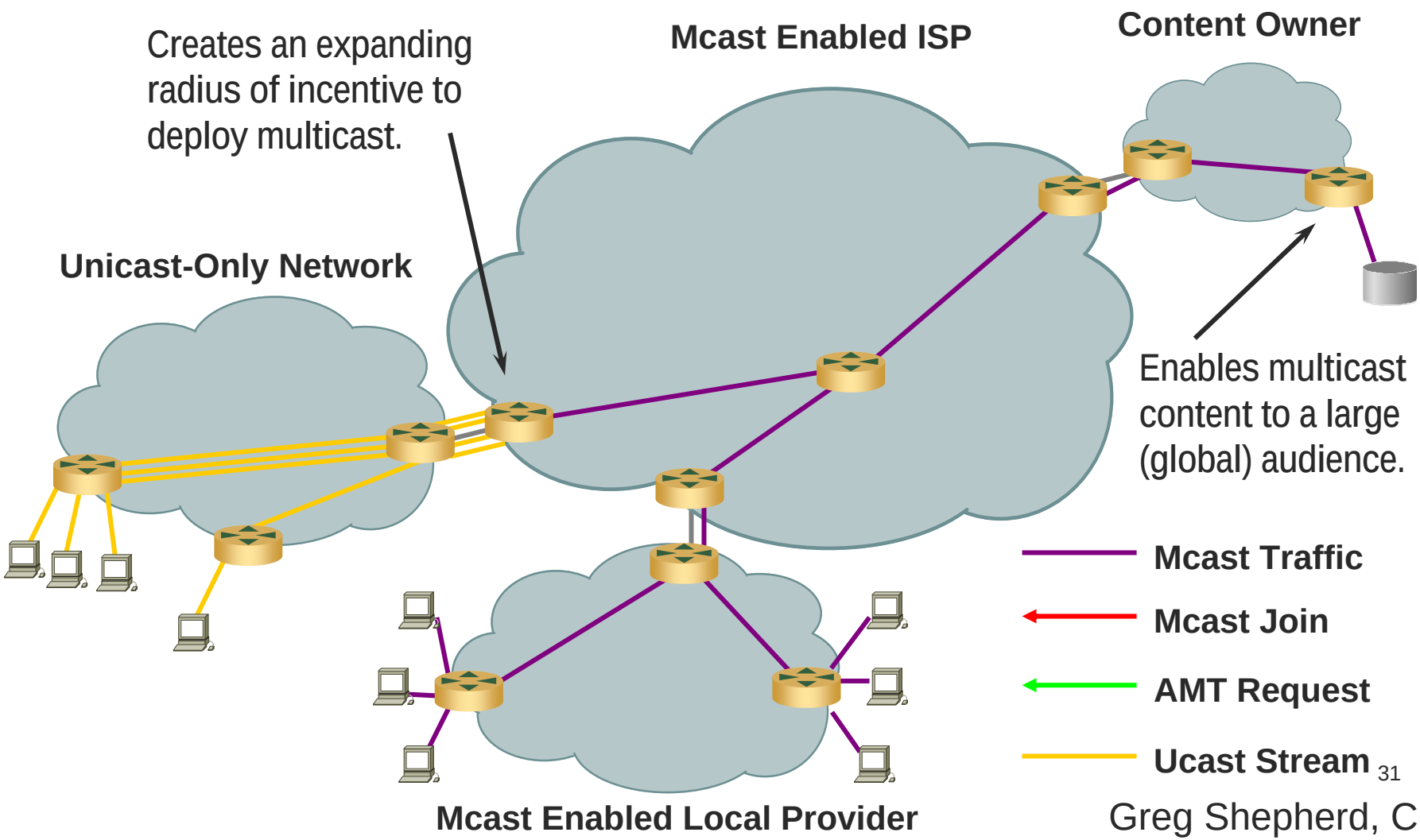
AMT


AMT Relay replicates stream on behalf of downstream AMT receiver, adding a uncast header destined to the receiver.
Mcast Enabled ISP
Content Owner
Unicast-Only Network
Mcast Traffic
Mcast Join
AMT Request
Ucast Stream
Mcast Enabled Local Provider

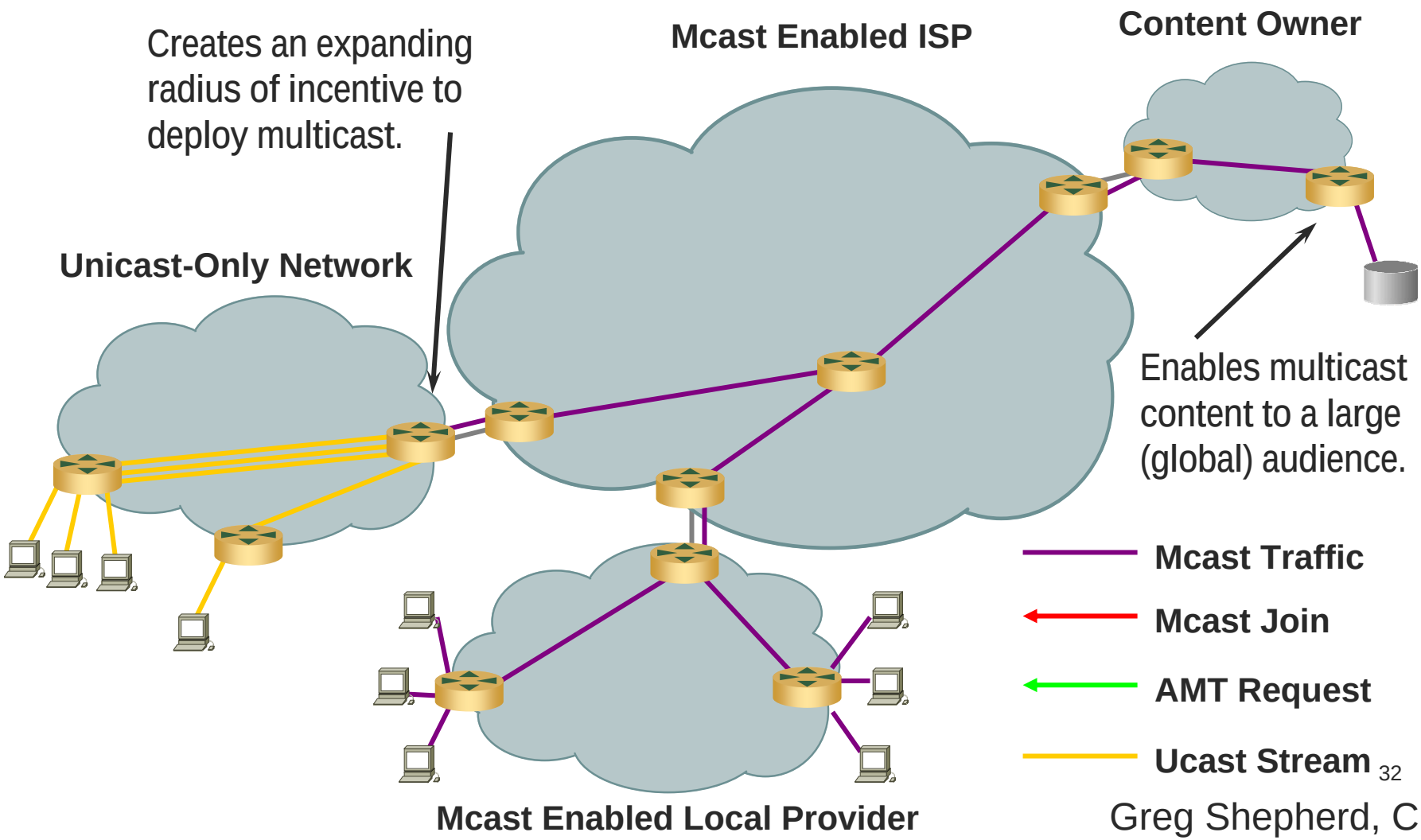
AMT


Additional recievers are served by the AMT Relays. The benefits of IPMulticast are retained by the Content Owner and all enabled networks in the path.
Mcast Enabled ISP
Content Owner
Unicast-Only Network
Mcast Traffic
Mcast Join
AMT Request
Ucast Stream
Mcast Enabled Local Provider

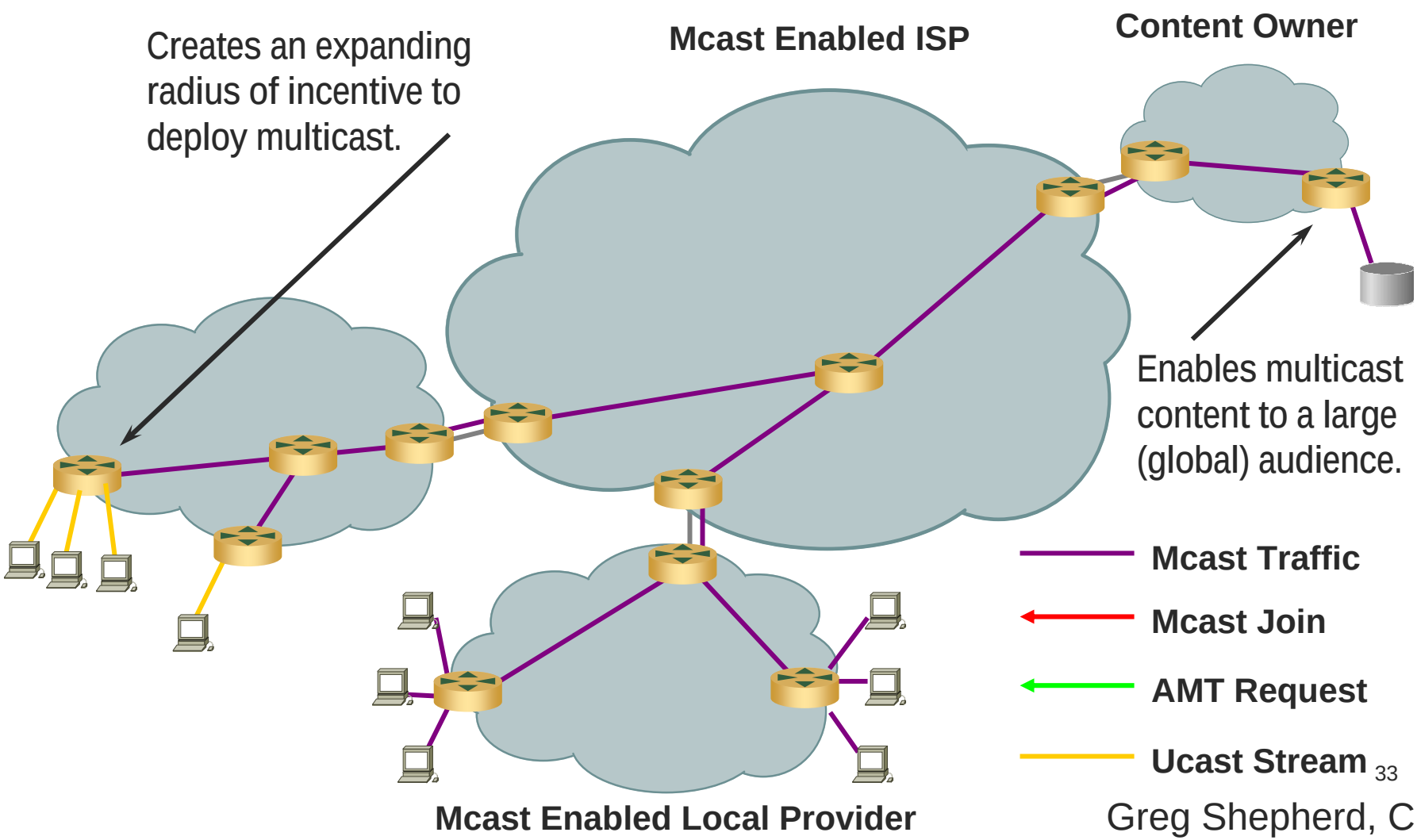
AMT


Creates an expanding radius of incentive to deploy multicast.
Mcast Enabled ISP
Content Owner
Unicast-Only Network
Enables multicast content to a large (global) audience.
Mcast Traffic
Mcast Join
AMT Request
Ucast Stream
Mcast Enabled Local Provider

AMT


Creates an expanding radius of incentive to deploy multicast.
Mcast Enabled ISP
Content Owner
Unicast-Only Network
Enables multicast content to a large (global) audience.
Mcast Traffic
Mcast Join
AMT Request
Ucast Stream
Mcast Enabled Local Provider

AMT


Creates an expanding radius of incentive to deploy multicast.
Mcast Enabled ISP
Content Owner
Enables multicast content to a large (global) audience.
Mcast Traffic
Mcast Join
AMT Request
Ucast Stream
Mcast Enabled Local Provider

Avoid Need for Universal Consensus

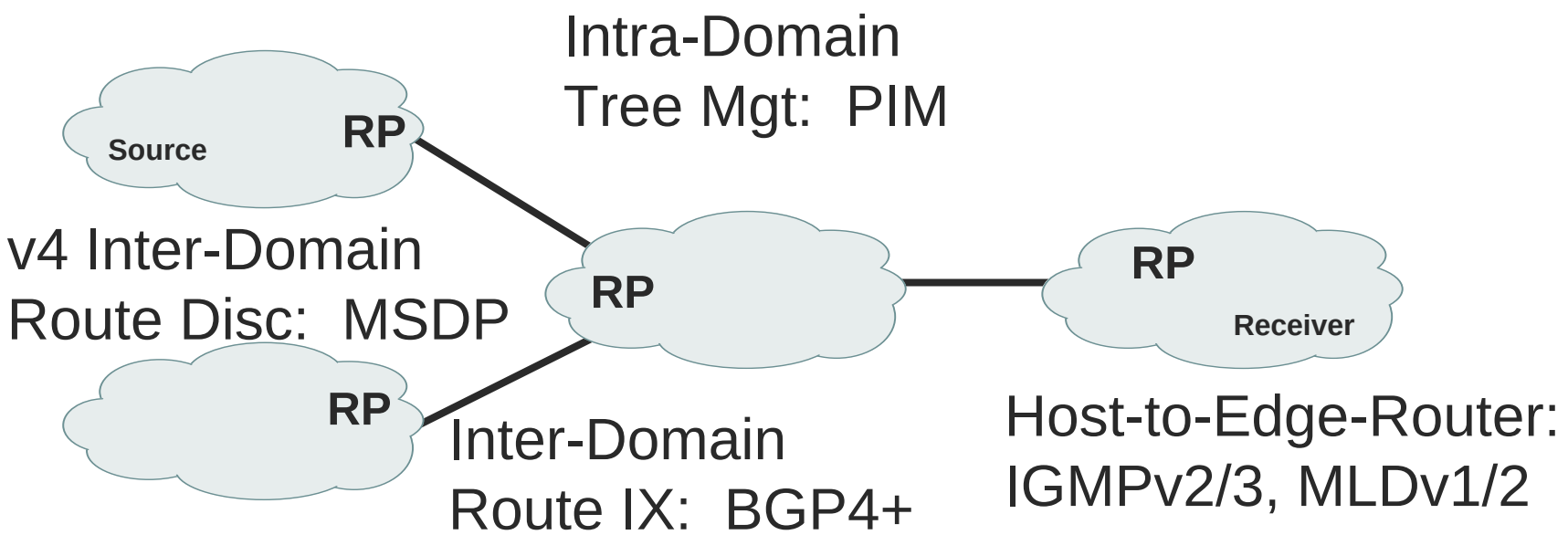
- There are multiple groups that need to participate:
  - Users
  - App developers
  - OS companies (socket interface)
  - Router vendors
  - Content providers
- The more a solution does not require the approval of multiple of these groups, the better
  - No solution is going to be universally approve and ubiquitously adopted

Conclusions

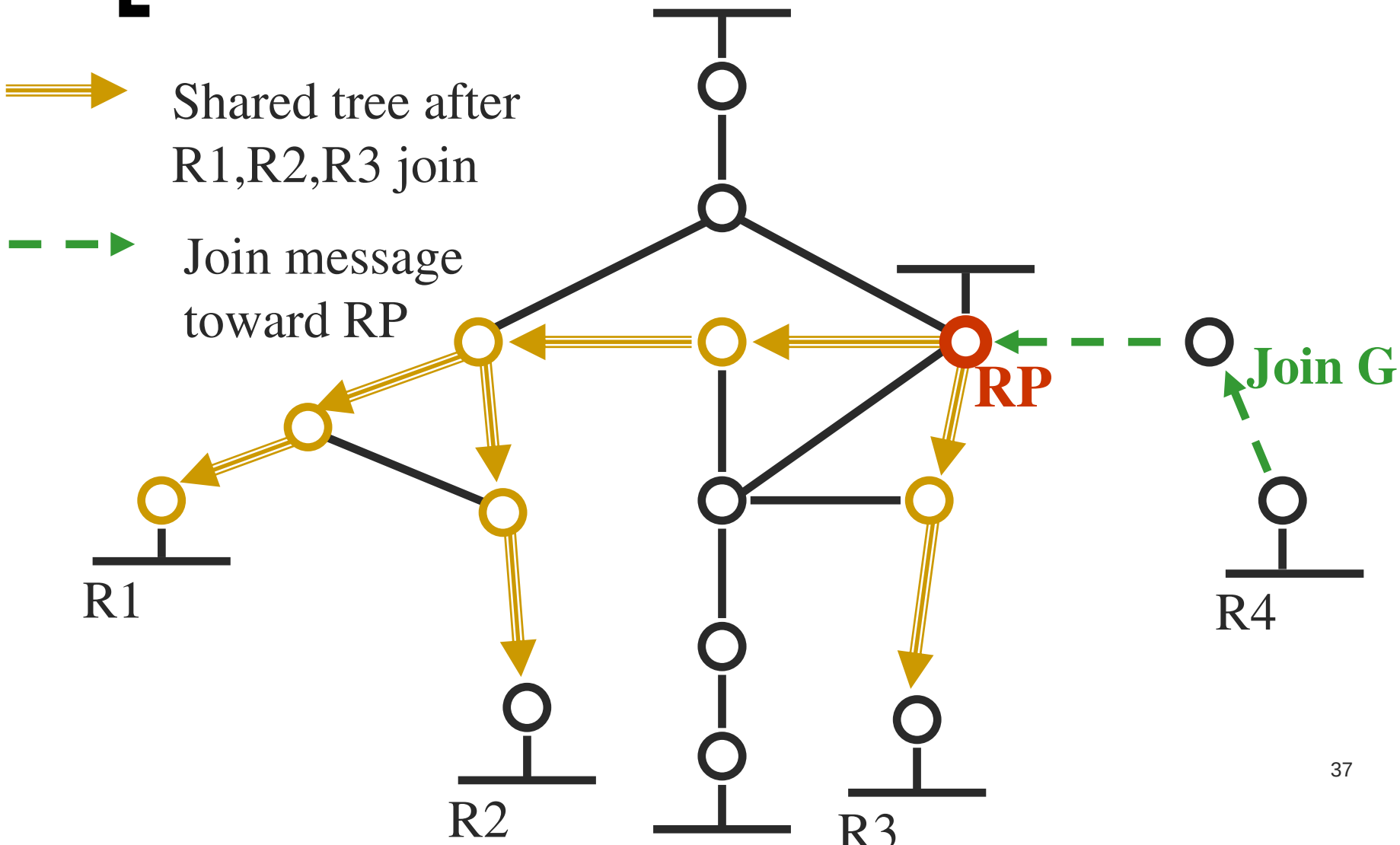
- Multicast has had a bumpy road…
- …but success is there if you look for it

- There are interesting challenges ahead…
- …but we need working solutions

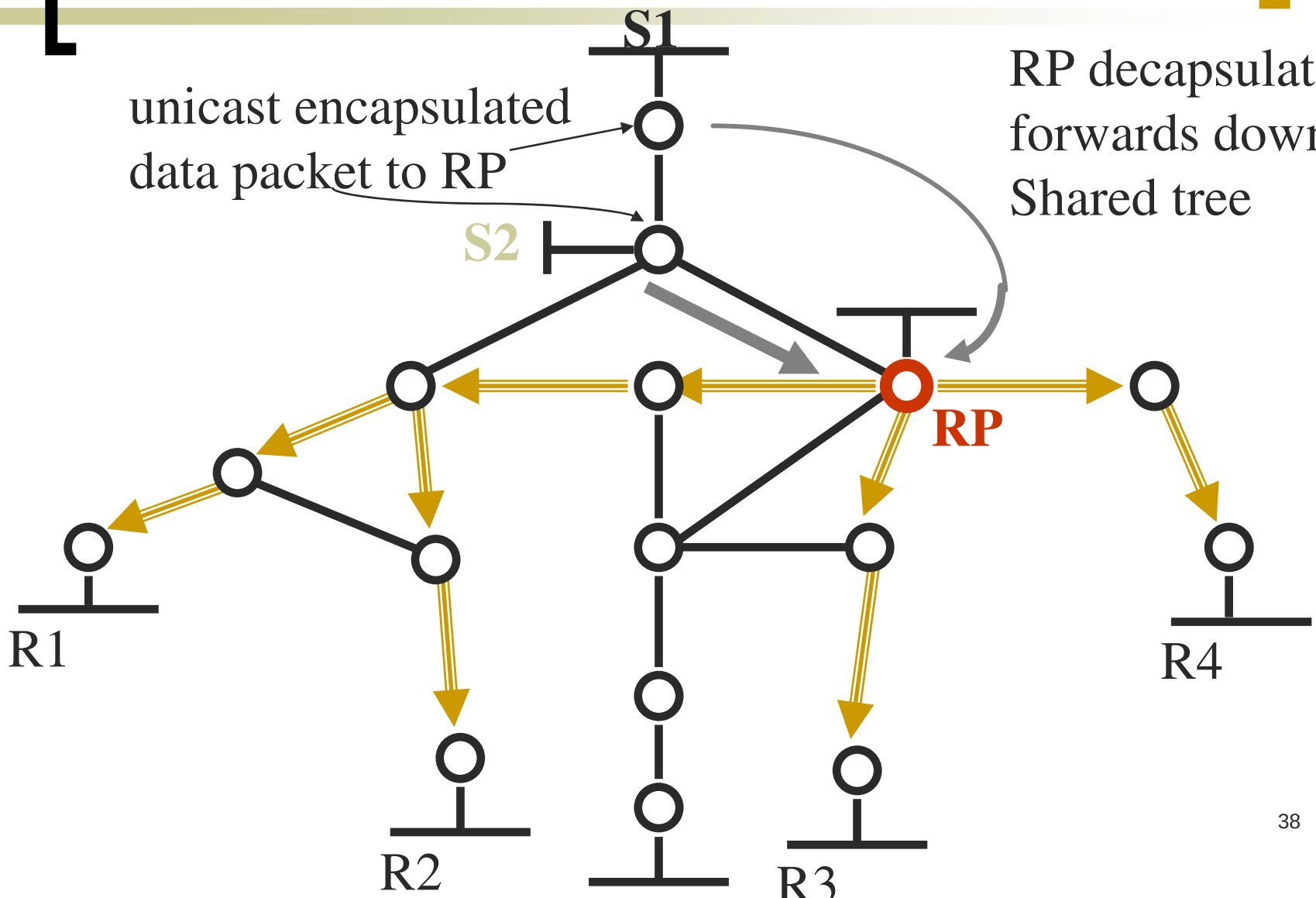
Multicast Protocols



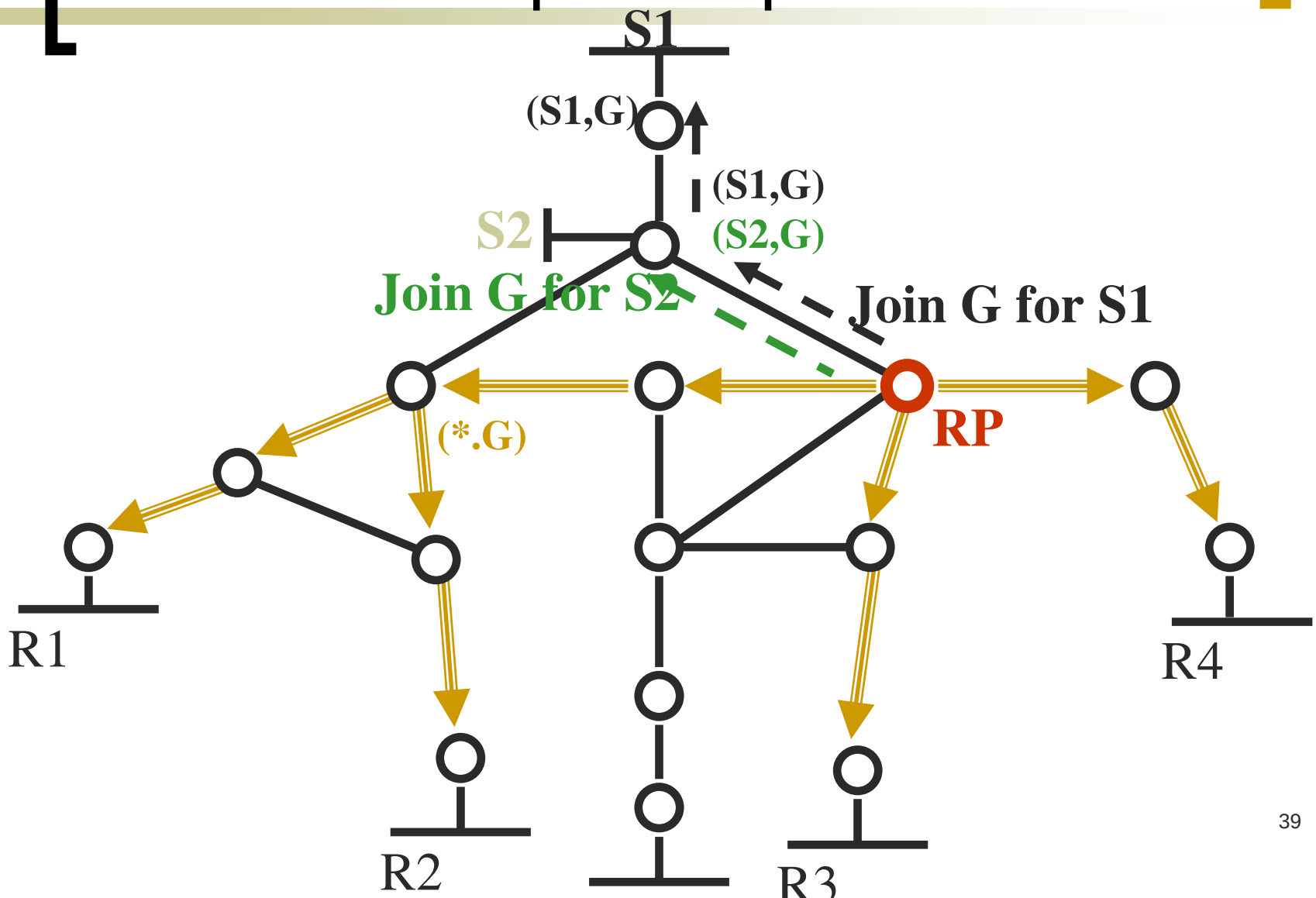
Phase 1: Build Shared Tree



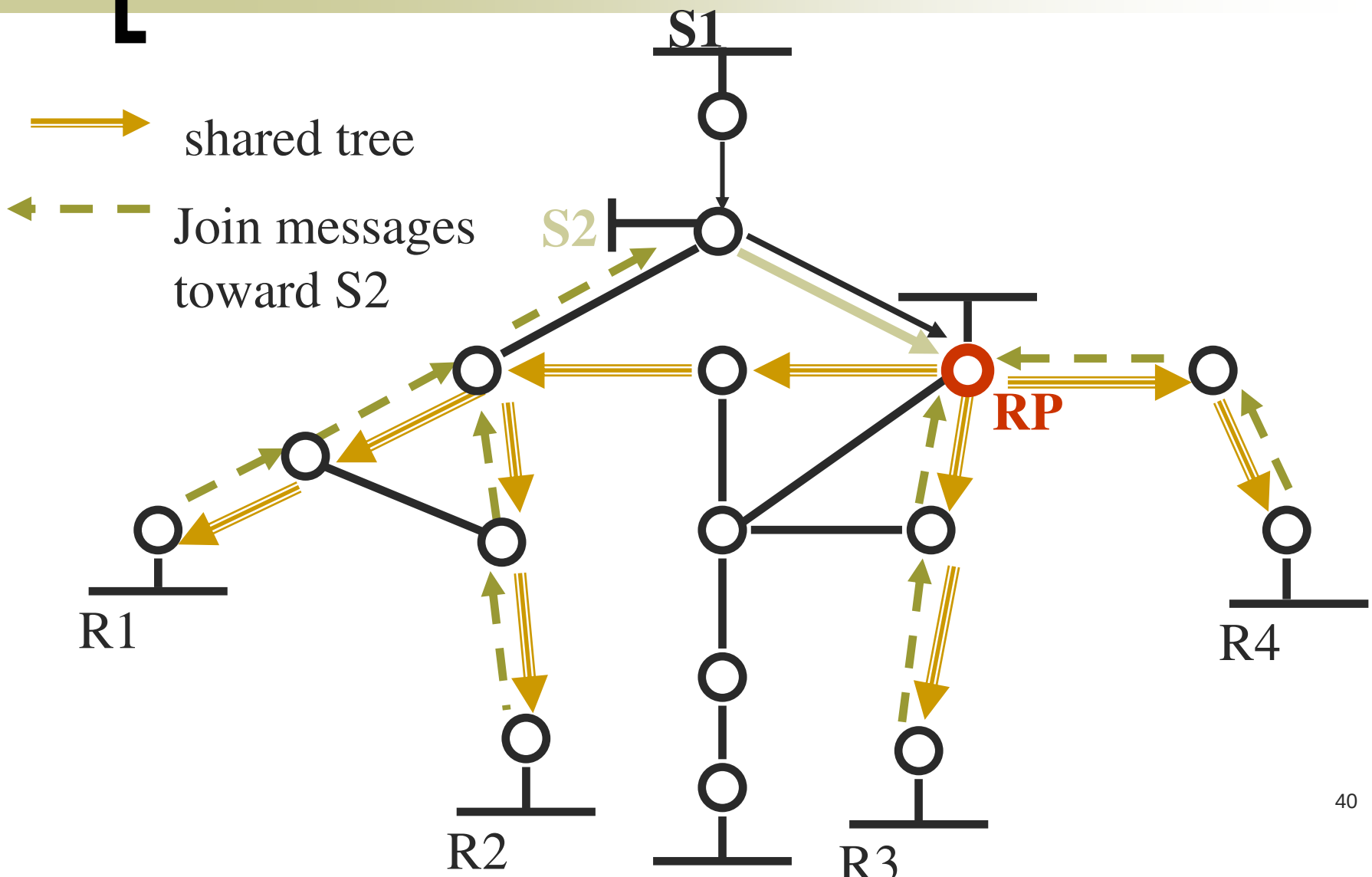
Phase 2: Sources Send to RP



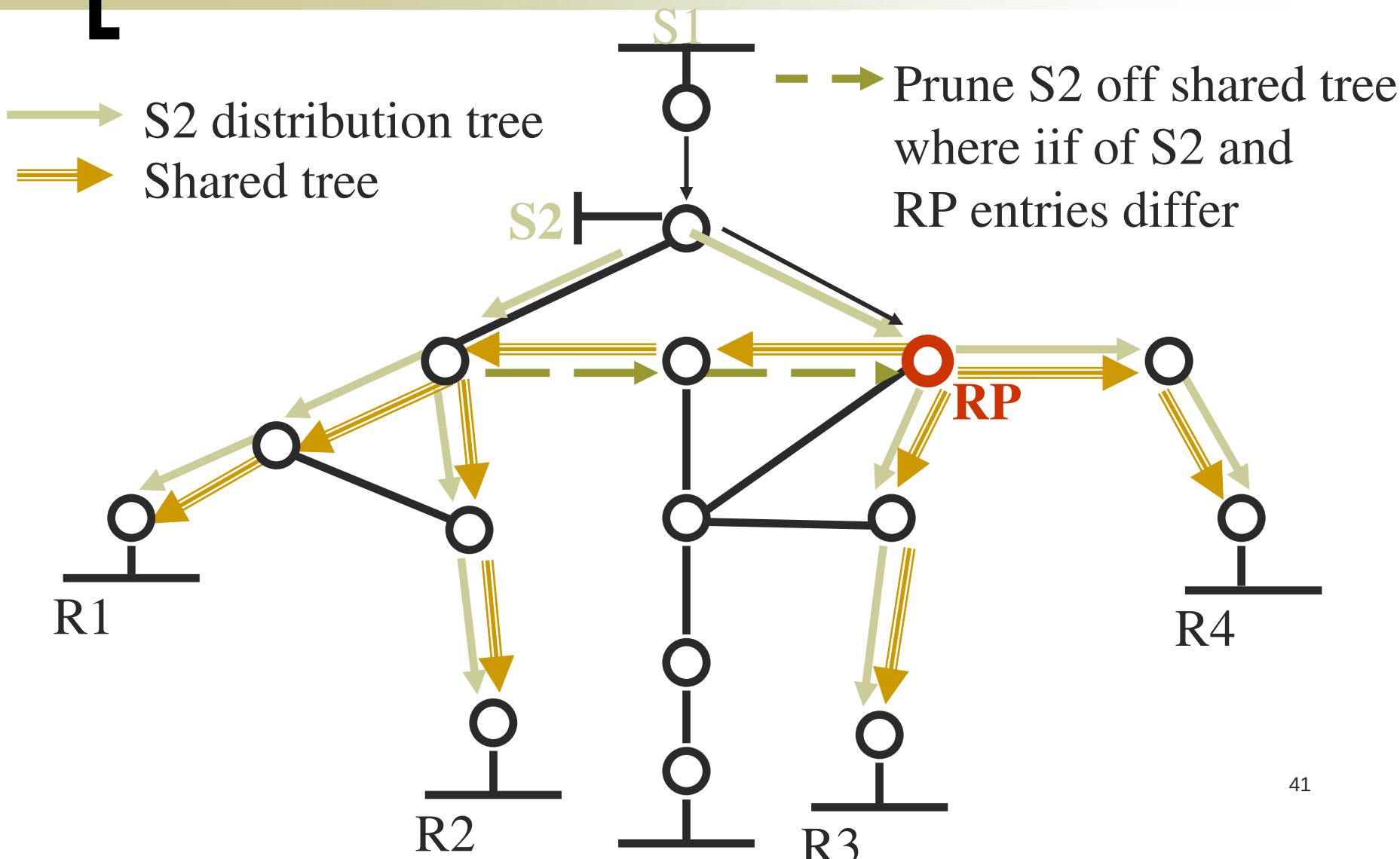
Phase 3: Stop Encapsulation



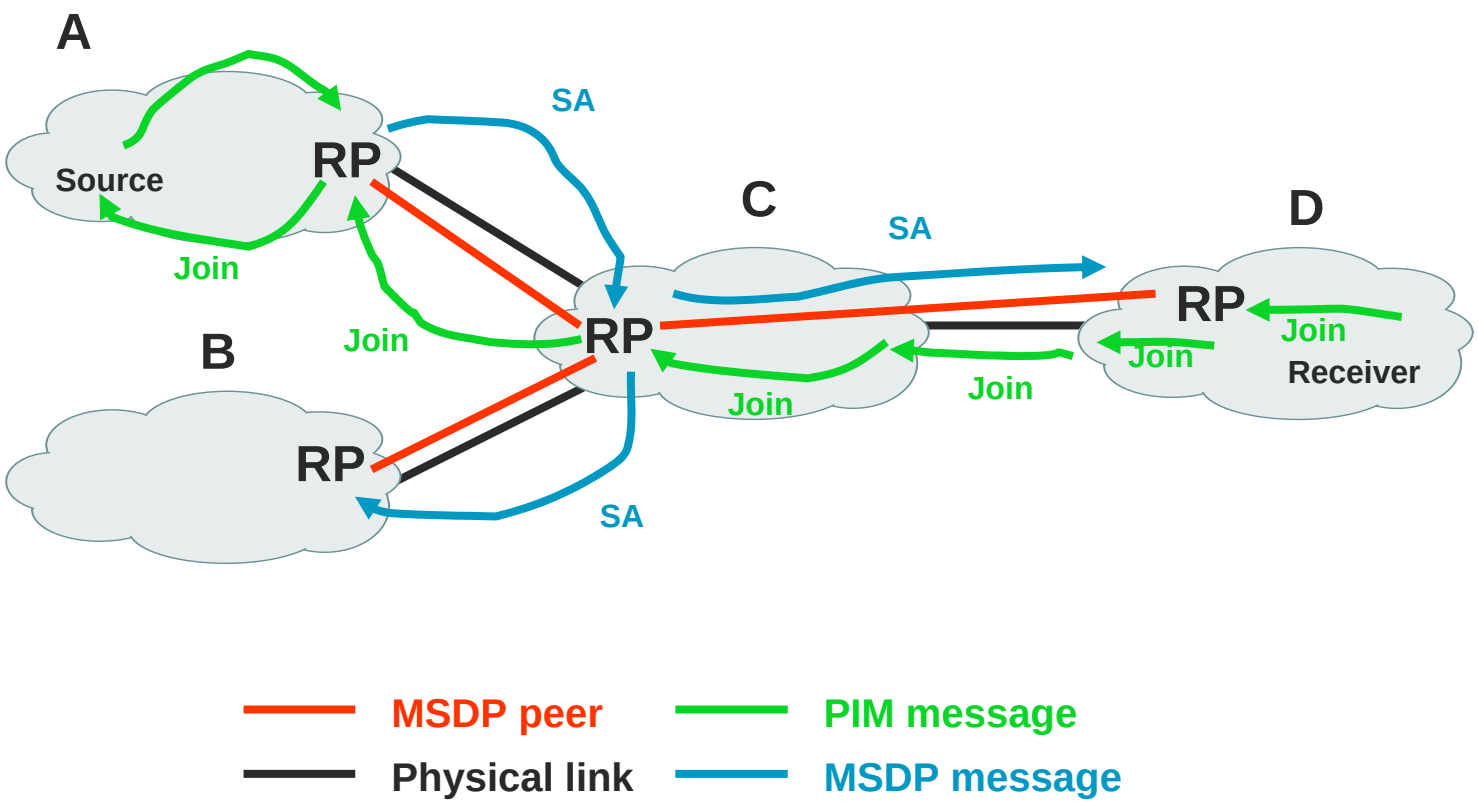
Phase 4: Switch to SPT



Phase 5: Prune S2 Shared Tree

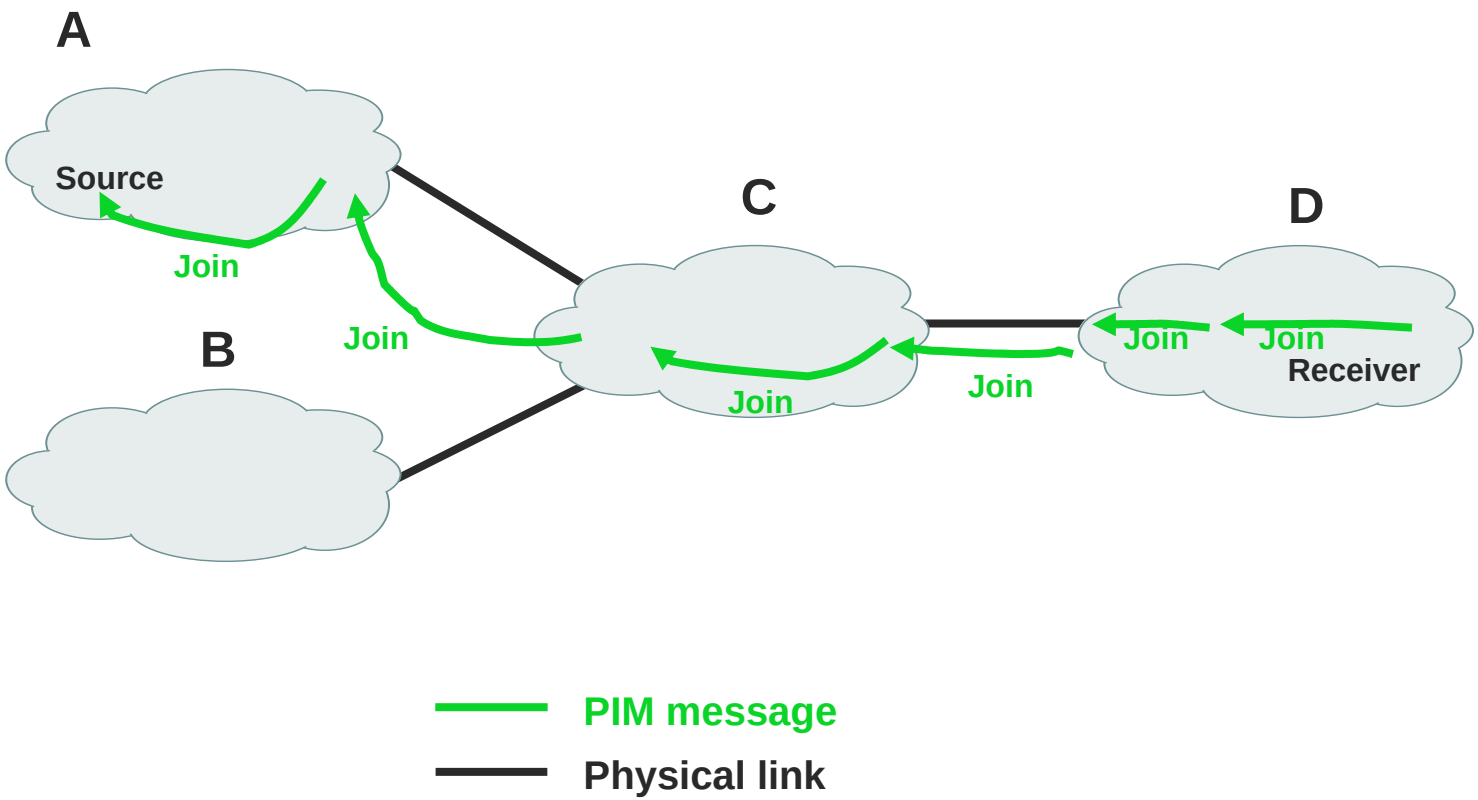


MSDP


A
Source
RP
SA
Join
B
Join
RP
C
SA
D
RP
Join
Join
Receiver
Join
Join
RP
SA
MSDP peer
Physical link
PIM message
MSDP message

SSM

Back


A
Source
Join
B
Join
C
Join
Join
D
Join
Join
Receiver
PIM message
Physical link