Introduction to Routing

How traffic flows on the Internet

Philip Smith pfs@cisco.com

RIPE NCC Regional Meeting,

Moscow, 16-18 June 2004

Abstract

- **Presentation introduces some of the terminologies used, and describes the constituent parts of ISP network infrastructure. The presentation looks at the routing design and operation of individual ISP networks, and how those interconnect to create what we know as the Internet today. It concludes by looking at some of the hot topics currently facing service providers.**

Agenda

- Topologies & Definitions
- Routing Protocols
- BGP
- Aggregation
- Current Hot Topics

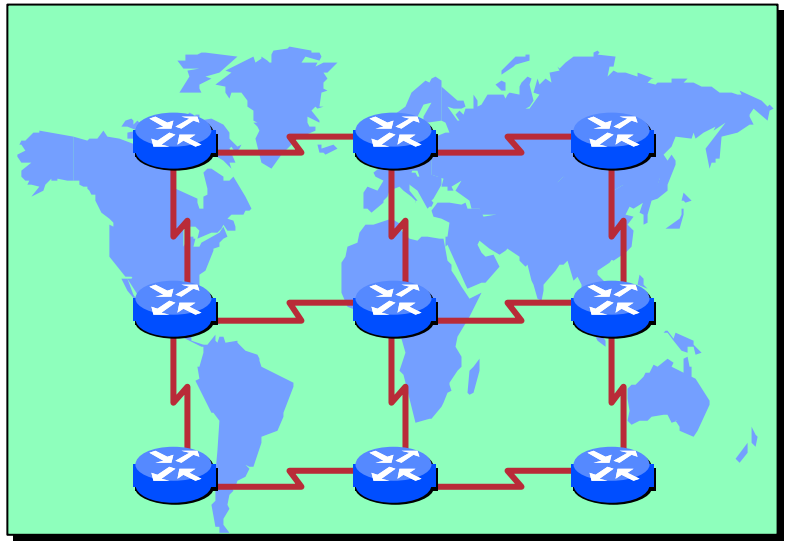
Topologies & Definitions

What does all the jargon mean?

Network Topologies

Routed backbone

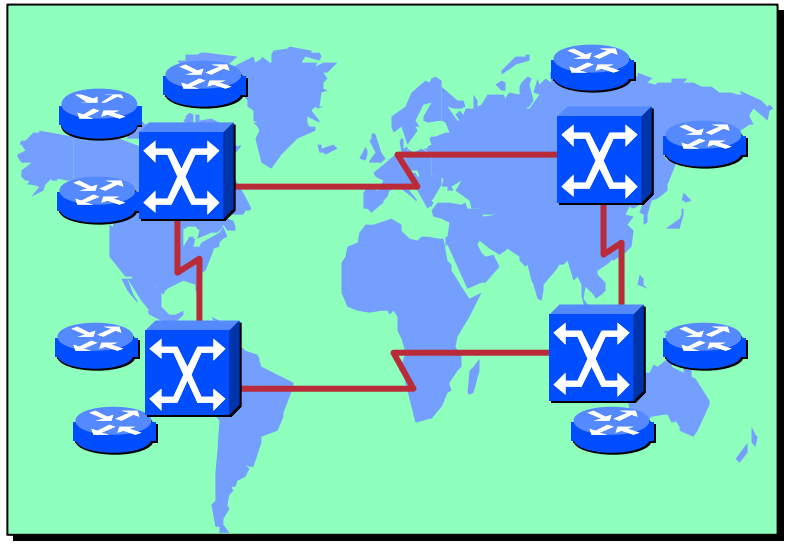
- Routers are the infrastructure
- Physical circuits run between routers
- Easy routing configuration, operation and troubleshooting



Network Topologies

Switched backbone

- frame relay or ATM
  - switches in the core
  - surrounded by routers
- Physical circuits run between switches
  - Virtual circuits run between routers
- more complex routing and debugging
- “traffic management”



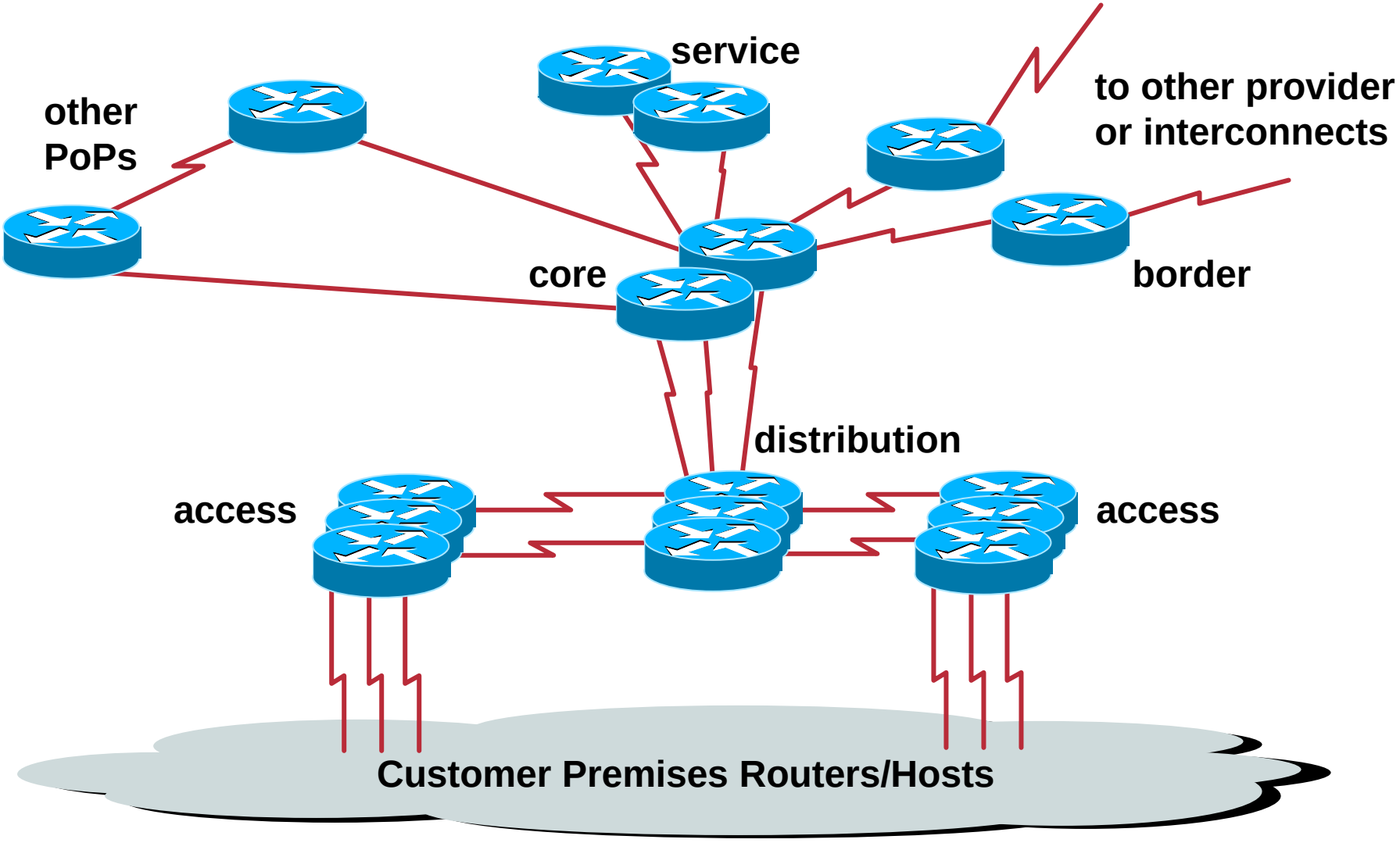
Definitions

- **PoP – Point of Presence**
  - **Physical location of ISP's equipment**
  - **Sometimes called a "node"**
- **vPoP – virtual PoP**
  - **To the end user, it looks like an ISP location**
  - **In reality a back hauled access point**
  - **Used mainly for consumer access networks**
- **Hub/SuperPoP – large central PoP**
  - **Links to many PoPs**

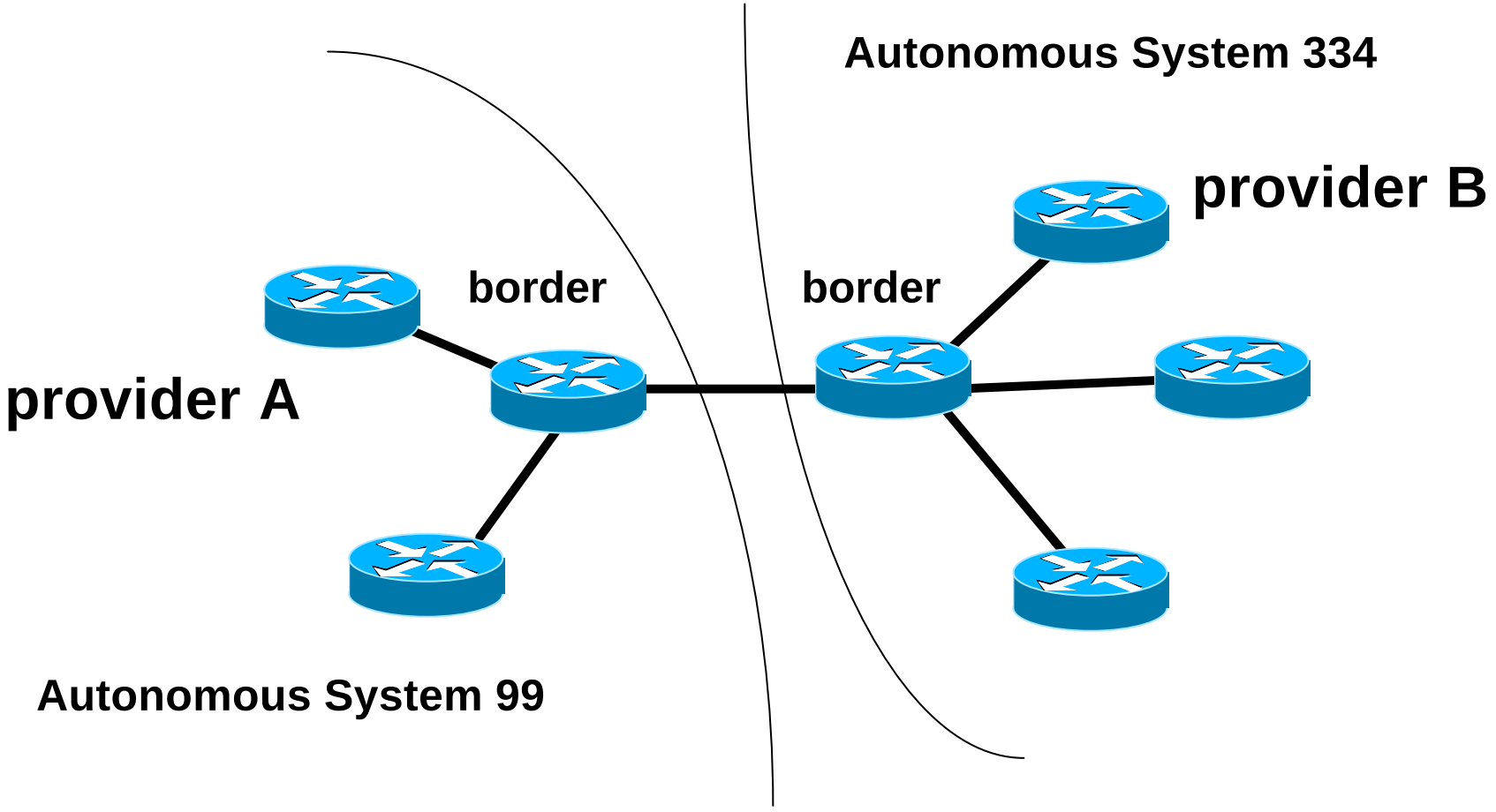
PoP Topologies

- **Core routers**
  - high speed trunk connections
- **Distribution routers**
  - higher port density, aggregating network edge to the network core
- **Access routers**
  - high port density, connecting the end users to the network
- **Border routers**
  - connections to other providers
- **Service routers**
  - hosting and servers
- **Some functions might be handled by a single router**

PoP Topologies



Private Interconnect



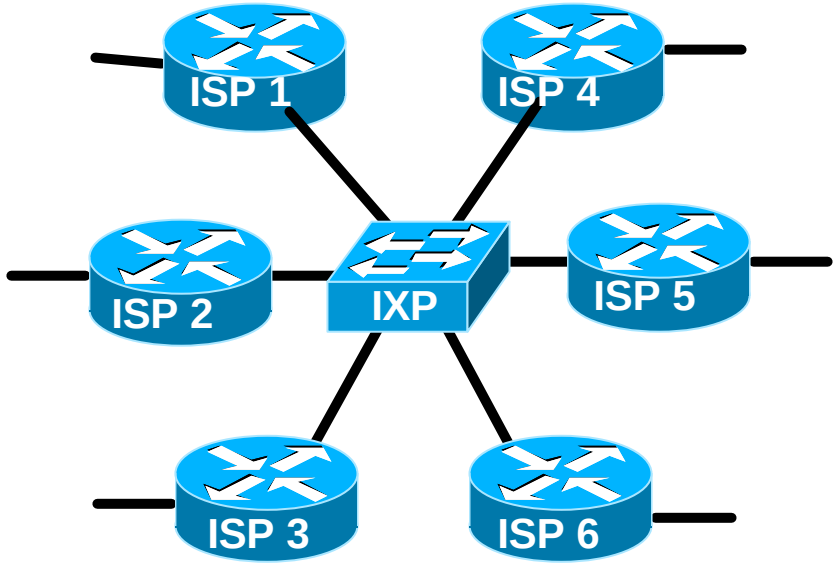
Two ISPs agree to have a private link to each other

Public Interconnect Point

- A location or facility where several ISPs are present and connect to each other over a common shared media
- Why?

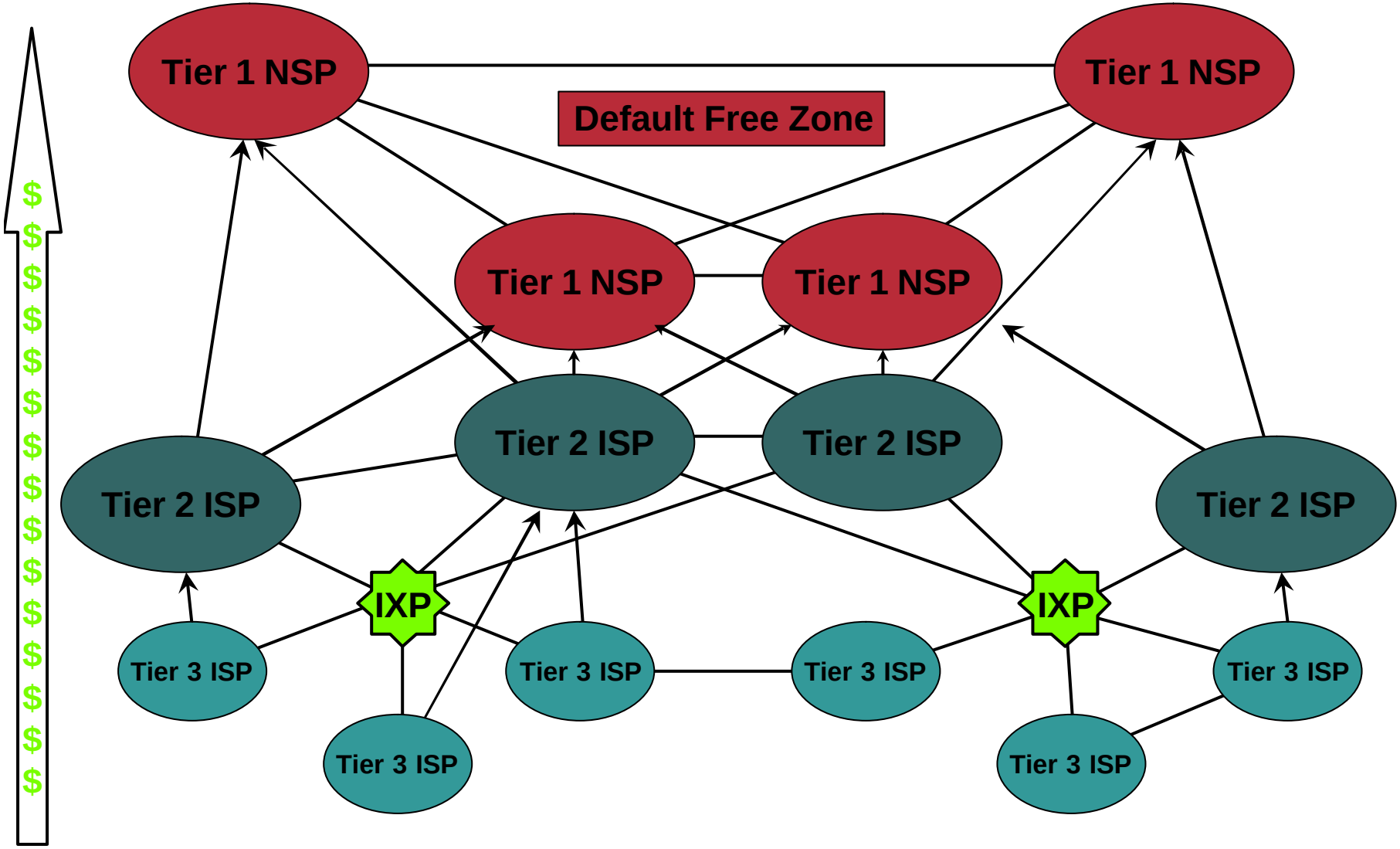
  To save money, reduce latency, improve performance,…

  Keeping local traffic local



- IXP – Internet eXchange Point
- NAP – Network Access Point

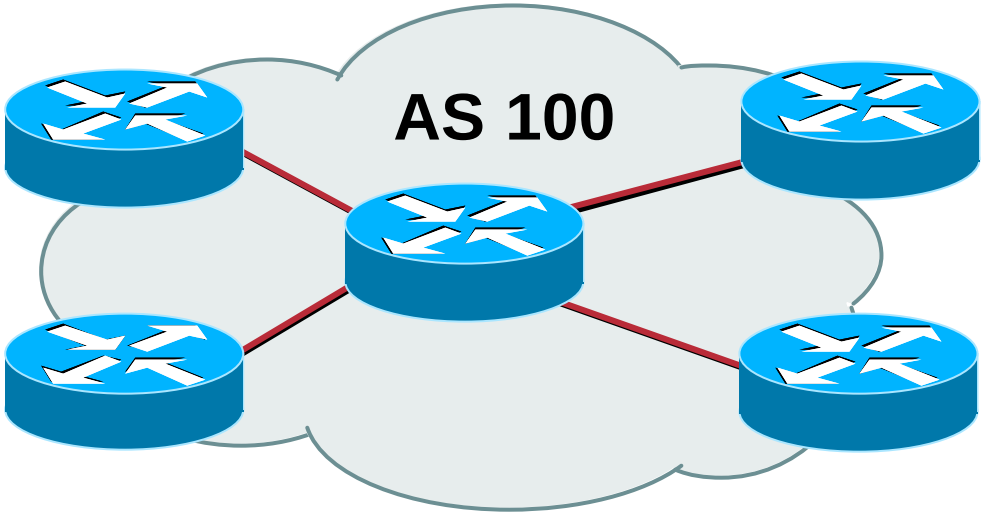
Categorising ISPs



Inter-provider relationships

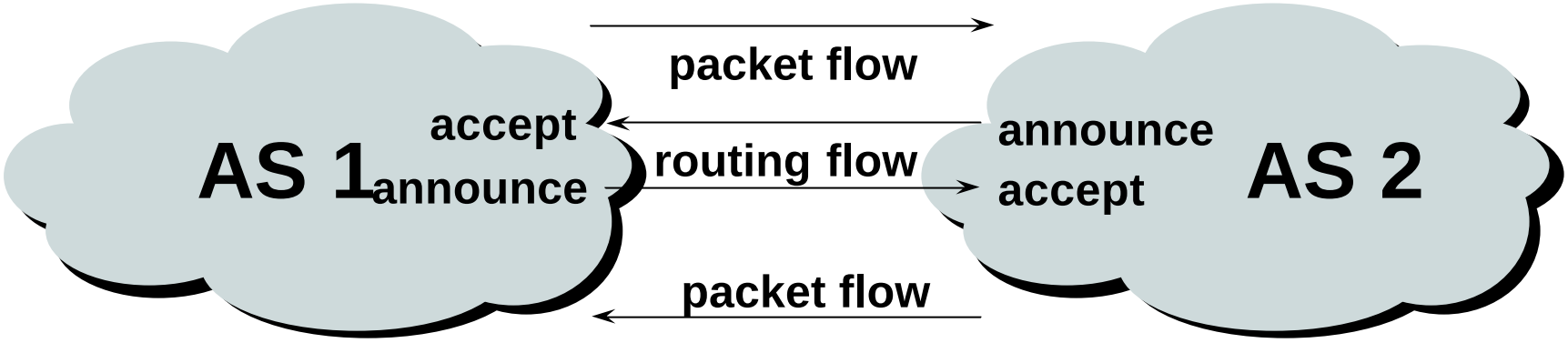
- **Peering between equivalent sizes of service providers (e.g. Tier 2 to Tier 2)**
  - **shared cost private interconnection, equal traffic flows**
  - **“no cost peering”**
- **Peering across exchange points**
  - **if convenient, of mutual benefit, technically feasible**
- **Fee based peering**
  - **unequal traffic flows, “market position”**

Autonomous System (AS)



- Collection of networks with same routing policy
- Single routing protocol
- Usually under single ownership, trust and administrative control
- Identified by 16-bit integer, of which ASNs 1-64511 are available for public use

Routing flow and Packet flow

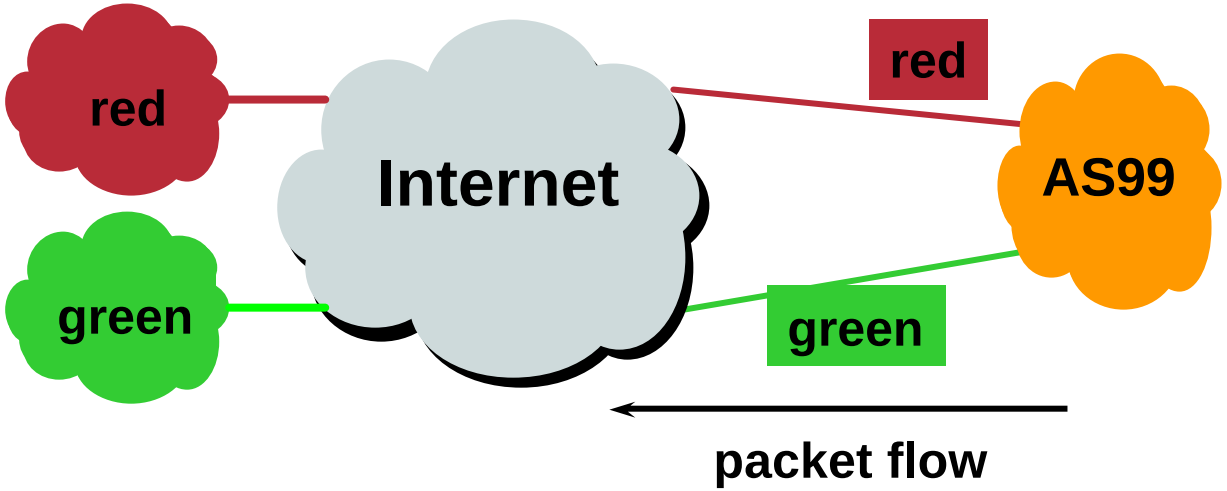


- For networks in AS1 and AS2 to communicate:
  - AS1 must announce to AS2
  - AS2 must accept from AS1
  - AS2 must announce to AS1
  - AS1 must accept from AS2
- Direction of Traffic flow is always opposite to the direction of the flow of Routing information

Routing Policy

- Used to control traffic flow in and out of an ISP network
- ISP makes decisions on what routing information to accept and discard from its neighbours
  - Individual routes
  - Routes originated by specific ASes
  - Routes traversing specific ASes
  - Routes belonging to other groupings
    - Groupings which you define as you see fit

Routing Policy Limitations

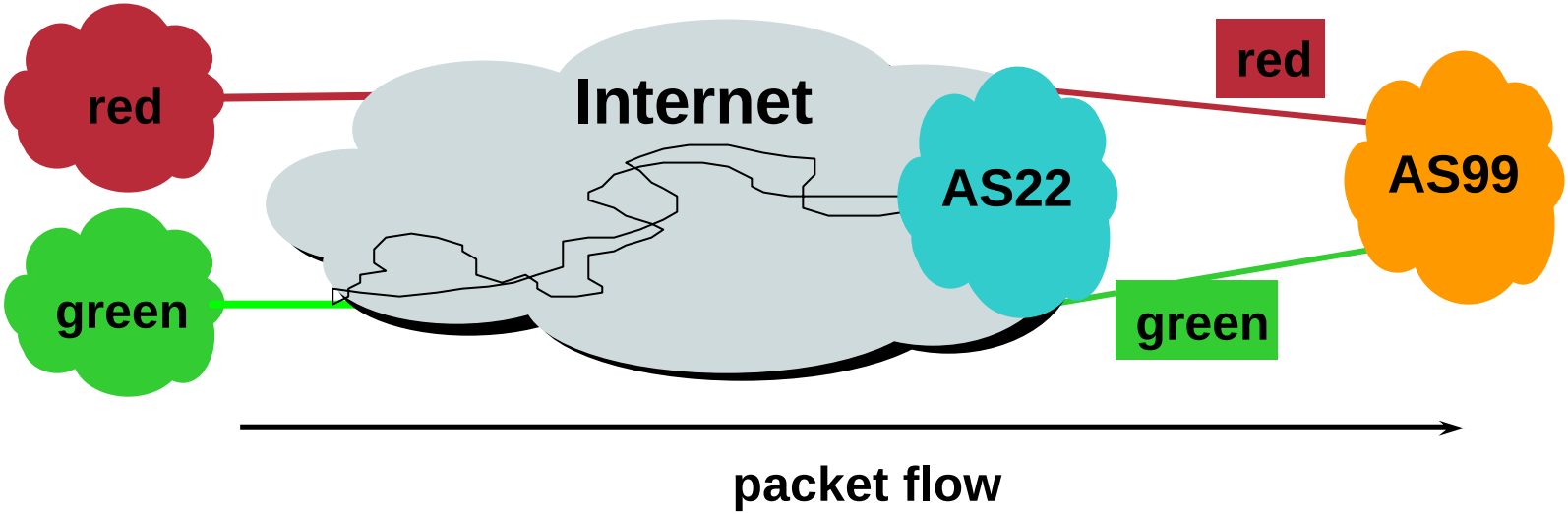


- AS99 uses red link for traffic to the red AS and the green link for remaining traffic
- To implement this policy, AS99 has to:

  Accept routes originating from the red AS on the red link

  Accept all other routes on the green link

Routing Policy Limitations



- AS99 would like packets coming from the green AS to use the green link.
- But unless AS22 cooperates in pushing traffic from the green AS down the green link, there is very little that AS99 can do to achieve this aim

Routing Policy Limitations

- **In the Internet today:**
  - **140000 prefixes (not realistic to set policy on all of them individually)**
  - **17500 origin AS's (too many)**
  - **Routes tied to a specific AS or path may be unstable regardless of connectivity**
- **Groups of ASes are a natural abstraction for filtering purposes**

Routing Protocols

We now know what routing means…

…but what do the routers get up to?

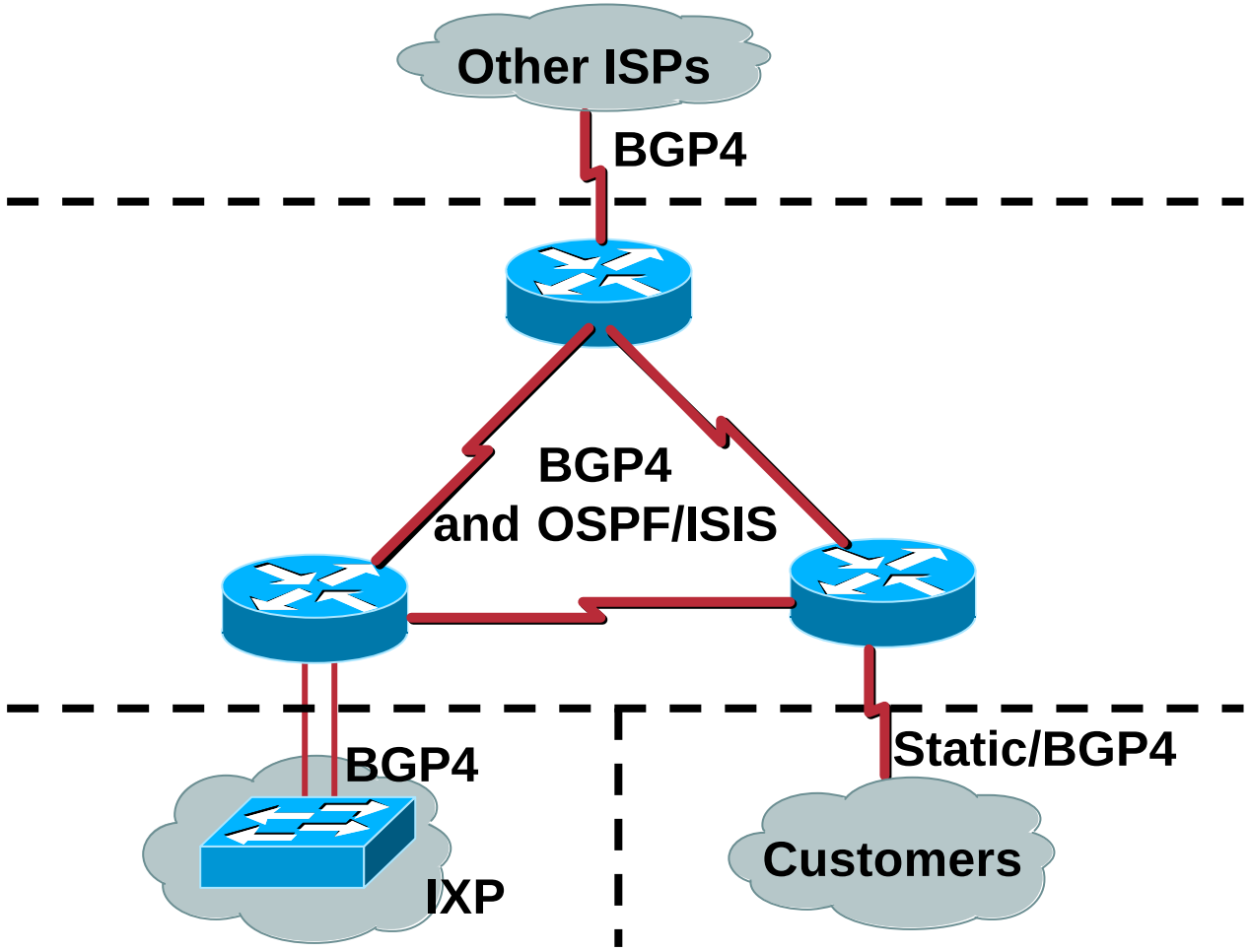
What Is an IGP?

- Interior Gateway Protocol
  - Used within an Autonomous System
  - Carries internal infrastructure prefixes
  - Examples – OSPF & ISIS
- Needed for scaling the ISP's backbone
  - Only used for ISP's infrastructure addresses, not customers
  - Design goal is to minimise number of prefixes in IGP to aid scalability and rapid convergence

What Is an EGP?

- **Exterior Gateway Protocol**
  - **Used to convey routing information between Autonomous Systems**
  - **De-coupled from the IGP**
  - **Current EGP is BGP4**
- **Allows scaling to a large network**
- **Defines administrative boundaries**
- **Used to apply Routing Policy**

Hierarchy of Routing Protocols



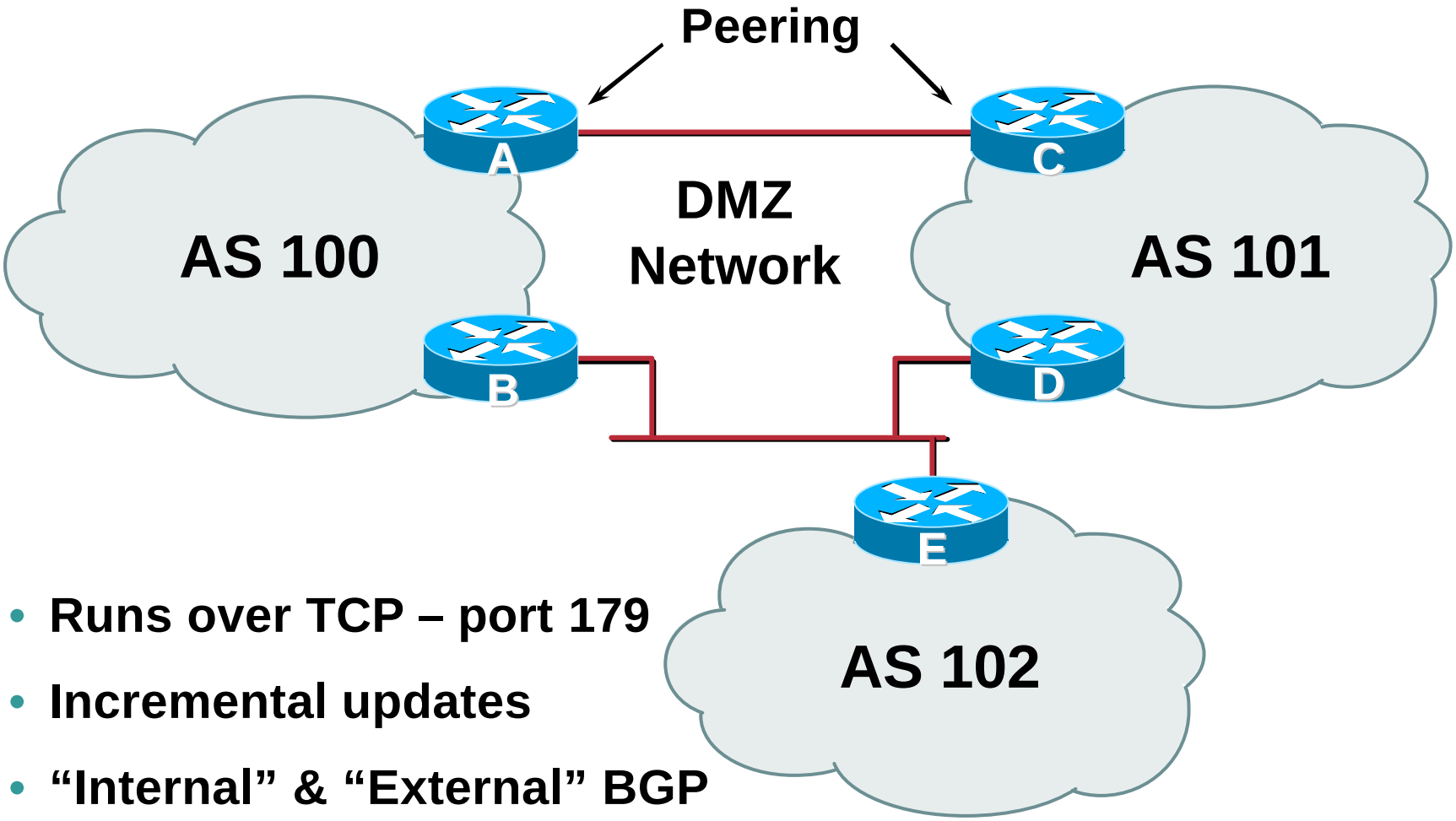
BGP Basics

People (and routers) talk about BGP – what is it?

Border Gateway Protocol

- **Routing Protocol used to exchange routing information between networks**
  - **exterior gateway protocol**
- **Described in RFC1771**
  - **work in progress to update**
  - **www.ietf.org/internet-drafts/draft-ietf-idr-bgp4-23.txt**
- **The Autonomous System is BGP's fundamental operating unit**
  - **It is used to uniquely identify networks with common routing policy**

BGP Basics



- Runs over TCP – port 179
- Incremental updates
- “Internal” & “External” BGP
- DMZ is shared network between ASes

BGP General Operation

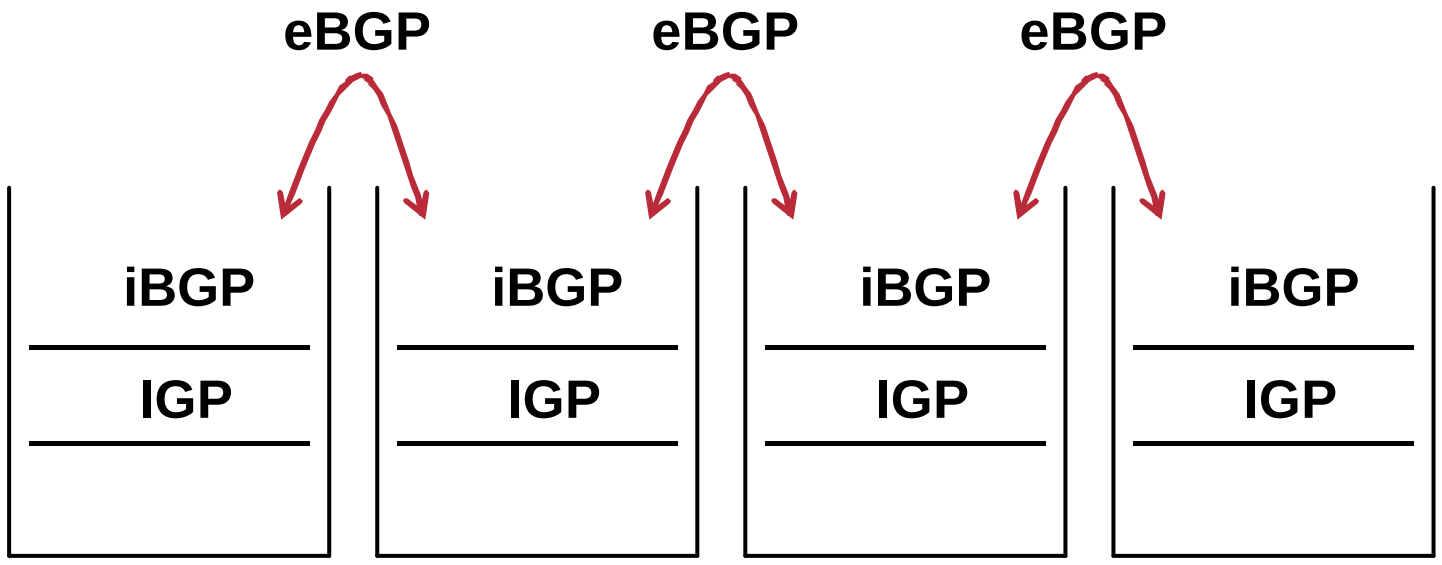
- Learns multiple paths via internal and external BGP speakers
- Picks the best path and installs in the forwarding table
- Best path is sent to external BGP neighbours
- Policies applied by influencing the best path selection

eBGP & iBGP

- **BGP has two applications:**
  - **Inside ISP networks – internal (iBGP)**
  - **Between ISP networks – external (eBGP)**
- **iBGP used to carry**
  - **some/all Internet prefixes across ISP backbone**
  - **ISP's customer prefixes**
- **eBGP used to**
  - **exchange prefixes with other ASes**
  - **implement routing policy**

BGP/IGP model used in ISP networks

- **Model representation**



Aggregation

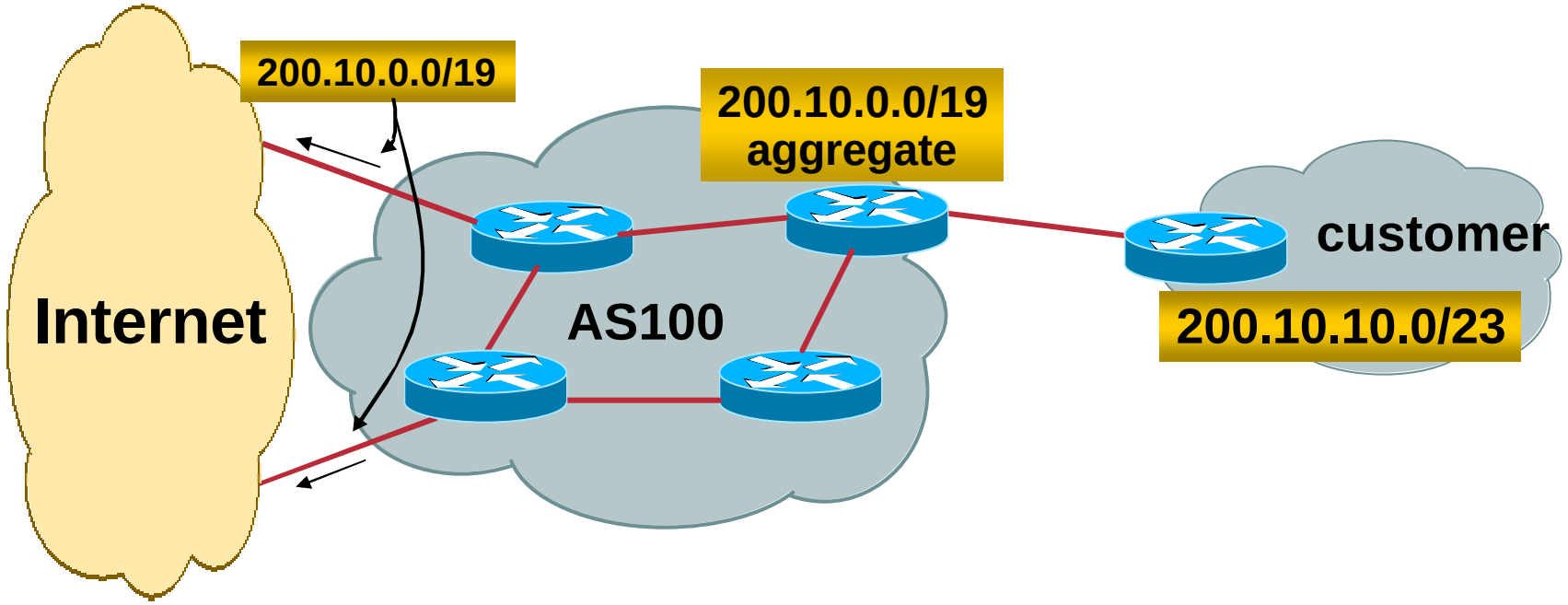
How to announce reachability information to the Internet

"Quality or Quantity?"

Aggregation

- **Aggregation means announcing the address block received from the RIR to the other ASes connected to your network**
- **Subprefixes of this aggregate *may* be:**
  - **Used internally in the ISP network**
  - **Announced to other ASes to aid with multihoming**
- **Unfortunately too many people are still thinking about “class Cs”, resulting in a proliferation of /24s in the Internet routing table**

Aggregation – Good Example



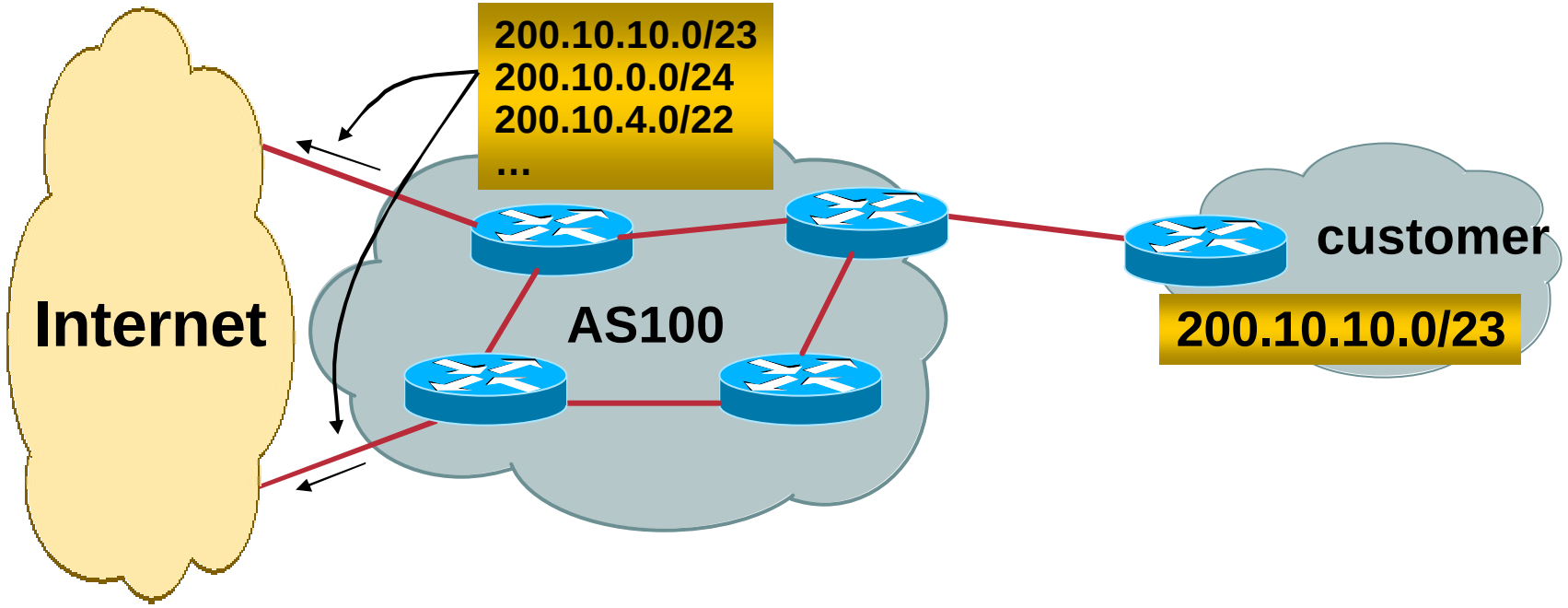
- **Customer has /23 network assigned from AS100's /19 address block**
- **AS100 announces /19 aggregate to the Internet**

Aggregation – Good Example

- Customer link goes down
  - their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- /19 aggregate is still being announced
  - no BGP hold down problems
  - no BGP propagation delays
  - no damping by other ISPs

- Customer link returns
- Their /23 network is visible again
  - The /23 is re-injected into AS100's iBGP
- The whole Internet becomes visible immediately
- Customer has Quality of Service perception

Aggregation – Bad Example



- Customer has /23 network assigned from AS100's /19 address block
- AS100 announces customers' individual networks to the Internet

Aggregation – Bad Example

- Customer link goes down
  - Their /23 network becomes unreachable
  - /23 is withdrawn from AS100's iBGP
- Their ISP doesn't aggregate its /19 network block
  - /23 network withdrawal announced to peers
  - starts rippling through the Internet
  - added load on all Internet backbone routers as network is removed from routing table

- Customer link returns
  - Their /23 network is now visible to their ISP
  - Their /23 network is re-advertised to peers
  - Starts rippling through Internet
  - Load on Internet backbone routers as network is reinserted into routing table
  - Some ISP's suppress the flaps
  - Internet may take 10-20 min or longer to be visible
  - Where is the Quality of Service???

Aggregation – Summary

- **Good example is what everyone should do!**
  - **Adds to Internet stability**
  - **Reduces size of routing table**
  - **Reduces routing churn**
  - **Improves Internet QoS for everyone**
- **Bad example is what too many still do!**
  - **Why? Lack of knowledge?**

The Internet Today (June 2004)

- **Current Internet Routing Table Statistics**

BGP Routing Table Entries	140396
Prefixes after maximum aggregation	84880
Unique prefixes in Internet	67713
Prefixes smaller than registry alloc	64097
/24s announced	76409
only 5529 /24s are from 192.0.0.0/8	
ASes in use	17405

Efforts to improve aggregation

- **The CIDR Report**
  - **Initiated and operated for many years by Tony Bates**
  - **Now combined with Geoff Huston's routing analysis**
  - **www.cidr-report.org**
  - **Results e-mailed on a weekly basis to most operations lists around the world**
  - **Lists the top 30 service providers who could do better at aggregating**

Efforts to improve aggregation
The CIDR Report

- Also computes the size of the routing table assuming ISPs performed optimal aggregation
- Website allows searches and computations of aggregation to be made on a per AS basis

  flexible and powerful tool to aid ISPs

  Intended to show how greater efficiency in terms of BGP table size can be obtained without loss of routing and policy information

  Shows what forms of origin AS aggregation could be performed and the potential benefit of such actions to the total table size

  Very effectively challenges the traffic engineering excuse

Status Summary

Table History

Date	Prefixes	CIDR Aggregated
17-05-04	134431	94339
18-05-04	134557	94505
19-05-04	134683	94655
20-05-04	134815	94861
21-05-04	134981	94909
22-05-04	135027	94796
23-05-04	135200	94941
24-05-04	136041	94926

Plot: BGP Table Size

AS Summary

17183	Number of ASes in routing system
6951	Number of ASes announcing only one prefix
1429	Largest number of prefixes announced by an AS AS7018: ATTW AT&T WorldNet Services
73561344	Largest address span announced by an AS (/32s) AS568: DISOUN DISO-UNRRA

Plot: AS count
Plot: Average announcements per origin AS
Report: ASes ordered by originating address span
Report: ASes ordered by transit address span
Report: Autonomous System number-to-name mapping (from Registry WHOIS data)

Aggregation Summary

The algorithm used in this report proposes aggregation only when there is a precise match using AS path so as to

Aggregation Summary

The algorithm used in this report proposes aggregation only when there is a precise match using AS path so as to preserve traffic transit policies. Aggregation is also proposed across non-advertised address space ('holes').

--- 24May04 ---

ASnum	NetsNow	NetsAggr	NetGain	% Gain	Description
Table	136002	94906	41096	30.2%	All ASes
AS4134	751	153	598	79.6%	CHINANET-BACKBONE No.31,Jin-rong Street
AS18566	704	163	541	76.8%	CVAD Covad Communications
AS4323	725	199	526	72.6%	TWTC Time Warner Telecom
AS9583	475	36	439	92.4%	SATYAMNET-AS Satyam Infoway Ltd.,
AS7018	1429	992	437	30.6%	ATTW AT&T WorldNet Services
AS6197	698	314	384	55.0%	BNS-14 BellSouth Network Solutions, Inc
AS7843	496	115	381	76.8%	ADELPH-13 Adelphia Corp.
AS701	1293	930	363	28.1%	UU UUNET Technologies, Inc.
AS22909	387	37	350	90.4%	CMCS Comcast Cable Communications, Inc.
AS6198	555	225	330	59.5%	BNS-14 BellSouth Network Solutions, Inc
AS22773	372	52	320	86.0%	CXAB Cox Communications Inc. Atlanta
AS27364	358	40	318	88.8%	ARMC Armstrong Cable Services
AS9929	334	33	301	90.1%	CNCNET-CN China Netcom Corp.
AS11172	355	55	300	84.5%	Servicios Alestra S.A de C.V
AS1239	940	644	296	31.5%	SPRN Sprint
AS17676	339	50	289	85.3%	JPNIC-JP-ASN-BLOCK Japan Network Information Center
AS4355	381	99	282	74.0%	ERSD EARTHLINK, INC
AS6140	386	121	265	68.7%	IMPSA ImpSat
AS6478	304	48	256	84.2%	ATTW AT&T WorldNet Services
AS6347	401	150	251	62.6%	SAVV SAVVIS Communications Corporation
AS1221	857	619	238	27.8%	ASN-TELSTRA Telstra Pty Ltd
AS209	735	502	233	31.7%	QWEST-4 Qwest
AS25844	243	16	227	93.4%	SASMFL-2 Skadden, Arps, Slate, Meagher & Flom LLP
AS14654	230	5	225	97.8%	WAYPOR-3 Wayport
AS3356	894	678	216	24.2%	LEVEL3 Level 3 Communications
AS4766	474	263	211	44.5%	KIX Korea Internet Exchange for "96 World Internet Exposition
AS9443	358	155	203	56.7%	INTERNETPRIMUS-AS-AP Primus Telecommunications
AS2386	427	240	187	43.8%	ADCS-1 AT&T Data Communications Services
AS5668	380	197	183	48.2%	CIH-12 CenturyTel Internet Holdings, Inc.
AS6327	208	28	180	86.5%	SHAWC-2 Shaw Communications Inc.
Total	16489	7159	9330	56.6%	Top 30 total

Current Issues

What's hot!

Hot Topics

- **“Internet Stability”**

  Safe network, router & BGP configuration:

  www.cymru.com/Documents/index.html

  www.cymru.com/Documents/secure-bgp-template.html

  www.cymru.com/Documents/bogon-list.html

- **“Security of the BGP session”**

  MD5 passwords on BGP sessions (?)

  BTSH – BGP TTL Security Hack

  www.nanog.org/mtg-0302/hack.html

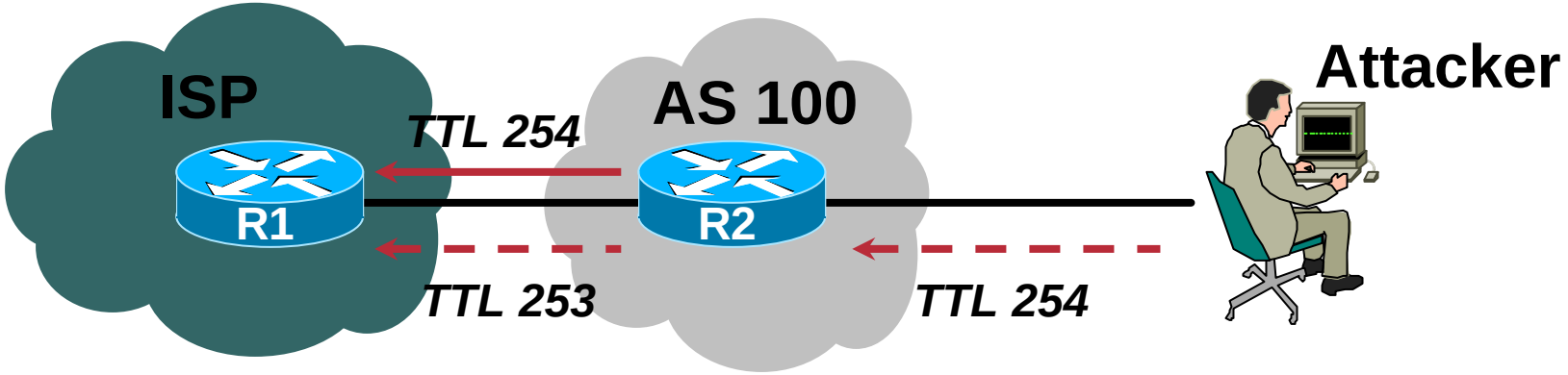
BGP TTL “hack”

- **Implement RFC3682 on BGP peerings**

  Neighbour sets TTL to 255

  Local router expects TTL of incoming BGP packets to be 254

  No one apart from directly attached devices can send BGP packets which arrive with TTL of 254, so any possible attack by a remote miscreant is dropped due to TTL mismatch



Summary

- Topologies, Definitions & Routing
- BGP
- Aggregation
- Current Hot Topics

Introduction to Routing

How traffic flows on the Internet

Philip Smith pfs@cisco.com

RIPE NCC Regional Meeting,

Moscow, 16-18 June 2004