Internetworking

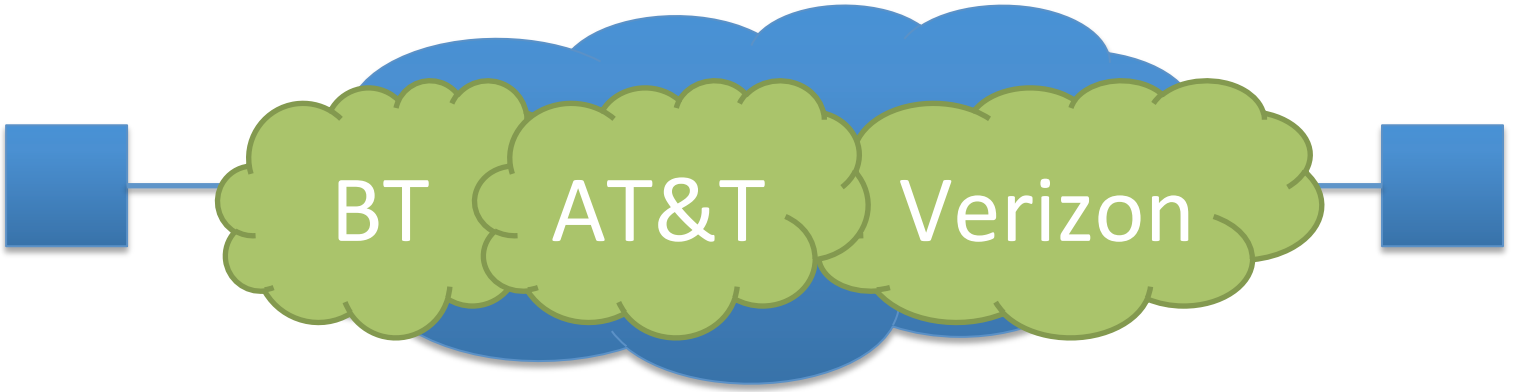
G54ACC

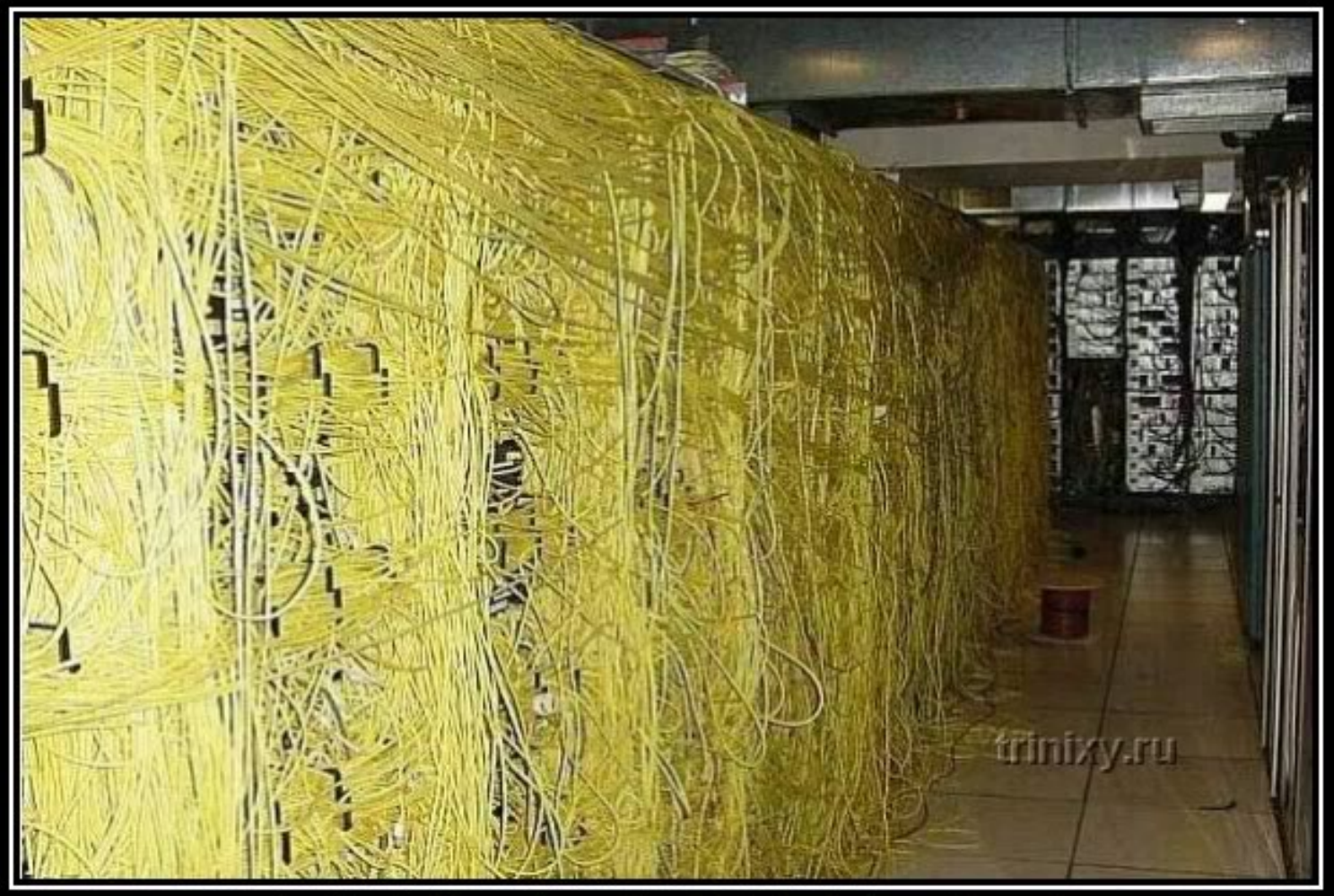
Lecture 14

richard.mortier@nottingham.ac.uk

Internetworking

- So far we have talked about:
  - Moving data between hosts
  - Moving data within a network (administrative domain)
- So what is the **Internet** then, really?





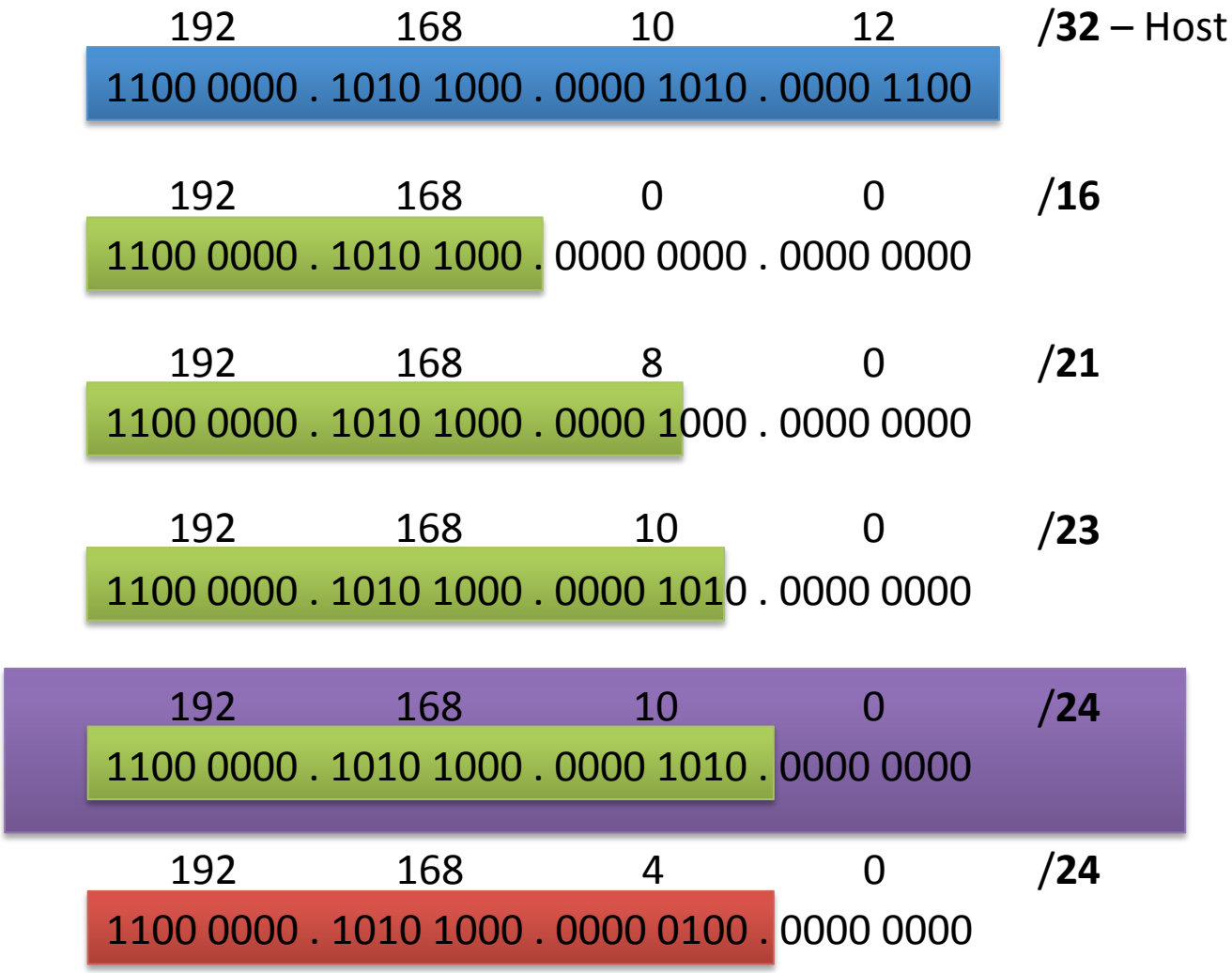
INTERNET

A series of tubes.

Recall: Routing *vs.* Forwarding

- Router receives an IP packet: what to do?
  - Drop or forward via an interface
- Deciding which interface is **forwarding**
  - IP bases this decision (almost) solely on the destination IP address
- Building up the information to do so is **routing**
  - Where are all the addresses at the moment?

Recall: Longest Prefix Matching



Contents

- Routing
- The Protocol
- Decision Process
- Operations

Contents

- Routing
  - Inter-domain Routing
  - BGPv4
  - Autonomous Systems
- The Protocol
- Decision Process
- Operations

Routing Protocols

- Distribute the data to build forwarding tables
- Examples we saw: OSPF, IS-IS, RIP
  - Link-state, Distance vector

- These are **intra-domain routing protocols**
  - Or **Interior Gateway Protocols**
  - Source and destination **inside** the same network

- What happens **between** networks?

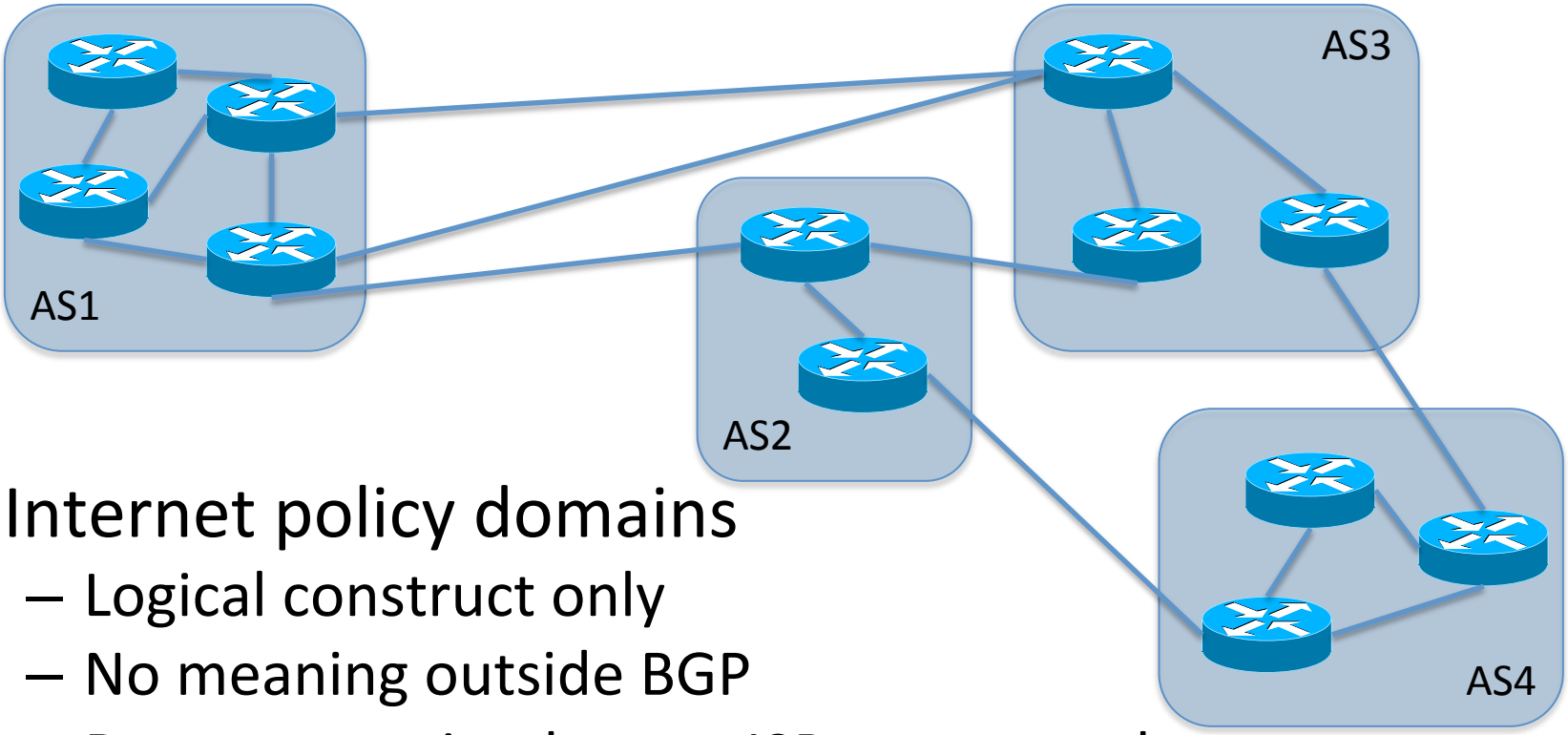
Inter-domain Routing

- An important distinction: local *vs*. global
  - Interior *vs*. Exterior Gateway Protocol (IGP, EGP)
  - Why is this important? Two reasons:
- Dynamics
  - Need to scope information propagation (why?)
- Protection (information hiding)
  - Competition: your goals are not your neighbours'

Border Gateway Protocol, BGPv4

- **The** Internet inter-domain routing protocol
  - RFC 4271, updating RFC 1771
  - Derives originally from GGP, EGP (1982)
  - Updated over time (RFCs 1105, 1163, 1267)
- Deals in IP prefixes and **Autonomous Systems**
  - Latter purely administrative
  - Only prefixes matter in the data-plane
- Purpose is to enable *policy* to be applied

Autonomous Systems, ASs



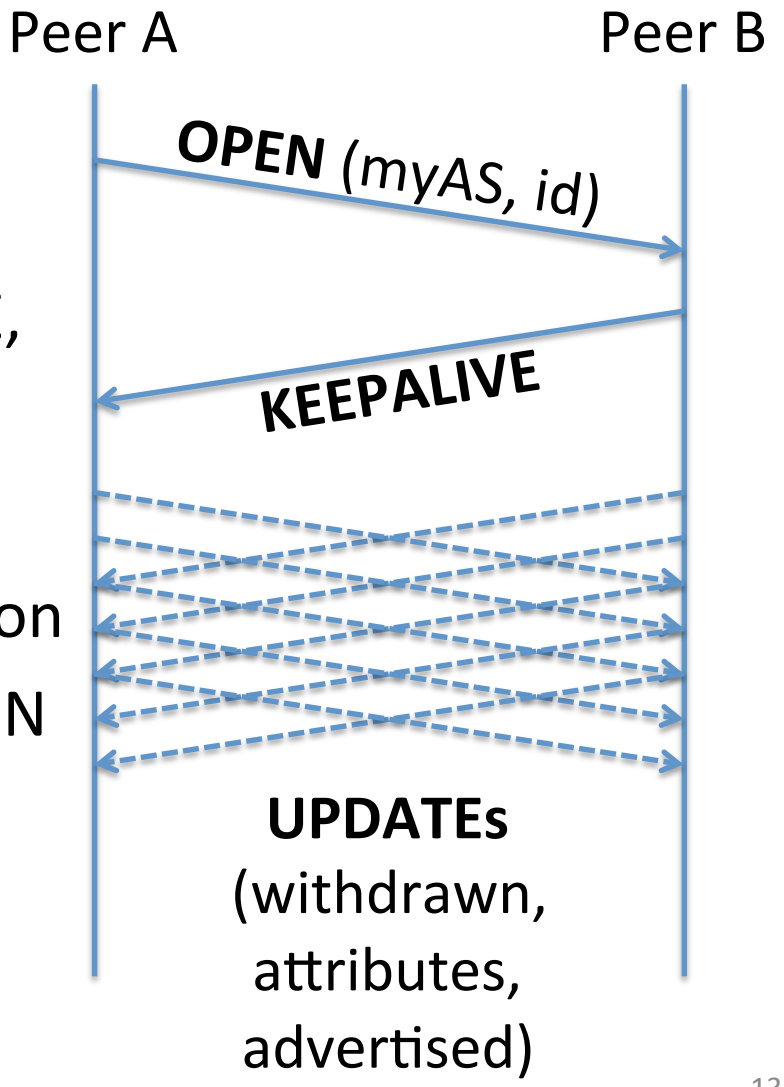
- Internet policy domains
  - Logical construct only
  - No meaning outside BGP
  - Do not map simply onto ISPs or networks
- Currently 410,000 prefixes, 40,000 ASs

Contents

- Routing
- The Protocol
  - Sessions
  - Updates
  - Path Attributes
- Decision Process
- Operations

A Very Simple Protocol

- Exchanges prefixes
  - Uses TCP/179 as transport
  - OPEN, UPDATE, KEEPALIVE, NOTIFICATION
- *Sessions* between *peers*
  - Simple capability negotiation
  - Manage simultaneous OPEN
  - Lose *everything* on failure (why?)



Sessions

- BGP peer typically has many sessions
  - 10? 20? 100s?
- Logically, *Adj-RIB-In*, *-Out* for each session
  - Advertisements received and to be sent
- Generate *Loc-RIB* from *Adj-RIB-In*
  - Routes to use and to distribute
  - Resolved into per-port forwarding tables
- Generate *Adj-RIB-Out* from *Loc-RIB* and policy

UPDATEs

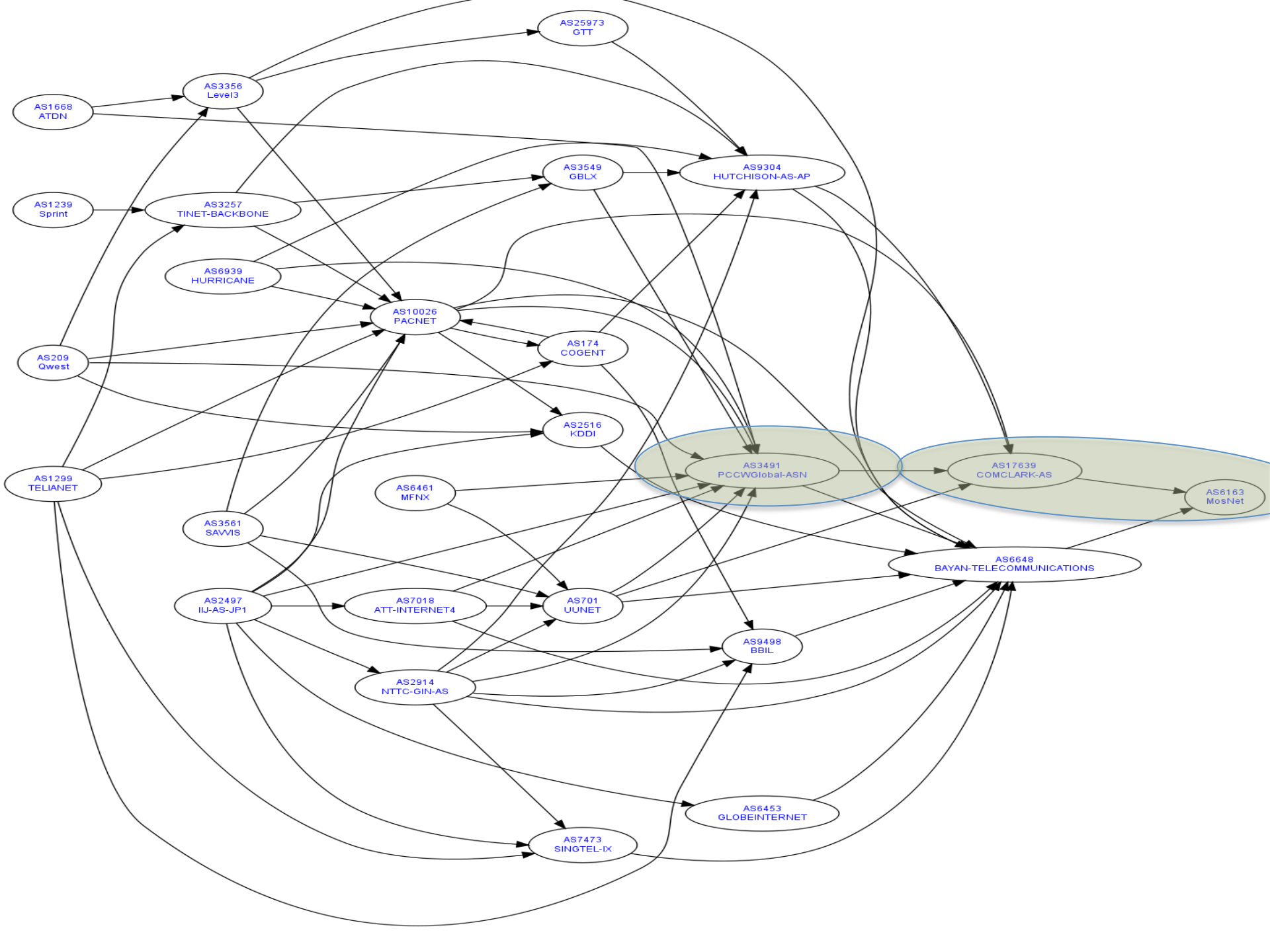
- Incremental – indicate *changes* to state
  - Withdrawn routes
  - Path attributes, common to all advertised routes
  - Advertised routes, known as NLRI
- There are ~27 path attributes defined
  - Perhaps a dozen or so are in common use
  - Communicate information about prefixes
  - Used to apply policy in BGP *decision process*

Path Attributes

- Well-known, Mandatory
  - **Next Hop**
  - **AS Path**
  - **Origin**
- Optional, Transitive
  - Aggregator
  - Community
  - **Extended Communities**
- Well-known, Discretionary
  - **Local Preference**
  - Atomic Aggregate
- Optional, Non-transitive
  - **Multi-Exit Discriminator**
  - Originator ID
  - …

An Example UPDATE

```
[ Thu Apr  1 04:26:25 2010 ]
MRT packet: len: 81, type: PROTOCOL_BGP4MP, subtype: MESSAGE
  AS(src): 39202, AS(dst): 12654
  ifc idx: 0, AFI: IP
  IP(src): 195.66.225.2, IP(dst): 195.66.225.241
  Update (len=65): unfeasible_len=0 path_attr_len=26
    UNFEASIBLE ROUTES:
    PATH ATTRIBUTES:
      ORIGIN: IGP [ transitive ]
      AS_PATH: (SEQUENCE)[ <- 39202 <- 3491 <- 17639 <- 6163 <- 6163 ] [ transitive ]
      NEXT_HOP: 195.66.224.167 [ transitive ]
    FEASIBLE ROUTES:
      1: 61.9.0.0/24
      2: 61.9.1.0/24
      3: 61.9.62.0/24
      4: 202.47.132.0/24
```


AS25973
GTT
AS3356
Level3
AS1668
ATDN
AS3549
GBLX
AS9304
HUTCHISON-AS-AP
AS1239
Sprint
AS3257
TINET-BACKBONE
AS6939
HURRICANE
AS10026
PACNET
AS174
COGENT
AS209
Qwest
AS2516
KDDI
AS3491
PCCWGlobal-ASN
AS17639
COMCLARK-AS
AS6163
MosNet
AS1299
TELIANET
AS6461
MFNX
AS3561
SAVVIS
AS6648
BAYAN-TELECOMMUNICATIONS
AS2497
IIJ-AS-JP1
AS7018
ATT-INTERNET4
AS701
UUNET
AS9498
BBIL
AS2914
NTTC-GIN-AS
AS6453
GLOBEINTERNET
AS7473
SINGTEL-IX

Contents

- Routing
- The Protocol
- Decision Process
  - Path Vectors
- Operations

Path Vectors – AS_PATH

- Distance vector – prefer lowest cost path
  - Need to break loops somehow (how?)
- **Path Vector**
  - How do we know if we've seen this advert before?
  - Store the list of ASs through which it reached us
  - The AS_PATH
- Loops can be broken:
  - If our ASN appears in a received AS_PATH, drop it

Decision Process

- Drop prefix if:
  - NEXT_HOP is unreachable via local routing table
  - Local AS appears in AS_PATH
- Then apply following preference:

1. Higher WEIGHT
   (local to this router)
2. Highest LOCAL_PREF
3. Shortest AS_PATH
   (leads to *AS padding*)
4. Lowest ORIGIN
5. Lowest MED
   if from same AS – why?
6. EGP to IGP
7. Shortest internal path
8. Prefer oldest route
9. Lowest Router-ID
   (usually, highest router IP)
10. Lowest interface IP address

Contents

- Routing
- The Protocol
- Decision Process
- Operations
  - Consistency
  - Scaling
  - Confederations
  - Route Reflectors

Consistency

- Learn external routes on EBGP sessions
  - Peers have different ASNs
  - Must ensure **every** router knows **all** external routes (why?)
- Redistribute external routes inside network
  - Via IGP – only in small networks (why?)
  - Via IBGP – gives full control over how
- What's the problem with IBGP?

Scaling

- Can't distribute IBGP routes on IBGP sessions
- Have to maintain N.(N-1)/2 IBGP sessions
  - Each carrying up to 410k routes x 2 tables
- Two solutions
  - Route reflectors:
    supernodes, readvertising IBGP routes
  - AS confederations:
    split AS up into mini-ASs
  - Both tweak decision process somewhat

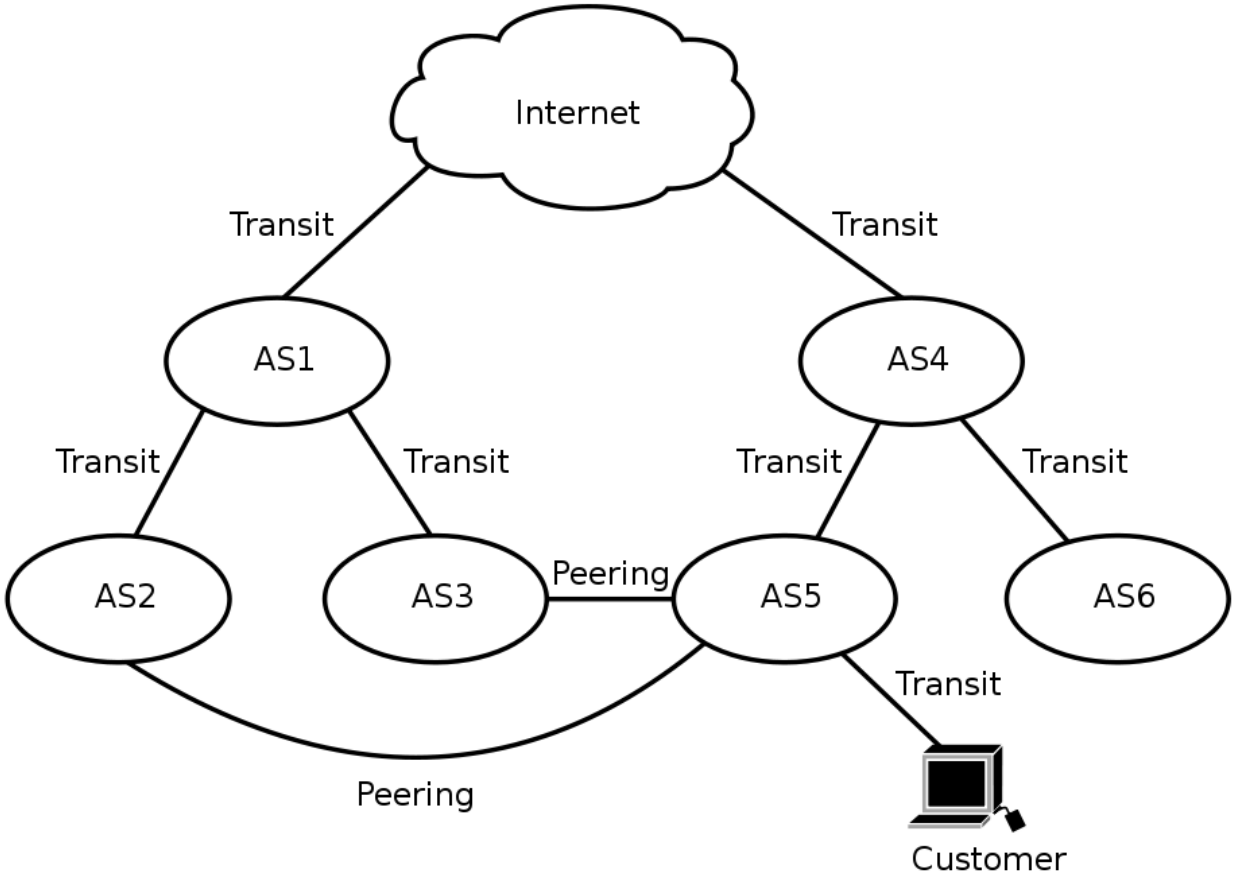
Operations

- Handle link failures
  - Bind to loopback
  - Flap damping (but can make things worse!)
- Process failures
  - Out of memory error due to too many routes
- Hijacking, intentional and unintentional
  - Don't believe everything you read
  - http://www.youtube.com/watch?v=IzLPKuAOe50
- *Anycast* (1:1-of-N)
  - Advertise same *prefix* in many places. Carefully.

Network Interconnection

- Networks interconnect via EBGP sessions
  - POPs, Points-of-Presence
  - IXs, Internet eXchanges
- Multi-homing
  - Note that this is all *logical* – what about *physical* diversity?
- How does this all fit together?
  - Public/Private Peering *vs*. Transit
  - Roughly hierarchical (this is changing)
  - Tier-1/core/backbone *vs*. the rest
- As ever, business and politics
  - E.g., Level3 *vs*. Cogent depeering

Simple Example of a Complex Graph



(Policy – example from Level3)

Contents

- Routing
  - Inter-domain Routing
  - BGPv4
  - Autonomous Systems
- The Protocol
  - Sessions
  - Updates
  - Path Attributes
- Decision Process
  - Path Vectors
- Operations
  - Consistency
  - Scaling
  - Confederations
  - Route Reflectors

Summary

- The Internet is inter-connected networks
  - The routing protocols are what hold it together
- BGPv4 is **the** inter-network routing protocol
  - All about application of policy
  - To meet business needs
- Simple protocol, can be arbitrarily complex
  - Many operational matters make this hard

Quiz (1)

1. What information needs to be exchanged between networks to route packets?
2. What constraints are different between an IGP and an EGP?
3. Why does BGP add path attributes to prefixes?
4. What is an AS?
5. Why is simultaneous open of BGP sessions an issue, and how is it resolved?
6. What might happen if the corresponding tables and routes were not removed on session failure?

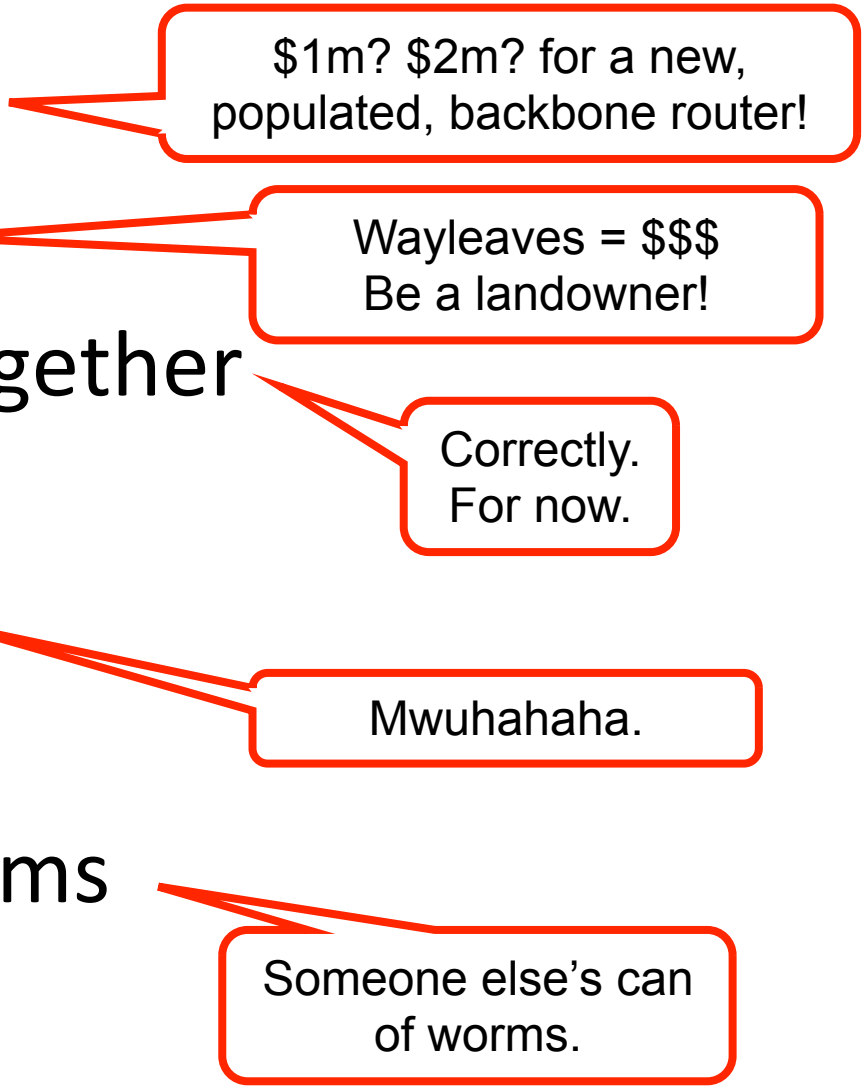
Quiz (2)

7. What are the 3 types of BGP table, and what are they for?
8. In what way(s) is BGP not a distance vector protocol, and why?
9. What are the different effects of the stages in the decision process on sl.21?
10. Why is redistributing BGP routes via the IGP a problem?
11. Draw two diagrams showing how AS confederations and route reflectors address IBGP scalability in different ways.
12. What is the difference between peering and transit?

Extras…

So, how do you build an IP network?

1. Buy (lease) routers
2. Buy (lease) fibre
3. Connect them all together
4. Configure routers
5. Configure end-systems



Multiple Router Flavours

- Core
  - OC-12 (622Mbps) and up (to OC-768 ~= 40Gbps)
  - Big, fat, fast, expensive
  - E.g., Cisco HFR, Juniper T-640
  - HFR: 1.2Tbps each, interconnect up to 72 giving 92Tbps, start at $450k
- Transit/Peering-facing
  - OC-3 and up, good GigE density
  - ACLs, full-on BGP, uRPF, accounting

Multiple Router Flavours

- Customer-facing
  - FR/ATM/…
  - Feature set as above, plus fancy queues, etc
- Broadband aggregator
  - High scalability: sessions, ports, reconnections
  - Feature set as above
- Customer-premises (CPE)
  - 100Mbps, maybe
  - NAT, DHCP, firewall, wireless, VoIP, …
  - Low cost, low-end, perhaps just software on a PC

Multiple Router Flavours

Cisco CRS-1
Multi-shelf system

Network Design

- Whose network?
  - ISPs, IXs, enterprise, campus
  - POPs, DCs
- Many designs:
  - Flat
  - Hierarchical
  - Hybrids
  - Multiple scales

Network Design Constraints

- Business
  - Backwards compatibility. Who to connect. Peering.
- Technology
  - Power – directly (24x7 operation) and indirectly (cooling)
  - Port density *vs*. raw bandwidth
  - Software reliability
  - Hardware/software capability
    - Addressing schemes for scalability, summarization
    - Can't run feature X with feature Y on vendor C in network size N
- Connectivity/resiliency
  - "All core routers connect to at least 2 other core routers"
  - "All edge routers connect to at least 2 core routers"

Router OS Configuration

- Initialization
  - Name the router, setup boot options, setup authentication options
- Configure interfaces
  - Loopback, Ethernet, fibre, ATM
  - Subnet/mask, filters, static routes
  - Shutdown (or not), queuing options, full/half duplex

Router Software Configuration

- Configure routing protocols (OSPF, BGP, &c)
  - Process number, addresses to accept routes from, networks to advertise
  - Access lists, filters, ...
    - Numeric id, permit/deny, subnet/mask, protocol, port
  - Route-maps, matching routes rather than data traffic
- Other configuration aspects: traps, syslog, &c
  - (Oh, and switch configuration is about as painful)

Router Configuration Fragments

```
hostname FOOBAR
!
boot system flash slot0:a-boot-image.bin
boot system flash bootflash:
logging buffered 100000 debu
logging console informationa
aaa new-model
aaa authentication login def
authentication login console
aaa authentication ppp defau
aaa authorization network ta
ip tftp source-interface Loo
no ip domain-lookup
ip name-server 10.34.56.78
!
ip multicast-routing
```

```
interface Loopback0
 description router-1.network.corp.com
 ip address 10.65.21.43 255.255.255.255
!
interface FastEthernet0/0/0
 description Link to New Yor
 ip address 10.65.43.21 255.
 ip access-group 175 in
 ip helper-address 10.65.12.
 ip pim sparse-mode
 ip cgmp
 ip dvmrp accept-filter 98 n
 full-duplex
```

```
router ospf 2
 log-adjacency-changes
 passive-interface FastEthernet0/0/0
 passive-interface FastEthernet0/1/0
 passive-interface FastEthernet1/0/0
 passive-interface FastEthernet1/1/0
 passive-interface FastEthernet2/0/0
 passive-interface FastEthernet2/1/0
```

```
access-list 24 remark Mcast ACL
access-list 24 permit 239.255.255.254
access-list 24 permit 224.0.1.111
access-list 24 permit 239.192.0.0 0.3.255.255
access-list 24 permit 232.192.0.0 0.3.255.255
access-list 24 permit 224.0.0.0 0.0.0.255
access-list 1011 deny   0000.0000.0000 ffff.ffff.ffff ffff.ffff.ffff 0000.0000.0000 0xD1 2 eq 0x42
                                                                    ff.ffff.ffff
```

```
tftp-server slot1:some-other-image.bin
tacacs-server host 10.65.0.2
tacacs-server key xxxxxxxx
rmon event 1 trap Trap1 description "CPU Utilization>75%" owner config
rmon event 2 trap Trap2 description "CPU Utilization>95%" owner config
```

Router Configuration

- Lots of large, fragile text files
  - 00s/000s routers, 00s/000s lines per config
  - Errors are hard to find and have non-obvious results
  - Router configuration also editable on-line
  - Order matters!
- How to keep track of them all?
  - Naming schemes, directory trees, CVS, ssh upload and atomic commit to router
  - Perhaps even a proper database
- State of the art is pretty basic
  - Few tools to check consistency, design goals
  - Generally generate configurations from templates and have human-intensive process to control access to running configs