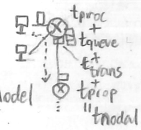
EIE4102 IP Network

Access Network
- Residential/Institutional/Mobile access network
- Digital Subscriber Line (DSL) = Use existing tel line to DSL Access Mux
- Cable net = to ISP router
- Home/Enterprise/Wireless Access network

Packet Delivery Model



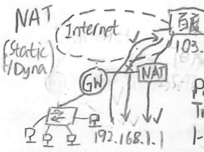
Intf Type by Spd
- NM Net Module 网络模块
- E Ethernet 以太网(10M)
- FE/ESW Fastethernet 快速以太网(100M)
- Gigabitethernet 千兆以太网 (1G)
- XGigabitethernet 万兆以太网 (10G)
- T Serial 串口

IP(v4) Network

(1~126, 129~223).X.X.X/X 可变长子网掩码(VLSM)
Pub IP with variable length subnet mask)
Classless Inter Domain Routing (CIDR) 超网
(224~239).X.X.X ClassD multicast (group)
(240~255).X.X.X ClassE military & research
127.X.X.X Loopback (Test NI(网卡))

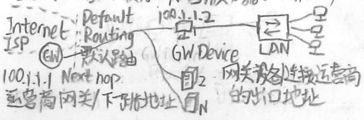
This net = Net.—00—
Broadcast to this net: Net.—FF—
Any address = 0.0.0.0 (0.X.X.X reserved)
Broadcast to Internet = 255.255.255.255
Priv IP: 10.X.X.X/8, 172.1X.X.X/12, 192.168.X.X/16

NAT (Static/Dyna)



Port Addr. Trans (PAT): 1-N Trans by Port

Company Internet Access 带宽接入
企业/大客户服务 N台服务器 (公司/IDC)

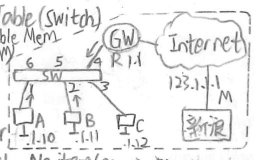


MAC Addr. Table (Switch)

Content Addressable Mem (CAM)

Intf	MAC
1	A
2	B
4	R

⊕... Pending for frame arrival



No item (eg. C): Flooding
Dynamic Table by self-learning
Table cleared when restart (can sol buffer OF)
SW without mem/table item = hub

A→新浪
SIP: 1.10 DIP: 123.1.1.1
SMAC: A DMAC: R
(out of LAN, to GW)
(上网必须设网关: 为了封装网关MAC)
路由器必须配路由
交换机插上就能用

ARP (Layer 2.5, Find MAC via IP)

eg. A→C
- A→ARP.1.11
  - IP S:1.10 D:1.12
  - MAC S:A D:FF
- B&GW receive but don't respond
- ARP resp ← C
  - IP S:1.12 D:1.10
  - MAC S:C D:A
- A add C to ARP Table

`>arp -a`
`$ arp -n`, `$ ip neigh`
`$ cat /proc/net/arp`

Eth. frame: Pre(7) | SFD(1) | Dest(6) | Src(6) | Typ(2) | Data(46~1500) | CRC(4)

0	8b	16b (IPv4=0800)	32b
HW Typ (Eth=01)		Protocol Typ	4B
HW len	Prot. len	Op. (Req=1, Resp=2)	8B
Src. HW Addr. (MAC) (flex len)			12B
		Addr. (IP)	16B
Src Prot.		(flex len)	20B
Dest. HW Addr. (0 in req)			24B
Dest. Prot. Addr			28B

Intf	IP	MAC	Type	State	Que	Attem	T
1:.1.10	.1.1	R	static	R			
1:.1.10	.1.12	C	Dyna	R			900
1:.1.10	.1.11		Dyna	P	1	2	
				F			

ARP Cache Entry: (Pending / Res ✓ / Free)
- Hit: go (no update)
- Miss: enqueue pkt, Add (P) & ARP req
- ARP Add/Update: Add/Update (T=900s)
- Every 60s: Update T, (P) Attempt+1 (>5 ⇒ F)

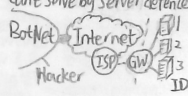
IP Pkt → Ctrl → Queue

RARP: ≈DHCP, use MAC to find IP, for network booting

TCP/IP Infra Attacks:
- Intranet: ARP Spoofing (ARP欺骗) (iframe挂马)
  Pheno: Intermittent, Laggy, No Internet, Insert Trojan
  eg. B as Attacker
  - A: ARP .1.1 (GW)?
  - B→A: .1.1 is Q(x)
  - A→GW ⇒ A→Q (x)

  sol:
  - Dyn ARP Insp. (DAI) 动态ARP检查
  - Static ARP Entry
  - ARP Limit
  - VLAN
  - 802.1X auth
  - Encrypt/VPN

  Pkt capture: B sends a lot of ARP resp msg
- Extranet/Internet: DDOS
  Backdoor → BotNet (肉鸡, 僵尸网络, eg. 灰鸽子)
  Can't solve by server defence / unplug cable



eg. DDoS → 2 ⇒ entire IDC ×, for core GW exit point blocked
sol.: contact upper ISP to create Empty Route to 2

DHCP (Layer 5, send in UDP packet, ServerPort=67, ClientPort=68)

UDP: Src Port | Dest Port | Len | Cksum | Data

DHCP 8b	16b	24b	32b	
OpCode	HWType	HWLen	HopCnt	4B
Trans ID (XID)				8B
Secs (from boot)		Flags		12B
Client IP Addr (Ciaddr)				16B
Your (Offered) IP Addr (Yiaddr)				20B
Server IP Addr (Siaddr)				24B
Gateway IP Addr (Giaddr)				28B
Client HW Addr (Chaddr) (+16B)				44B
Server Name (Sname) (+64B)				108B
Boot File Name (File) (+128B)				236B
Option (text) (+64B)				300B

OpCode
1 DISCOVER
2 OFFER
3 REQ
4 DECLINE
5 ACK
6 NAK
7 RELEASE
8 INFORM

Flag: 0: 0 unicst, 1 bcst; 1~15: rsvd

(All by bcst)
DISCOVER (SIP=0) →
← DHCPOFFER
DHCPREQ →
← DHCPACK

Renew: ACK(t) [BOUND] T1 = 0.5t → REQ (unicst) [RENEWING] T2 = 0.875t → REQ [REBINDING] t → DISCOVER [INIT]

Large Net: Mult DHCP Server
- Relay Agent (≈ Sub server)
- Mult balancer (must sync)
  Otherwise IP collision (FCFS)

DHCP Spoofing (DHCP欺骗)
eg. 1 DHCP Server, 3 Attacker
- 2 DISCOVER
- 3→2 .10.1
- 2 .10.1 rej by GW

sol. 802.1X auth; DHCP snooping (conf SW trusted intf)

ICMP v4

IP: Head | Data

0	8	16	32b
Type	Code	Cksum	4B
(Qry:) ID		Seq	
(Err:) Rest of Hder			8B
Data			

Type
0: PING resp.
3: Dest. Unreachable
4: Src. Quench (源关闭) 阻塞
8: PING req.
13: Timestamp req.
14: Timestamp resp.

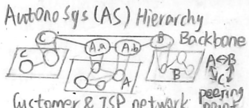
Code (Type 3)
0: Net × 网络
1: Host × 主机
2: Prot × 协议
3: Port × 端口
4: Frag required, remove DF 要求分段并取消DF标志
5: Src routing fail 源路由失败
6: Dest Net Unknown 未知的目标网络
7: Dest host unknown 未知的目标主机
8: Src host isolated 源主机隔离
9: Proh. Dest Net 禁止访问的网络
10: Proh. Dest Host 禁止访问的主机
11: Spec TOS Net × 特定TOS网络不可达
12: Spec TOS Host × 特定TOS主机不可达
13: × Admin filter 由于过滤网络流量被终止
14: ×: 主机越权 host preced. vio
15: ×: preced. cutoff 优先权终止生效

5: (Route) Redirect
Code 0: Net route
1: Host route
2: Net (spec TOS)
3: Host (spec TOS)
9: Router Advertise
10: Router Solicit
11: ICMP timeout
0: TTL 1: Frag Assem
12: IP Param Problem
code
0: Err field (with 8b pntr)
1: Missing required field
2: Unsupported len

IP Routing & Forwarding

逐跳查询转发
Forwarding: time-critical. Table lookup + forwarding (per hop) ⇐ forwarding table: Addr matching: best possible (Longest Pre Match LPM)
Routing: less time-critical. Find + setup correct & fast routes ⇒ routing table

Autono Sys (AS) Hierarchy



Customer & ISP network; peering point (peer 对等体)

Routing table

Type	Dest(seg)	Mask	GW/Nxt Hop	intf	HopCnt/Cost
S[1]	0.0.0.0	/0	192.168.1.1	f1/0	20
C[0]	203.1.1.112	/32	On Link	f2/0	276
B[ ]	154.2.4.0	/24	100.1.1.250	f3/0	306
R[ ]	154.2.4.0	/24	100.2.1.100	f3/0	334

default (0.0.0.0/0) Single GW ⇒ Nxt Hop Fixed
Loopback (127.0.0.0/8) create VNIC (虚拟网卡)
To virtualize a net seg. eg. VMware

Admin Dist (AD 管理距离): Static=1, RIP=120, EIGRP(D)=90, Conn=0, ISIS=115, OSPF(D)=110, IBGP=200, EBGP=20
`>route print`
`$ route` | `$ ip route show`
`# show ip route (sh ip r)`
`# show ip protocols`

Example Algo
, 1; 01*, 10*;
001*, 101*;
1001*, 0110*;
1011*; 11101*
(5 bit Addr)

reTRIEval
⇒ * → 001*, 01*, 1* → 0110*, 10*, 11101* → 1001*, 101* → 1011*

Bin Search on Prefix length
HT search tree
5b S=2 S=1; S=3 → 1 → 2; 3 → 5 → 4
(HT of 3b, 5b, 4b)

3b HT:

Pre	F (cont)	LSF (long so far)
001	Off	–
111	On	1
100	On	2
011	On	2
101	On	3

Bin Search Table

Bin	>	=
00000	*	*
00100	001	001
01000	01	01
01100	0110	0110
01101	01	0110
01111		01
10000	10	10
10010	1001	1001
10100	101	101
10110	1011	1011
10111	1	1011
11101	1	11101
11111	–	1

Cisco 3640 Router Internetworking OS (IOS)

- ROM: bootstrap 引导系统 (加载主板驱动)
- RAM: #copy running-config startup-config
- NVRAM: conf 配置 ← #write memory
- FLASH: IOS 应用系统 升级清空: #format flash: / erase startup-c

Router> user 用户 (有限度查看配置)
Router# privilage 特权 ⇄ enable (en) / disable (disa)
`#show ip interface brief (sh ip int br)` 查接口IP表
`#show running-config (sh run)` 查所有配置
`#show line (all) (sh lin)`
Ctrl+6/C: Abort

Config mode 配置模式
`#config terminal (conf t)`
`(c)# hostname R1 (host R1)`
1) 工程3组 接口配IP (业务接口)
`(c)# interface f0/0 (int f0/0)`
`(c-if)# ip address IP MASK (ip add A.B.C.D A.B.C.D)`
`(c-if)# no shutdown (no sh)`
`(c)# no ip domain lookup` 关闭命令域名解析 `(no ip do lo)`
`(c)# enable password/secret XXX` 小心
`(c)# no enable password/secret` 空格

(控制接口)
2) `(c)# line console 0 (lin co 0)`
`(c-line)# no exec-timeout (no exec-t)` 关闭对话超时
3) `(c-line)# logging sync (logg s)` 打开信息同步
`(c)# line vty A B` (telnet A~B 个进程)
`(c-line)# password XXX`
`(c-line)# login` 开启远程
`(c)# user XX pass XXX`
`(c)# user XX pri N`

静态路由 Static Routing
`(c)# ip route DEST MASK NXT`
`# no ip route`
`# sh ip r static`
建立虚拟网段 Vir Net Seg
`(c)# interface loopback 0 (int lo 0)`
`(c-if)# ip add IP MASK`
`(c-if)# no sh`

Dyn Routing 动态路由: Dist Vec (DV) 距离矢量 / Link State (LS) 链路状态
IGP Int GW Prot. / EGP Ext GW Prot.

RIP (Routing Info Prot.) v2 路由信息协定: DV based

Periodic complete info ex (25~35s, avg 30s, not Hello) + triggered update
RIPv2 = Grpcast Update @ 224.0.0.9
No neigh relation, no Hello pkt
Use info ex. for confirm alive
180s timeout (possi down) + 60s GC
HopCnt as measurement (max 15) (same: load balance)
Count to inf. 路由环路 (can't solve)

sol:
- split horizon 水平分割 (info no send to src.)
- poison rev. 毒性逆转 info ∞ send to src (work as ACK)

req=1, resp=2; Prot. TCP/IP=2; in IP Pkt

0	8	16	32b
Com	Ver	Rsvd	
Family		Tag (AS)	
Network Addr.			×N
Mask			
Nxt-hop Addr.			
Distance			

`(c)# router rip`
`(c-r)# version 2 (ver 2)`
`(c-r)# no auto-summary (no au)` 关闭网段自动汇总 (根据地址主类) may cause false forward
`(c-r)# network IP (net IP)` 宣告网段
`# show ip rip database (sh ip rip dat)`

B→C; B start AD→B; C joined → B

EIGRP (Enhanced Interior GW RP) 增强型内部网关路由 DV based + LS feature

Privately owned by cisco, fast convergence, widely used in financial institutions
Triggered update, support VLSM, EIGRP and IGRP with same AS are compatible

3 Tables 表
- Route 路由: final routing
- Topology 拓扑: for route selection
- Neighbors 邻居: Hello-pkt test
  (5s interval on multicst, else 60s; Hold time 持续时间 = 3×interval)

Reliable Transp Prot. 可靠传输协议
@224.0.0.10, retrans based on RTO, retry 16

`(c)# router eigrp AS`
`(c-r)# no au`
`(c-r)# net IP INVMASK`
`# show ip eigrp neighbors`
`# show ip eigrp topology`
`# show ip route eigrp`

Neighbor Discovery 邻居发现
1) Hello/Ack, 2) same AS, 3) same metrix → (BW, delay, reliability, load, MTU)

For R1:
- Feasible Dist. (FD) 可行距离 = 5: Total Dist from this to dest
- Ad. Dist (AD) 通告距离: NxtHop to Dest = 1
- Successor 后继站 (继任者) = Nxt Hop = R2
- Feasible Successor 可行~: Back Up = R4 (AD of FS must < FD)

Diff Update ALgo (DUAL) 弥散更新算法
- Topo. maintain all route of neigh
- Fast query if no FS

eg. For R1, R3 is not FS as AD > best FD
If R1-R2, R4-R5 break ⇒ no FS ⇒ fast query R3
SIA: Stuck in active (Hidden danger)

OSPF (Open Shortest Path First) 最短路径优先 LS based

Link State Adv. Update (LSA Update) (LSU) 链路状态通告更新 (triggered + 30 min itemwise update) In IP Pkt = OSPF Hder | Data

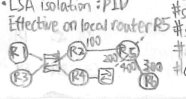
3 Tables 表
- Route 路由: 30 min period, 60 min timeout
- LSDB 链路状态数据库 (拓扑表)
- Neighbor 邻居: Hello pkt test
  (10s interval on multicst, else 40s; Hold time = 4× interval)
  - 邻居 2-WAY: Hello (among DROTHERS)
  - 邻接 FULL: Hello + LSA (with DR)

OSPF Pkt
- Hello
- DBD 数据库描述: LSDB summary
- LSR LS请求: req specific items
- LSU LS更新: send reqed items
- LSAck

Typ of link in OSPF
- P2P: 2 router direct conn, R2-5
- Transient: ≥3 router net, R1234
- Stub: Net with 1 router
- Virtual: Used when link broken

Save LSA BW (ASBR, ABR)
- Area
  LSA Typ 1,2 trans within area
  Cross-area share ABR's LSDB (区域边界路由器)
  - backbone (area 0) 骨干区域: all area must dire comm to 0
  - Standard 标准区域
  - Stub 末梢区域 × Typ5 LSA
  - Totally Stubby Area 完全末梢区域 × Typ 3,4,5 LSA
  - Not-So SA (NSSA) 非纯末梢区域 ✓ Typ 7, non-flood Typ5
  - Totally NSSA 完全非纯末梢区域 × Typ 3,4,5 ✓ Typ 7

- DR/BDR/DROTHERS
  Designated Router (DR) 指定路由器
  Backup DR (BDR) 备份~
  Collect and disseminate LSA
  router-id: 1) max local lookback IP 2) max local physical IP
  elect DR/BDR: 1) Priority (Pri) max 2) Time from start 3) router-id max
- LSA isolation: PID
  Effective on local router R5



LSA Type & ID
1. Router link: Router IP
2. Net link: IP of DR
3. Sum link to net: Net Addr (by ABR)
4. Sum link to AS boundary R: IP of ASBR (by ABR)
5. Ext link: Ext Net Addr, other AS (by ASBR)
6. Grp Member (MOSPF)
7. NSSA Ext (by ASBR in NSSA)

0	8	16	32
Ver	Typ	MsgLen	
Src Router IP			
Area ID			
Cksum		AuthTyp	
Auth			

`(c)# router ospf PID`
`(c-r)# router-id IP (rou A.B.C.D)`
`(c-r)# net IP INVMSK area N`
`# show ip ospf nei`
`# show ip ospf data`
`# show ip route ospf`
`# show ip ospf int fA/B`
`# clear ip ospf process`

- Down: no Hello
- Init: sending Hello
- 2-Way: Comm start, electing DR/BDR
- Exstart: prepare to exchange
- Exchange: DBD exchanging
- Loading: LSA LSU exchanging
- FULL: ✓

EGP: BGP (Border GW Prot 边界网关协议). Work in Application layer (TCP Port 179), DV based, Triggered Incremental Update Only (TCP reliable, 触发增量更新)

- Enterprise: Datacom among areas (OSPF, BGP)
- IDC: Routing selection, Advertise servers (Access 接入, Distribution, Core; OSPF ← → BGP)
- ISP: InterAS Routing (BGP: Cross-hop 路由选择; OSPF: lower layer comm)

very costly (CNY ¥) { AS(2+) 10k/yr/AS; Member fee (CNNIC) 10k/yr; BGP line: 100k+/mo/line }
(ASBR maintain route of entire Internet ≈ 1M)

IBGP: Within same AS; EBGP: Cross AS (maxhop=255) (maxhop=1)
3 Table: Routing Table, Neighbor Table, Topology Table (= BGP Table)
Keepalive pkt: maintain TCP session with neighbors (60s interval, 180s timeout)
Open pkt: establish neighbor (peer 对等体)
Update + ACK: Routing table (with prop. net info)
Notification: Report err
Stub AS (only 1 conn.); Multihome AS
✓ Transit AS allowed (穿越AS) for ISP
✓ Multiple links from one AS to another
✓ Traffic routing policy available
✗ Only one GW link to Internet (Def routing)
✗ Low Mem/CPU resources, Low BW between AS
✗ Lack of Understanding to BGP policies

(c)# router BGP AS (this 本AS)
(c-r)# nei A.B.C.D remote-as AS
(c-r)# net IP mask MASK
show ip bgp summary (sh ip bgp su)
sh ip bgp nei/rou
(c-r)# nei IP ebgp-multihop MAXHOP
(c-r)# nei IP update-s IP/lo A
(c-r)# nei IP next-hop-self
R1: 30.2 NxtHop R4 (→R2)
clear ip bgp * soft (BGP 软刷新)
(c-r)# sy (BGP load to routing table only when received IGP)
no sy: R4→R3, R1→R2; sy: R4→R1 ✓, R1→R4 ✗ (Mem hungry on all rou!)
Sol: MPLS Mul-Prot Label SW 多协议标签交换 (FW Pkt based on incoming port)

Anti loop:
General (Split Hor/Poison Rev)
iBGP: no entry forward from nei unless RR 路由反射器
eBGP: contain AS-PATH, reject if containing this AS

Neighbor Establish States
- Idle: Route Lookup, Timer RST
- Connect: Establish TCP (MaxHop, No Route, No TCP)
- Active: Retry TCP (Addr/AS mismatch, Wrong route, firewall)
- Opensent: Send & wait Open (Auth fail, ver mismatch, AS wrong)
- OpenConfirm: Send & wait keepalive
- Establish: ✓ Exchange route info

Routing selection
1 Weight (highest)
2 LOCAL PREF (highest) (global within AS)
3 ORIGIN (from local R) (nxt Hop = 0.0.0.0)
4 AS-PATH (shortest)
5 ORIGIN (code i < e < ? (static))
6 MED (lowest) (Mul-Exit Discriminator)
7 EBGP path > IBGP path
8 closest IGP nei (IGP cost)
9 (EBGP path:) oldest route
10 nei BGP R router-id (lowest)
11 nei BGP R IP (lowest)
8+ max-path (load balance)
10+ shorter cluster-list (RR)
RR antiloop: rej if have my RID

Routing Entries → Import Policy (accept, deny, set pref) → BGP Table (decision) → IP Rou Table; Export Policy (FW, no FW, set MEDs) → Other Routers

	This routes	learned from peers & providers
To peer & provider:	✓	✗ (unless backbone)
To Customers:	✓	✓

Multicast-Groupcast 组播: To save BW and optim processing, when 1) sending the same data to multiple dest, or 2) receiver unicast addr unknown. ⚠: UDP, unreliable

(224~239).X.X.X Class D Groupcast
224.0.0.(1~255) Link Local 本地链路 (effective only on local link, one hop)
(224~231, 233~238).X.X.X 公有任意源组播 Public, ASM (any-source multicast) (*,G)
232.X.X.X 公有指定源组播 Public, SSM (specific-src multicast) (S,G)
239.X.X.X private groupcast

224.0.0.1 All devices
224.0.0.2 All routers
~.4 All DVMRP routers
~.5/6 All OSPF routers
~.9 All RIPv2 routers
~.10 All EIGRP routers
~.13 All PIMv2 routers
L2 Switch MAC (Last 23 bit of grpcst IP): 0100.5E00.0000 (1 MAC ⇒ 2^5 = 32 grpcst IP)

Enforce URPF (Unicast Rev Path Forwarding)
Server Originally = Anti IP-spoof; 单播逆向路径转发 AntiRing 防环
Auto enforced, check if grpcst pkt (src=S) come from correct port by unicst routing table, multicst route static, MBGP (Mulcst Addr Family)
1. running mroute 2. nearest (cost) 3. LPM (Longer IP) 4. larger IP
Only one RPF port/nei for each Grpcst src even having mul candidates
RPF by unicst fail; Tunnel cannot dyn adver (route jitter)
Sol: RPF by mulcst static

PIM, MOSPF (no prot), IGMP, DVMRT, CBT
v1 / v2 + leave msg; v3: allow src selection within grpcst IP (SSM)
- Last Hop Router / Leaf Router 最后一跳路由器/叶子路由器
- First Hop Router 第一跳路由器
- 1st Hop-leaf: 组播分发树 (Distribution Tree)
- No load-balance in grpcst, select only best route (SPF route) (SPT) 最短路径树
- Routing among AS: MBGP, MSDP

Multicast 多播	Broadcast 广播	Groupcast 组播
Router Transit?	✗ LAN only	✓ (IP Mroute)
Process Decision?	✗ (rely on upper)	✓ (Layer 3)
Anti Ring?	✗ (rely on STP)	✓ (RPF)

IGMP (work between leaf R and grp mem)
Host: Mem report (dst = grp IP); leave msg (dst = 224.0.0.2)
Router: Query (general & specific)
General: every 60s periodic, all grp (0.0.0.0)
Host mem report suppress enable (first join: 1~2 times no suppress)
dst = 224.0.0.1 (mul-GW election) (elect by smallest IP, 120s timeout)
max resp time = 10s (grp mem rand 0~10s) (first query 0.25 × max resp time)
no max resp time in IGMPv1 (10s timeout)
Specific: only when receive leave msg, specific grp, max resp time = 1s

CGMP (Cisco Grp Mgmt Prot) (Between R and SW)
Original: SW MAC CAM record src only, and broadcast when grp mem send mem report
Problem: incoming grpcst always broadcst
Sol: CGMP: record both src port and grp from router (grpcst MAC table)
RGMP ≈ CGMP (but is public prot.)
IGMP snooping: CGMP without R (require L3 SW)
GMRP (GARP General Attr Reg Prot Mulcst Reg Prot) 组播注册协议: between SW and grp mem

Routing & Dist. Tree Model
- Dense Mode 密集模式 (DM) DVMRP, MOSPF, PIM-DM
Active push, implicit join, Routing flood + prune per 3 min (泛洪+剪枝)
Based on Src Tree (Shortest path 树)
Tree SPT 最短路径树 Indi shortest
✓ Min transmission delay, suitable when most of hosts are grp mem
✗ Every grpcst router need many memory to maintain mroute as num of src ↑
- Sparse Mode 稀疏模式 (SM) CBT, PIM-SM
Passive Push, explicit join
Leaf R "Pull" from meeting point (Rendezvous Point RP 汇合点) or from src
Based on SPT + RP + RPT (Grp shared Tree 组共享树 Min cost Tree 最小开销树 (overall))
Only based on (*,G), no (S,G)
✓ Less res required to store mroute, suitable when only some hosts are grp mem
✗ Individual suboptimal path 次优路径
{(*,G) In = null, RPF nei = 0.0.0.0, Out = xxx; (S,G) In = RPF, RPF nei = xxxx, Out = xxx}

DM: (*,G) in null; (S,G) in ... out ...
- Mroute table DM specific rule
- Out (OIL) of (*,G) maps 1) existing PIM-DM nei 2) direct conn to grp mem
- OIL of (S,G) marked "Prune/Dense" when pruned
- OIL reset "Forward/Dense" when new nei added
- In PIM, (*,G) created when join, (*,G)(S,G) created when pkt transit, no mroute exchange
- Mroute table SM specific rule
- (*,G) created when explicit join, (S,G) created when 1) receive (S,G) join/prune, 2) prev hop change to SPT, 3) unexpected (S,G), 4) RP receive reg msg
- OIL add when 1) join msg, 2) direct conn grp mem
- OIL delete: intf without direct conn receive prune
- Intf timer reset 3min: 1) join, 2) IGMP mem report
- Send (S,G) RP prune msg to Grp Shared Tree if RPF nei of (S,G) differ from (*,G)
- RPF of (S,G) is determined by src IP, unless when setting RP (use IP of RP)
- Mroute Table General Rule
- (*,G) is (S,G)'s parent and must exist
- RPF and In intf refresh every 5s
- OIL of (S,G) always sync from (*,G)
- In intf: (S,G) = Src, (*,G) = RP, (*,G) of RP or in DM = null
- In intf of any entry cannot appear in OIL

SM: Static RP, Auto RP, BSR (bootstrap RP); DM SPT : SM RPT; (*,G) in null, out 1; 0 2; 0 1

IPv6
Limitation of IPv4
- Address shortage (32 bit only + improper addr. allocation by IANA)
Asia Pacific: run out on 2011/1/31
S. America: 2014/6/10, N. Ame 2015/9/24
Africa: 2017/4/21, Europe 2019/11/25
China: ≈36BM Public IPv4 in 2023
- Complex IPv4 pkt structure (12 field opt)
- Complex Host & router config if no DHCP
- Don't allow dual IP (⇒ jitter)
- Non-native IPSec, Multicast, Mobile IP
- Bad traceability (1 Pub - N Pri NAT)

IPv6 tech stack
128 bit addr format
Routing: RIPng, EIGRPv6, OSPFv3, BGP, static
IPv6 to IPv4 traffic: NAT-P, ARP ⇒ NDP (nei disco)

Ver	Traffic Cls	Flow Label	
Payload Len		Next Head	Hop Limit
Src Addr (16B)			
Dest Addr (16B)			

Advantage
Global readability & flexibility
End to end without NAT
Better net aggregation / summary
Multihome (One IP ⇒ mult device)
Easier renumbering (mul-IP)
Auto config, plug and play
Simplified head (40B fixed size)
Easier QoS (ToS ⇒ Traffic Class)
No fragment, no checksum
No broadcast (⇒ any/grpcst)
(✗) IHL, ID, F, FragOffset, Cksum, Opt

IPv6 Addr (X:X::X to omit cont. 0)
X:X:X:X:X:X:X:X / Port / mask (64)
- Unicast: 2000::/3 (≈ IPv4 Public)
2001::/16 Internet, 2002::/16 6to4 tunnel
FE80::/10 Link Local (+64 b EUI-64 Intf ID) MAC AABB : CCFF:FE ... editable, work as nxthop in routing
FEC0::/10 Site Local (≈ IPv4 Private IP) revoked, replaced by FC00::/7 (Uniq Local)
::/0 Unspecified Addr (eg. DHCP)
::/128 Loopback Addr
- Anycast 任意播 (eg. dist DNS) when a unicast addr assigned to mul intf, for load balance, transparency, anti DDOS

Typ	Prov	Subscr	Subnet	Node

- Grpcast: FF00::/8
0 = Perm (Pub); 1 = Temp (Pri); Scope: 1 = Intf-Local, 2 = Link Local, 3 = Realm Local (Subnet), 4 = Admin Local, 5 = Site Local, 8 = Org Local, E = Global
FF02::1 All device
~::2 All routers
~::5 All OSPFv3 R
~::6 All OSPF v3 DR
~::9 All RIPng R
~::A All EIGRP R
~::16 All PIM R
FF02::1:FF xx:xxxx (last 24b of intf unicst IP)
Nei Discovery Prot (NDP); Dup Addr Detect (DAD)

Mobile IP
Retain IP (keep connection) and timely update routing, L3 prot. (can't purely rely on routing!)
Home Net, Home Agent, Perm Addr; Correspondent; Foreign Agent, Visited Net; Care-of Addr (CoA); indirect; direct
RFC3220 MobileIP: Indirect routing, agent discovery, home agent reg.
Mobility Agents { Home Agent ⇒ mobility binding table (Home, Care-of, lifetime); Foreign Agent: Visitor List (Home, Home Agent, MAC, lifetime) }
Agent Discovery: Determine home/foreign net by Agent Ad (AA) from mobility agents; Detect net move by 1) AA lifetime 2) AA net ID
Registration: HA:H, FA (CoA) F, MA (IP under HA, Perm) = M
Tunnel encap: CoA = F, HA = H, MA = M, Lifetime = t_1, ID
In-Service (✓) (HA → FA: Encap Tunneling)
Tunnel encap: HA = H, MA = M, Lifetime = t_2, ID

Transportation Layer 传输层 TCP, UDP Socket (套接字) = IP:Port (0~1023 well-known, 1024~49151 ICANN, 49152~65535 Dynamic)

UDP: Hder | Data
Src Port	Dest Port
Total len	Total Cksum
(4B, 8B)

TCP: Hder | Data (Hder 20~60B)
Src Port	Dest Port
Seq num	
Ack num	
Hlen / 000 / flags	Window Size (B)
Tot Cksum	Urgent Ptr
Options and Padding	

NS - ECN; CWR (Cong Wd Reduce); ECE (ECN-Echo); URG, ACK, PSH, RST, SYN, FIN; NS-ECN (Explicit Congestion Noti); Reset, Finish
HLEN (4b): ×4B
Timestamp; Max Segment Size; Window Scale Factor (2^B)

CLOSED → (Active open) SYN → SYN-SENT; LISTEN (Passive open); SYN-RCVD; SYN+ACK; ACK; ESTABLISHED; [Data]; RcvBuffer; BufferedData; FIN; ACK; FIN-WAIT-1; FIN-WAIT-2; CLOSE-WAIT (half-close); LAST-ACK; TIME-WAIT; CLOSED

Return to CLOSED direct: Timeout or RST
CLOSING state (seldom): Send and rcvd FIN, wait ACK (simultaneous close)
All SYN, FIN carry 1 seqNo
ACK and error control
✓: Ack throttled 500ms (or 2seg)
Out-of-order or gap: dup ACK, Seq = Nxt expected
(Partial) gap fill: imme ACK
3-dup ACK: retransmit

Congestion Control (rcv (PURAM)) (net BW)
sender window Size = min(Rwnd, Cwnd)
Slow start (e^x) → Conge Avoid (+); Retrans timeout; Fast Retrans (ssthresh = cwnd/2, cwnd = ssthresh + 3 dup ack)
Cwnd: ssthresh, 8 (reno), (tahoe); 1 2 3 4 5 6 7 8 9 10 11

Retrans timeout (RTO)
- Karn/Partidge: no measure RTT (RTT_M) $RTO^* = 2$ for each retrans (expo backoff)
- Jacobson/Karel Algo: $RTT_S(1) = RTT_M$, $RTT_D(1) = 0.5 RTT_M$
$RTO = RTT_S + 4RTT_D$
$RTT_S(n) = (1-\alpha)RTT_S(n-1) + \alpha RTT_M$
$RTT_D(n) = (1-\beta)RTT_D(n-1) + \beta|RTT_S(n) - RTT_M|$

Latency
Cwnd · S/R ≥ RTT + S/R; continuous RTT; 1.5RTT + F/R + 1.5RTT; batch-by-batch; $\sum (RTT + S/R) + \ldots + ( )$

Applications (Server-Client, Peer-to-Peer)
Domain Name System (DNS 域名系统) UDP:53
Server-client distributed DB offering domain-IP map
DNS Servers
Root domain Server (13) 根域名服务器 backbone, All top-level domain & IP (a~m).root-servers.net
TLD Server 顶级域名服务器
All 2LD & IP or DNS IP
Authoritative DN Server 权威域名服务器
Zone (E) under a DN
Hierarchical Tree: Root; .com .cn .org; 163 cctv mail; Top level domain (TLD), 2nd ... Leaf
Resolver; Local DNS
Query
- iterative
- recursive

FTP (File Transfer Prot) / TFTP
interactive file access supporting heterogeneous (异构) arch, TCP
Main Process (new requests) + 1 ctrl, 20 data subprocess (individual reqs)
Can only access whole file ⚠
Sol: NFS (Net File Sys) allow partial file access
TFTP (Trivial FTP) UDP:69 unreliable non-intractive file trans
USER, PASS, ACCT (info), REIN (git), QUIT, ABOR (undo); RNFR, RNTO; CWD (cd), CDUP (../); LIST, NLIST (ignore attr), MKD/RMD; SMNT (mount), DELE, PWD (./)

TELNET 远程终端协议
TCP:23 (SSH:22) 终端仿真协议
Client: Format I/O & key stroke transparent (NVT)
Data: ascii | Control: ascii

WWW, HTTP (TCP 80, trans-oriented, stateless in HTTP/1.0, stateful from HTTP/1.1)
SYN, SYN+ACK, ACK + HTTP req, HTTP resp
req: Hder | Body — Method | URL | Ver | CRLF; field : value CRLF; CRLF
GET/PUT/POST/DELETE, HEAD/TRACE/OPTIONS, CONNECT (rsvd)
User-Agent; Accept { charset, encoding, language }; Host; Authorization; Date; If-modified-Since (cond GET); Upgrade; Cookie
Status: Hder | Body — Ver | Status | Phrase
1xx: Noti; 2xx: ✓; 3xx: Redirect; 4xx: Client ✗; 5xx: Server ✗
400 Bad Req; 401 Unauth; 403 Forbid; 404 Not Found; 405 Method ✗
500 Internal server error; 501 Not Implemented; 502 Bad GW; 503 Service Unavail

Email (IMAP 143 / POP3 110/995, SMTP 25/587, MIME)
SMTP → SMTP(S) → SMTP(C); SMTP TCP; POP3/IMAP → POP3/IMAP/WWW
username@domain.com
Multipurpose Internet Mail Extension (MIME)
Content Type: Text/(Plain, HTML); Multipart/ (Mixed (ordered), Parallel (no order), Digest/Alternative); Message (报文)/RFC822; Image/(JPEG, GIF); Video/MPEG; Audio/Basic; Application/
Encoding Type: 7-bit (ASCII, 100%, <1000 char per line); 8-bit (ASCII + special latin, 100%, <1000 char per line); Binary (Raw Bytes); Base64 (A-Za-z0-9+/, 128); Quoted-Printable (most ASCII), nonascii: =XX (UTF-8)

P2P - Distributed Hash Table (DHT, Chord Algo)
{ Resource: Key KID; Node: NID } scale 2^m
Finger Table 指针表/路由表/查找路由表
eg N2: N2+1 N3; N2+2 N7; N2+4 N7