IP Telephony and SIP – IP convergence for integrated voice, video and data networks

Henning Schulzrinne
Dept. of Computer Science
Columbia University
New York, New York
(sip:)schulzrinne@cs.columbia.edu

Worldbank IT Department, Washington, DC

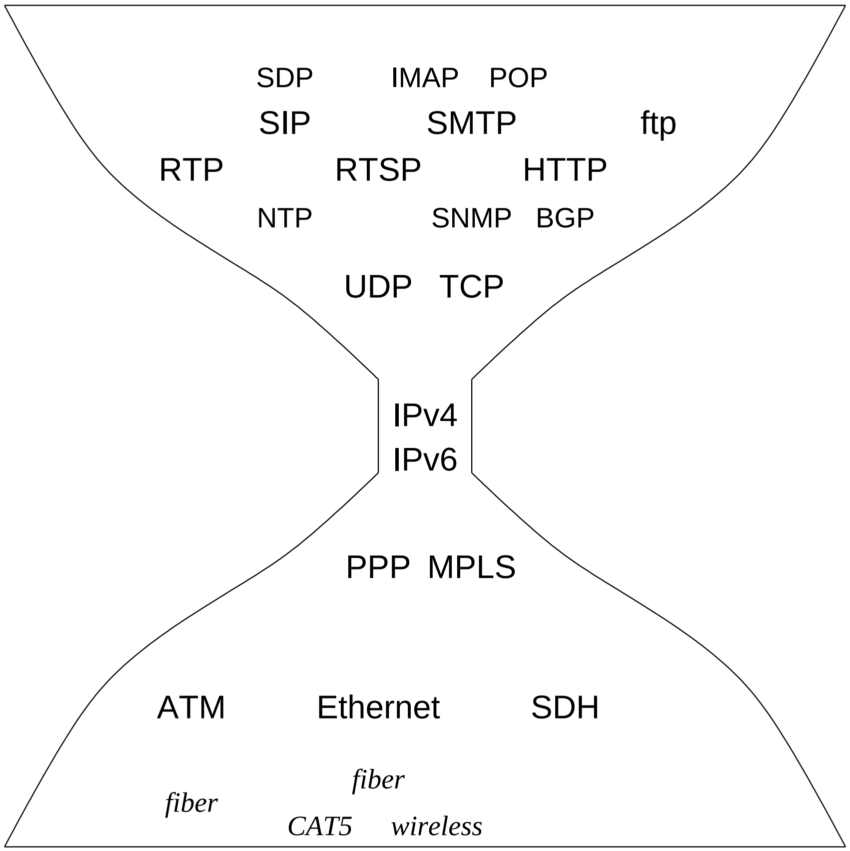
May 14th, 2001

Overview

- Motivation for integration
- Difficulties: security, QOS, reliability
- SIP

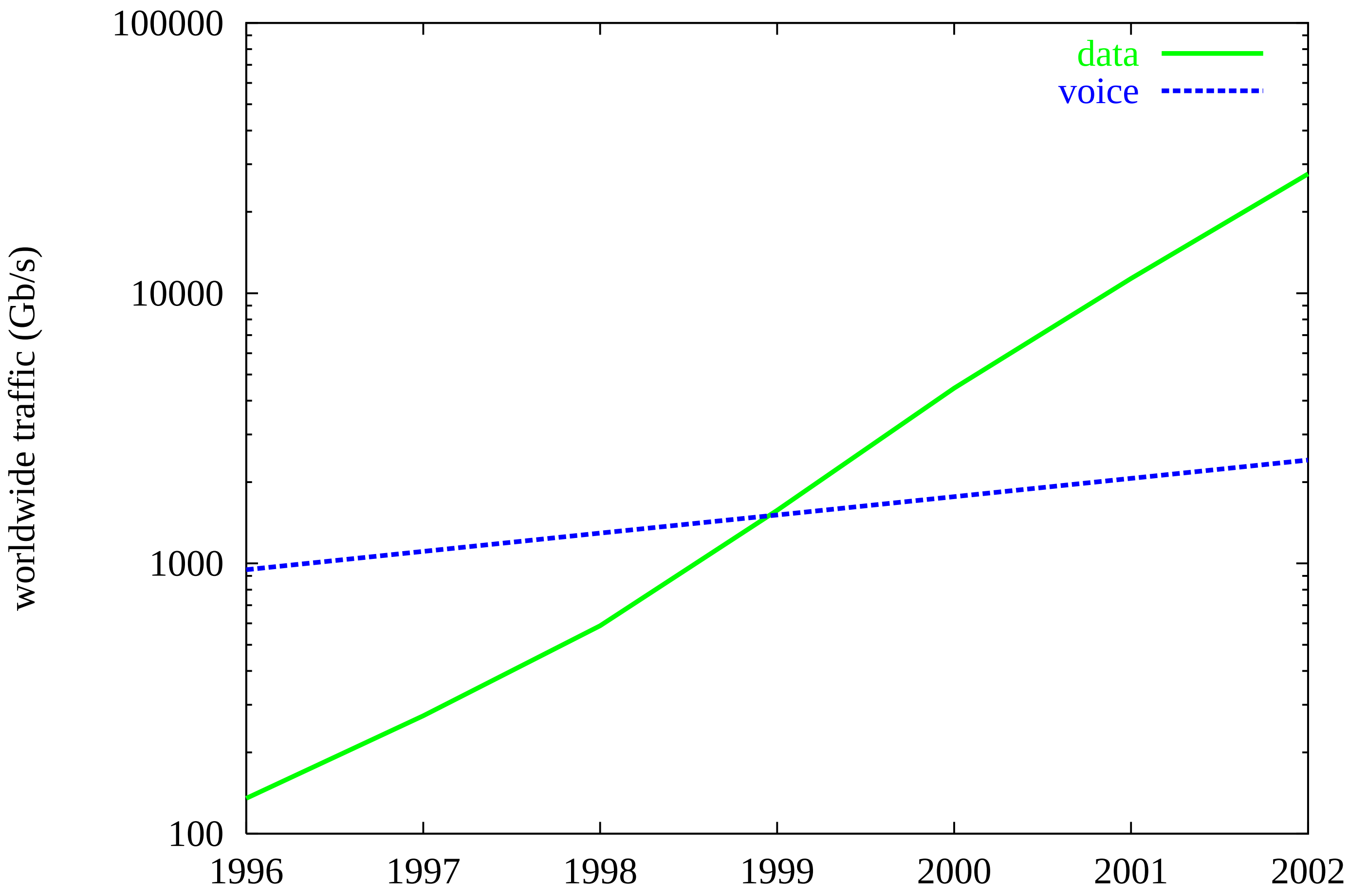
Integrated networks

Hourglass model:



Typically, same wiring infrastructure.

Voice and data traffic



The phone works — why bother with VoIP?

user perspective	carrier perspective

user perspective

- variable compression: tin can to broadcast quality
- security through encryption
- caller, talker identification
- better user interface
- internat. calls: TAT transatlantic cable = \$0.03/hr
- local calls: possibly cheaper (local access fees)
- easy: video, whiteboard, . . .

carrier perspective

- silence suppression ➠ traffic ↓
- shared facilities ➠ management, redundancy
- advanced services (simpler than AIN and CTI)
- separate fax, data, voice
- cheaper switching
- better management platforms

Bandwidth advantages

- cost($2B$) $<$ 2 cost(B)
- B(voice + data) $<$ B(voice) + B(data)

Pricing

- initial motivation for Internet telephony was transport pricing
- however, *services* (like CLID, *69) have margins of 75%
- Ameritech: service revenue of $1b/yr
- transport: leased local wire, everything else flat or volume-based (service-independent)

Internet telephony as PBX replacement

global Internet not quite ready ➠ try as PBX

- have mission-critical LAN, PCs anyway
- usually ample (if switched) bandwidth, low latency
- packet switching is cheaper
- network PCs $\stackrel{\$}{=}$ ISDN phones
- no need for billing
- new services easy to build:
  - voice mail $\longrightarrow$ email
  - calendar integration: *Mr. Jones is in a meeting. Please call back at 3:30 pm*
  - logic: if insurance agent calls, forward call to dial-a-joke

Traffic (1998)

Measured in Dial Equipment Minutes (DEM) or bandwidth:

	GDEM	bandwidth (Gb/s)
Local	2986	364
Intrastate toll	422	51
Interstate toll	555	68

PBX: typically, about 10% utilization per phone ➠ 6.4 kb/s per employee (128 Mb/s for 20,000 person campus)

Switching costs

switching method	ports	capacity (Gb/s)	cents/64 kb/s	$/interface
10/100BaseT Ethernet hub	24	2.40	0.6	10.00
100BaseTX Ethernet switch	24	2.40	0.9	14.60
PBX	256	0.02	218.	140
Lucent 5ESS local (no AIN)	5,000	0.32	469.	300
Lucent 5ESS local (AIN)	20,000	1.28	273.	175
Lucent 4ESS toll	100,000	6.40	7.8	

Internet telephony problems

- reliability
  - power
  - denial-of-service
- QoS
  - delay
  - local area network
  - access network
  - wide-area network
- architecture
- address space
  - NAT
  - IPv6

Why aren't we junking switches right now?

What made other services successful?

email: available within self-contained community (CS, EE)

web: initially used for local information

IM: instantly available for all of AOL

All of these . . .

- work with bare-bones connectivity ($\geq$ 14.4 kb/s)
- had few problems with firewalls and NATs
- don't require a reliable network

Why aren’t we junking switches right now?

Telephone services are different:

- reliability expectation 99.9% $\nearrow$ 99.999%
- PC not well suited for making/receiving calls – most residential handsets are cordless or mobile
- business sets: price incentive minor for non-800 businesses
- services, multimedia limited by PSTN interconnection
- initial incentive of access charge bypass fading (0.5c/min.)
- international calls only outside Western Europe and U.S.

Reliability

- phone switch: downtime 120 seconds/year
- AOL: 88 hours/year for 1996
- ANS: 44 hours/year promised
- Ethernet switch: MTBF $\approx$ 20 years
- router configuration, route flap
- on-line software upgrades
- end-system auto-configuration (already easier than ISDN. . .)

Reliability: power

- more decentralized ➠ harder to provide power coverage
- need power for Ethernet switches, phones – ≈ 7W/phone (48V)
- Ethernet powering (spare pairs), tandem or integrated into switch
- also useful for wireless base stations
- Columbia approach: separate power circuit for wiring closets

Reliability: denial-of-service

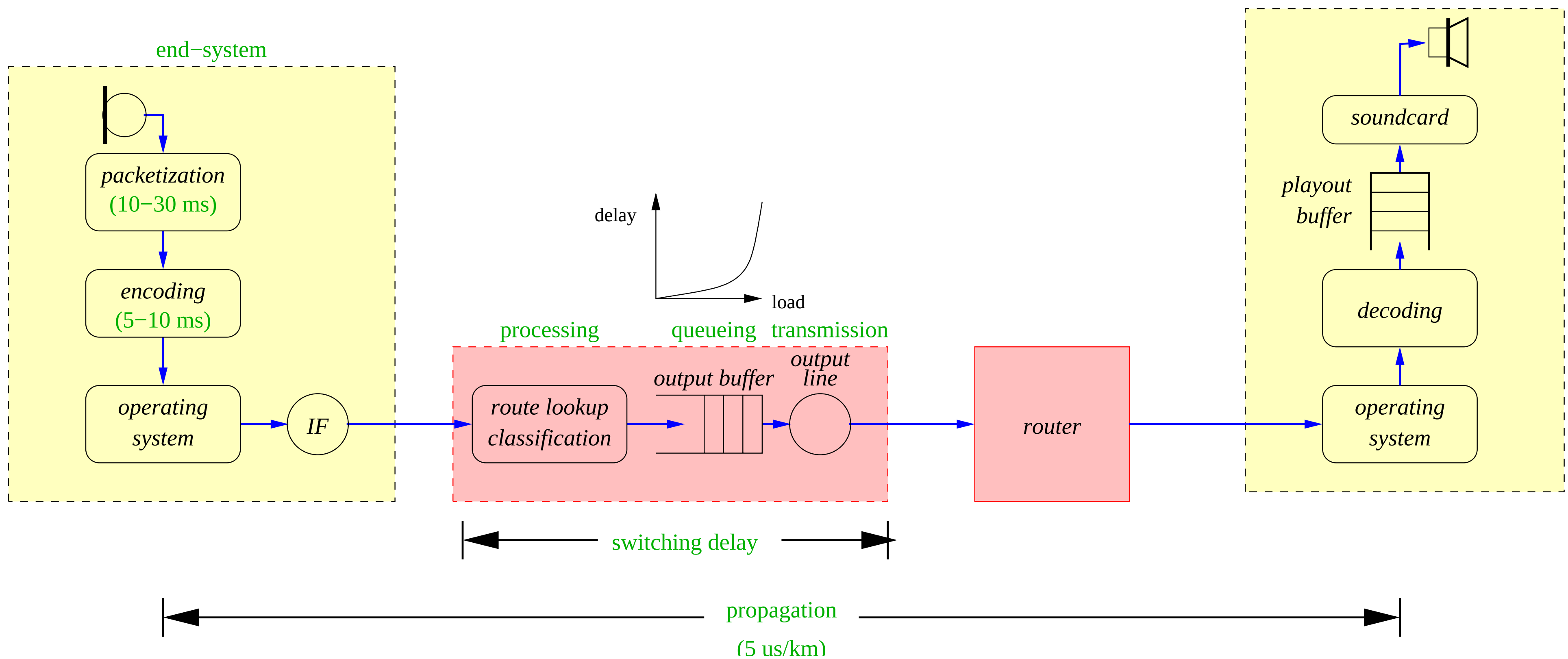
- denial-of-service and attacks more likely than with traditional phones
- but traditional phones (including 800#) also subject to auto-dialers
- different scenarios:
  - external attack ➠ can be filtered
  - internal compromise ➠ spoof DiffServ, RSVP
- disadvantage of integration: no secondary channel
- thus, maybe keep authorized RSVP “circuits”

Quality of service issues

Three types of traffic:

	loss	delay	bandwidth
(Web) data	$<$5% (bursts ok)	not critical	peak
Streaming audio/video	$<$5% (random)	not critical	avg.
Voice-over-IP	$<$10% (random)	$<$ 150 ms	avg.

Sources of delay

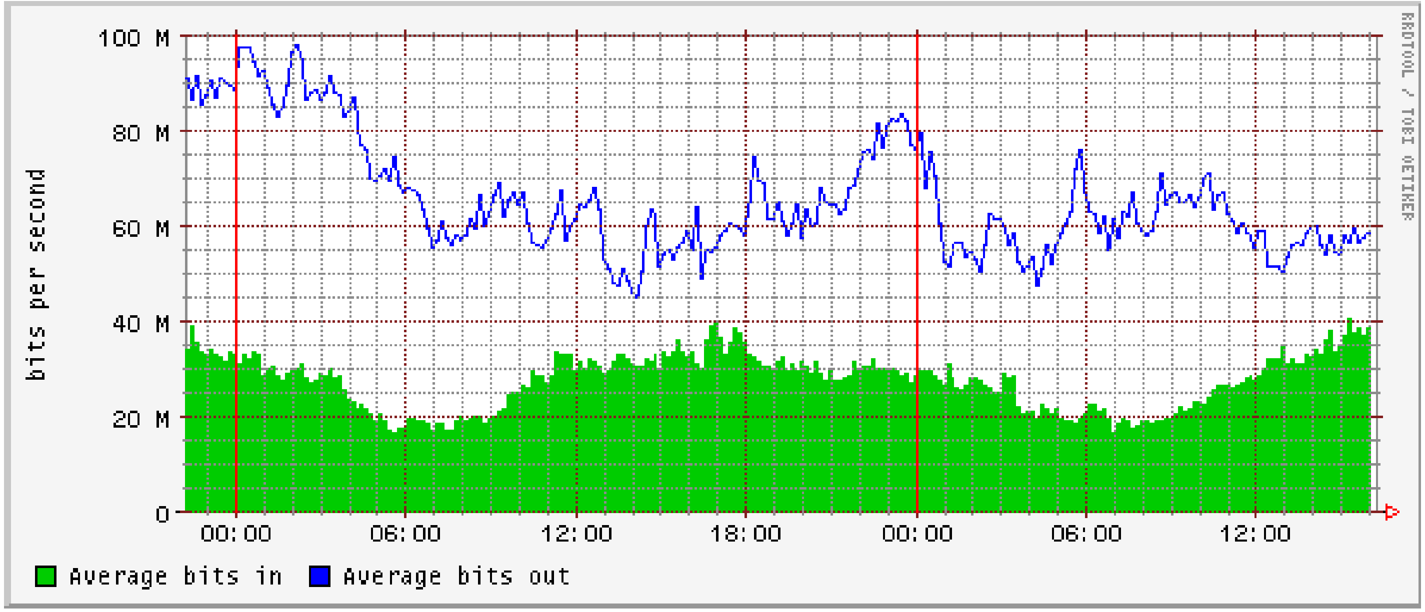


QoS: local area network

- typically, very low average utilization (few %)
- very little packet loss (a few packets a day)
- but long delay spikes (300 ms) due to Ethernet collisions if heavy file transfer
- ➠ avoid hubs
- ➠ Ethernet prioritization

QoS: access network

- usually, bottleneck (1:10 concentration)
- usually, asymmetrically loaded, depending on web traffic
- solution: TOS marking (supported by most phones)



QoS: wide-area network

- on *average*, enough bandwidth to most places
- however, bursts of loss ➠ interruptions
- likely solutions:

  DiffServ: works well for small number of predictable classes

  IntServ: (RSVP) ➠ interdomain difficult, security

  MPLS: only single provider, additional complexity

Architecture

- “classical” applications (web, file servers, SMTP): client-server
- client inside network, server often outside
- VoIP: every phone is a server
- classical applications: mostly single stream (except ftp)
- VoIP: control + data

Address space

- about half the IPv4 address space is allocated
- ARIN hands out additional space if existing allocation is more than 80% used
- minimum of 25% initial usage and 50% within a year
- minimum for direct assignment is /20
- cost is about $0.30/year/host

Network address translation

- commonly used for DSL, possibly multiple stages
- but: work well only for client-server, with server on fixed IP address
- need application-layer-gateway for each new service (or constrain new services) ➠ break service neutrality of Internet
- *not* a security mechanism
- makes VoIP deployment brittle
- makes network-merging difficult

Regulatory issues

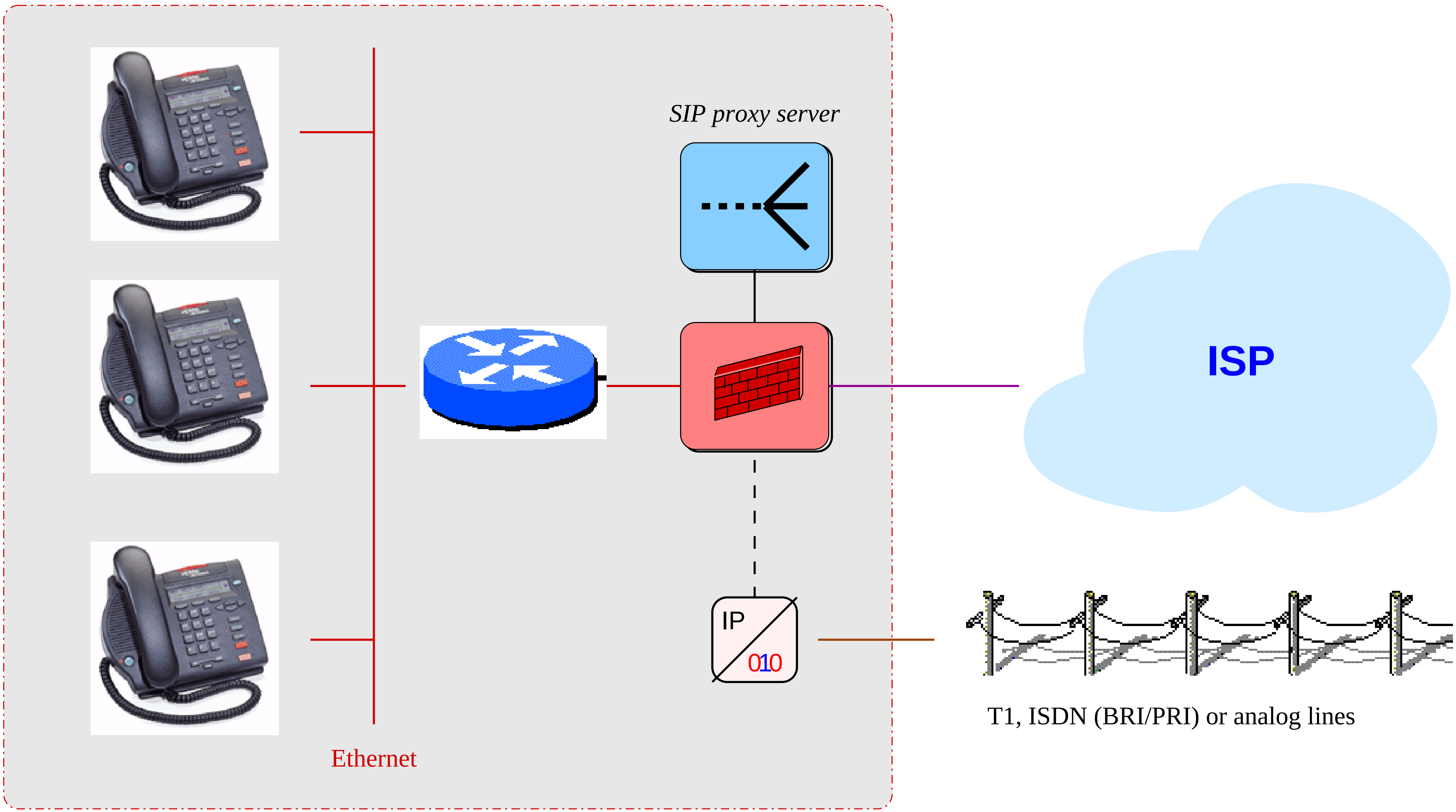
- E911 service: where is the IP address located?
- billing long-distance vs. local service ➠ local service has to be self-supporting
- universal service = support for rural areas?
- infrastructure support fund?
- distinction of TV vs. telephone licensing and regulation?

Internet telephony service models

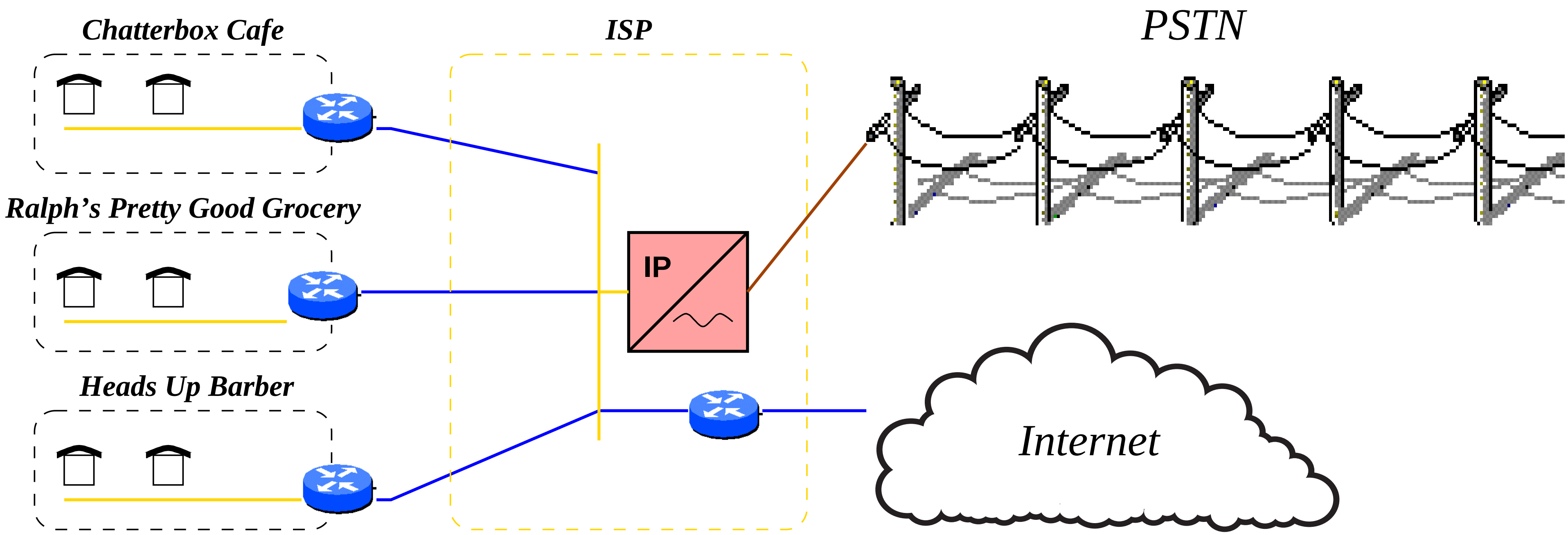
- Internet “PBX”
- Internet Centrex
- Internet Carrier

➠ same basic equipment, but size of gateway varies

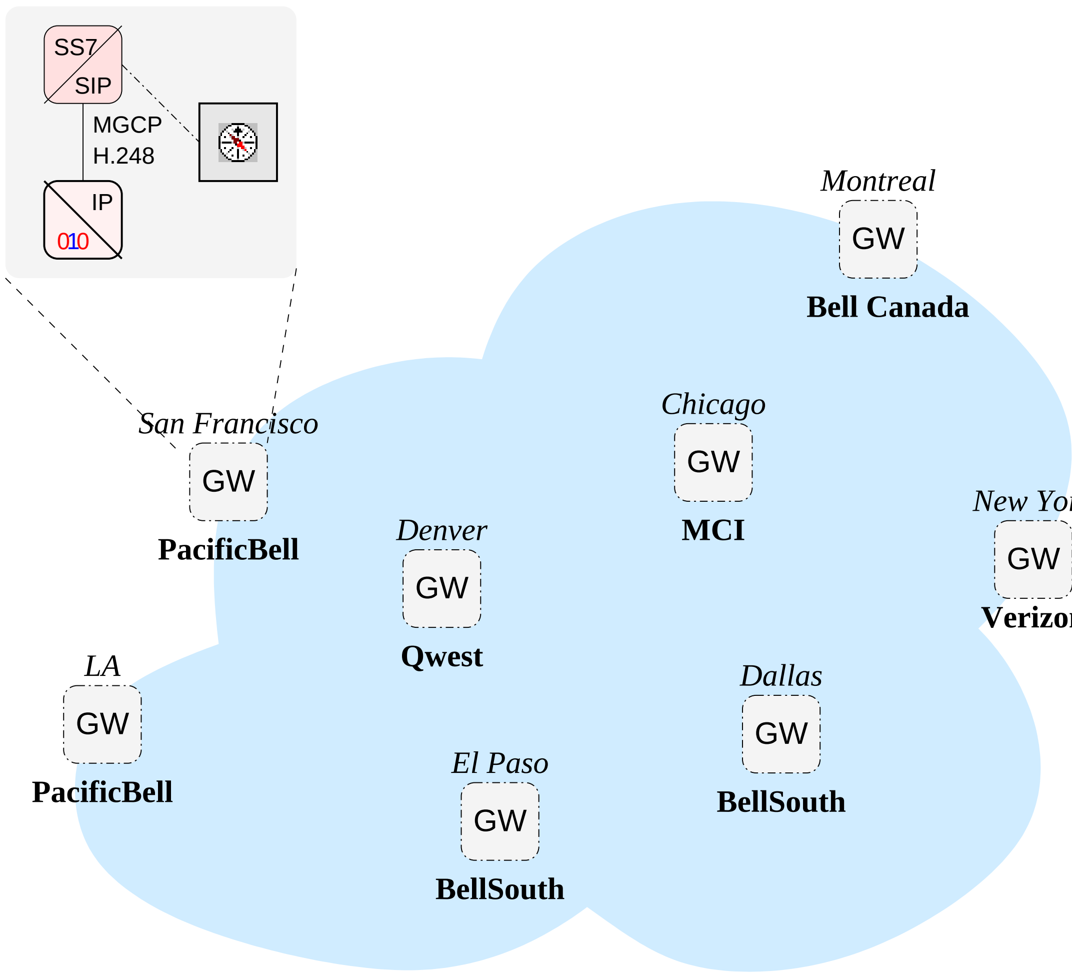
Internet PBX



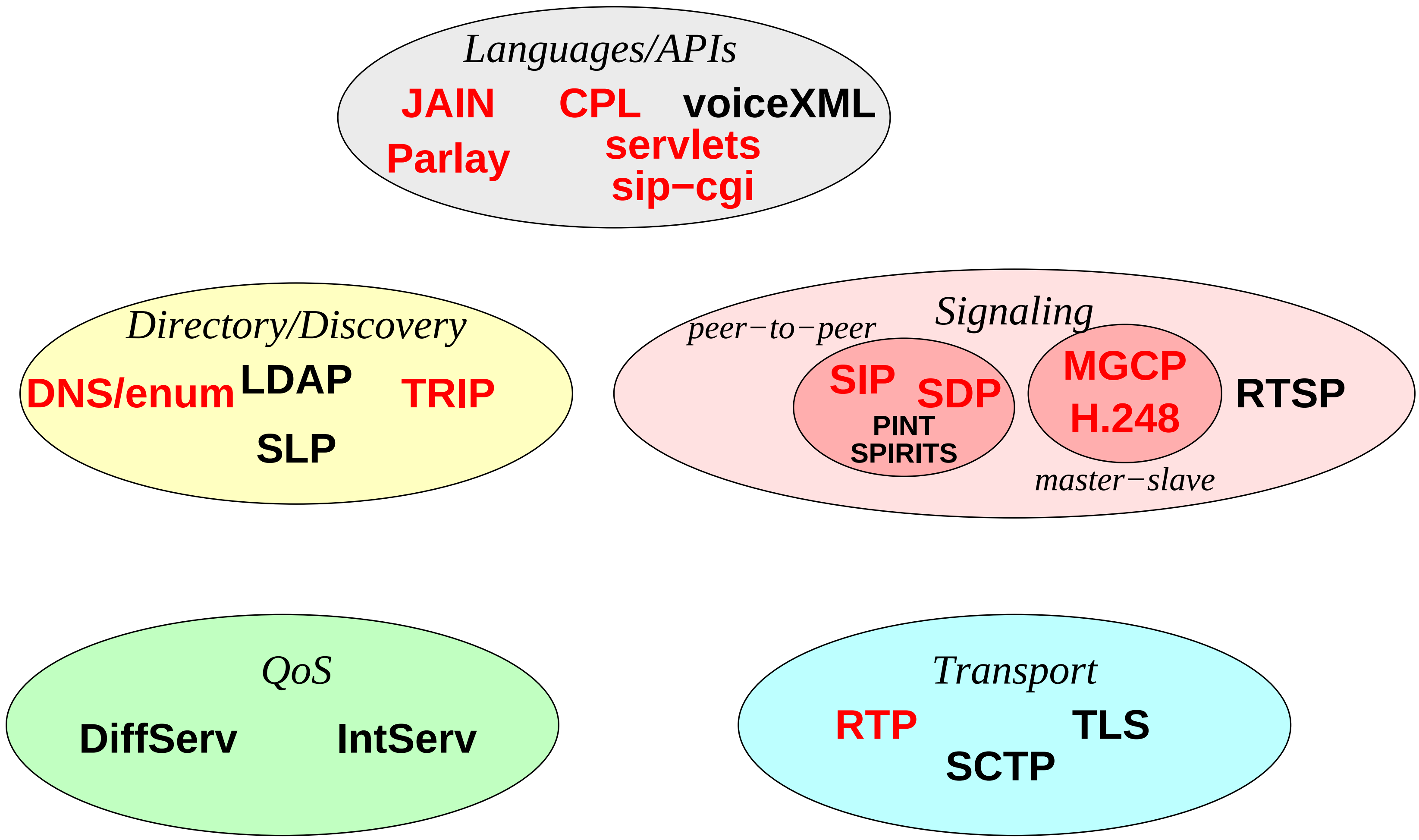
IP Centrex


Chatterbox Cafe
ISP
PSTN
Ralph's Pretty Good Grocery
IP
Heads Up Barber
Internet

IP Carrier


SS7
SIP
MGCP
H.248
IP
010
Montreal
GW
Bell Canada
Chicago
GW
MCI
San Francisco
GW
PacificBell
Denver
GW
Qwest
New York
GW
Verizon
LA
GW
PacificBell
Dallas
GW
BellSouth
El Paso
GW
BellSouth

IETF VoIP Protocol Architecture



IETF VoIP Protocol Architecture: Goals

- Leverage content-neutrality of Internet ➠ more than just voice and legacy services ➠ video, shared applications, multi-party text chat
- Imperceptible transition between communication modalities
- Extensible to presence, instant messaging and event notification
- Centrex lesson: user-controlled services
- Allow services in end systems and network servers
- Multiple levels of security: IPSec, TLS, application-layer

Differences: Internet Telephony ↔ POTS

- separate control, transport (UDP) ➠ no triangle routing
- separate connectivity from resource availability
- separate services from bit transport
- datagram service ➠ less bootstrapping
- in-band signaling ➠ higher speed
- features network $\longrightarrow$ end system: distinctive ringing, caller id, speed dialing, number translation, ... ➠ scaling
- features: intra-PBX = inter-LATA and general
- protocols: user-network = network-network signaling

PSTN Legacies to Avoid

- E.164 numbers – might as well wear bar codes
- tones and announcements
- in-band signaling for features (DTMF)
- systems with user-interface knowledge (12 keys, voice)
- voice-only orientation (BICC, MGCP/Megaco)
- integration of bit transport and services
- service-specific billing ➠ separate signaling & billing
- trusted networks without crypto

➠ confine PSTN knowledge to edge of network

Invisible Internet Telephony

“VoIP” technology will appear in

- Internet appliances
- home security cameras, web cams
- 3G mobile terminals
- fire alarms and building sensors
- chat/IM tools
- interactive multiplayer games
- 3D worlds: proximity triggers call

Carrier and Enterprise VoIP

Traditionally,

- separate signaling: ISDN, CAS vs. ISUP
- service restrictions, e.g., CF inefficient

Now, largely the same:

- hosted ("ASP"), run own servers or combinations
- carrier: multiple domains per server
- if not outsourced, TRIP for gateway selection

Peer-to-Peer Architecture

- "IP telephones", gateways, PCs with software = IP hosts
- *may* use servers (H.323 gatekeepers, SIP proxy servers)
- end system fully state-aware
- protocols for call setup: H.323 or SIP
- more flexible user interface

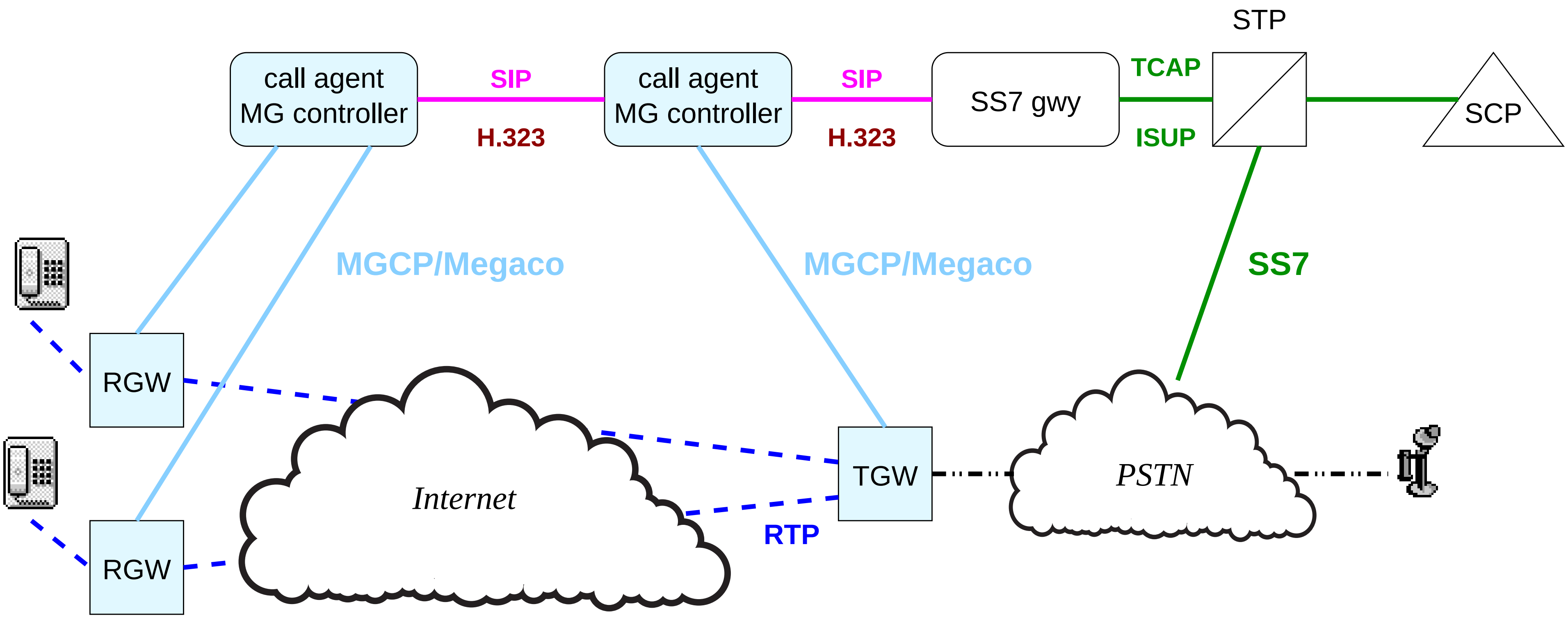
Implementing Services

	end system	server
caller id	x	–
call forwarding, follow me	x	x
three-way calling	x	–
distinctive ringing	x	–
69	x	?
no solicitation	x	x
do not disturb	x	x
call curfew	?	x

Master-Slave Architecture

- master-slave: MGC controls one or more gateways
- allows splitting of signaling and media functionality
- “please send audio from circuit 42 to 10.1.2.3”
- uses MGCP (implemented) or Megaco/H.248 (standardized, but just beginning to be implemented)
- gateway can be residential
- basis of PacketCable NCS (network control system) architecture
- service creation similar to digital PBX or switch
- $\longrightarrow$ can charge for caller id, call waiting

MGCP Architecture



- for all but small system, need peer-to-peer!
- MGCP system can call SIP or H.323 end system
- all use RTP to transfer data

Deployment scenarios

Inside-out: IP as transmission medium ➠ transport between switches

Outside-in: IP in corporate networks, with circuit-switched access to PSTN

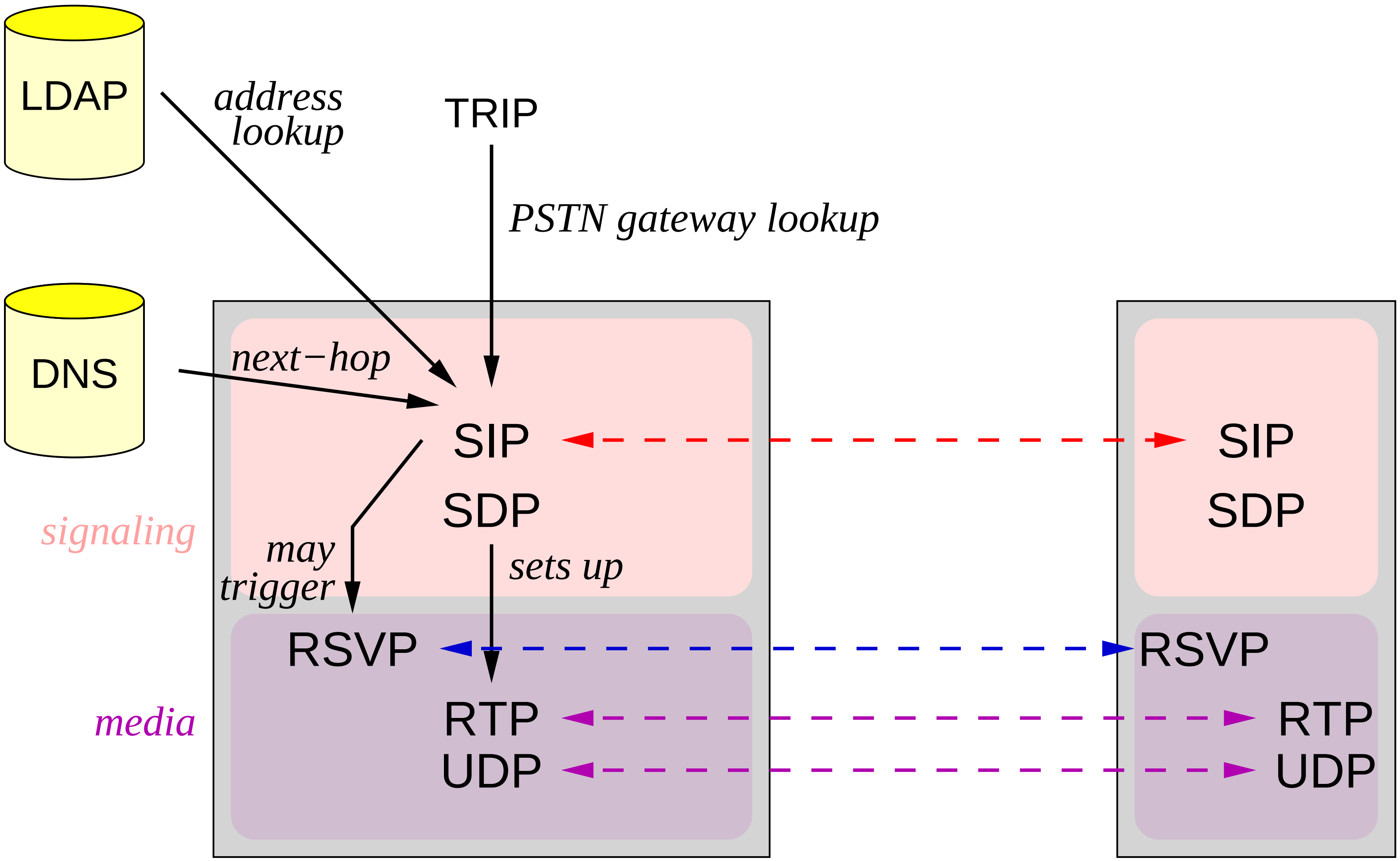
Wild card: 3G wireless

SIP 101

- SIP = signaling protocol for establishing sessions/calls/conferences/...
- session = audio, video, game, chat, ... – described by SDP carried in SIP message

1. called server may map name to *user@host*
2. callee accepts, rejects, forward (→ new address)
3. if new address, go to step 1
4. if accept, caller confirms
5. ... conversation ...
6. caller or callee sends BYE

SIP stack

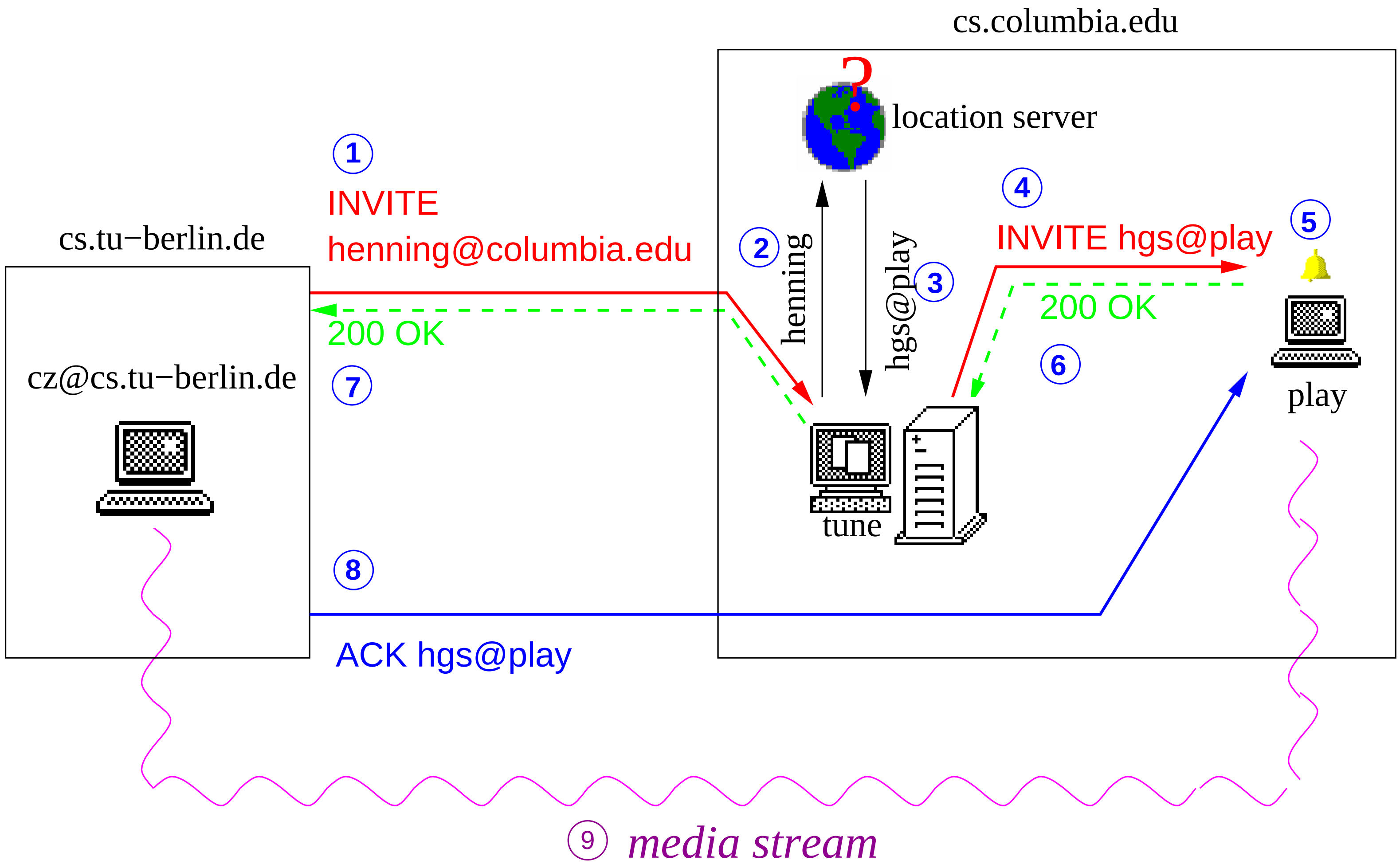

LDAP
address lookup
TRIP
PSTN gateway lookup
DNS
next−hop
SIP
SDP
SIP
SDP
signaling
may trigger
sets up
RSVP
RSVP
media
RTP
UDP
RTP
UDP

SIP Components

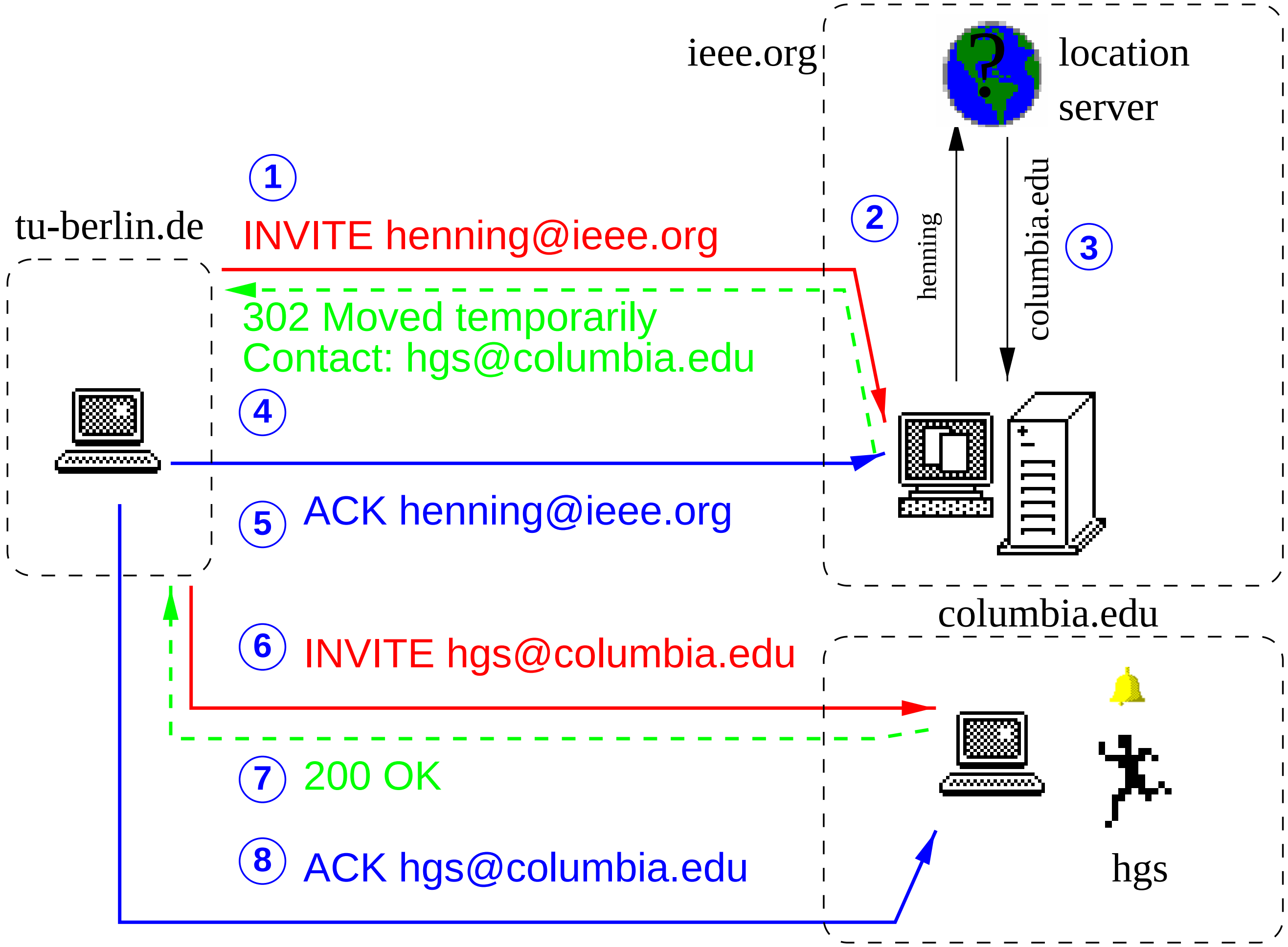
entity	does	examples
proxy server	forward calls	firewall controller, “call router”
redirect server		“application server”
user agent	end system	SIP phone, gateway, “softswitch”
registrar	location mgt.	mobility support

Roles are changeable, on a request-by-request basis

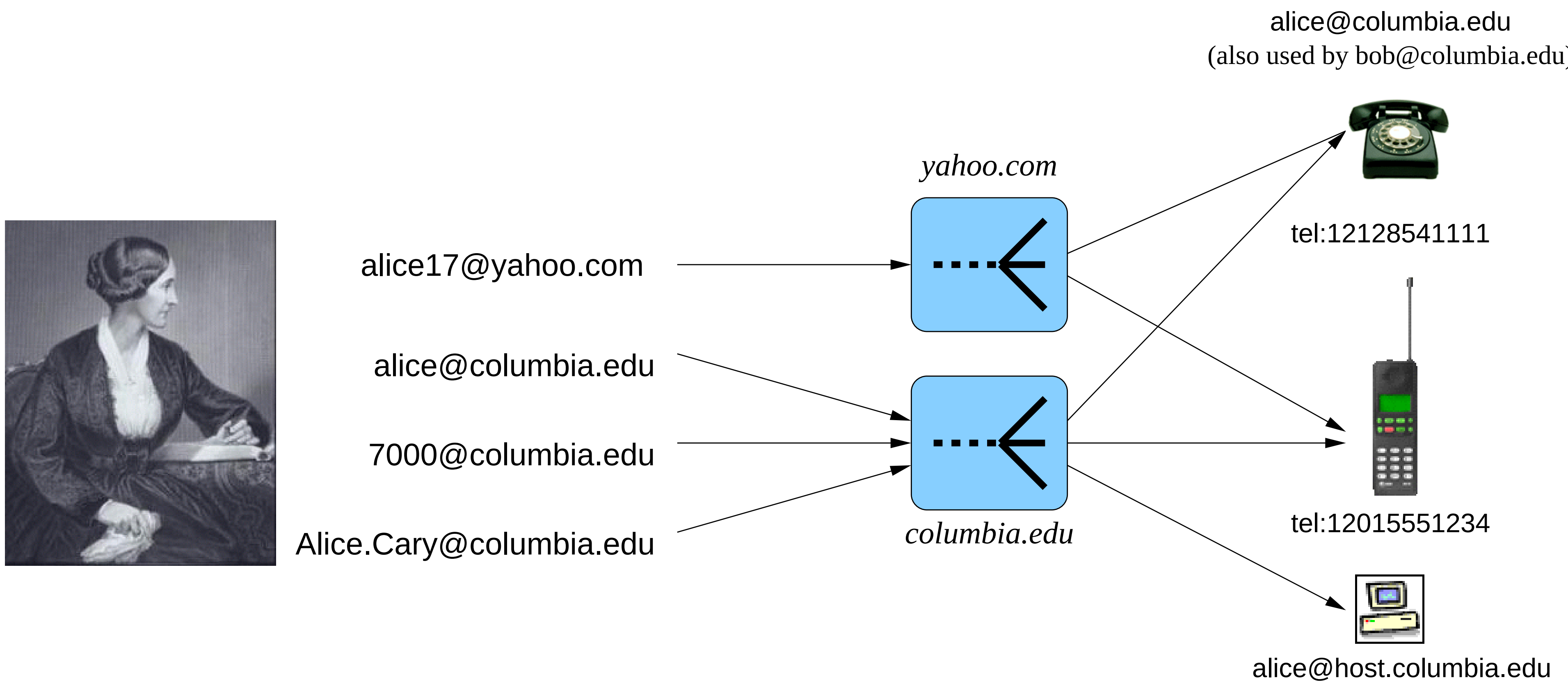
SIP Operation in Proxy Mode



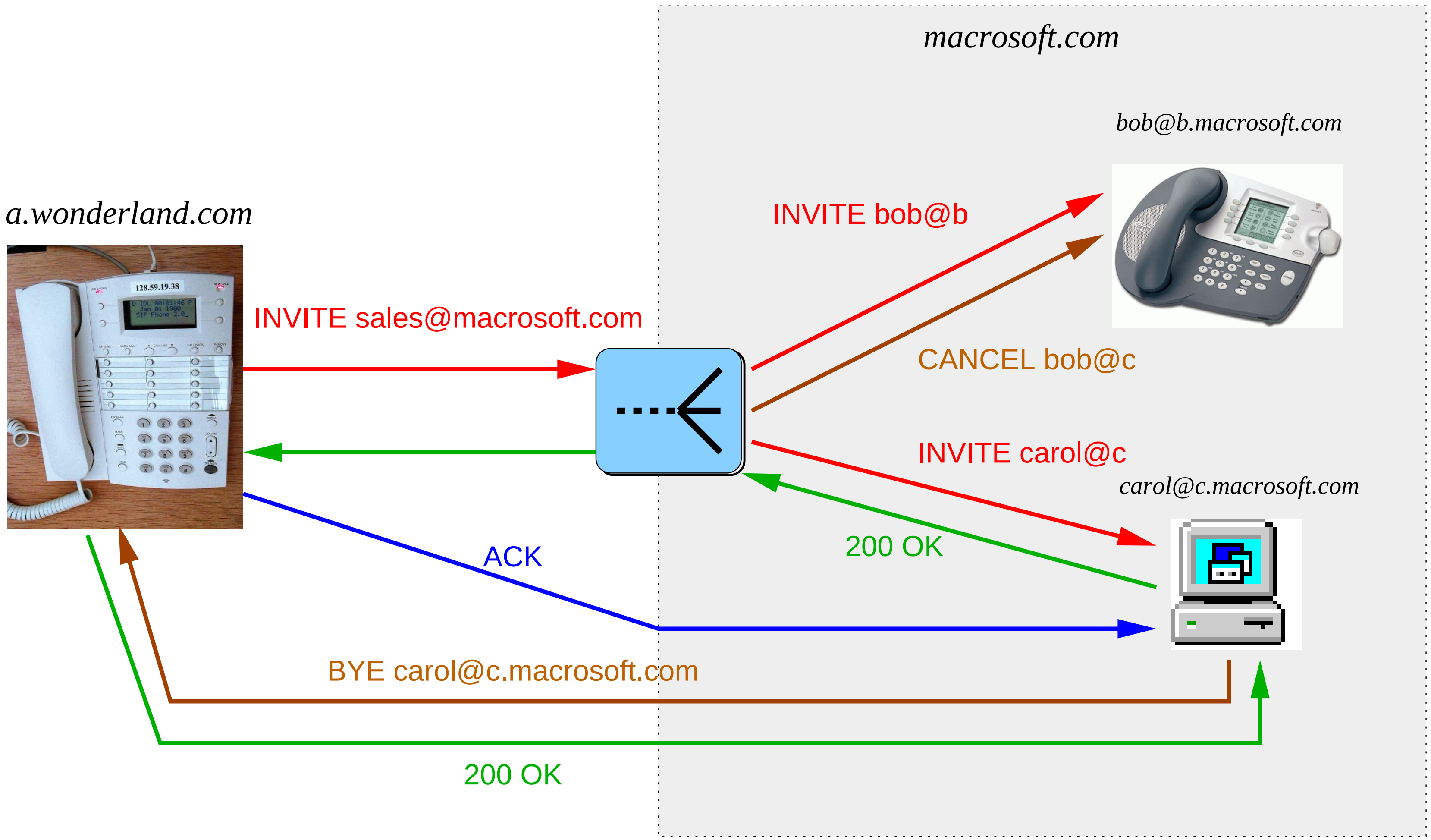
SIP Operation in Redirect Mode



SIP Personal Mobility



SIP Forking Proxies



SIP Advanced Features

- forking
- extensibility: new headers, methods, bodies
- security: web-like, PPP/CHAP or PGP
- multicast-capable
- support for personal, session, terminal, service mobility
- caller preferences: direct calls based on properties

More SIP Internet Telephony Services

- camp-on without holding a line
- short message service (“instant messaging”)
- schedule call into the future
- call with expiration date
- add/remove parties to/from call ➠ mesh
- “buddy lists”

Internet Telephony – as Part of Internet

- universal identifier: email address = SIP address = IM address
- SIP URLs in web pages
- forward to email, web page, chat session, . . .
- include web page in invitation response (“web IVR”)
- third-party control of calls via scripts,
- include vCard, photo URL in invitation
- user-programmable services: CGI (RFC 3050), CPL, servlets

Example: Pingtel SIP phone

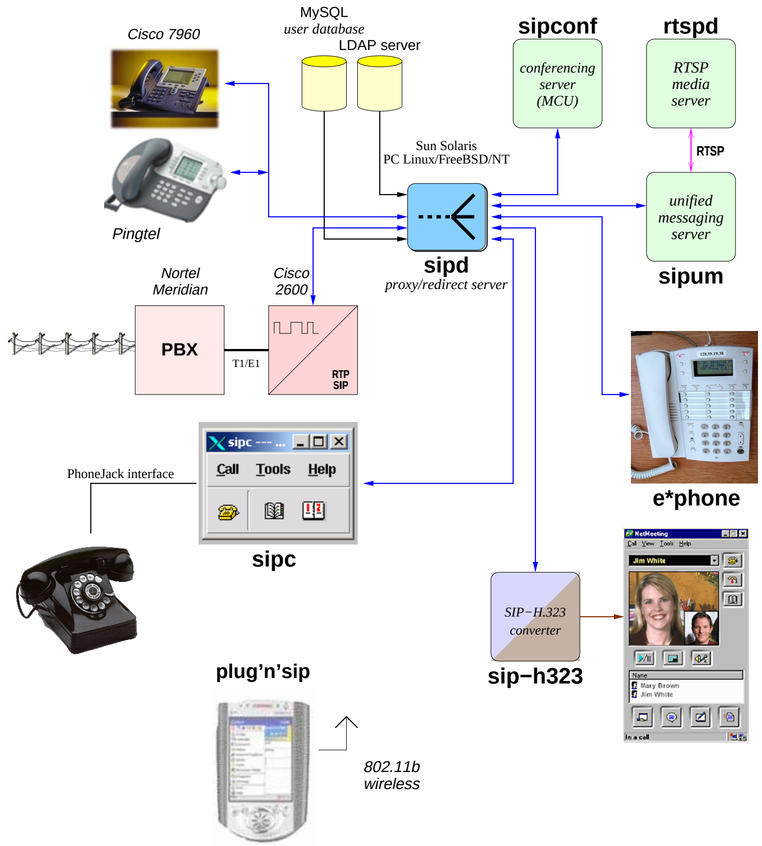
Example: Cisco and 3Com SIP phones

Cisco

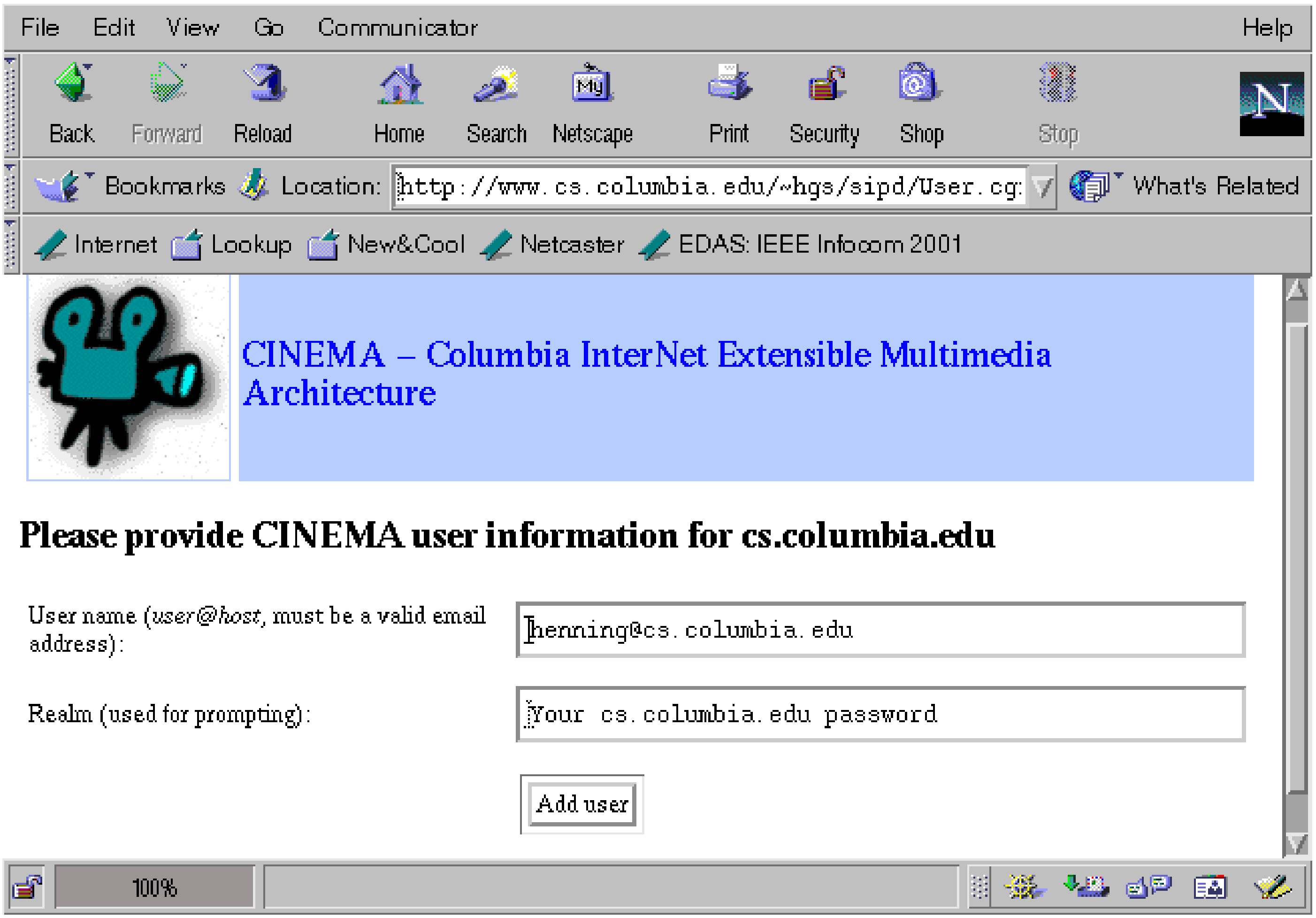
3Com ($395 list)

Example: Columbia CS Phone System

Expand existing PBX via IP phones, with transparent connectivity



sipd single sign-on for account creation and modification



sipd contact management

CINEMA – Columbia InterNet Extensible Multimedia Architecture

Contacts for User hgs@cs.columbia.edu

Deleted contact **hgs@erlang.cs.columbia.edu**.

Contact		Preference	Expires	Action	Last modified	
<mailto:hgs@cs.columbia.edu>		0.1	01 Dec 2001 00:00	Proxy	12 Oct 2000 18:41	Change
<sip:hgs@muni.cs.columbia.edu>				Proxy	15 Dec 2000 18:06	Change
<sip:hgs@128.59.19.205>		1.0	05 Jan 2001 00:49	Proxy	04 Jan 2001 18:49	Change
<sip:hgs@128.59.19.216:5060>		1.0	05 Jan 2001 00:53	Proxy	04 Jan 2001 18:53	Change

Programming Internet Multimedia Services

Primarily, creation, forwarding, proxying, rejection of calls

APIs (Parlay, JAIN): protocol-neutral (SIP, H.323, ISUP), but may be least common denominator

SIP CGI: use Perl and other scripting languages; easy to learn

Servlets: Java only; faster than cgi; limited functionality

CPL: = XML-based language for *user* service creation; portable across providers, but not all services

- Protocol-neutral: Parlay, JAIN, CPL
- Call creation: Parlay, JAIN
- VoiceXML is for voice-service creation *after* call setup

Conclusion

- integrated networks: why? why not?
- Internet telephony: architecture and operational issues
- SIP for creating enhanced services

For more information. . .

SIP: http://www.cs.columbia.edu/sip

RTP: http://www.cs.columbia.edu/~hgs/rtp

Papers: http://www.cs.columbia.edu/IRT