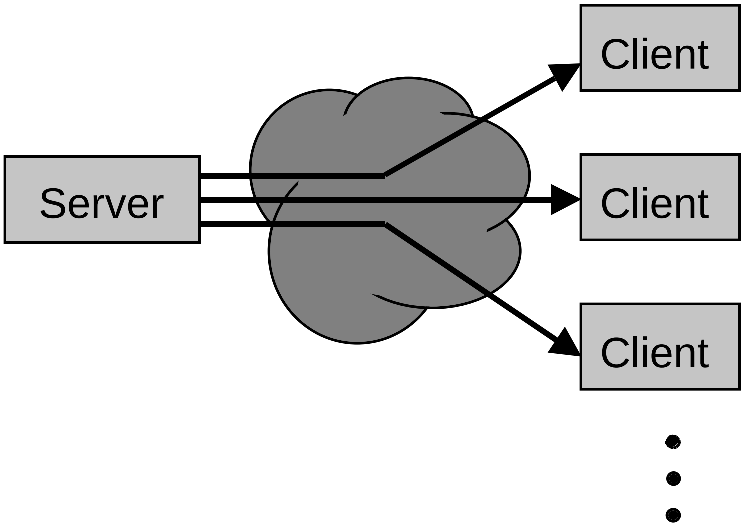
THE Problem

SCALABILITY is the number one problem in networking…

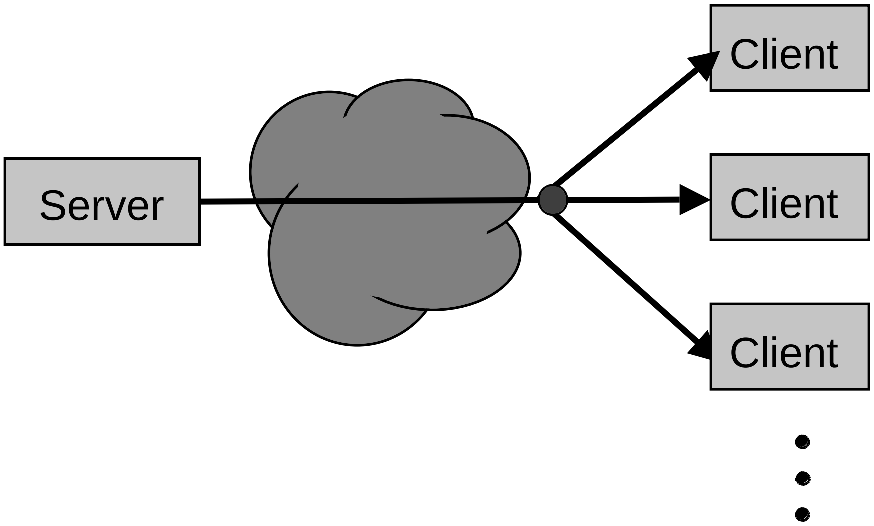
Everything else is secondary.

Phil Rosenzweig
Director
Boston Center for Networking
Sun Microsystems Laboratories

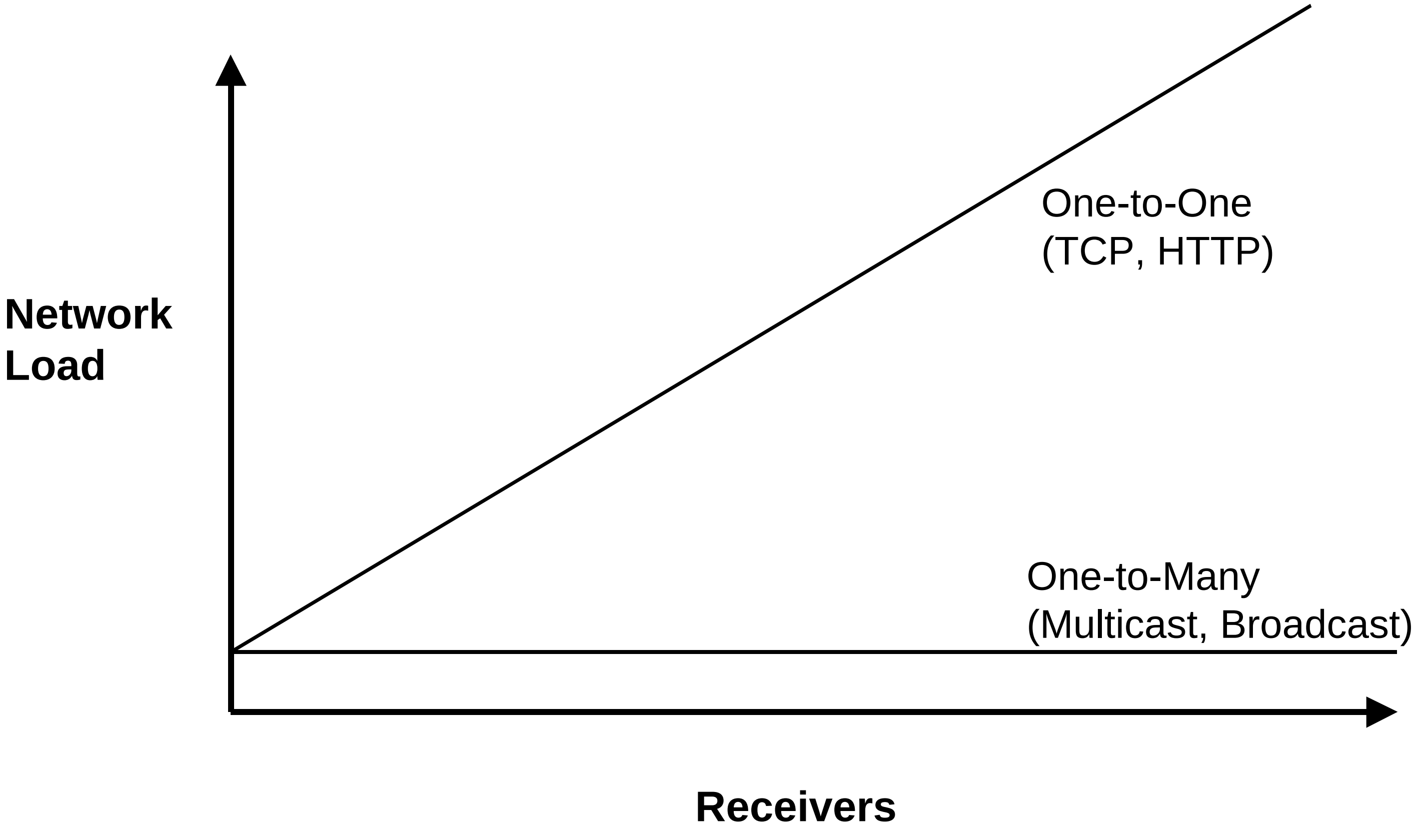
Conventional Reliable Transport



Multicast



Multicast Scales Well



Broadcasting and Multicasting

- There are three *kinds* of IP addresses
  - Unicast
  - Broadcast
  - Multicast
- A unicast address specifies a single interface
- A broadcast address specifies all interfaces
- A multicast address specifies some of the interfaces

Types of IP Broadcasts

- Limited broadcast
  - – 255.255.255.255
  - – Appears only on the local cable
  - – Never forwarded by a router
- Net/Subnet directed broadcast
  - – *Netid*.255.255.255 (host portion all 1’s)
  - – All machines on the specified network
  - – Forwarded by routers (can be disabled)

The Required Pieces

- Three pieces are required for a multicast system
  - A multicast addressing scheme
  - A notification and delivery system
  - An inter-network forwarding facility

IP Multicasting

- IP Multicasting provides two services for an application
  - Delivery to multiple destinations
  - Solicitation of servers by clients
- Class D IP addresses are used for multicast

1110	Multicast group ID

Host Group

- The set of hosts listening to a particular IP multicast address is called a *host group*
- A host group can span multiple networks
- Membership in the host group is dynamic
  - Hosts may join and leave at will
- No restriction on the number of hosts in a group
- A host can simply listen in on a group

Permanent Host Groups

Address	Description
224.0.0.1	All systems on this subnet
224.0.0.2	All routers on this subnet
224.0.1.1	NTP
224.0.0.9	RIP-2
224.0.1.2	SGI Dogfight
224.0.1.84	Jini Announcement
224.0.1.85	Jini Request

Host Multicast Support

- A host participates in IP multicast at one of three levels

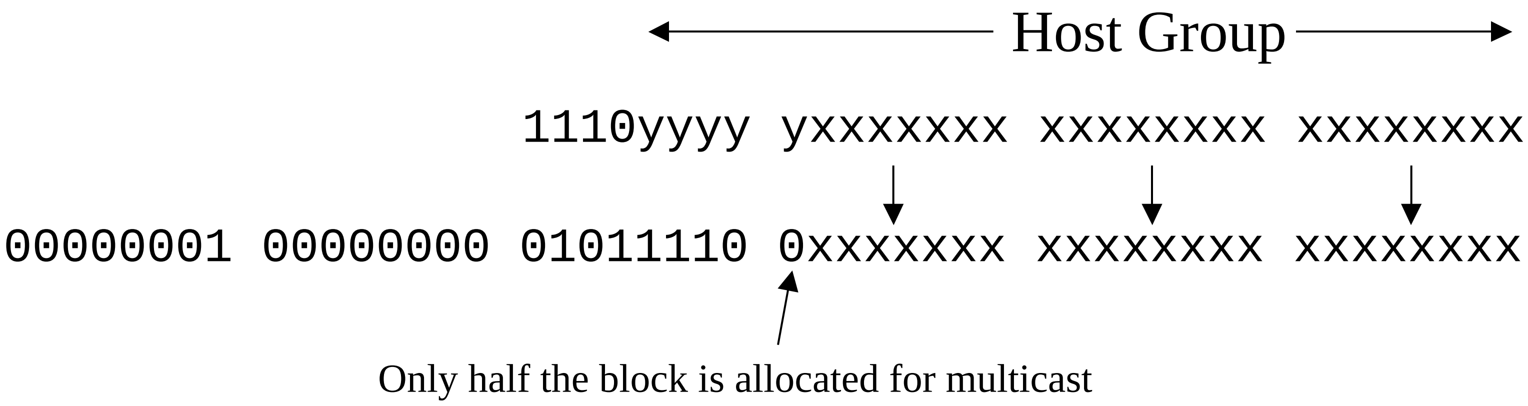
Level	Meaning
0	Host can neither send nor receive IP multicast
1	Host can send but not receive IP multicast
2	Host can both send and receive IP multicast

Multicast on a LAN

- Ethernet supports multicasting
  - The first byte of an Ethernet multicast address is 01
- LAN cards come in two varieties
  - Multicast filtering is done based on the hash value of the multicast hardware address
  - The card contains room to store a small, fixed, number of multicast addresses to listen for

MAC to Multicast

- IANA owns the Ethernet block
  - 00:00:5e:xx:xx:xx
- The addresses 01:00:5e:xx:xx:xx are used for multicast



Host Group

```
                  1110yyyy yxxxxxxx xxxxxxxx xxxxxxxx

00000001 00000000 01011110 0xxxxxxx xxxxxxxx xxxxxxxx
```

Only half the block is allocated for multicast

Example

- IP multicast address 224.0.0.2 becomes
  - `11100000.00000000.00000000.00000010`
  - `      e0.00.00.02`
  - `      00.7f.ff.ff`
  - `01.00.5e.00.00.02`
- IP multicast address 225.0.0.2 becomes
  - `11100001.00000000.00000000.00000010`
  - `      E1.00.00.02`
  - `      00.7f.ff.ff`
  - `01.00.5e.00.00.02`

Beyond a Single Network

- Clearly the IP to MAC scheme only works for a single physical network
- How is the mapping done when machines from different networks are part of a host group
- The IGMP protocol is used provide multicasting between networks

IGMP

- Internet Group Management Protocol (IGMP)
  - Defined in RFC1112/RFC2236
  - Considered to be part of the IP layer
  - Messages sent in IP datagrams
  - Has a fixed-size message with no optional data

IGMP Message

<table>
<tr><td rowspan="2">8-bytes</td><td>4-bit version</td><td>4-bit type</td><td>unused</td><td>16-bit checksum</td></tr>
<tr><td colspan="4">32-bit group address (class D IP address)</td></tr>
</table>

- The Current IGMP Version is 2
- IGMP Type
  - 1 is a query sent by a multicast router
  - 2 is a response sent by a host

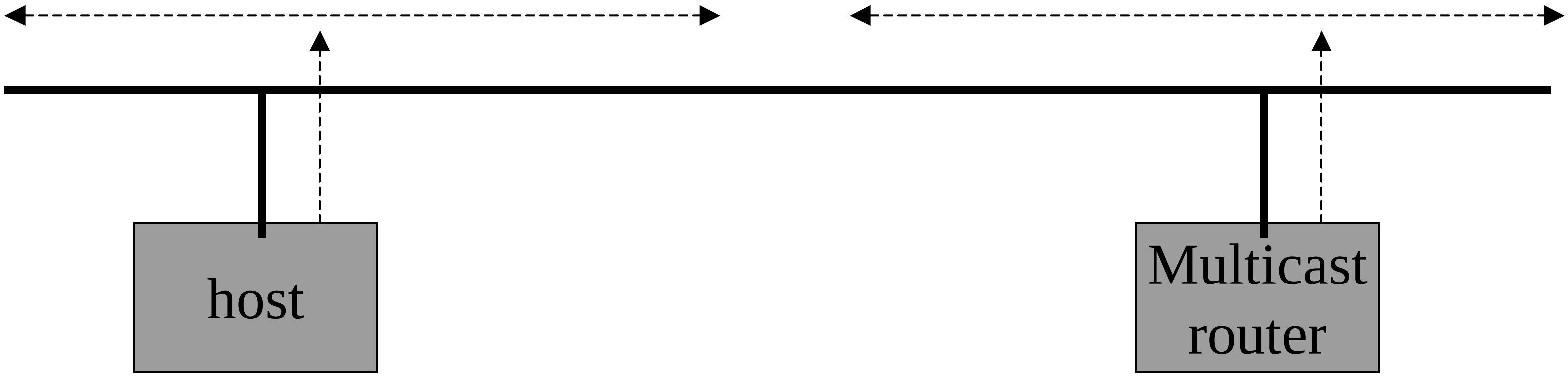
IGMP Rules

- Basic rules
  1. A host sends an IGMP report when a process first joins a group
  2. A host does not send a report when processes leave a group (even when the last process leaves a group)
  3. A multicast router sends an IGMP query at regular intervals to see if any hosts have processes belonging to any groups
  4. A host responds to a query by sending one IGMP report for each group that still has members

IGMP Reports and Queries

IGMP report, TTL =1,
IGMP group addr = group addr
Dest IP addr = group addr
Src IP addr = host’s IP addr

IGMP query, TTL =1,
IGMP group addr = 0
Dest IP addr = 224.0.0.1
Src IP addr = router’s IP addr



My groups are…

Identify each group…

Implementation Details

- There are several ways that IGMP minimizes its effect on the network
  - All communication between hosts/routers use multicast
  - A single query to request group information is sent to all groups (default rate is 125 seconds)
  - If multiple routers are on the same network, one is selected to poll membership
  - Hosts do not respond to the router's IGMP query at the same time
  - Hosts listen for responses from other hosts in the group, and suppresses unnecessary response traffic

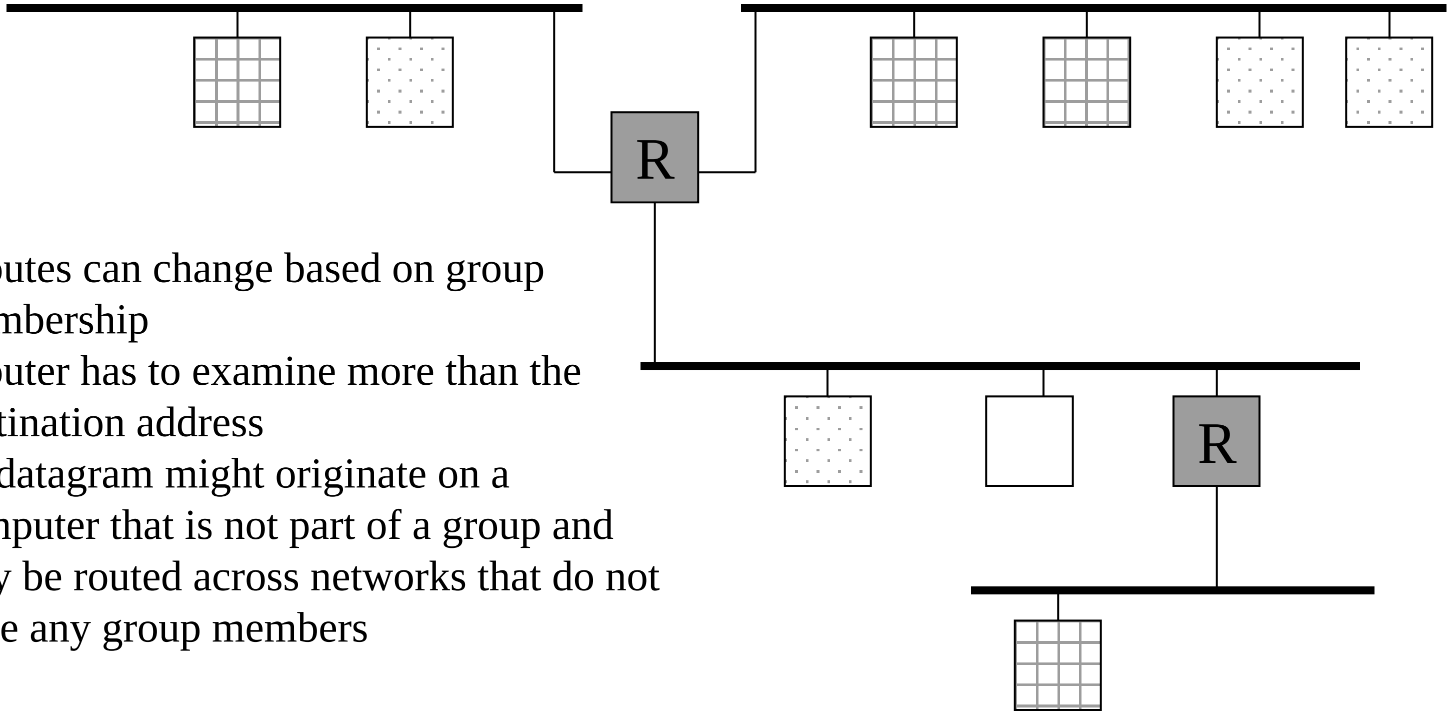
Multicast Scope

- The scope of a multicast address refers to the range of its group members
  - All members on the same physical network
  - All members lie within a single organization
- Multicast datagrams have a scope which is the set of networks over which the datagram will be propagated
- Informally a datagram’s scope is referred to as its range

Controlling Scope

- Two techniques are used to control scope
  - The TTL field is used to limit the range of a multicast datagram
    - Control messages must have a TTL of 1
    - Two applications on the same host use TTL of 0
    - Some router vendors suggest configuring routers to restrict datagrams from leaving the site unless the TTL is 15 or larger
  - Administrative scoping
    - Reserves parts of the address space for groups that are local to a given site or local to a given organization
      - 239.192.0.0 – 239.251.255.255 restricted to one organization
      - 239.252.0.0 – 239.252.255.255 restricted to one site

Multicast Routing



- •Routes can change based on group membership
- •Router has to examine more than the destination address
- •A datagram might originate on a computer that is not part of a group and may be routed across networks that do not have any group members

Multicast Routing

- What information does a multicast router use when deciding to forward a datagram?
  - An optimal forwarding scheme will reach all members of a group without sending a datagram across a network twice
- To avoid routing loops, multicast routers rely on the datagram’s source address

Reverse Path Forwarding

- To use RPF
  - Multicast router must have a conventional routing table
- When a datagram arrives
  - Router extracts the source address
  - Looks up the address in the routing table and determines the interface, *I*, that leads to the source
  - If the datagram arrived on *I* it is forwarded to each of the other interfaces, otherwise it is discarded.

Consequences of RPF

- Since the datagram is sent across every network in the internet, every host in the group will receive a copy
- Wastes bandwidth by transmitting multicast datagrams over networks
  - That do not have group members
  - That do not lead to group members

Truncated RPF

- TRPF avoids propagating datagrams where they are not needed
- Routers need two pieces of information
  - Conventional routing table
  - A list of multicast groups reachable through each interface
- To route datagrams
  - Follows the basic RPF scheme
  - IF RPF says to forward, check the list to make sure the group can be reached on *I* before sending it

The Current State of Multicast

- Most routers, switches, NICs and TCP/IP stacks support multicast
- MBONE operational on the Internet
  - http://www.lbl.gov/WWW-Info/MBONE.html
- Significant work in IETF on standardization
- Reliable multicast is a research topic in the IRTF

Reliable Multicast

- Multicast solves many problems
  - Bandwidth crisis
  - Timely Delivery
  - Latency Control
- Most applications need reliability
  - Or at least *partial* reliability

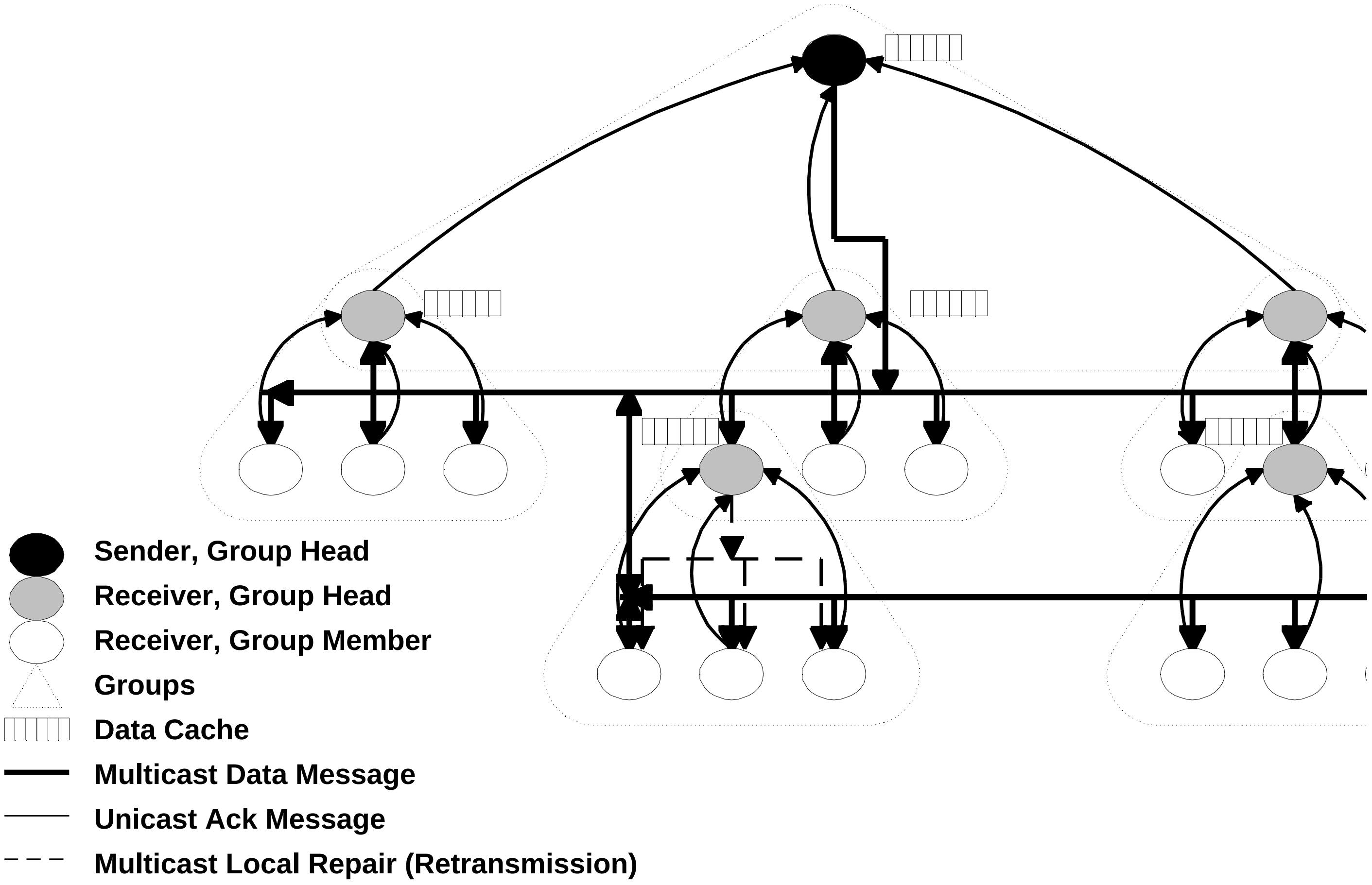
Terminology

- Multicasting is centered on groups
  - Single/Multiple Senders
- Dynamic Group formation/management
  - Joins
  - Late Joins
  - Leaves
- Error Recovery
  - Full/Partial Repair
  - No Repair

TRAM

- A tree-based reliable multicast protocol
  - Sender and receivers dynamically form repair groups
  - Repair groups are linked together to form a tree
- TRAM has been kept as lightweight as possible

Basic TRAM Model



Sender, Group Head

Receiver, Group Head

Receiver, Group Member

Groups

Data Cache

Multicast Data Message

Unicast Ack Message

Multicast Local Repair (Retransmission)

Automatic Tree Formation

- The tree
  - Each receiver is associated with a repair head
  - Be able to add new receivers to the tree at any time
  - Recover from head failure through re-affiliation
- What is a suitable repair head?
  - Shortest TTL distance
  - Eagerness to be head
  - Head experience
  - Repair data availability

TRAM Features

- Reliable
- Avoids ACK implosion
- Local Repair
- Rate based flow control and congestion avoidance
- Feedback to sender
- Scalable

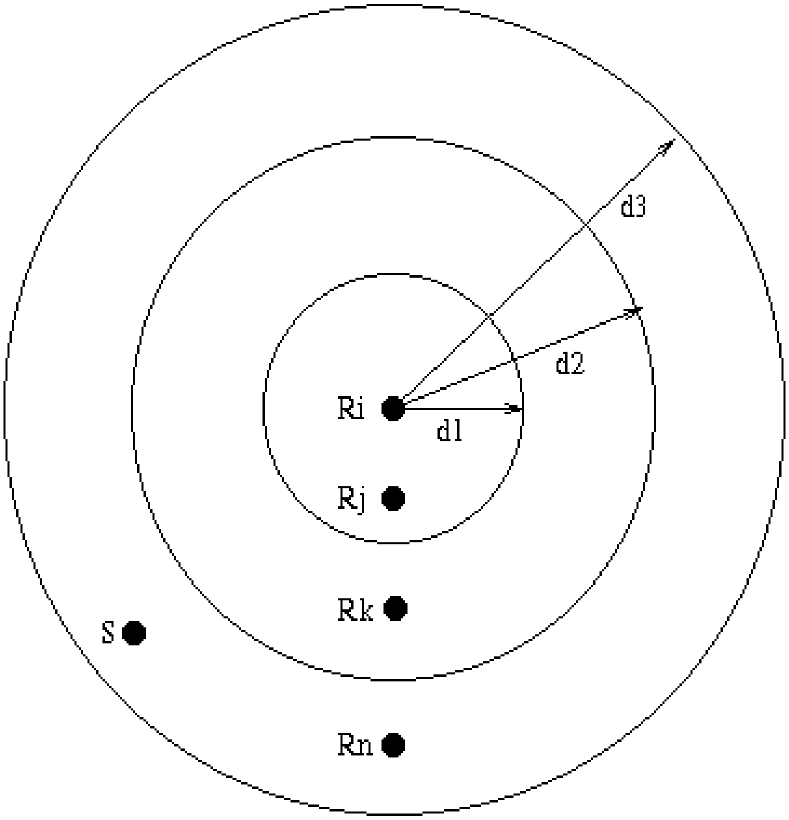
LRMP

- The Light-Weight Reliable Multicast Protocol
  - Guarantees sequenced and reliable delivery
  - Places no restrictions on receiver’s membership
  - Allows multiple senders
  - Light-weight in terms of protocol overhead and simple in control mechanisms

Random Expanding Probe

- Would prefer the repair information be as close to the receiver as possible
- REP consists of three steps
  - – Divide a multicast session into hierarchical subgroups
  - – Report errors to a subgroup
  - – Send repairs to a subgroup

Hierarchy of Subgroups



LRMP

- Normal Operation
  - A source multicasts a set of data packets
    - Transmission is controlled by a transmission interval
  - A receiver detects packet loss using sequence numbers
- LRMP makes no effort to handle full repairs for late joining members

Error Reporting in LRMP

1. Set the number of NACK request N = 0 and the domain level i = 1
2. Schedule a random timer and wait.
3. When the timer expires check
    1. If the lost packets have been received, repair terminates
    2. Otherwise if no NACK was received, send a NACK to the domain D_i
4. If D_i is not the highest level, then i=i+1; otherwise N=N+1
5. If N < Max, go to step 2

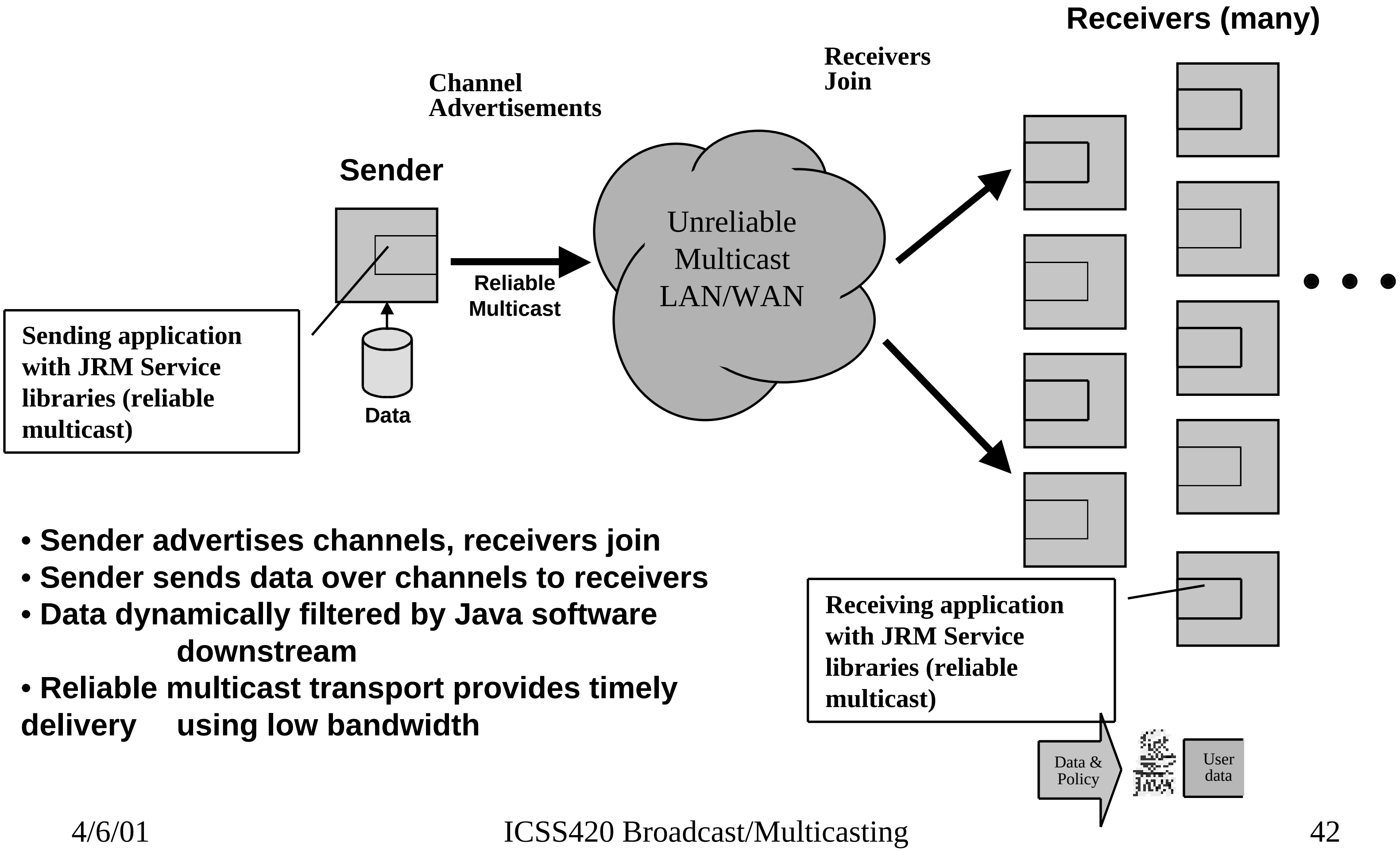
LRMP Features

- Suitable for bulk data transfer
- Provides support for multiple senders
- Congestion control
- Distributed Control

JRMS

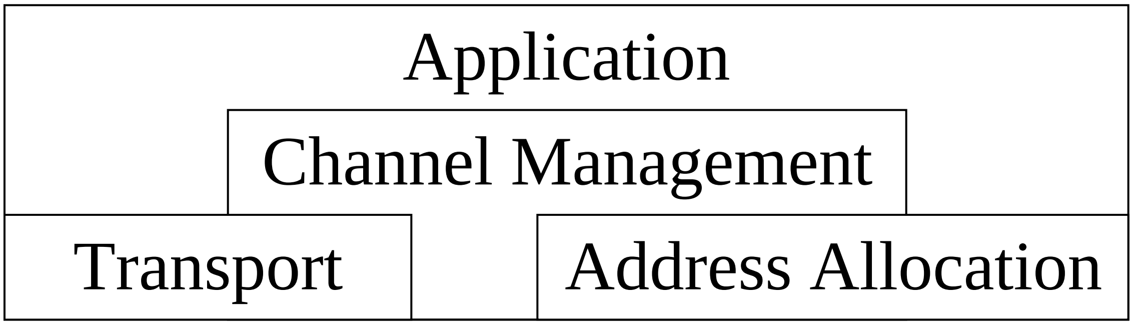
- The Java Reliable Multicast Service
  - Enables building applications that multicast data from “senders” to “receivers” over “channels”
- Organized as a set of libraries and services for building multicast applications
- Functional components:
  - A common API which supports multiple concurrent reliable multicast transport protocols
  - Services for multicast address allocation and channel management

JRMS Data Flow Model



- Sender advertises channels, receivers join
- Sender sends data over channels to receivers
- Data dynamically filtered by Java software downstream
- Reliable multicast transport provides timely delivery using low bandwidth

JRMS Service System

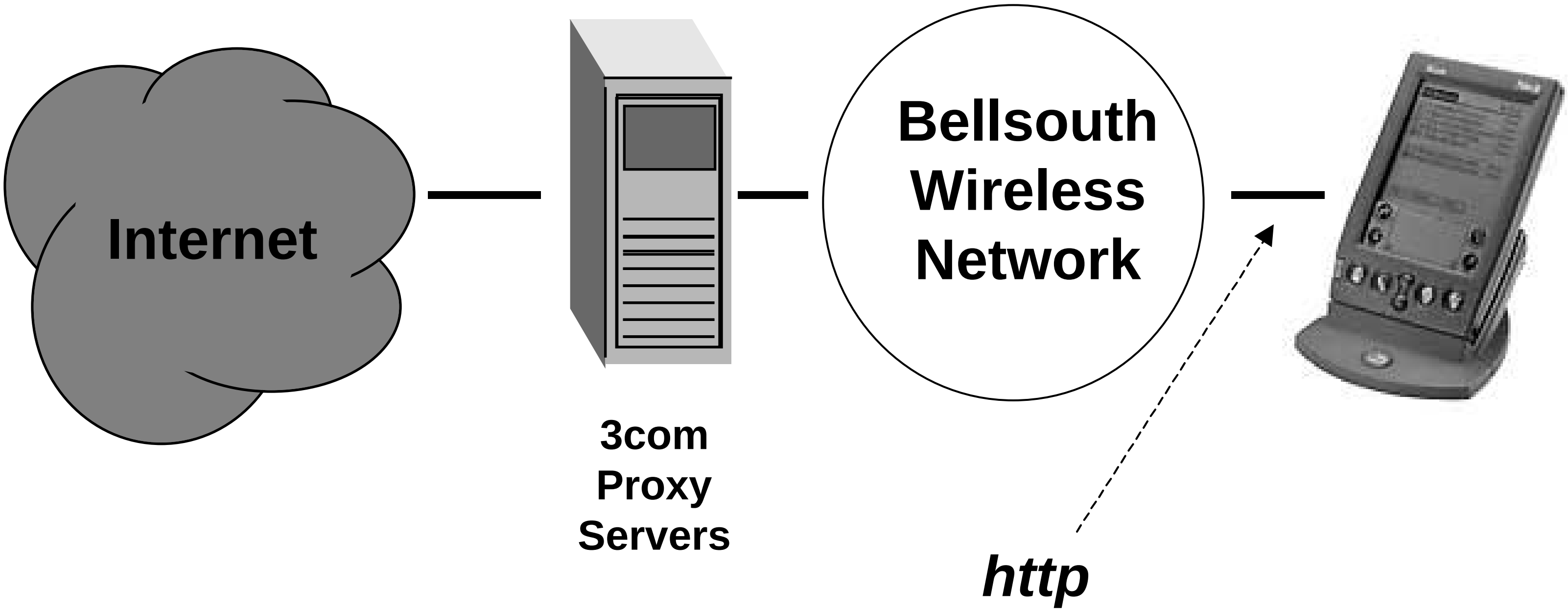


- Transport system transports data over multicast
- Address allocation system manages multicast address allocation
- Channel management system manages channels.

The Palm VII

- A Palm pilot with a built in wireless receiver/transmitter.
- The wireless network is not active all the time
  - Power concerns
  - Impossible to *call/page* the Palm Pilot
- Programming
  - PQAs
  - Java (Spotless/Kjava)

Bellsouth Wireless Network



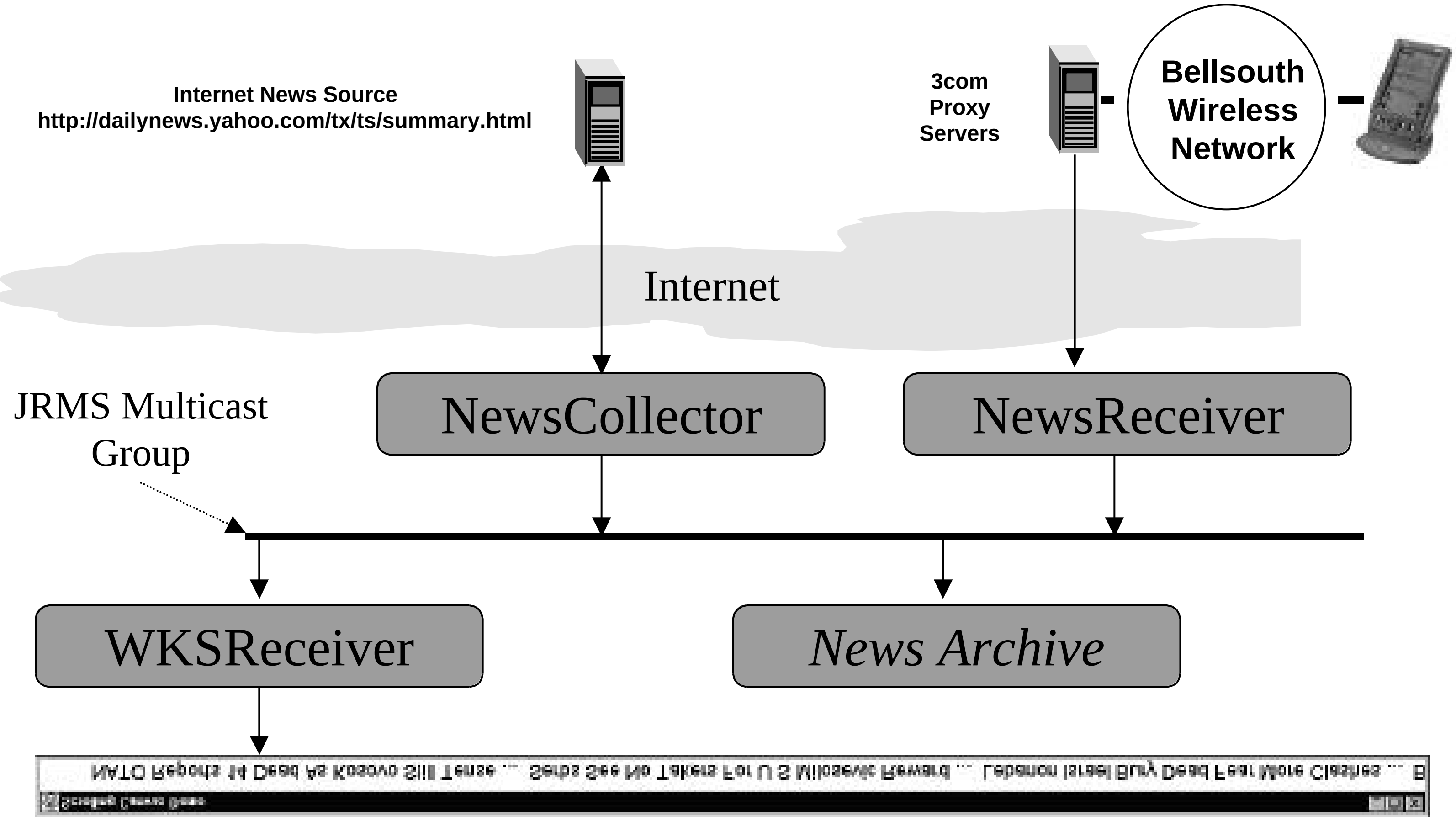
Limitations

- Currently libraries exist only for wireless transfer via http
  - Sockets are possible when using the cradle
- Palm based network applications must *pull* data, data cannot be pushed to the palm
- PQAs are the primary means of web access
- No debugging support for Java on the Palm

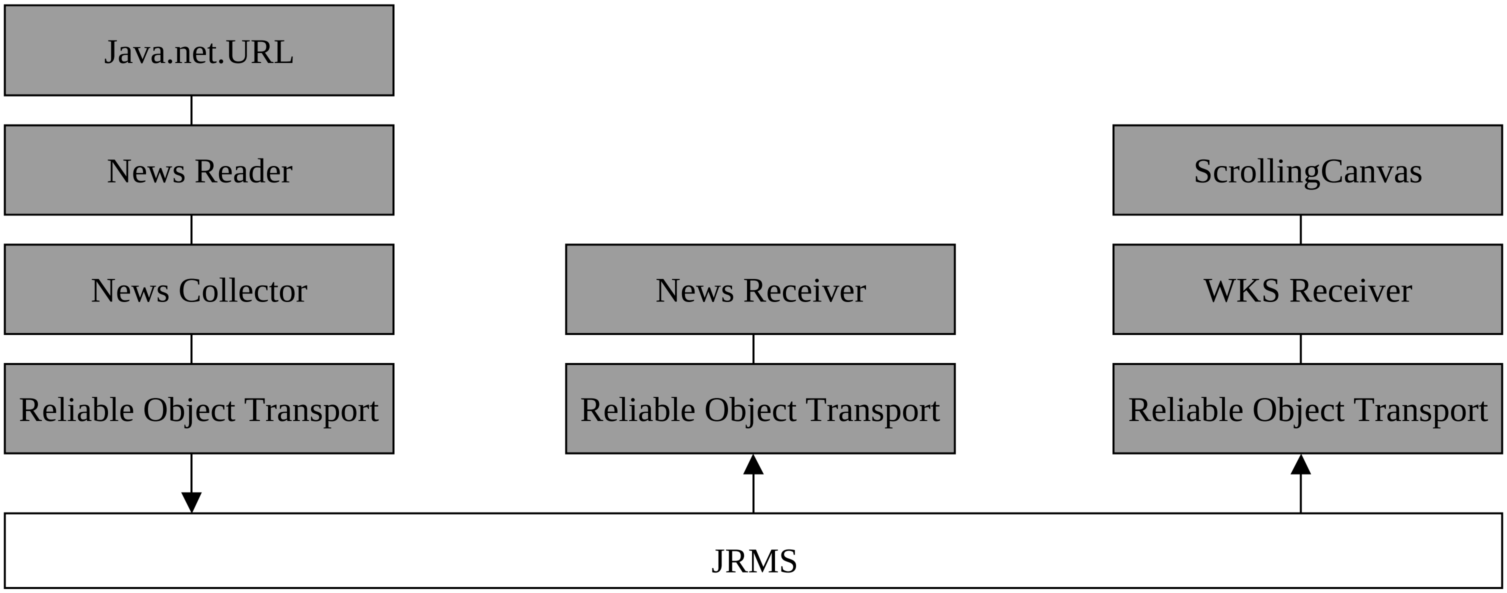
RIT/SUN Project

- Applied research collaboration between SUN Microsystems and RIT
  - Part of the RIT First in Class Initiative
- Primarily interested in wireless networking and IP multicast applications
- Consists of RIT CS faculty students, and staff from SUN

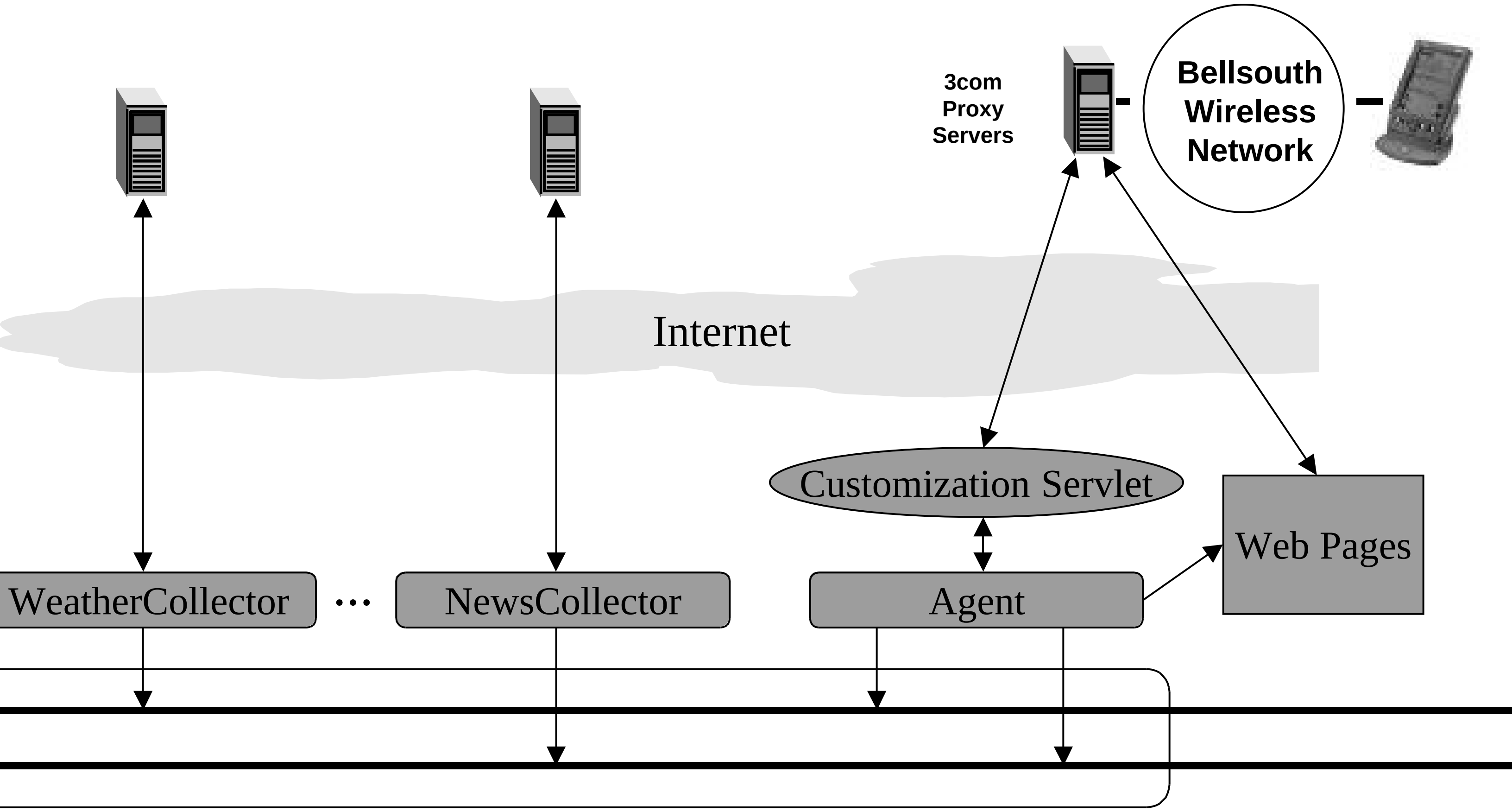
Headline News Application


Internet News Source
http://dailynews.yahoo.com/tx/ts/summary.html
3com
Proxy
Servers
Bellsouth
Wireless
Network
Internet
JRMS Multicast
Group
NewsCollector
NewsReceiver
WKSReceiver
News Archive
NATO Reports 14 Dead As Kosovo Still Tense ... Serbs See No Takers For U S Milosevic Reward ... Lebanon Israel Bury Dead Fear More Clashes ... B

Software Architecture



Agent Architecture



Current Projects

- Multimedia Conferencing
- Student Registration Information
- SunSpot/Kjava Debugger
- JRMS Stress Testing

References

- P. Rosenzweig, M. Kadansky, S. Hanna, *The Java Reliable Multicast Service: A Reliable Multicast Library*, Sun Microsystems Laboratories, SMLI TR-98-68, September 1998.
- D. Chiu, S. Hurst, M. Kadansky, J. Wesley, *TRAM: A Tree-based Reliable Multicast Protocol*, Sun Microsystems Laboratories, SMLI TR-98-66, July 1998.
- M. Kadansky, D. Chiu, J. Wesley, J. Provino, *Tree-based Reliable Multicast (TRAM)*, draft-kadansky-tram-01, Internet Draft, September 1999.
- T. Liao, *Light-Weight Reliable Multicast Protocol*, http://webcanal.inria.fr/lrmp.