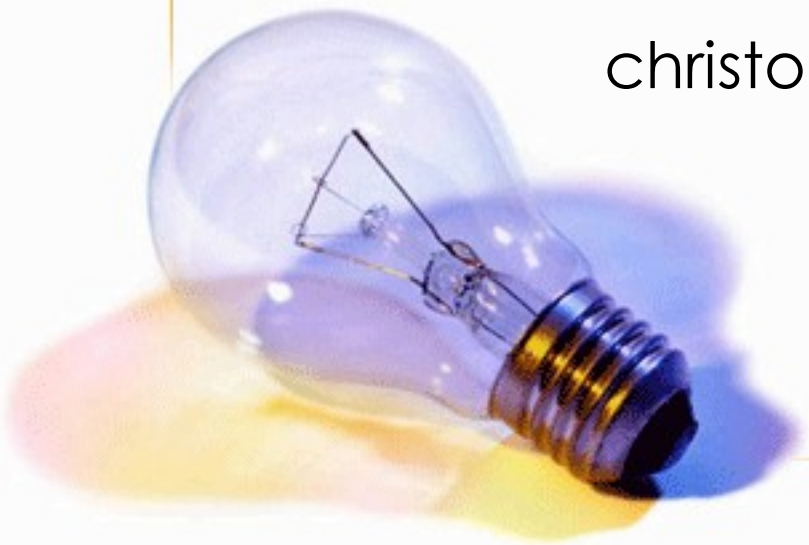
Multicast issues in Infiniband and Fast Ethernet Technology Applications

Christoph Lameter

christoph@graphe.net

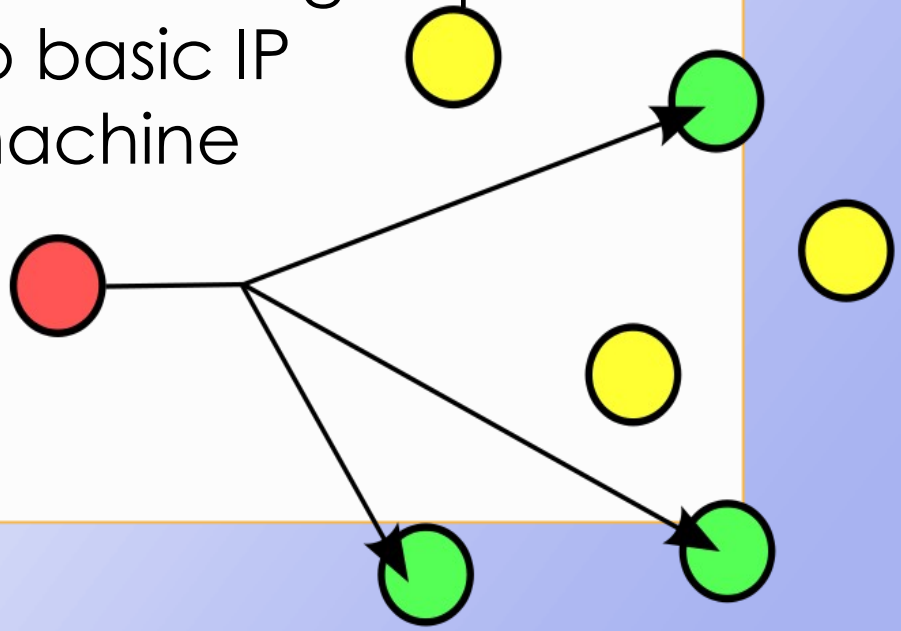


Agenda

- Overview
- Multicast uses in HPC and HFT
- Linux Network Stack issues
- 1G hardware issues
- Infiniband Limitations
- Router / Switch support for Multicasting
- PGM support

Broadcast, Unicast, Multicast

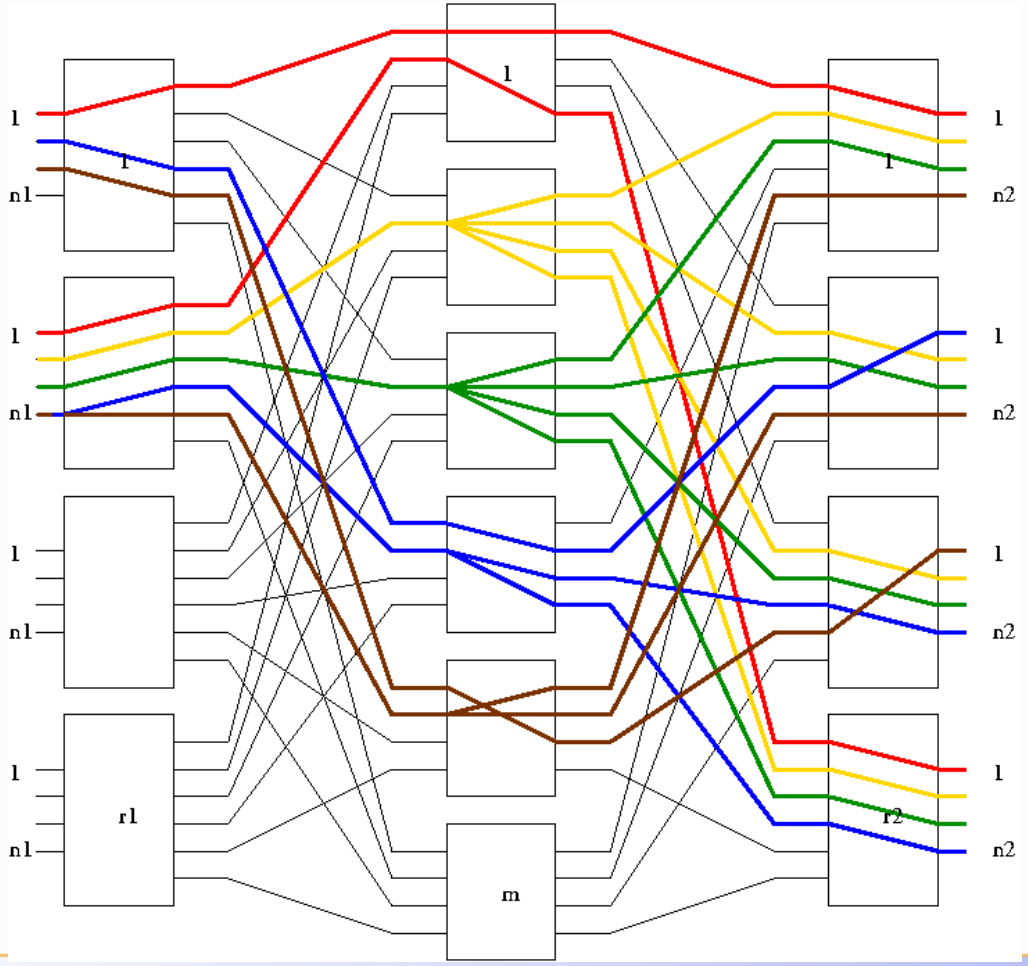
- **Unicast**: one sender, one receiver
- **Broastcast**: One sender, all receiving
- **Multicast**: Receivers opt in
- Must opt-in: Join the Multicast group
- Typically UDP but also basic IP
- Receivable by any machine
- Security issues



Why Multicast?

- Replication by network infrastructure.
- Single message reaches large amounts of receiver.
- Latency (sending to each costs time)
- Simplify configuration: Independent of IP addresses.
- Discovery of network services (UpnP, Bonjour etc.)
- Audio / Video streaming
- Event notifications

Multicast and a CLOS-3 switch fabric



Multicast Basics

- IGMPv2/3 support in Linux, switches routers.
- Special MAC addresses (L2)
- Must use unreliable transport since 1-1 tcp style congestion control not possible.
- “Middleware” to implement “reliability” through requests for retransmission (NAK).
- UDP is protocol of choice for Middleware vendors due to Linux lack of PGM support.
- Retransmission is a band aid. Causes of drops that result in retransmission must be avoided. For that it needs to be possible to detect the reasons for packet drops.

Multicast use in HPC and HFT

- HPC
  - Data broadcast to nodes
  - Job setup / Data setup
  - Service discovery
  - Synchronization point
- HFT
  - Event notification
  - Service discovery.
  - Control / Firing.

Linux UDP/Multicast issues

- Broken flow control to NIC. Network stack may drop UDP packets due to internal congestion.
- Dropped packets were not accounted until 2.6.32. Counter update for drops on sent was broken in 1999.
- Vanishing UDP packets on IP o IB due to overrun issues.
- Linux cannot sent UDP at line rates unless special measures are taken.
- Fix: Rely on throttling through SO_SNDBUF (socket output buffer). If SO_SNDBUF < size of data in packets bufferable by device then packet loss will not occur.
- Keep SO_SNDBUF small (<20k) in order to avoid network stack dropping packets if bursts occur.

Linux IP/PGM Multicast issues.

- No PGM support in Linux IP stack
- Routers require PGM use in order to do NAK suppression at the network layer.
- Existing software (openpgm) must emulate a protocol in user space to make effective use of NAK suppression.
- Emulation is slow therefore Middleware vendors use PGM over UDP which means that Linux cannot use router NAK suppression.

Linux Network Stack TBD

- Full accounting for lost packets.
- Lets not trip over ourselves on outbound packets.
- Tracking causes of packet loss.
- We need PGM support.
- The way of accounting drops in various layers of the OS (socket layer, device layer, qdisc layer and NIC layer) is a bit strange to follow.

Hardware issues

- Ethernet NICs
  - 1G
  - 10G
- Infiniband
  - Switches / Fabric
  - IP gateways
  - HCAs (NIC)

1G NIC issues

- Inconsistent accounting of packet drops between vendors. Counters do not give clear indication of the reason for the drops.
- NIC buffer reconfiguration issues.
  - Ethtool -g NIC buffer sizes
    - Broadcom et al reduced the size of receiver buffer to accommodate multiple queues?
    - Multiqueue support reduces the individual queue size leading to increase of overruns.
- Limited packet rate. Cannot use full 1G rate with small packet sizes.
- Some NICs cannot keep up at full 1G rate without transmission errors or sync loss with switches.

10G NICs and technology

- Inconsistent accounting of packet drops.
- Must use multiqueue to be able to handle traffic load. However, multiqueue support is not mature yet.
- System API has difficulties keeping up with traffic consisting out of small packets.
- No standardized API to bypass IP stack packet processing overhead (But Mellanox is now allowing to operate IB NICs in 10G mode. Then IB offload techniques can be used).

Infiniband Fabric / Switches

- Limitations on the # of MC groups
- Subnet manager configures static routes for multicast traffic around a calculated “network center”.
- Specific loads can cause overload.
- Credit System can bring fabric to a halt with a slow receiver on a multicast group.

Infiniband IP gateways

- Gateways subscribe to all traffic and thereby cause useless replication.
- Gateways are complex to configure.
- Load balance issues. IP gateways cannot discover local nodes but round robin through available gateways.
- Single Vendor(?)

Infiniband HCAs

- No accounting of UD drops(per IB spec!)
  - QP overrun by UD packet causes silent packet drops.
  - QP are used f.e. for IPoIB
  - Multicast packets are silently dropped.
- Ability to overrun fast because receiver speed issues.
- Trouble controlling scheduler and application overhead
- Context switch overhead.
- IBVerbs interface difficult to program.

PGM Pragmatic General Multicast (RFC 3028)

- NAK suppression essential to MC delivery reliability
- Commercial routers support NAK suppression for native PGM.
- Linux Middleware vendors use PGM over UDP.
- Available open source PGM implementation (openpgm) supports native PGM by emulating PGM protocol in user space. The only kernel implementation is in MS-Windows.
- I am working on a PGM implementation for the Linux IP stack.
- Basic agreement on socket API exists.
- Implementation of /proc /sys interfaces.
- Native PGM support will allow use of NAK suppression and interaction with commercial and open source PGM implementations.

Future

- Consistent reporting of ethernet statistics
- Accounting for different causes of drops.
- Latency measurement infrastructure.
- Hardware fixes.
- PGM implementation at the network layer.