LNet Mysteries

Isaac Huang
LNET Architecture

Generic LNET

LNET users

LND

LND

LND

LNET API

LND API

Network – Specific APIs
LNet Semantics

Connection less:
  PTLRPC service connection based.
  LNDs connection based.

Unreliable: upper layer resend

End-to-end delivery, out of order
  LNDs reorder messages: separate channels
  Routers reorder messages
  Channel Bonding
  ZC on receive side
Sample Config

Options ko2iblnd peer_credits=8 credits=256 concurrent_sends=8 peer_buffer_credits=16

Options kptllnd peercredits=8 credits=256 peer_buffer_credits=32
LNet Peer TX Credits

Peer TX credits: # outstanding outgoing messages
  Outstanding: passed to LND, SENT event pending
  Accounted per NID, instead of per PID

What it is: fairness control
Not end-to-end: peer is next hop
Not flow control: peers not involved
LNet NI TX Credits

# outstanding outgoing messages per interface
LNet messages need both credits
Fairness
Credits exhaustion for shared FS
The SENT event
Time to reuse buffer safely
When LNDs deliver SENT event
  OFED: IB and iWarp
  SockInd: short messages
The story of sockInd zero-copy and SENT semantics.
LND Peer TX credits

# outstanding messages per peer
   LND messages instead of LNet messages
   Per-NID or per-PID (kptllnd)

Flow-control: credits returned by peers
   Not end-to-end

peer_credits for o2iblnd and ptlllnd
   Socklnd is different
   No way to change old o2iblnd LND tx credits

sdf
Router buffer credits

# router buffers a sending peer can use

Src, instead of dst

Per NID, not per PID

Previously default to peer TX credits

Bad for userspace clients

Bad for short bursts of data
Summary
No worry unless routers or WAN
Window size determined by both:
Router buffer credits:
  Userspace peers
  Double TX to smooth jitter
IB Multipath and Multirail

Multiple path in the fabric; multiple ports

Multipath: timeout and resend

Multirail: status quo and future
  Static link aggregation, see Lustre Operations Manual
  Failover
  The Channel Bonding Project
Protocol Compatibility

Multiple Layers: PTLRPC service, LNet, LNDs

PTLRPC: negotiation at connect

LNet: connectionless, only one protocol version

LNDs: negotiation at connect
Thank You