Arguments for Standard 3m no-FEC DAC Solution

MIKE ANDREWARTHA, BRAD BOOTH

MICROSOFT

8/12/2015

Supporters

Tom Issenhuth – Microsoft
Eric Baden – Broadcom
Rob Stone – Broadcom

Topics

- Latency matters
- Management implications
- Value to broad market potential of standard solution
- Other factors

Latency matters

Who Cares? - Latency Sensitive Application Spaces

- High Performance Computing
- Financials – High Frequency Trading
- New apps using RDMA are emerging in storage, virtualization, etc.
  - See Open Fabric Alliance Developers Workshop or User Group papers for examples

Why do they care?

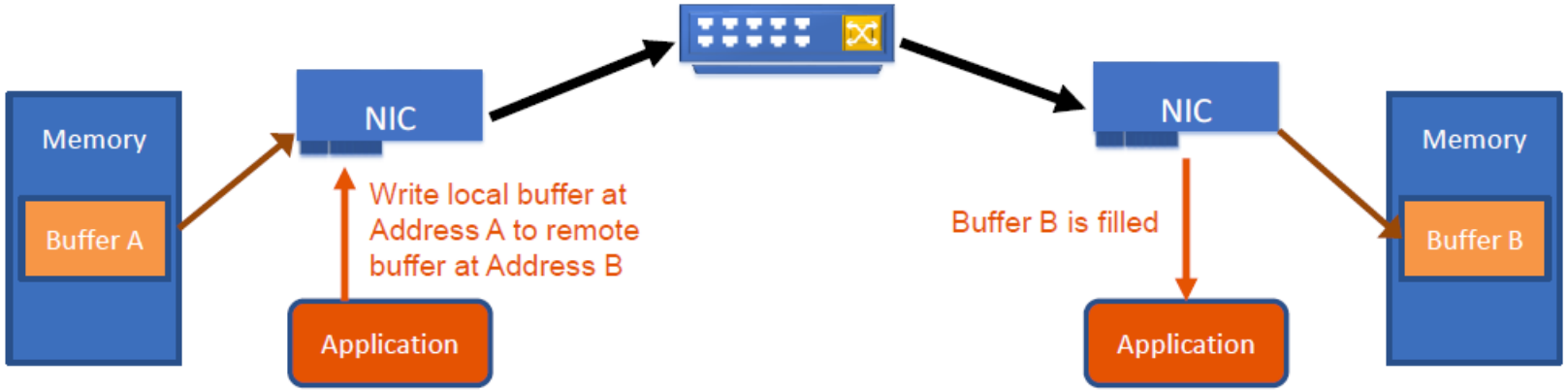
- Latency can be limiting factor in scaling of parallel applications
- Latency is visible to customers using standard benchmarks
- Competitive pressure on HPC/cloud providers to offer lowest latency option
  - Even if it takes an engineered solution

How large is the impact of adding Base-R FEC at 25 Gbps?

- Baseline for latency without FEC
- Impact of adding FEC

Baseline for no-FEC latency

- No one wants to reveal details of his implementation
- No published 25G performance data -> use 40 GbE for baseline
- Multiple published claims of < 2 us End to End latency – keywords RDMA, RoCE, iWARP, OFED



- Remote DMA primitives (e.g. Read address, Write address) implemented on-NIC
  - Zero Copy (NIC handles all transfers via DMA)
  - **Zero CPU Utilization at 40Gbps** (NIC handles all packetization)
  - <2µs E2E latency

Source: https://www.openfabrics.org/images/eventpresos/workshops2015/DevWorkshop/Monday/monday_15.pdf

Estimating latency impact of Base-R FEC

Extra latency encountered at each sender to encode & each receiver to decode/correct

Sender encode latency
- no extra blocking required
- Encode time - implementation dependent but likely small

Receiver decode/correct latency
- requires time to receive full block – 2112 bits x 40 ps = 84.48 ns
- decode/correct time - implementation dependent

Short packets see added latency
- minimum packet size set by 2112 bit encoding block (256 Bytes payload).
- Many RDMA apps use smaller packets for synchronization/control. Single byte and 64B benchmark results are common.

For estimation purposes use 100 ns for combined per hop incremental delay through sender encode + receiver block time + receiver decode/correct time

@ 2 us E2E, 200 ns incremental delay adds 10%.

10% is approximate lower bound on latency penalty: lower E2E and/or higher implementation delays increase impact

Management Implications

3m reach is required in some applications

- Enterprise: see http://www.ieee802.org/3/by/public/July15/goergen_3by_02a_0715.pdf
- Cloud: see http://www.ieee802.org/3/by/public/Jan15/andrewartha_3by_01a_0115.pdf

D2.0 requires both ends of link to agree to not request FEC to auto-negotiate no-FEC operation on the link.

- Endpoint has to decide whether to request FEC based on cable type connected and a-priori knowledge of host losses

Don’t want to operate some server links with FEC and others without on same top of rack switch

Value to Broad Market Potential

Common, standard cable spec is good for everyone

- Highest volume/lowest cost from shared solution – avoid splintering market with engineered solutions
- Less confusion among end users
- Fewer combinations for manufacturers to test/qualify
- Manufacturers build & users buy to a standard spec rather than multiple proprietary specs for engineered solutions

Consistency with emerging multi-lane standards creates a larger market

- 50G Ethernet @ 2x 25G – latency penalty is 2x single lane
- No reason for cable performance specs to be different

Other Factors

Feasibility of interoperable standard solution

- Subject of multiple other presentations
- Baseline assumption is any solution can't change compliance of NICs & switches that also support 802.3bj 100GE

Power consumption & Implementation overhead (gates/logic)

- Impact is implementation dependent but is non-zero in all cases.
- Logic implementation is required to be compliant as Base-R FEC is mandatory

Your mileage may vary ☺

Thank You!