Feasibility & Rationale for 3m no-FEC server-switch DAC

MIKE ANDREWARTHA, BRAD BOOTH – MICROSOFT

CHRIS ROTH - MOLEX

9/14/2015

Supporters

John D’Ambrosia – Dell
Vittal Balasubramanian – Dell
Rob Stone – Broadcom
Tom Issenhuth – Microsoft
Mike Dudek – Qlogic
Piers Dawe – Mellanox

Topics

- Latency matters
- Management implications
- Value to broad market potential of standard solution
- Other factors
- Possible path forward for server-switch links

Latency matters

Who Cares? - Latency Sensitive Application Spaces

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

Why do they care?

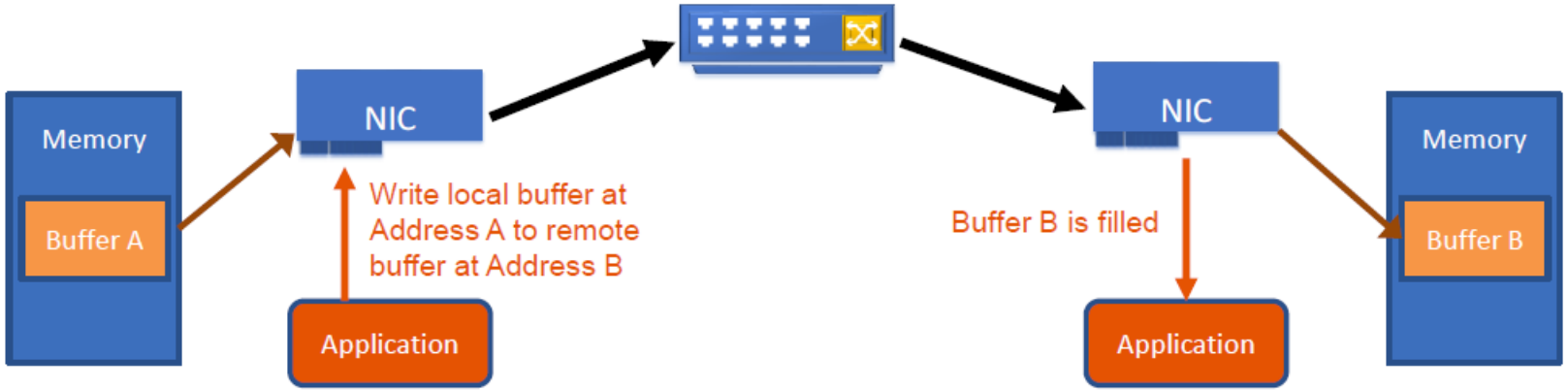
- Latency can be limiting factor in scaling of parallel applications
- Latency is visible to customers using standard benchmarks
- Competitive pressure on HPC/cloud providers to offer lowest latency option
  - Performance metrics influence customer behavior => providers motivated to optimize performance metrics
  - Opens the door for a non-standard, proprietary 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 µs 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 FEC

Base-R FEC is lower latency than RS-FEC so focus discussion on Base-R

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 µs 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 many intra-rack 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

- Drives end users to consider engineered/proprietary solutions if latency is a competitive disadvantage.

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
- If IEEE 802.3by doesn't agree on a standard solution, industry likely will

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 ☺

A Path Forward for server-switch links?

Observations

- Server add-in card NIC designs often don't need 6.26 dB host PCB loss assumed in COM.
  - andrewartha_3by_01a_0315
- 3m cables can be made using 26 AWG twinax with insertion loss < 16 dB
  - tracy_3by_01_0715
- Many suggestions for tweaking COM parameters yield small improvements
- Switch to switch links likely to be 100 GE (802.3bj) and/or use FEC for longer reach between racks

Revisit asymmetric host loss budget for server-switch links

- Can combination of reduced server host loss = 4 dB & ~16 dB cable can meet 3 dB COM without modifying COM parameters?

Asymmetric server-switch simulations

Channel

- TX -> 4 dB server loss -> SFP28 -> cable -> QSFP28 -> standard switch host loss -> RX
- RX <- 4 dB server loss <- SFP28 <- cable <- QSFP28 <- standard switch host loss <- TX
- Approximate 4 dB server loss by:
  - setting Z_b (TX) = 84 mm for SFP driving case and Z_b (RX) = 84 mm for QSFP driving case
  - Setting Z_b (NEXT) = 40 mm for SFP driving case and Z_b(FEXT) = 40 mm for QSFP driving case

Cable models

- Molex 25 AWG QSFP28 – 4x SFP28 – lengths range from 3.0m to 3.4m in 0.1m increments
- TE QSFP28 – QSFP28 contributed channels near 16 dB described in tracy_3by_01_0715
- FCI 26 AWG QSFP28 – 4x SFP28 contributed channels near 16 dB at 25C described in zambell_090215_25GE_adhoc-v2

Asymmetric Simulation Results

COM results for test 2 (worst case)

- Good lanes can meet 3 dB with margin but few with IL ≥ 16 dB
- Bad lanes distinguished by higher crosstalk

Cable (pair)	Length	Min COM	Avg COM	Avg Cable Assy IL	Max Cable Assy IL	Cable Assy IL @ min COM	IL_dB_at_Fnq
Molex 25 AWG (all pairs)	3.0m	2.837 dB	3.098 dB	15.25 dB	15.626 dB	14.736 dB	25.02 (avg)
Molex 25 AWG (all pairs)	3.1m	2.027 dB	2.951 dB	15.413 dB	15.89 dB	15.836 dB	25.20 (avg)
Molex 25 AWG (all pairs)	3.2m	2.908 dB	3.090 dB	15.4 dB	15.716 dB	15.716 dB	25.15 (avg)
Molex 25 AWG (all pairs)	3.3m	2.612 dB	2.874 dB	15.93 dB	16.278 dB	16.278 dB	25.62 (avg)
Molex 25 AWG (all pairs)	3.4m	2.613 dB	2.844 dB	16.035 dB	16.64 dB	16.64 dB	25.82 (avg)
TE 25 AWG 15.35 dB (P1 TX4)	3.0m	2.819 dB	2.819 dB	15.35 dB	15.35 dB	15.35 dB	24.829
TE 26 AWG 15.96 dB (P1 TX3)	3.0m	2.191 dB	2.191 dB	15.96 dB	15.96 dB	15.96 dB	25.501
TE 26 AWG 15.99 dB (P2 TX1)	3.0m	2.419 dB	2.419 dB	15.99 dB	15.99 dB	15.99 dB	25.713
FCI 26 AWG 25C (Pr 6 to Pr 14)	3.0m	2.937 dB	2.937 dB	15.52 dB	15.52 dB	15.52 dB	25.245
FCI 26 AWG 25C (Pr 10 to Pr 2)	3.0m	2.336 dB	2.336 dB	16.02 dB	16.02 dB	16.02 dB	26.107

Proposal

- Plan A: agree on a set of changes to COM model & parameters that meet 3m w/o FEC for all cases
- Plan B: document informative guidelines to meet 3 dB COM for 3m no-FEC for special case(s)
  - Use informative annex to avoid changing base spec
  - Asymmetric server host loss approach
    - Don’t change receiver or transmitter specs
    - Reduce server side host loss limit to 4 dB (Z_bp = 84 mm)
    - Define CA-N cable max insertion loss = -16 dB
    - Adjust receiver test parameters accordingly
    - Still need minor(?) corrections/improvements to COM parameters to allow for manufacturing tolerances
  - Other approaches could also be documented
  - Develop complete concrete proposal for annex if Plan A fails & Plan B has broad support

Thank You!