Coherent Ethernet for the AI Era

Marco Mascitto, Nokia

Version 1.2

IEEE SA Standards Board Bylaws & Disclaimer

From the IEEE SA Standards Board Bylaws

5.2.1.6 Official statements, and political and legislative activity

A statement, written or oral, that is not processed in accordance with subclause 5.1.3 of the IEEE SA Standards Board Operations Manual shall not be considered the official position of the IEEE Standards Association or any of its committees and shall not be considered to be, nor be relied upon as, a formal position of IEEE. At lectures, symposia, seminars, or educational courses, an individual presenting information on IEEE standards shall make it clear that his or her views should be considered the personal views of that individual rather than the formal position of IEEE.

IEEE SA volunteers and participants in IEEE standards activities are not permitted to state or in any way imply that they represent IEEE or IEEE SA in any public statements or political and legislative activity, including lobbying, unless authorized to do so. Volunteers shall comply with the Guidelines Regarding Political and Legislative Activity by IEEE Standards Association (IEEE SA) Volunteers and Staff.

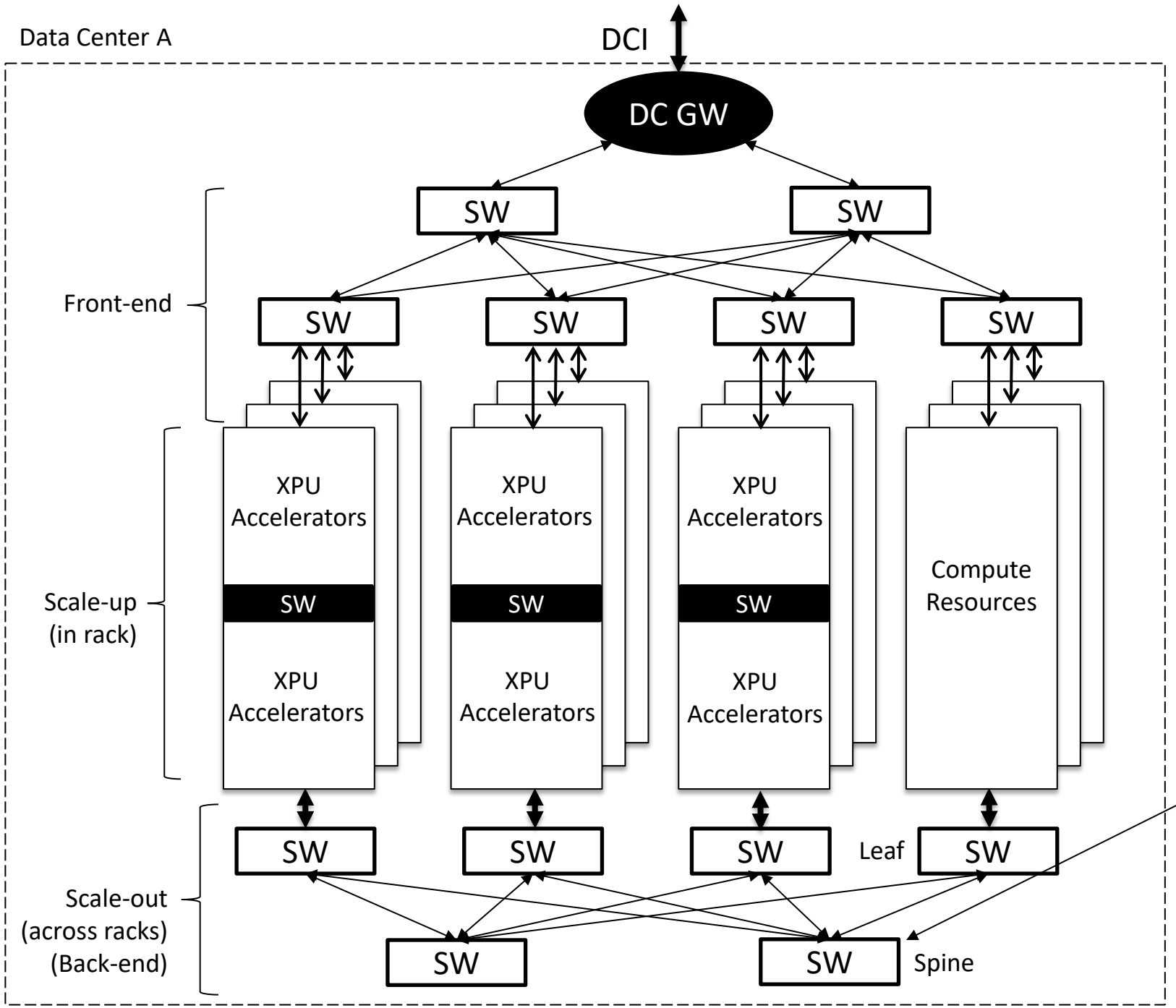
Disclaimer

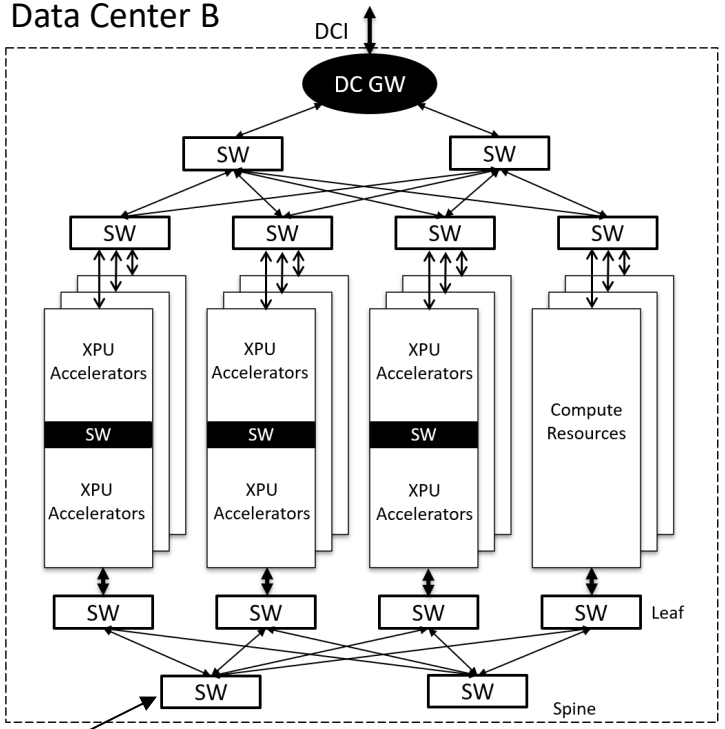
This contribution presents the views of the author and is not a formal position of the IEEE.

Coherent Ethernet for the AI Era

INTRODUCTION

AI Data Center Architecture



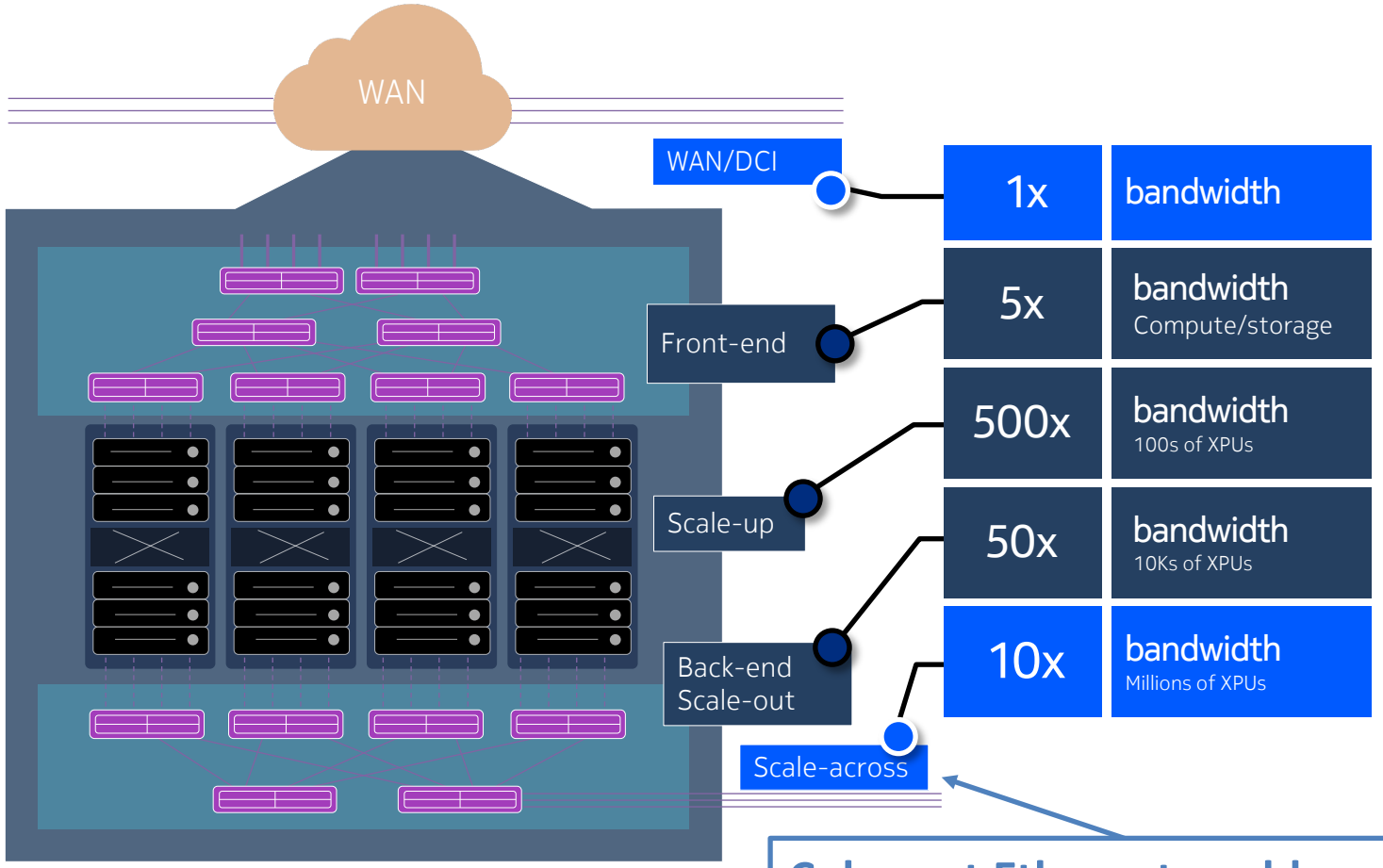


Scale-across (across buildings)

Coherent Ethernet for Super-Clusters

- To reduce training times, hyperscalers are deploying increasingly large XPU clusters
  - Million-XPU-scale systems with millions of connections!
- Space and power constraints make it impractical to deploy super-clusters in a single building
- Scale-across networking addresses this challenge
  - Interconnects XPU clusters in different buildings to create a “cluster of clusters”
  - Spans multiple power domains for resiliency
- Coherent Ethernet offers low latency, high-capacity links that make efficient use of fibers in the outside plant

Scale-across: extending back-end AI fabrics



1x	bandwidth
5x	bandwidth Compute/storage
500x	bandwidth 100s of XPUs
50x	bandwidth 10Ks of XPUs
10x	bandwidth Millions of XPUs

Increasing speeds

Electrical going to optical

Increasing power optimization

Coherent Ethernet enables creation of XPU super-clusters

Coherent Ethernet for fiber efficiency & reach

- AI campus are predicted to scale to 5M fibers, with the largest campuses exceeding **20M fibers**
  - Extremely challenging for Hyperscale construction projects and fiber management technology
  - Maximizing the transmission capacity of each strand is key!
- Shorter reach scale-up/out connections addressed by 800G IMDD FR4/LR4 today
    - Enables 800Gb/s transmission over a pair of SMF using WDM
- 1.6Tb/s transmission rate is the next capacity step
  - IMDD at 400G/lane is constrained by CD, OSNR, penalties at very high symbol rates – perhaps 2km will be feasible?
  - Coherent Ethernet will provide 1.6Tb/s transmission over a single pair of fibers to cover the LR applications (10 km)
- Coherent + WDM should be considered to maximize fiber utilization

IEEE 802 ITU

800G+ Interfaces & Applications

Application	Technology	Unamplified	Amplified	Single Span	Multi Span
Scale-up, Scale-out	Ethernet IMDD	X		X	
DCI, Scale-across (campus, metro)	Ethernet Coherent	X		X	
DCI, Scale-across (campus, metro)	OIF ZR		X	X	
DCI (regional, long-haul)	OIF ZR+		X		X
Telco DCI services: campus, metro, regional, long-haul	ITU-T OTN	X	X	X	X

IEEE 802 ITU

Coherent Ethernet for the AI Era

P802.3dj COHERENT PMDS

P802.3dj Coherent PMD Overview

PMD	Bit Rate	Reach	Modulation	~λ	~Baud Rate	FEC Code	FEC Type
800GBASE-LR1	800Gb/s	10 km	DP-16QAM	1312 nm	124 GBd	RS(544,514)+BCH(126,110)	Concatenated
800GBASE-ER1-20	800Gb/s	20 km	DP-16QAM	1548 nm	118 GBd	OFEC BCH(256,239)	Segmented
800GBASE-ER1	800Gb/s	40 km	DP-16QAM	1548 nm	118 Gbd	OFEC BCH(256,239)	Segmented

- 800GBASE-LR1 PHY composed of the following sublayers:
  - 800GBASE-R PCS sublayer
  - LR1 Inner FEC sublayer
  - PMD sublayer
- 800GBASE-ER1/ER1-20 PHY composed of the following sublayers:
  - 800GBASE-R PCS
  - 800GBASE-ER1 FEC
  - 800GBASE-ER1 PMA
  - PMD

TIES users TD link:
[316-GEN] LS/i on IEEE P802.3dj draft D3.1

Figure 184-2 on page 597

800GBASE-LR1 Inner FEC

- Simple and lightweight
- Per flow BCH(126,110) encoding of 32 RS-FEC encoded flows
- Also contains the functions traditionally associated with a PMA
- *See Cl184 for more info*

800GBASE-LR1 Optical Specifications

TIES users TD link:
[316-GEN] LS/i on IEEE P802.3dj draft D3.1

Figure 185-5 on page 633
Table 185-5 on page 630
Table 185-6 on page 632

- Provides a fiber optic cabling model to define channel characteristics
  - Makes use of “black link” methodology that was developed initially in Q6/15
  - Optical channel characteristics, including operating distance, insertion losses, dispersion characteristics, DGD max, min return loss
  - Optical fiber and cable characteristics, including nominal fiber specification wavelength, max attenuation, zero dispersion wavelength, max dispersion slope
- Transmitter optical specifications
- Receiver optical specifications
- *See Cl185 for details*

TIES users TD link:

[316-GEN] LS/i on IEEE P802.3dj draft D3.1

Figure 186-3 on page 649

800GBASE-ER1 FEC sublayer

- Based on ITU-T FlexO-8e-DO developed in Q11/15
- RS-FEC is removed and a stronger FEC is applied to mitigate transmission errors over the fiber
  - The 32 PCS lanes at the FEC service interface are aligned and reordered to form 2 PCS flows with 16 lanes each.
  - Each PCS flow undergoes a RS(544,514) decode
  - A single serial stream of 257-bit blocks is created
  - Distributed to 8 FEC flows that are GMP mapped into tributary slot to create 800GBASE-ER1 frame (similar to FlexO frame format)
  - 8 flows are interleaved and encoded with an extended BCH(256,239 code), i.e. OFEC
- Result is presented to the PMA sublayer
- *See Cl186 for more details*

TIES users TD link:
[316-GEN] LS/i on IEEE P802.3dj draft D3.1

Figure 186-13 on page 670

800GBASE-ER1 PMA

- Polarization distribution
- Inserts FAW, TS, PS and RES symbols
- DAC of I/Q components of X-pol and Y-pol
- *See Cl186 for details*

800GBASE-ER1/ER1-20 Optical Specifications

TIES users TD link:
[316-GEN] LS/i on IEEE P802.3dj draft D3.1

Figure 187-5 on page
Table 187-5 on page
Table 187-6 on page

- Channel model and characteristics, Tx specs, and Rx specs defined as for LR1
- ER1-20 and ER1 differ only in launch power, Rx sensitivity, and channel characteristics (due to different operating distances)
- *See Cl187 for details*

New! APSU for Coherent PMDs in D3.1

- Normative Annex 178B defines an Autonomous Path Startup
  - ILT function: a training protocol and training frame format for IMDD PMDs
  - RTS function: coordinates the switch to DATA mode for all interfaces in a path
- If one or more interfaces requires 178B in the path, all ISLs need to support the RTS function
- D3.1 adds APSU support for LR1 and ER1 coherent PMDs

TQM for Coherent

- Quantify how much penalty is introduced by the transmitter if its signal is received by an ideal receiver
  - A single figure of merit that attempts to encompass multiple impairments
  - The goal is to allow vendors to make their own trade-offs in how they design their transmitter
  - IMDD used TDECQ, which has a similar objective
- The Extended Transmitter Constellation Closure (ETCC) metric is used to define the quality of 800Gb/s DP-16QAM transmitters
  - Data shows good correlation (see maniloff_3dj_01a_2411)
  - Significantly helps with interoperability
  - Is designed as a test that can be used to qualify transmitters in minutes (Annex 185A)

Coherent Ethernet for the AI Era

WHAT'S NEXT FOR COHERENT ETHERNET?

Future Coherent Ethernet PMDs to Consider

- 2km to 40km reach 1.6Tbps PMDs will be a must!
  - Scale-out, scale-across, and metro DCI applications
  - Should we consider something < 2km to 40km?
  - Consider working with loss budgets instead of reach
- n x 1.6T PMDs for scale across
  - Optimize efficiency of fiber utilization in the outside plant
  - FR4/LR4 (8?) type PMDs in the O-band, WDM, concatenated FEC
  - Should we consider MCF medium?
  - Channel model needs to be developed
- We will need a 3.2T rate (MAC) by 2030
  - Driven by 400G/lane maturity

Summary

- Coherent Ethernet at 800G will play a significant role for metro DCI and emerging scale-across applications
- Fiber efficiency is a top concern for AI architects and coherent Ethernet offers the most efficient use of each strand
- Coherent Ethernet outperforms IMDD in reach and in data transfer rate per fiber, allowing it to “fill the gap” in scale-out networks as IMDD moves to 400G/lane


IEEE
802


ITU