Proposal to adopt additional objectives to better support HPC/AI/ML applications - Part 1

Kent Lusted, Intel Corporation
David Ofelt, Juniper Networks

Supporters (of Part 1 and Part 2)

- John Johnson, Broadcom
- Gary Nicholl, Cisco
- Jeff Maki, Juniper Networks
- Zvi Rechtman, NVIDIA
- Eugene Opsasnick, Broadcom
- Karl Muth, Broadcom
- Peter Winzer, Nubis
- Brian Welch, Cisco
- Tony Chan Carusone, Alphawave Semi
- Vasu Parthasarathy, Broadcom
- Michael Xin, CIG
- Scott Sommers, Molex
- Massimo Sorbara, Global Foundries
- Piers Dawe, NVIDIA
- Ernest Muhigana, Lumentum
- Phile Sun, Credo
- Paul Brooks, Viavi Solutions
- Mabud Choudhury, OFS
- Drew Guckenberger, Maxlinear
- Reza Eftekhar, Amazon Web Services
- Jamie Gaudette, Microsoft
- Cathy Huang, Intel
- Sheng Zang, Broadcom
- Vipul Bhatt, Coherent
- Adee Ran, Cisco
- Howard Heck, Intel
- Jon Lewis, Dell Technologies
- Roberto Rodes, Coherent
- Shimon Muller, Enfabrica
- David Piehler, Dell Technologies
- Samuel Kocsis, Amphenol
- Shawn Nicholl, AMD
- Mark Sikkink, HPE
- David Malicoat, Malicoat Networking Solutions
- Karen Liu, Nubis
- Atul Srivastava, NEL
- Chris Doerr, Aloe Semiconductor
- Rick Rabinovich, Keysight
- Sridhar Ramesh, Maxlinear
- Chongjin XIE, Alibaba
- Craig Thompson, NVIDIA
- Sam Sambasivan, AT&T
- David Chen, AOI
- Eric Maniloff, Ciena
- Hacene Chaouch, Arista
- Frank Chang, Source Photonics
- Matt Brown, Alphawave Semi
- Nathan Tracy, TE Connectivity
- Tom Palkert, Samtec/MACOM
- Adam Healey, Broadcom
- Jason Chan, Arista
- Jim Theodoras, HGGenuine
- Jim Weaver, Arista
- Cathy Liu, Broadcom
- Chris Lyon, Amphenol
- Ted Sprague, Infinera
- Mike Li, Intel
- Yi Sun, OFS
- Rich Mellitz, Samtec
- Priyank Shukla, Synopsys
- Jeff Rahn, Meta

Outline

- Market requirements
- Current Task Force Status
- Proposed Path Forward
- Proposed Objectives
- Next steps

Market Requirements

- Multiple markets require 200 Gb/s-based technologies to be defined, to become available for early deployment and to interoperate
- The recent rise in importance of HPC/AI/ML clusters based on Ethernet technologies creates an additional prioritization of a few specific Physical Layer specification characteristics
  - Distinct from the equally important traditional Ethernet networking applications
  - For example, HPC/AI/ML workloads are known to be power and latency sensitive; creating an opportunity to define distinct Physical Layer specifications that are better aligned to that application
- Supporting network operational considerations is always a path to success
  - Clear definitions, naming, usage, ability to manage all factor in to solution
  - Stakeholders includes procurement, qualification, test & validation, network deployment, network debug

HPC/AI/ML - At a High Level

- Modern HPC/AI/ML is all about disaggregated, but high-throughput, tightly-coupled computing
  - Interconnect must be very cost effective, very low power, and very high throughput
- HPC/AI/ML machines (clusters) commonly reside inside data centers
- Data center facilities are power limited
- High machine density presents power problems
  - Goal is to maximize completed work in the available power
- Power (and cooling) are high cost

HPC/AI/ML Relevance to P802.3dj

- Key value target for AI/ML deployments is to maximize throughput and reduce power
- Maximizing the throughput of the AI/ML deployment depends on many factors:
  - Cluster quantity - facility limitations
  - Cluster capacity - therefore higher bandwidth interconnect, switching capacity and GPU/NIC throughput is important
  - Cluster performance and flexibility to workloads - architecture and latency are considerations
  - Workload size and granularity

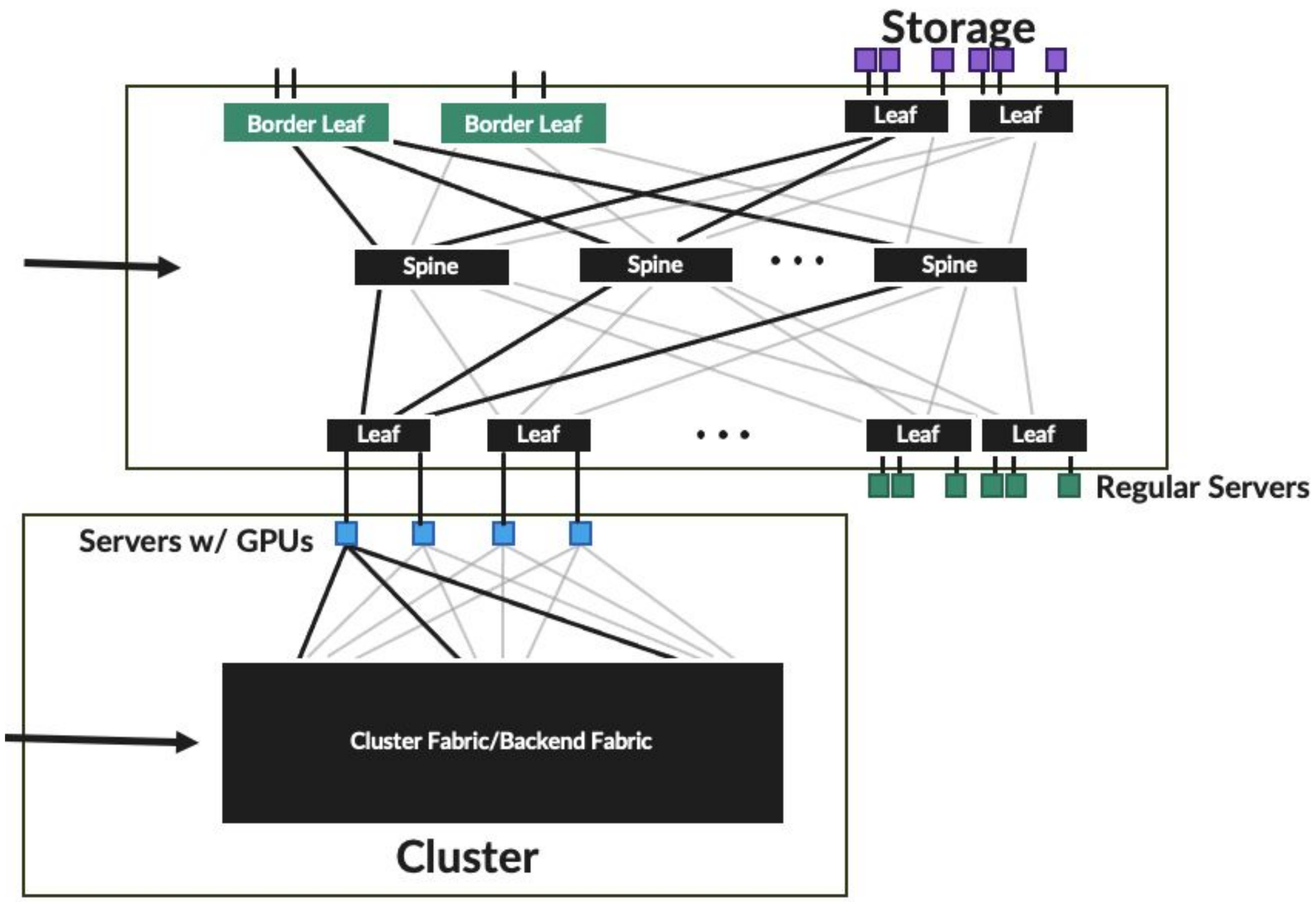
Why Low(er) latency?

- The lower the latency, the easier it is to get higher performance with:
  - Smaller workloads
  - Irregular workloads
  - Smaller units of computation
  - Smaller unit of communication
  - More frequent communication
  - More independent compute resources
- Latency predictability (aka “long tail”) is also very important
  - Can frequently hide latency with pipelining if things are predictable
- Note that low(er) latency needs to be balanced with all our other economic factors - power, cost, etc.
- IEEE P802.3dj should not specify latency

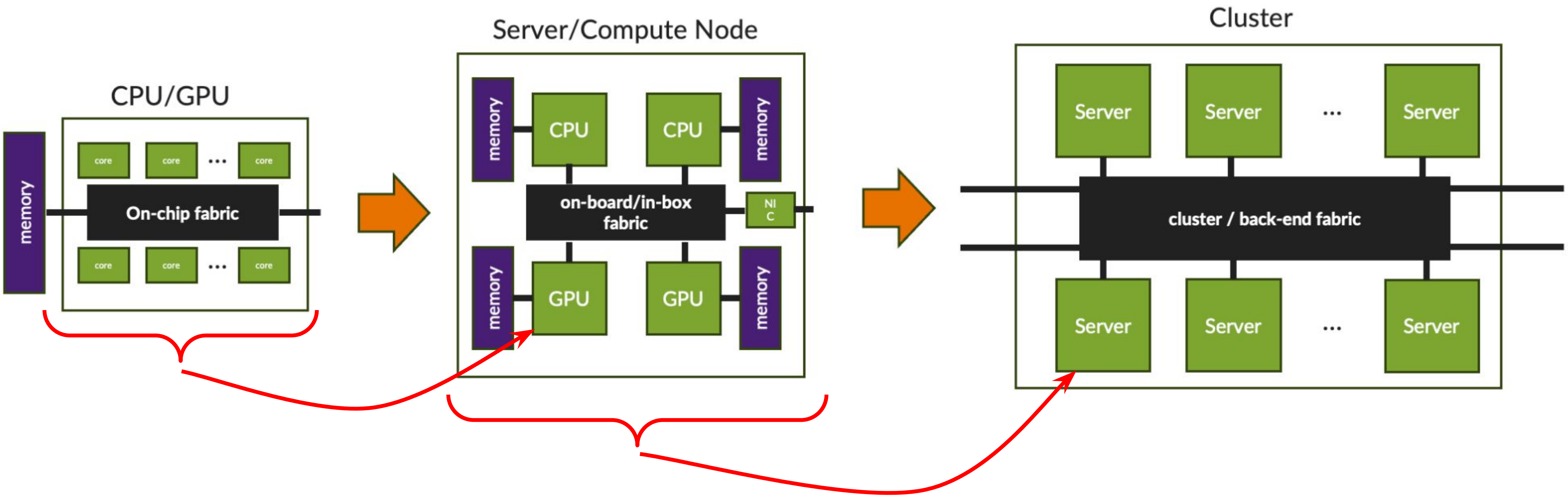
An Example Modern Datacenter

Front-end Fabric

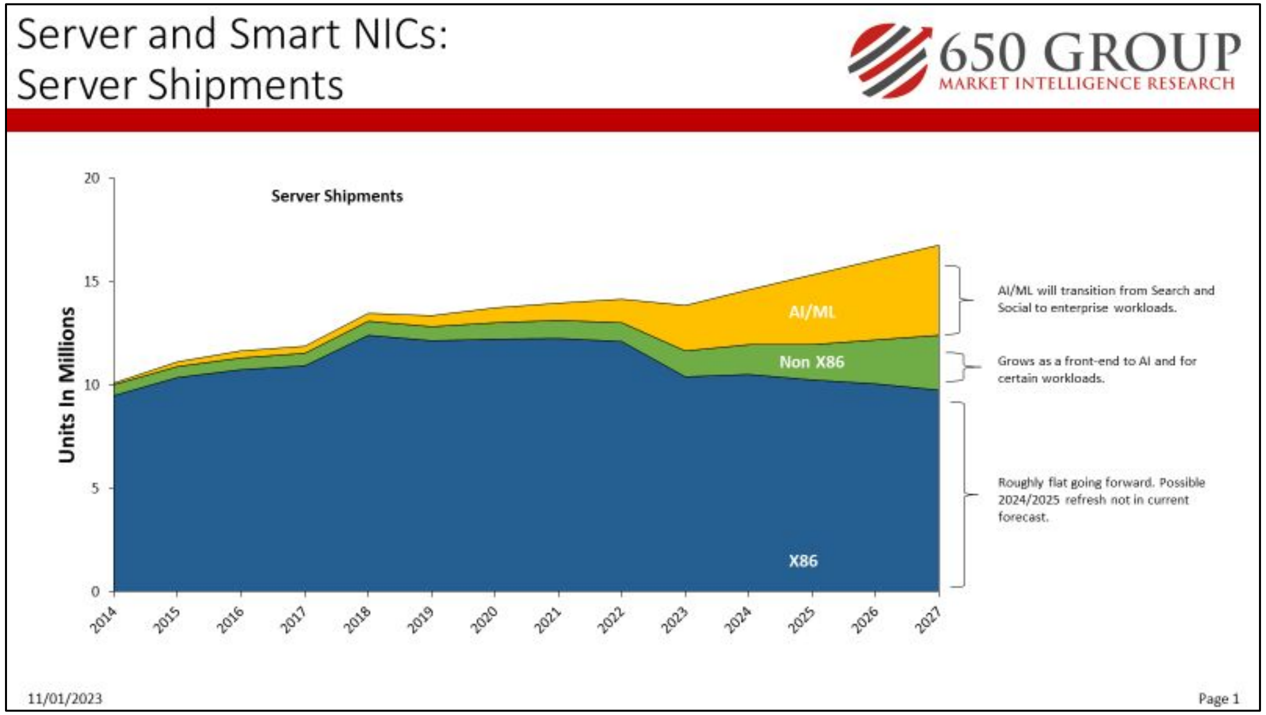
Back-end Fabric



Example Compute Hierarchy



Growing Importance of AI/ML on Ethernet



Server and Smart NICs:
Server Shipments

650 GROUP
MARKET INTELLIGENCE RESEARCH

AI/ML will transition from Search and Social to enterprise workloads.

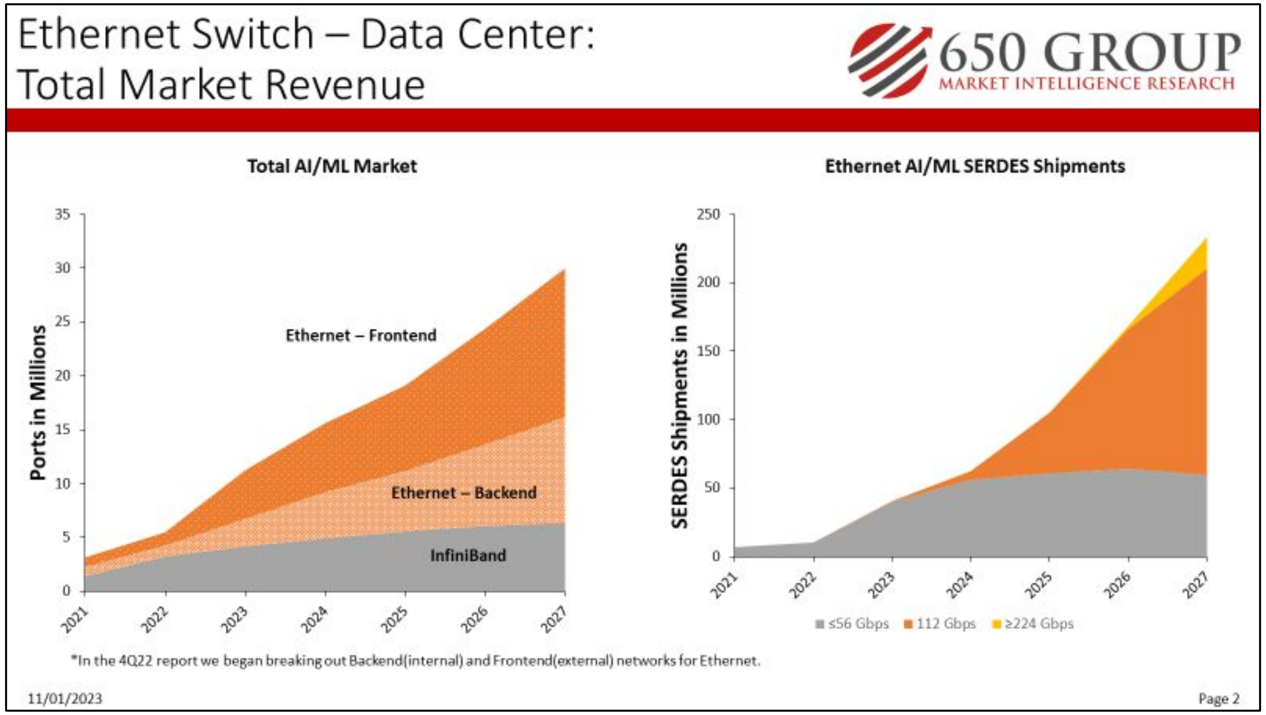
Grows as a front-end to AI and for certain workloads.

Roughly flat going forward. Possible 2024/2025 refresh not in current forecast.

11/01/2023

Page 1

Courtesy - Alan Weckel, 650 Group



Ethernet Switch – Data Center:
Total Market Revenue

650 GROUP
MARKET INTELLIGENCE RESEARCH

*In the 4Q22 report we began breaking out Backend(internal) and Frontend(external) networks for Ethernet.

11/01/2023

Page 2

Courtesy - Alan Weckel, 650 Group

Task Force Considerations

Observations To Date

- Consensus in supporting both FECo and FECi by the Task Force has been established
  - Mode_FECo: Optical link runs with RS(544,514) FEC protection.
  - Mode_FECi: Optical link runs with RS(544,514) FEC protection operating as an outer code, supplemented by Hamming(128,120) FEC protection operating as an inner code.
  - FECi logic already adopted
- Currently no consensus on how to achieve the goal of supporting both
- The inability to achieve consensus on an approach is potentially due to:
  - A difference in understanding of market focus
  - A difference in perspective on timing (immediate implementation concerns vs future implementation possibilities)
- Additional PMD technical analysis is not a considerable factor affecting this debate:
  - Baselines have been proposed for all options (rates/reaches/fiber variant)
  - Contributed technical results always help to shore up technical feasibility aspects.
- Need to define a path forward where everyone can be successful

Paths Forward

Option A: "Single PHY" approach

- Define a single physical layer solution with both FECo and FECi
- Problems to solve: what is mandatory vs. optional within the solution? What is the performance of the solution in each FEC mode? How does the end-user have certainty in the operational FEC mode? How to switch between FEC modes? General concern that this is stretching the precedent of a single Physical Layer specification
- No consensus within Task Force has been achieved around a way to define this

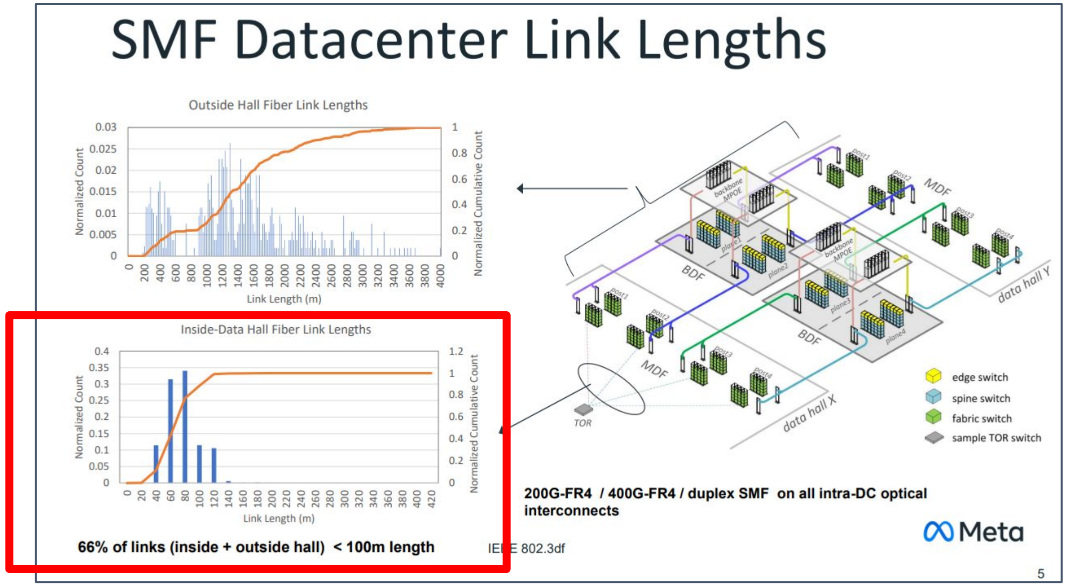
Option B: This proposal (previously known as "Two PHY" approach)

- Add new objectives that address the specific market requirements of HPC/AI/ML
- Define new Physical Layer specifications for the new objectives
- For the currently adopted objectives, proceed with FECi-only based architecture/logic and PMD proposals
  - It is a product implementation choice to include either or both solutions into a common design
- Problems to solve: define and adopt the new objectives and solutions for them

Proposed Additional Objectives

Additional Objectives Proposal

- The current adopted solutions are not well suited for HPC/AI/ML applications, yet important for front-end Ethernet networking use cases
- Need new objectives for Physical Layer specifications to address the unique HPC/AI/ML applications
- The new objectives need to be distinct
- This will require 5 new PHYs/objectives with shorter reach:
    - 200 Gb/s 1 pair – 250m reach
    - 400 Gb/s 2 pair – 250m reach
    - 800 Gb/s 4 pair – 250m reach
    - 800 Gb/s 4 λ – 250m reach
    - 1.6 Tb/s 8 pair – 250m reach



SMF Datacenter Link Lengths

66% of links (inside + outside hall) < 100m length

200G-FR4 / 400G-FR4 / duplex SMF on all intra-DC optical interconnects

IEEE 802.3df

https://www.ieee802.org/3/dj/public/23_09/welch_3dj_02a_2309.pdf

New proposed objectives

1. Define a physical layer specification that supports 200 Gb/s operation:
   - over 1 pair of SMF with lengths up to at least 250 m
2. Define a physical layer specification that supports 400 Gb/s operation:
   - over 2 pairs of SMF with lengths up to at least 250 m
3. Define a physical layer specification that supports 800 Gb/s operation:
   - over 4 pairs of SMF with lengths up to at least 250 m
4. Define a physical layer specification that supports 800 Gb/s operation:
   - over 4 wavelengths over a single SMF in each direction with lengths up to at least 250 m
5. Define a physical layer specification that supports 1.6 Tb/s operation:
   - over 8 pairs of SMF with lengths up to at least 250 m

Next Steps

Next Steps

- Review proposed updates to CSD. See lusted_3dj_06_2311
- Task Force consider adoption of new objectives and modified CSD
- If approved, Task Force leadership will progress the procedural work with necessary approvals
- Task Force needs to consider technical proposals to address new objectives and eventually adopt something
  - see next slide

Potential Solutions for the New Objectives

Option 1: Base any adopted solutions on bypassing the FECi Convolutional Interleaver

- known reduction of latency

Option 2: Base any adopted solutions on FECo

- known reduction of latency
- known reduction of power
- demonstrated technical feasibility

Option 2+n: Something else?

Other Task Force considerations after this proposal?

- Explore a need for and method to switch between optically compatible PHYs
  - Building upon ghiasi_3dj_01a_2309.pdf, mehta_3dj_01_2309.pdf, and brown_3dj_01_2311
  - Task Force should continue to review
- New nomenclature will be needed for additional Physical Layer specifications

Summary

- Trying to find a consensus path to move the Task Force forward
- There is a market need for Ethernet better suited for HPC/AI/ML applications
  - Current objectives are important for front-end Ethernet networking use cases
- Additional distinct objectives are proposed for the HPC/AI/ML use cases
- Choose a solution for the new objectives after objectives are adopted

Thank you!