# Data Center Ethernet
# Call for Interest

Jonathan Thatcher

16 March 2004

# Data Center Ethernet -- CFI

New applications of Ethernet-interconnected cluster and grid computing present requirements for improvement in performance, efficiency, and reliability. Over the last decade, performance improvement simply meant increasing Ethernet's bandwidth by a factor of 10 every four years -- a dramatic outpacing of Moore's Law. 10-Gigabit Ethernet crashed into that wall in 2003 and is only now beginning to enter the market in a serious way. The time is opportune to investigate means to advance Ethernet's performance beyond what is possible at either 1-Gig or 10-Gig through protocol efficiency enhancements. It is proposed that a Study Group be created to investigate means to reduce latency, decrease congestion and inevitable packet loss, and provide features that allow upper layers to more effectively utilize Ethernet networks and thus enable an IEEE 802.3 compliant means to build next generation systems.

# Agenda

- Introduction
- Facilitation of Enterprise Clustering
  - *David Flynn (Linux Networx)*
- Ethernet for High Performance Computing
  - *Moray McLaren (Quadrics); presented by Jonathan*
- Streaming Media
  - *Mike McCormack (3COM)*
- Market Trends
  - *Brian MacLeod (Wildwood Associates)*
- Study Group Purpose & Direction
  - *Jonathan Thatcher*

# “Call for Interest”

## 4. 802.3 Study Groups

### 4.1 Function

- The **function of a Study Group is to complete a defined task with specific output and in a specific time frame** established **within which they are allowed to study the subject**. Once this task is complete the function of the SG is complete and its charter expires.
- The normal function of a 802.3 Study Group (SG) is to draft a complete PAR and five criteria (see 7.2) and to gain approval for them at WG 802.3, 802 EC, IEEE New Standards Committee (NesCom) and the IEEE Standards Board.

### 4.2 Formation

- **A SG is formed when enough interest has been identified for a particular area of study within the scope of WG 802.3.**

# The Request

- **Defined Task → Study the Subject**
  - ***Brainstorm, Study, & Recommend improvements to increase Ethernet performance, efficiency and effectiveness for 802.3***
- **Specific Output**
  - ***Specific project recommendation***
  - ***Obtain authorization***
    - ***IEEE 802.3 -- Objectives***
    - ***IEEE 802 EC -- 5 Criteria***
    - ***NesCom and IEEE-Standards board -- PAR***
- **Specific Time Frame**
  - ***Ready for PAR approval in Nov '04***

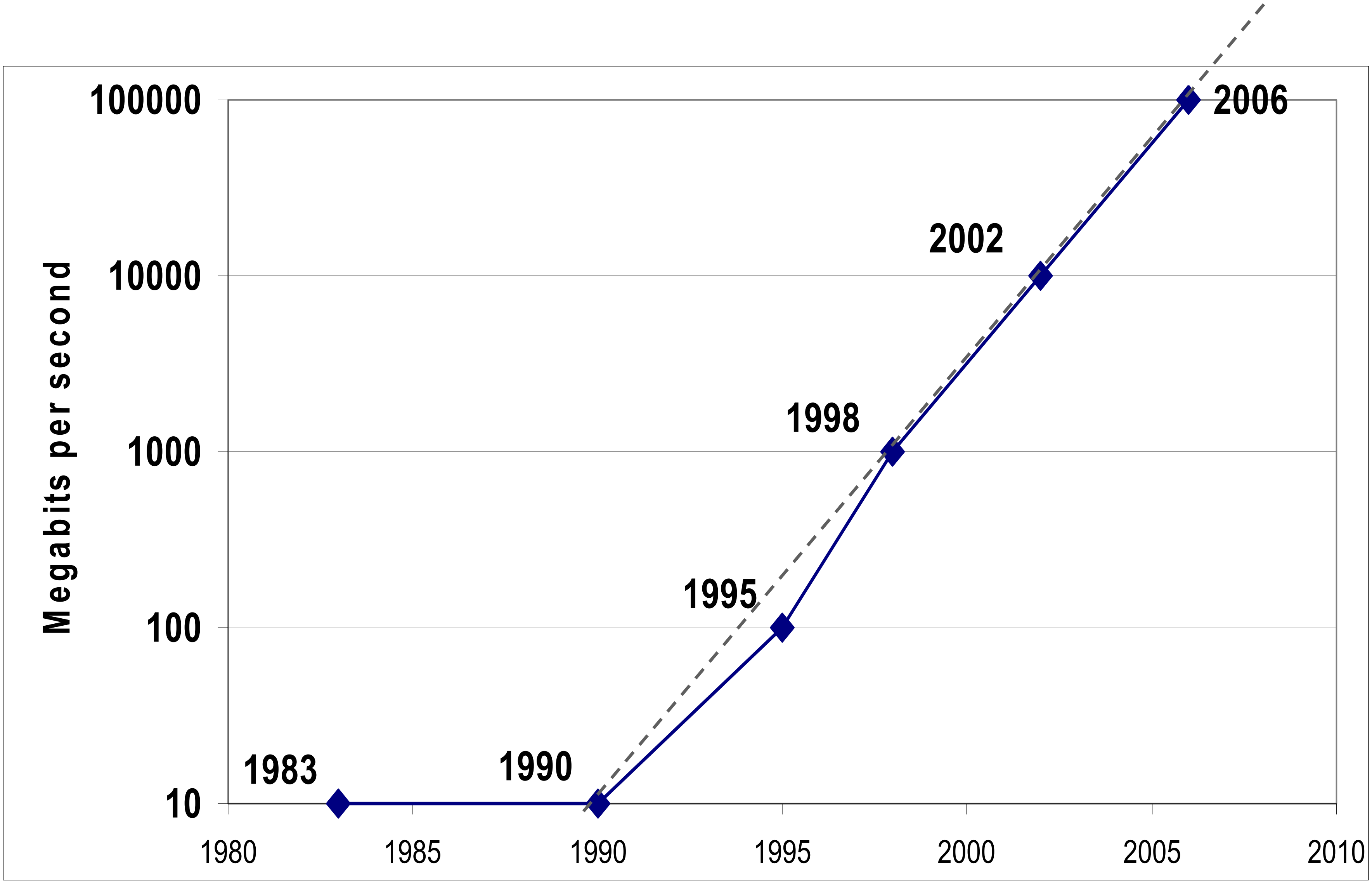
# Data Center Ethernet CFI Isn’t

- **Isn’t ramrod of some predefined solution**
  - ***Err on side of openness and ambiguity***
  - ***Trust the process; trust 802.3 membership***
- **Isn’t brought by some secret cartel**
  - ***No support has been requested***
    - ***Exception: some experts have been told that their skill and expertise would be helpful***
  - ***Political risk? Yes -- But, see first bullet***
- **Isn’t an attempt to grab 802.1 turf**
  - ***Over time, 802.1 participation and perhaps even a supporting project may be essential***

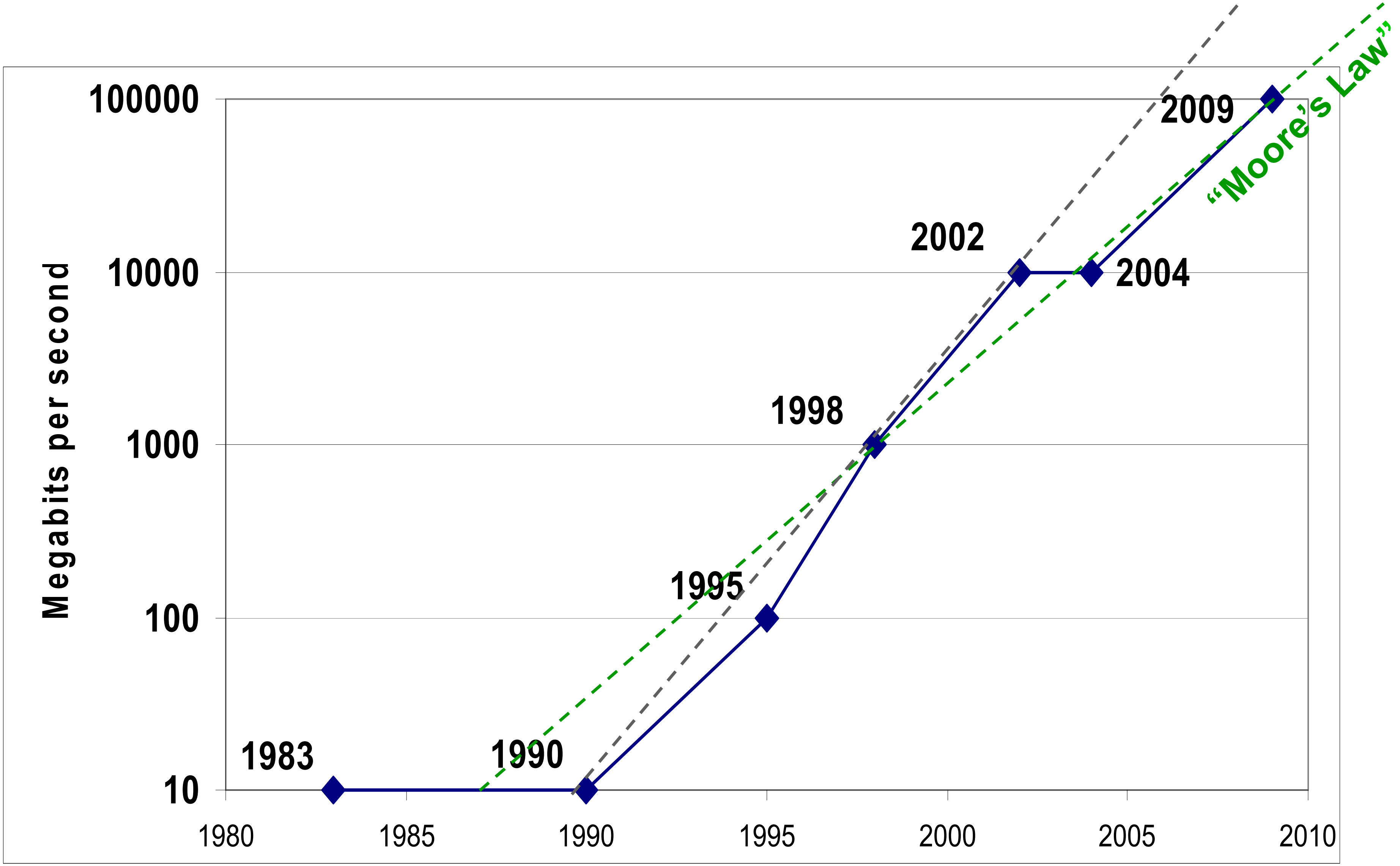
# Motivation(s)

- Customer demand for Ethernet connectivity that meets performance requirements for growing cluster and blade market is increasing rapidly.
  - *Other technologies (most proprietary) attempting to solve problem (nature abhors a vacuum)*
- Can’t just turn up the dial by 10X again
  - *Brute force has crashed into “Moore’s Law”*
  - *Need to be more efficient going forward*

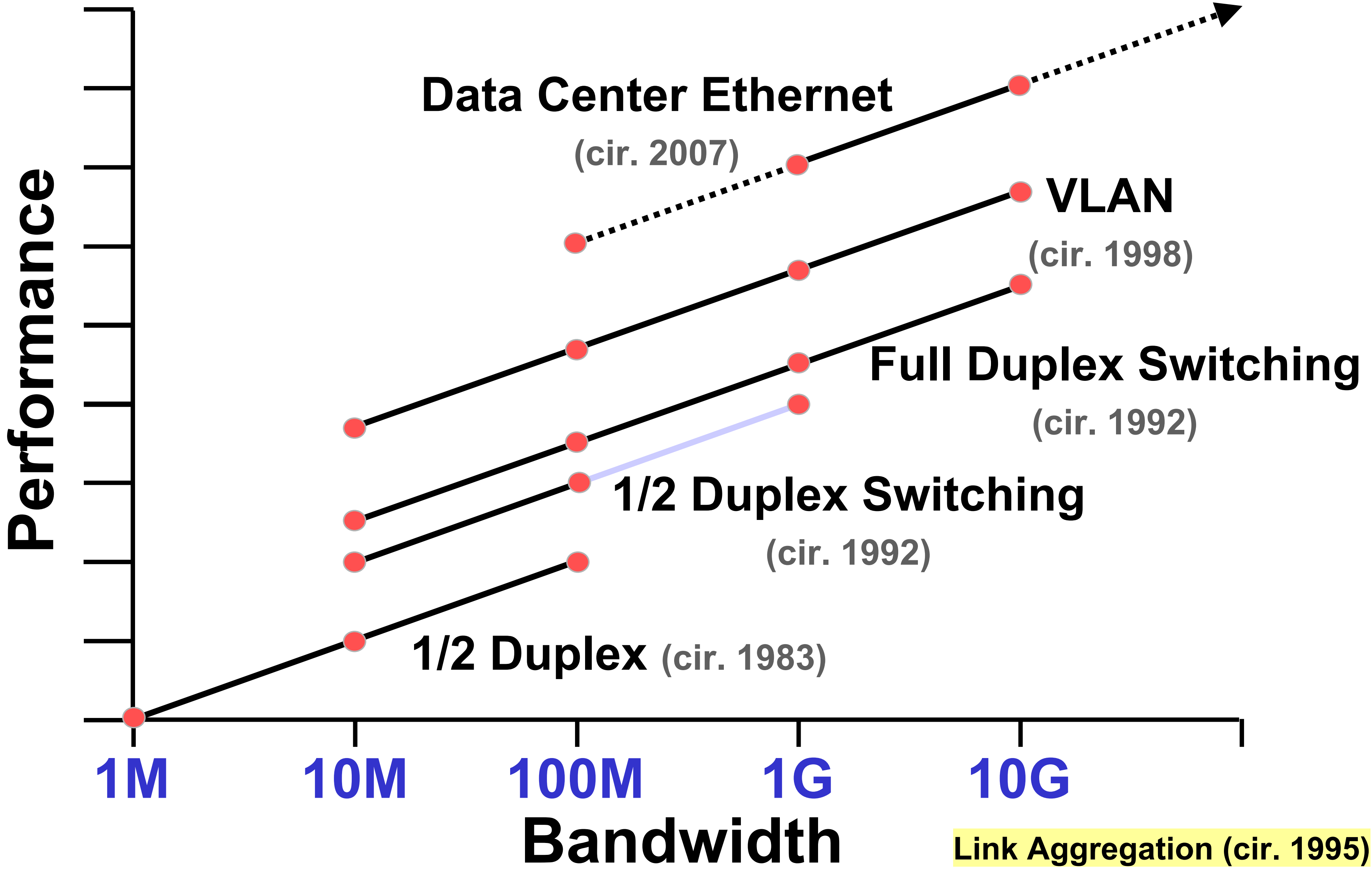
# Ethernet Speed Progression



# Real Ethernet Progression?



# Ethernet Performance Upgrades



Link Aggregation (cir. 1995)

# Presentations

# Example Areas for Exploration

- Enhanced flow control
- Higher reliability (availability)
  - *Data*
    - *Guaranteed Delivery (ACK/NACK)*
  - *Rapid Failover*
    - *Trigger adjustable per application requirement*
- Assist higher layer protocols
- Preemption
  - *Lower latency*
  - *Higher efficiency*

# Flow Control

- **Ideally, flow control is done with combination of techniques:**
  - ***End-to-end***
    - ***Traffic management***
    - ***Not done by 802.3***
  - ***Link-by-link***
    - ***Congestion management***
    - ***Done by 802.3***
    - ***Done better if more granularity***
      - ***Yes, this implies segregation of buffer space***

# Flow Control Options

- Flow control by:
  - *Link (PAUSE)*
  - *Priority*
  - *VLAN*
  - *DA*
  - *SA*
  - *Other*
  - *Combination*
- Important consideration: relationship to requests from / mapping to upper layers

# Conflict Space

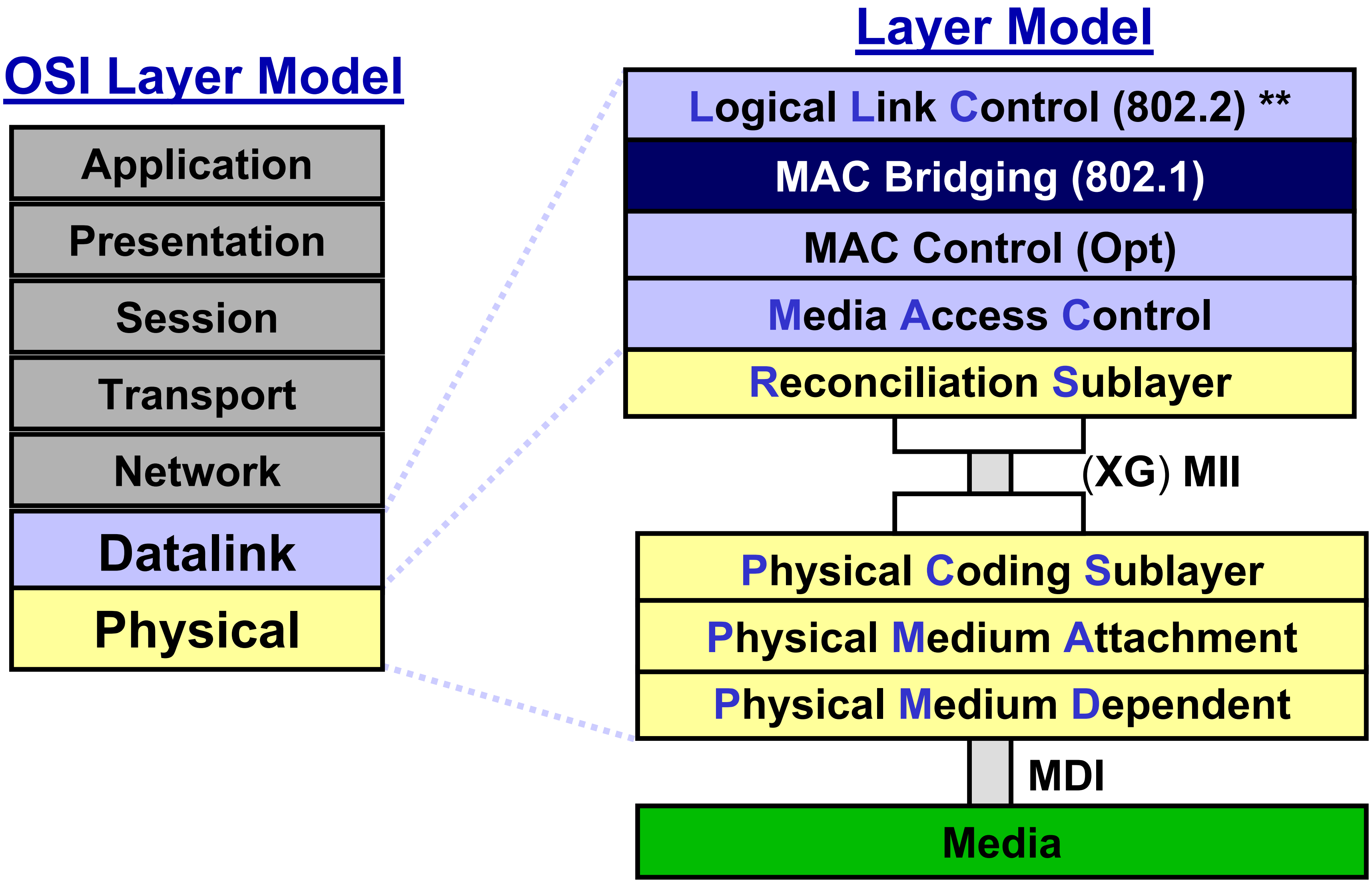
## Technical

- *Efficiency → (super) jumbo packets*
- *Latency → smaller packets, cut through*
- *Reliability →*
  - *Link-by-link ensured delivery → Cost*
  - *Rapid failover → Hair trigger vs robust link*
- *Cost → single network*
- *Etc.*

## Organizational

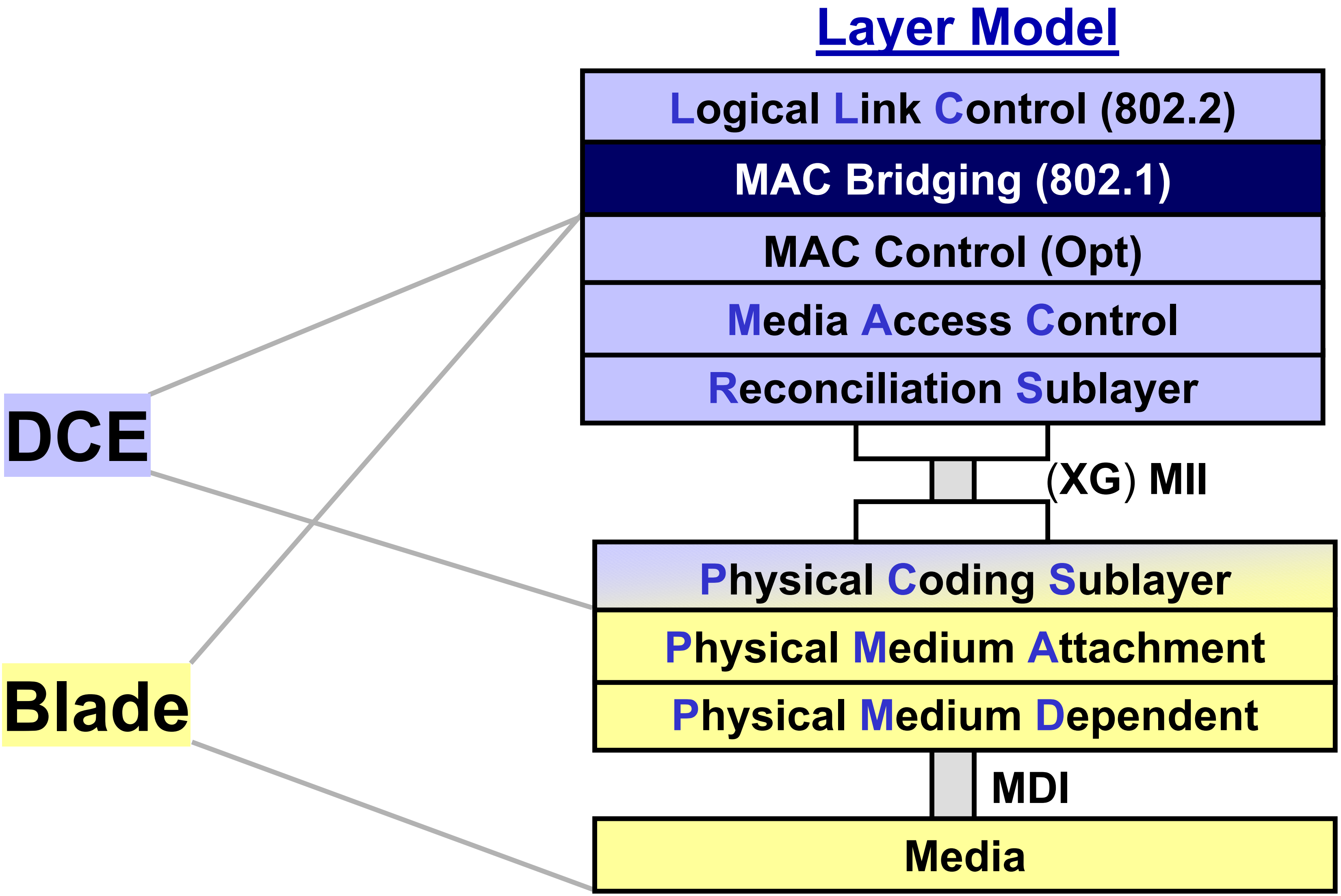
- *Blade vs DCE*
- *802.1 vs 802.3*

# Ethernet Sublayers

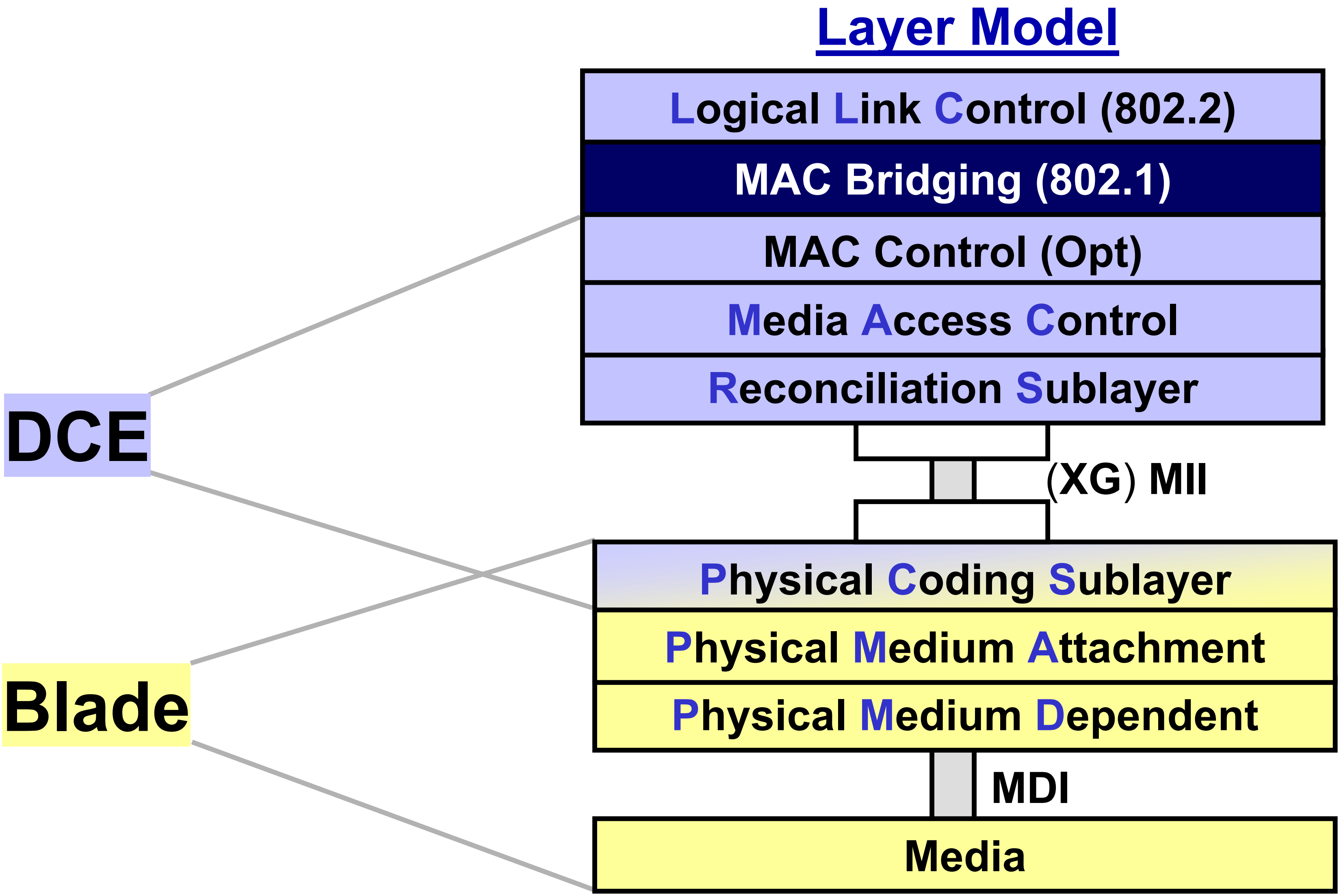


** This is technically not a correct placement of 802.2, but useful as a placeholder

# DCE & Blade Coordination?



# DCE & Blade Coordination!



# Relationship to Backplane SG

- What is important to Data Center Ethernet
  - *General logic solution, not Backplane optimized*
- What is important to Backplane:
  - *Move forward this week*
    - *PAR to NESCOM*
    - *5 Criteria to 802 Exec*
    - *Objectives to 802.3*
  - *Need to have "flow control problem" solved*
    - *PAUSE not good enough*
    - *Some desire to not be dependent on DCE*
- Possible path forward
  - *No Flow Control (logic) in Blade PAR*
  - *Tentatively place objective within Blade*

# Study Group Bounds

- Backward compatibility with existing Ethernet
- Stay within scope of 802.3
  - *No transaction layer added to 802.3*
    - *May investigate 802.3 functions that assist higher layer(s)*
    - *Potential close working relationship(s) with outside organization(s) – includes 802.1*
- Principal focus on 1GbE and 10GbE
  - *Freedom to look at 100 MbE*
- No new PMA/PMD
  - *No increased speed (e.g. 40 G or 100 G)*
  - *No new media (e.g. parallel optics)*
- Minimal changes to PCS & Clause 4 (4A)
  - *Avoid if reasonable*

# Straw Poll

- **Individuals that will participate in Study Group:**
- **Companies that will participate in Study Group:**
- **IEEE 802.3 should study means to improve Ethernet performance**
  - ***All***
  - ***802.3***
  - ***802.1***