Initial thoughts on EEE for 400 GbE
P802.3bs 400 Gb/s Ethernet Task Force

Steve Trowbridge
Alcatel-Lucent

Historical Implementations of EEE

- IEEE Std 802.3az-2010 specified EEE for 100BASE-TX, 1000BASE-T, 10GBASE-T, 1000BASE-KX, 10GBASE-KX4, 10GBASE-KR, and the XGXS extension for 10 Gb/s PHYs. Original mechanism now known as “deep sleep”
- IEEE P802.3bj defines EEE for 100GBASE-KR4, 100GBASE-CR4, and 100GBASE-KP4 interfaces and adds EEE capability for electrical PHYs and internal electrical interfaces added by IEEE Std 802.3ba-2010. Distinction first introduced between “deep sleep” and “fast wake” modes of operation.
- IEEE P802.3bm defines EEE for 100GBASE-SR4 and 40GBASE-ER4, and adds EEE capability for optical interfaces added by IEEE Std 802.3ba-2010 and IEEE Std 802.3bg-2011. This was the first application of EEE to optical interfaces, which use only the “fast wake” mode of operation

Should EEE for 400GbE follow the historical path?

- Classical approaches to EEE involve entering a “Low Power Idle” (LPI) state during periods of link inactivity
- Since the P802.3bs only has objectives to develop new optical PHYs, the rationale previously used to decide that only “fast wake” mode would apply
- But for a 400 Gb/s Ethernet PHY, is “periods of link inactivity” really the “fat rabbit” opportunity for reducing power usage?

Observations about 400 GbE Traffic Patterns

- The majority of Ethernet PHYs at lower rates (10/100/1000 Mb/s) are used to connect to end stations including laptop computers, point-of-sale terminals, etc.
- The higher the rate, the less likely it is that an interface connects to an end station. Some 10 Gb/s may connect to a server, moving to 40 Gb/s for the highest-end servers, but essentially all 100 Gb/s interfaces are carrying large aggregates of lower-rate services. In P802.3ba, aggregation was considered to be the “killer app” justifying the 100 GbE development
- 400 GbE surely will continue the trend, with the overwhelming majority of the links used for aggregation, with even a higher ratio of the PHY rate to the average service rate and a larger number of services per interface.
- The higher the aggregation ratio, statistically, the lower the percentage of time one will find for a period of “link inactivity” as the opportunities to enter a LPI state and save power

So where else could power be saved?

- Some 400 Gb/s PHYs will be installed because the application requires 400 Gb/s
- Many more 400 Gb/s PHYs will be installed because the application requires more than 100 Gb/s and LAG is not satisfactory or desired for the application.
- Some 400 Gb/s PHY applications may be in network scenarios where there is a known limit on the traffic less than 400 Gb/s: for example, a 200 Gb/s router-to-tranport link where the transport traverses a 200 Gb/s DP-16QAM optical carrier. With that network configuration, it is known that this particular link will never exceed 200 Gb/s

Opportunity for 400 Gb/s Power Savings

- Like all rates starting with 40 Gb/s, 400 Gb/s is expected to be comprised of parallel physical and logical lanes
- An 400 Gb/s PHY that is expected to carry less than 400 Gb/s of traffic could turn off a subset of the lanes
- The lane counts for all 400 Gb/s PHYs may not be the same (16, 8, or 4 physical lanes are likely candidates). Presumably one doesn't want a different capacity granularity for different PHYs
- Depending on the architecture, there may be a modularity to the interface and not all physical or logical lanes would be independent. For example, gustlin_400_02a_1113.pdf would have a modularity of 100G, being comprised of four 100G FEC instances
- Could consider turning off lanes or groups of lanes in a granularity or modularity that is common to all 400 GbE PHYs according to the architecture

Proposal

- Consider providing a EEE mechanism for 400 Gb/s PHYs based on turning off a subset of the lanes for links that use less than 400 Gb/s of traffic. For example, a 200 Gb/s link could use approximately half the power of a 400 Gb/s link.
- Granularity should be according to the architectural modularity common to all interfaces
- Use cases to be considered:
  - Interfaces that have a known maximum rate less than 400 Gb/s
  - Interfaces where the amount of traffic may vary – is there a need to adjust the capacity of an individual link in service? Loose analogy – a 10/100/1000 Mb/s Ethernet PHY will negotiate a rate with its partner when the link is brought up, but once negotiated the rate never changes in service

THANKS!