System Vendor Concerns

Jeff Maki – Juniper Networks
Gary Nicholl – Cisco Systems
David Ofelt – Juniper Networks
Rob Stone - Broadcom
802.3ck ad-hoc call 2018-08-15

Outline

Evaluating Tradeoffs

Module Thoughts

Evaluating Tradeoffs

Background

During the July 802.3ck meeting, we had a series of straw polls where the group was asked to pick between different approaches.

Many found it hard to give a meaningful answer, due to not having enough information to evaluate

- *"We are being presented with a menu with no ~~prices~~costs"* – Adam Healey

The straw polls did lead to some good discussions!

We want everything (and more)!

For each new generation of technology, both the system vendors and end users:

- Want everything they had in the previous generation
- Don’t want to have to change anything

But– the new stuff should to be:

- Cheaper
- Lower power
- Denser
- More reliable
- Simpler

Reality

The reality is 100Gb/s SERDES are going to be difficult and we are not going to be able to do everything we did at 50Gb/s for the same cost.

- Ex: Classic 1m PCB-based backplanes have already been voted off the island

Exactly what we give up, and how we manage the various facets of cost is complicated and may be different for each person/company.

Knowing how much “better” an idea makes things is not enough- we need to know how much the improvement costs.

If there are multiple routes to getting something to work, which we pick can be a complicated decision and may be very vendor/end-user specific.

What is “cost”?

It isn’t just money…

- It is anything that goes into the economics of products based on 100Gb/s SERDES
- Frequently includes both “Economic Feasibility” and “Broad Market Potential”

Cost is at least:

- Monetary cost
- Complexity
- Power
- Area
- Physical Volume
- Manufacturability
- End-user acceptance
- Reusability
- Reliability
- Development Cost
- Opportunity Cost
- Management Cost
- Interoperability Cost
- Etc.

Request

For analysis work going forward that is evaluating tradeoffs

- Clearly state what is the base case (this has value 1.0)
- Show deltas from that base case and show relative “cost”

It is best if everyone (eventually) uses the same base cases.

- Many are pretty standard or obvious
- Over time, folks will settle on some common ones
- When in doubt- ask around

If only relative numbers are used, it can be hard to compare across vendors

- System vendors and end users can get full details under NDA and compare

Note- comparing actual monetary cost, even in a relative way, is tricky

- When in doubt – ask for help

Module Thoughts

Expectations

My read of the end-user expectations:

- For pluggable modules is that they are expecting the density improvements seen over the past many years.
  - QSFP – 40Gb/s -> 100Gb/s -> 200Gb/s -> [400Gb/s]
  - QSFP-DD -> 200Gb/s -> 400Gb/s -> [800Gb/s]
- Cost per Gb/s goes down over time
- Total bandwidth goes up and so does radix when looking at faster rates.

When this can happen will be different for:

- Different vendors
- Different PMDs
- Different systems

Why denser?

Fixed form-factor systems

- These can usually get larger
  - Ex: a 1U TOR moving to 2U
- At some point, it gets harder and/or more expensive to build

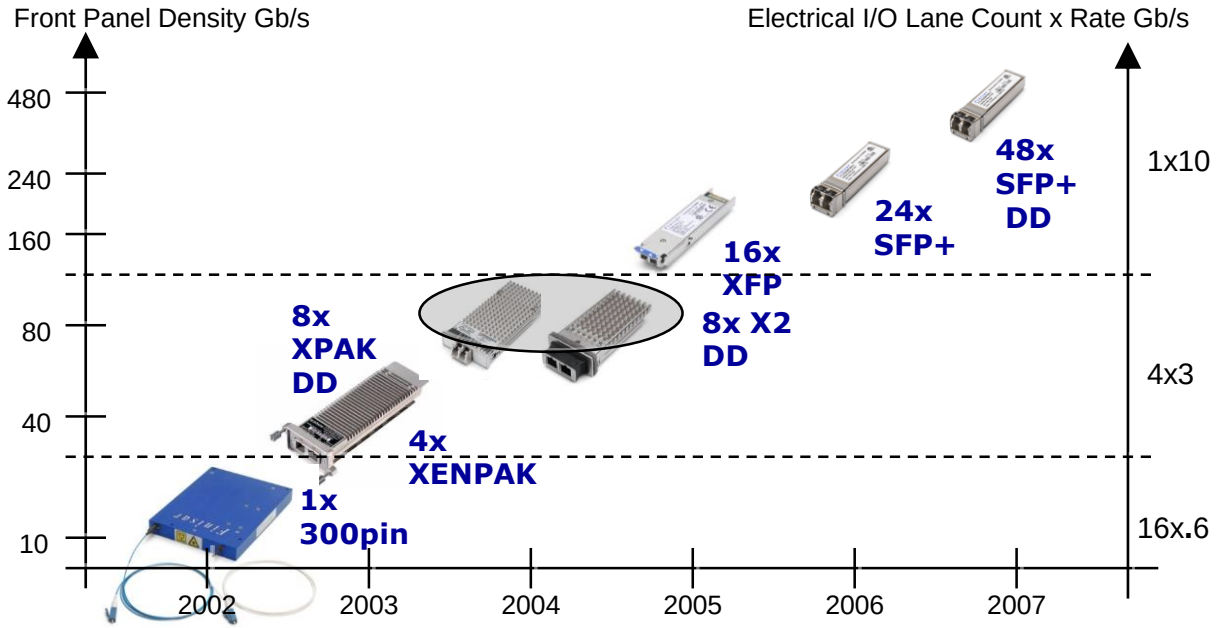
Line-card based systems

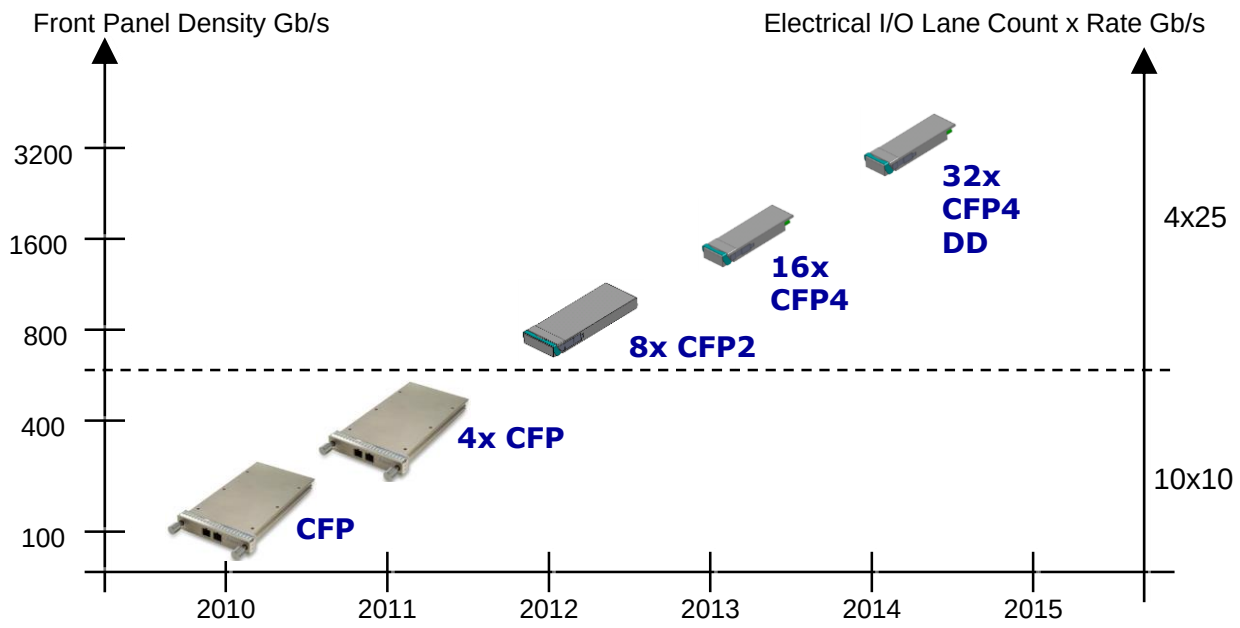
- Fixed line card pitch
- Once the system is face-plate limited, denser modules are needed
  - Everyone is face-plate limited!
- Do these continue to be practical to build in their current form?

At some point, we'll have >400Gb/s Ethernet rate

10GbE & 100GbE Module History

Don’t underestimate the march of technology progress!





Source: Chris Cole, Finisar

Module Power

Module power is definitely a concern

That said- it isn't a fixed constraint-

- Improvements are constantly being made
  - Have to keep EF and BMP in mind
- Different vendors and/or markets can have different capabilities
  - Ex: systems that meet NEBS can have a harder time than those that don't need to

Many ways to deal with hot modules- Ex:

- Improved TIM
- Liquid Cooling
- Careful vendor selection
- Limitations in how many of certain modules
- Fan improvements
- Etc.

QSFP-DD/OSFP with 4x100Gb/s electrical

There has been discussion of a QSFP-DD/OSFP module using 4x100Gb/s AUI, not 8x100Gb/s AUI

Allows high-radix single-chip switch chips based on 100Gb/s SERDES to avoid 1:2 demux parts

These systems would not be able to use first-gen 8x50Gb/s based modules

- Not enough links without board-level 1:2 demux
- Can use QSFP modules

4x100Gb/s capable modules could also support 8x50Gb/s

- This works with 50Gb/s or 100Gb/s based switches
- Useful PMDs would need to be re-released with this capability

End-user acceptance may differ by market

- Hyperscale my be fine – largely homogeneous, single point in time deployment
- Core/Edge/Enterprise more complicated – more desire to have pay-as-you-grow and very heterogeneous

800Gb/s Modules

Premise of 100Gb/s C2M is that we'll (eventually) get to 800Gb/s QSFP-DD/OSFP modules

- …And 400Gb/s QSFP modules

Industry is assuming this will be possible at some point

- Might take a while
- Different PMDs will happen at different points in time
- Not all technologies/vendors may be relevant
- If not possible – the pluggable module era is over – we need to start preparing end-users ASAP
- If not possible or it takes too long or is too expensive - new technologies will emerge

Conclusions

- Need to establish some kind of cost comparison for the menu of system options being discussed to assess relative merits of solutions
  - Keep within IEEE guidelines on cost discussions
- Many varied end user applications for this technology, BMP should be assessed by application
  - General case “one size fits all” may end up fitting none
- Keep in mind the standard needs to be relevant over a long time period, technologies will evolve and improve
  - Example: power reduction via silicon process improvement

Thanks!