Low power MMF objective for High Performance Computing and End-of-Row applications

Piers Dawe	IPtronics
Brad Booth	Dell
Oren Sela	Mellanox

Supporters

- Marc Verdiell — Samtec
- Shimon Muller — Oracle
- Petar Pepeljugoski — IBM
- Phil McClay — TE Connectivity

Contents

- Inevitable circumstances
- HPC*, EoR and other applications
- Options for this future
- Low power MMF objective

* List of abbreviations on last slide

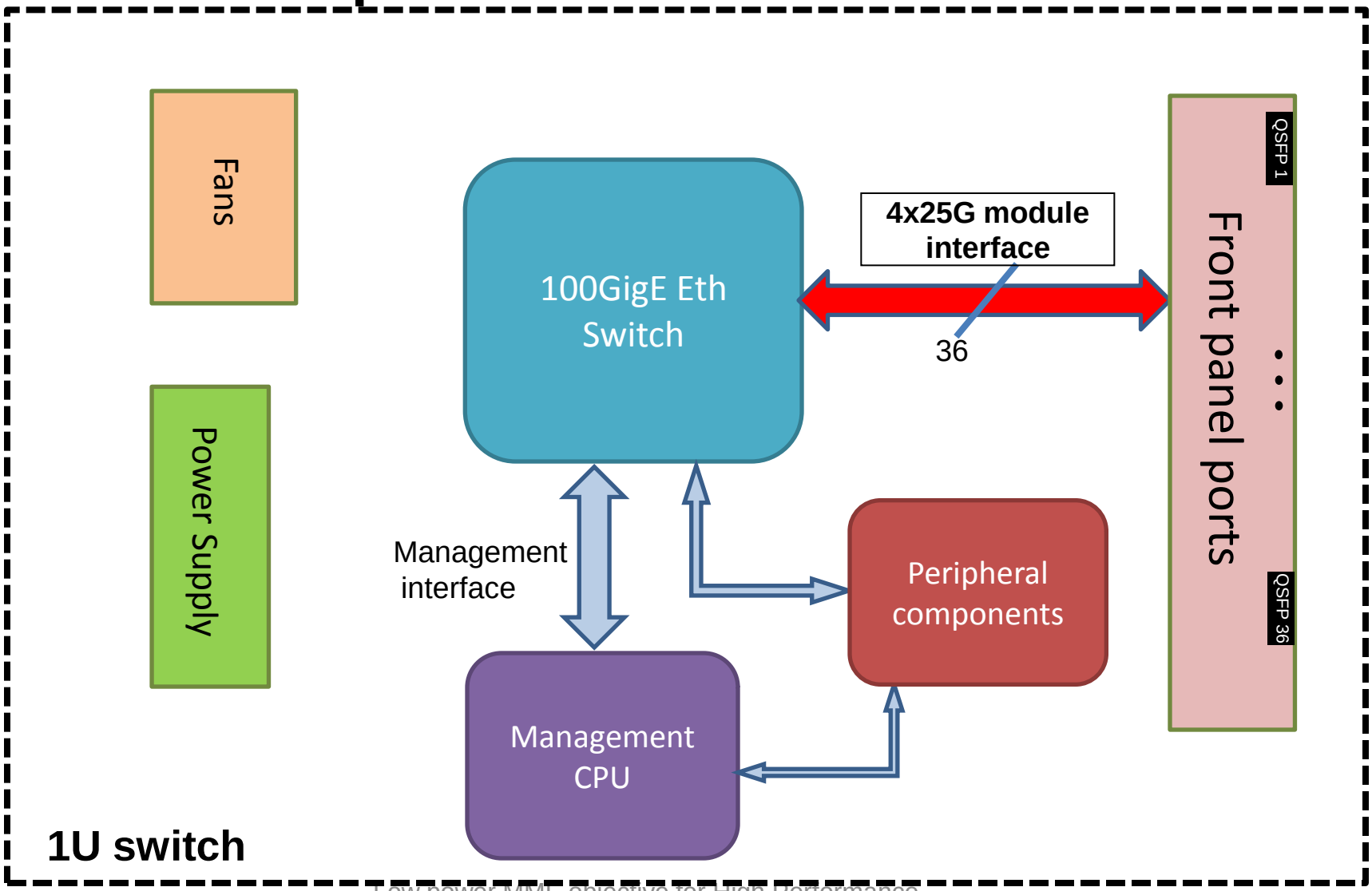
Inevitable circumstances

- HPC and data centers won't tolerate unnecessary power
  - Will not pay a tax to benefit others
- Most links are short
  - More so in future, with denser computing
  - For many projects, all links are known to be short
- So the low power thing will get designed and made
  - Not just HPC for power and reach
  - Large and very-large data centers (Web 2.0)
- The low power thing will be suitable for Ethernet use
- It will get used for Ethernet
  - **There is a broad market potential for high density, low power shorter reach MMF**
- It will be **QSFP+ sized** or smaller (SFP++)**, not CFP/CFP2/CFP4 sized**

How much is low power?

- QSFP+ has power levels 1.5, 2, 2.5, 3.5 W
- However that's too much for a card with 16 or 18 or 36 QSFP+ ports, as is expected
- See 802.3bj sela_01a_0112 and following slides
- **1.5 W to 2 W per module**, about the same as for 40G
- Higher power would translate into lower port density which means higher cost

Top-of-Rack switch



Top-of-Rack switch

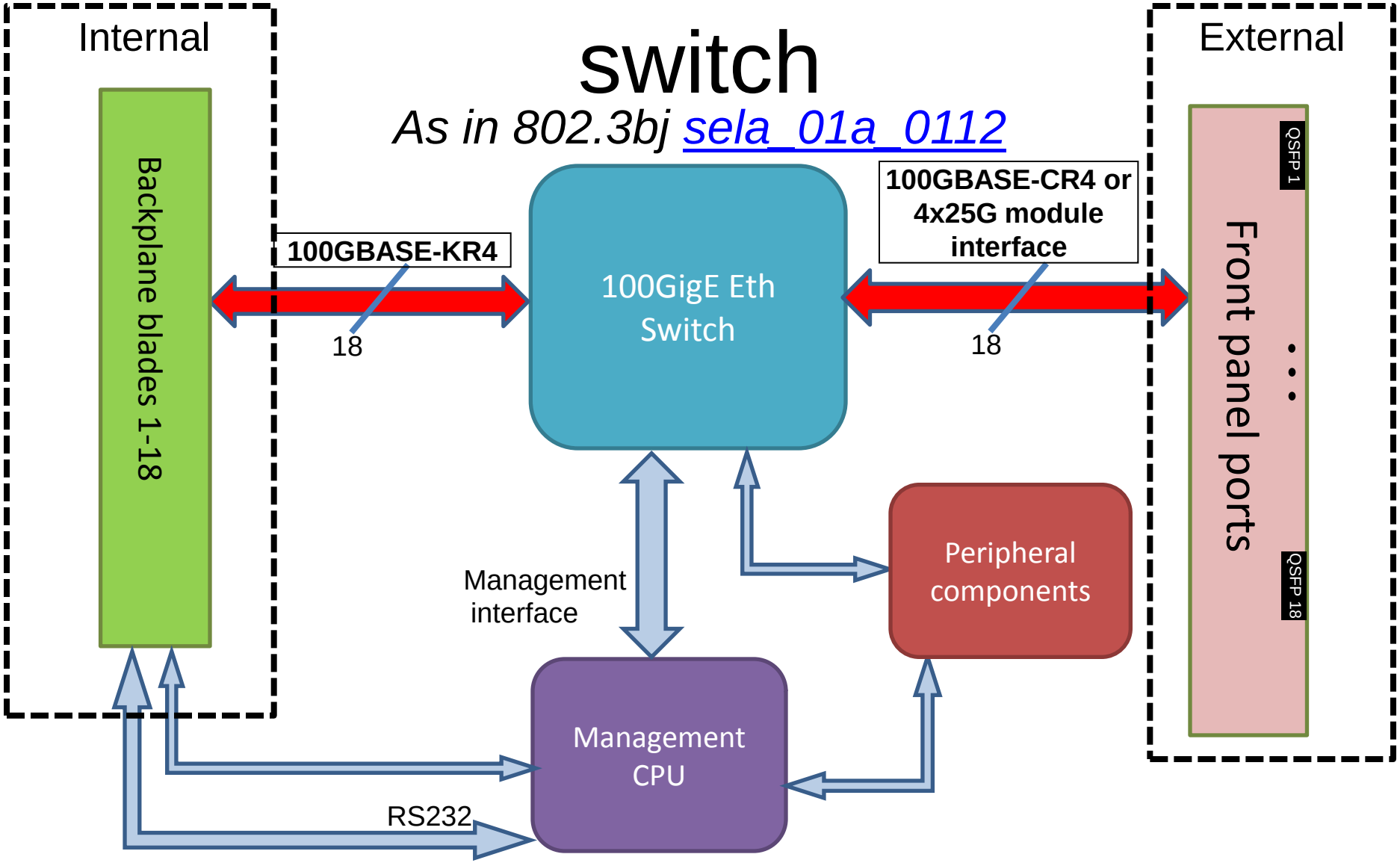
- Density
  - 40GigE dense switches are 36 ports
  - Density reduced from the 1GigE or 10GigE – can't reduce the port count further
- Total Power Budget is 250 W to 280 W
  - Thermal limitations
  - Power supply limitations
- Analysis based on 40GigE ToR switch
  - Some component's power may increase for 100GigE
  - No external memory for the switch - more power
  - Without FEC
- For 250 W budget-
  - Power consumption excluding the optics = 144 W
  - Max power for optics < (250-144)/36 = **3 W**
- **Module power determines number of ports, hence cost**

Component	**40G**	**100G**
Switch ASIC (include 36 4x25G ports)	75	75
Fans [1]	3-24	3-25
Management CPU	10	10
Misc	9	9
Power supply (in)efficiency	10% or 25	10% or 25

[1] Fan power consumption increases with fan speed

Core switch – modular leaf switch

As in 802.3bj [sela_01a_0112]



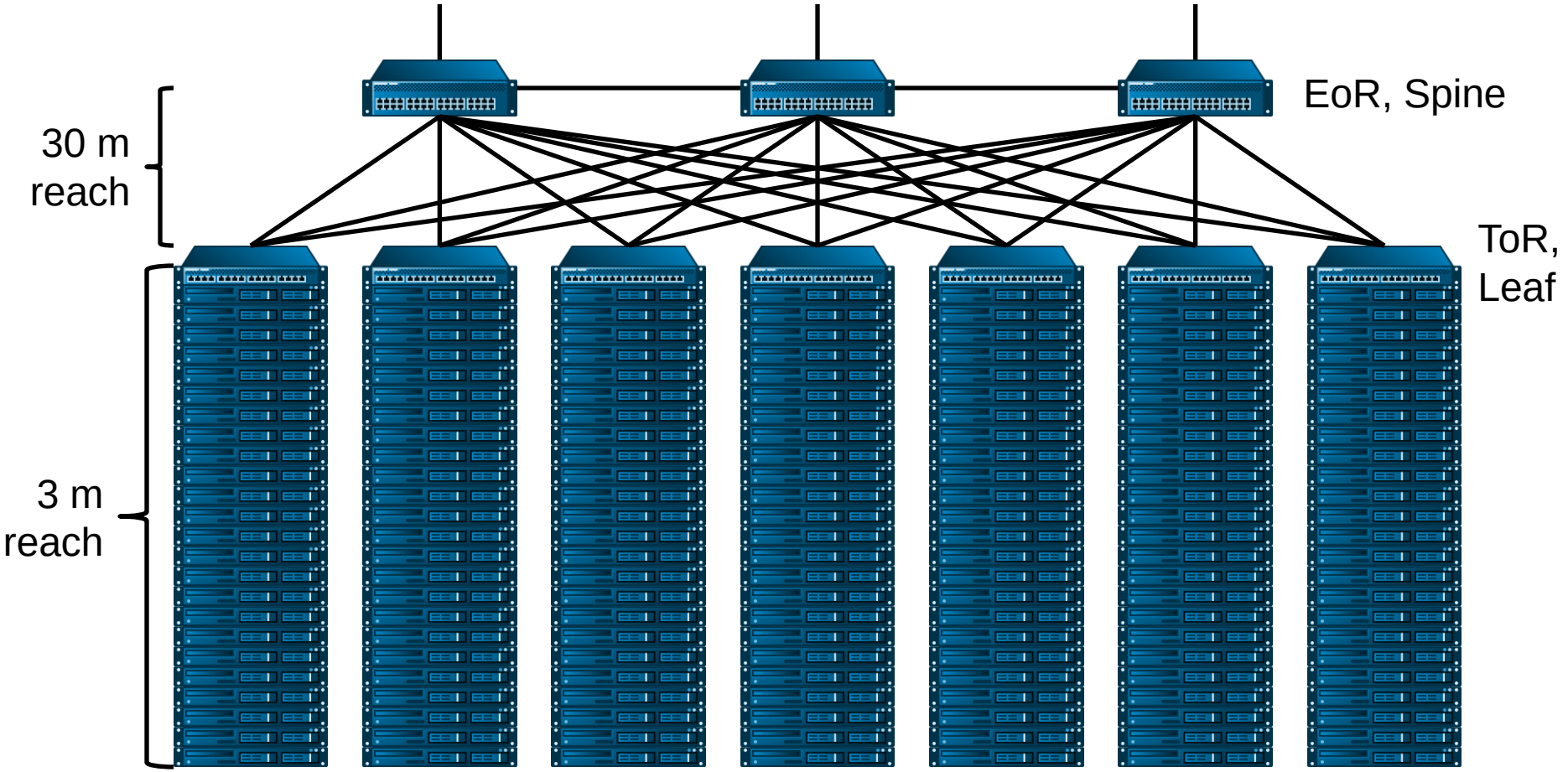
Core switch – modular leaf switch

- 18 port leaf switch
  - Fully non-blocking – 18 internal ports to spine
- Total Power Budget 150 W to 160 W
  - Thermal limitations
  - Other limitations may reduce this towards 140 W
- Analysis based on 40GigE modular leaf switch
  - Some component's power may increase for 100GigE
  - No external memory for switch – more power
  - No external PHY for the backplane – more power
  - Without FEC for external ports, with FEC for backplane
- For 160 W budget-
  - Power consumption excluding the optics is 110 W
  - Max power for optics < (160-110)/18 = **2.7 W or less, e.g. 2 W depending on power supply**
- **Module power determines number of ports, hence cost**

Component	**40G**	**100G**
Switch ASIC (include 18 KR4 and 18 4x25G ports)	85	85
Fans [2]	0	0
Management CPU[2]	0	0
Misc	9	9
Power supply (in)efficiency	10% or 15	10% or 15

[2] Fans are powered from the Chassis, CPU management is done for the chassis of the core switch

EoR, Leaf-Spine

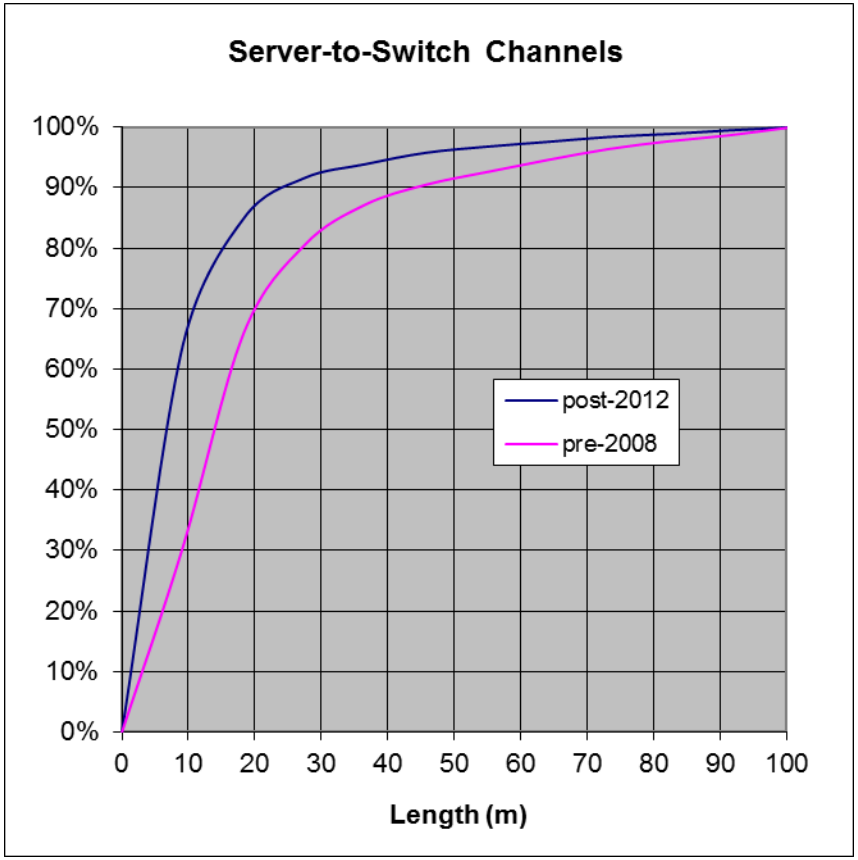


EoR, Leaf-Spine (cont)

- Servers to ToR is typically 3 m
  - Copper technology is commonly used
  - 10G in the current generation, 40G support by 4x PCIe gen3
- ToR to EoR (or leaf to spine)
  - Full mesh network to create a fabric with short path bridging, TRILL, etc.
  - Multiple connections with less than 30 m reach
  - No 100G copper technology to support
- If ToR is missing, links are to EoR switch

Supporting Data

- From Kolesar_Kalculator_12_01_25.xls
- Post 2012, server-to-switch channels are >90% of the links
- ~30% switch-switch links are <= 30 m
- Link cost needs to be small % over server cost
  - Equal to server cost is bad



Options for this future - summary

1. Two module types
2. A single module type, with options within or outside the module
3. Low power thing is active optical cable (AOC), never pluggable optics
4. 802.3 could moderate its MMF reach and concentrate economies of scale on the low power thing

- If there are two variants, 802.3 could specify both together, one after the other, or only one
- Variants might be PMDs or might not
- Variants can be interoperable
- See backup for more

What to do now?

- The Study Group does not need to answer these questions
  - If a variant that satisfies a longer reach objective also satisfies the cost and power requirements associated with the 30 m reach, we can all support a single variant
- The presenters and supporters just want the 30 m reach application space to be a consideration in the Task Force meetings, and the only way it can be considered is through an objective

What does the SG need to decide?

- We are here to write PAR, responses to 5 criteria, and some objectives
- Objectives are a project's contract with the Working Group
  - To keep the project working in the agreed direction
  - Should be a measure of success that cannot be subverted easily
  - Need to focus on the right thing
    - E.g. 802.3ae chose and met its 300 m objective but missed the main point of low end 10G fibre optics, which is data centers not campus
- The Study Group is not the Task Force
  - It is here to decide what problems are to be solved, not choose how to solve them
  - Clearly, one "problem" is the need for low power
  - **It needs an objective**

In other words...

- Explore a 4x25G parallel optical solution that targets 0-(20-30) m with MMF optimized for cost/power (SR-lite), in parallel with the expected SR 100 m solution. If the exploration shows that there are major cost/power gains to be achieved, consider a separate SR-lite specification.

What objectives ?

- For some, the promise of longer runs of fibre for 4-wide 100G is all that matters
- For others, cost and power on shorter runs is all that matters
- For many, no one criterion excludes all others from consideration
- It appears that there are **two markets, that need two objectives**
- Depending what we find when we get down to the detail in Task Force, one or two versions might be optimum
- "Known unknowns": we know we can't predict reach vs. complexity well at present
  - A reach-only objective is a resolution that we will spend power and money like water as needed to achieve it, because they don't matter
- We cannot find the answers in SG without spending the time a TF would take (another year?)
- The TF should be allowed to do its work; the SG should write appropriate objectives and not try to do the TF's job

Low power MMF objective

- **Define a power-optimized 4-lane 100 Gb/s PHY for operation over MMF**
  - (See other presentations for reach-oriented objective)
- Possible bridging objective:
- **The two MMF objectives *n* and *m* are to be achieved either while enabling interoperability between two sets of specifications, or by one set of specifications**
  - Alternative wording: The two MMF objectives *n* and *m* are to be achieved by specifications that are interoperable and may be the same

Backup

- Detailed discussion of possible future scenarios
- Abbreviations

Options for this future 1

1. Two module types
2. A single module type, with options within or outside the module
3. Low power thing is active optical cable (AOC), never pluggable optics
4. 802.3 could moderate its MMF reach and concentrate economies of scale on the low power thing

Options for this future 2

A. Just the low power thing, specified by 802.3
B. One of these things gets specified elsewhere then adopted by 802.3
C. One of these things gets specified elsewhere, used for Ethernet but not adopted by 802.3
   - Either way round, e.g.
   - 10GBASE-ER and 10GBASE-ZR
   - 10GBASE-SR and "SR lite"
D. 802.3 specifies both low power and high power versions

Discussion 1 of 6
1. Two module types

- Most component parts of the module are common
- Two types interoperable on the easier channel, if we so choose
- Two PMD types, two port type names
  - Might make a dual spec module like 10/100BASE-T, but not required to
- Differentiating feature might be FEC, EQ, or CDR
- Tx and/or Rx side optical and/or electrical specs might differ
- Use the power-hungry port type when the channel needs it

Option 2. A single module type, with options in or outside the module

- Single module type
- One or two PMD types
- One or two port type names
- Optional feature might be FEC, EQ, or CDR
- Tx and/or Rx side optical and/or electrical specs might differ
- Use the power-hungry mode when the channel needs it
- Possibility for interoperability but expect both ends of link would use same mode
- One mode could be optional, or both required

3. Low power thing is AOC, never pluggable optics

- AOC optimises cost and performance
  - Lowest power
  - Unretimed (limiting) interface expected with AOCs first
  - No optical connectors
- Some handling issues depending on length and ducting type
- Not for passive patch panels but could use active cross-connect panels
- Some issues connecting between different brands of equipment
- Inventory: need to stock a separate AOC for every brand, and for every cable length
- Two port types, one "unofficial" but can have standard specs
  - Might have same specs at electrical interface as pluggables
- Horses for courses: both AOCs and pluggables will be used
- Furthermore, low power pluggables are required by some markets
- **Conclusion: this option not as attractive as previously thought**

Option 4. 802.3 moderates its MMF reach Option A. Just the low power thing in 802.3

- Leave some links to be served by SMF or link extenders, or positioning boxes or patch-panels differently
- The proportion of links not served by MMF might increase (more cross-building 100G links) or decrease (new equipment more compact, new layouts take reach capability into account, faster Ethernet speed may be introduced)
  - Expect a net decrease

B. One of these things gets specified elsewhere then adopted by 802.3

- Costly overhead of two projects
- Likely that first project will make decisions that impede later project
  - E.g. 802.3ae nearly sunk SFP+; future projects will leave less margin for recovery on the table
- Incumbents of first version might be tempted to obstruct the second version's development of Ethernet
- A smaller group might be able to focus better on the technical issues
- If the two projects are some years apart, some technological progress might appear, but
- HPC will be moving to 25G/lane within the lifetime of this project

C. One of these things gets specified elsewhere, used for Ethernet but not adopted by 802.3

- Which one? The low power thing or the other thing?
  - In the past it has been the extended-reach thing (1000BASE-ZX, 10GBASE-ZR), *or* the low end thing (SR lite). Should 802.3 make official the higher volume one?
- Specified once elsewhere (e.g. by another industry body) or multiple variants (e.g. the various SR lite or 10GBASE-ZR like things)?
- Does it need e.g. MDIO registers that only 802.3 can sort out?
- Host silicon will need specs for electrical interfaces that might differ
- **At worst, leaves a broad market left in disarray trying to use "non-standard" solutions**

D. 802.3 specifies both low power and high power variants

- Can design for commonality where it makes sense
- Can be interoperable
- Decisions on what's mandatory and what's optional, port names, MDIO registers, any mechanism for interoperability, and robust specification all get public scrutiny

Abbreviations

- AOC active optical cable
- CDR clock and data recovery
- CFP/CFP2/CFP4 Pluggable form factors for 100 Gb/s
- EoR end of row
- EQ equalizer or equalization
- FEC forward error correction
- HPC high performance computing
- MMF multimode fibre
- PCIe gen3 Third generation Peripheral Component Interconnect Express
- SMF single-mode fibre
- ToR top of rack
- TRILL Transparent Interconnect of Lots of Links (an IETF standard)
- QSFP+ Enhanced quad small form factor pluggable module
- PMD Physical Medium Dependent
- SG Study Group
- TF Task Force
- PAR Project Authorization Request
- TF Task Force