Short Host Channel System Implications

Rob Stone

May 2018, Pittsburgh

BROADCOM®

Supporters

- Ali Ghiasi, Ghiasi Quantum
- Christophe Metivier, Arista
- Brian Welch, Luxtera
- Eric Baden, Broadcom

BROADCOM®

Background

- Several presentations have been made on SI budget partitioning to enable a 2m DAC use case
- This material examines the system trade-offs associated with this approach, as well as anticipated market trends which should be considered

BROADCOM

Use of DAC
Current generation technologies

- Server (or more generally endpoint) to ToR connections
  - Low cost, low power, used in close to all use cases today
- Fixed box “virtual chassis”
  - Example is Facebook “Fabric Aggregator” with DAC based “sideplane”

From: FacebookFabricAggregatorOCP_Spec_v1.0 (OCP)

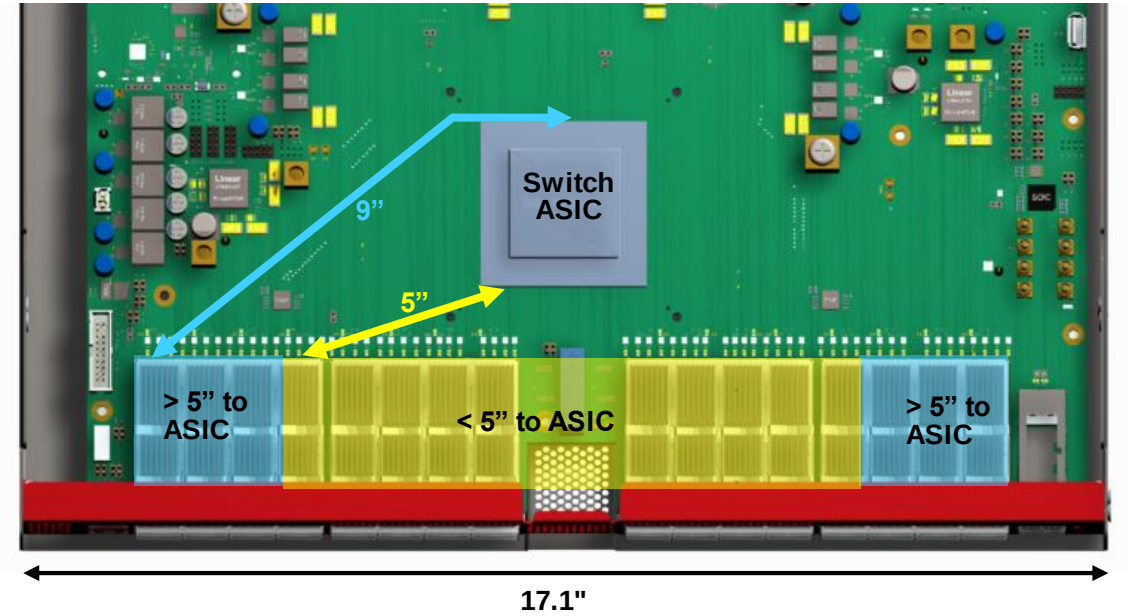
- Relevant attributes of these use cases for this discussion:
  - All IO is from the front of the box (no backplanes)
  - Long PCB traces from the most distant ports to the switch (~ 9”)
- Will this scale to 100G / lane DAC?

BROADCOM

Architectural changes to ToRs due to reduced physical VSR reach

- Hypothetical Example:
  - 25.6T, 256 x 100G
  - 1RU box, Single ASIC (ToR design profile, also used as virtual chassis, aka “Fixed Box”)
  - Can be used with all optical IO in a spine application (common practice today in hyperscale datacenters)
  - 32 x 800G module cages, all front panel IO

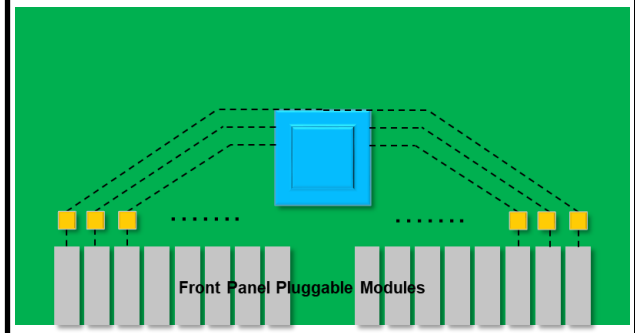
- Using Rosemont budget proposal from Jane Lim:
  - http://www.ieee802.org/3/100GEL/public/18_03/lim_100GEL_01b_0318.pdf
  - [~ 5” Host trace supported for VSR channels]
  - Approximately 12 / 32 module cages cannot accommodate the proposed host budgets (VSR or CR), requiring either intermediate retimers, or intra-box cabling



BROADCOM

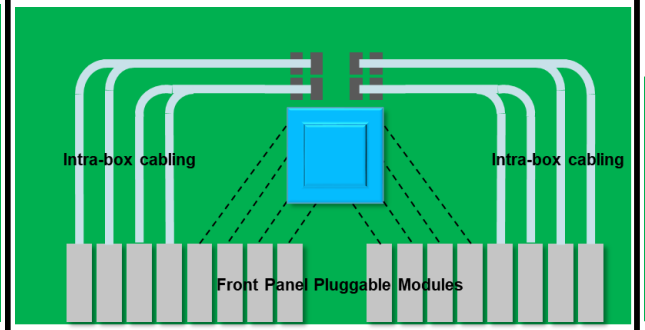
How Shorter Host Loss Maps to Possible "Universal Port" Solutions

Add retimers



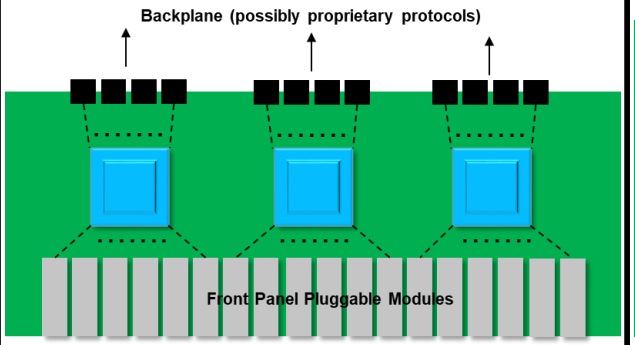
- Middle ports within proposed VSR budget do not require additional retimers
- Edge ports use additional retimers (shown in yellow) to enable longer overall host channels
- **Pros:** similar architecture to prior generation systems
- **Cons**: Cost and power of additional retimers

Intra-box cables



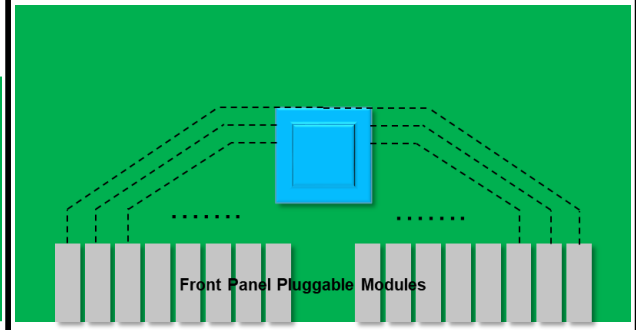
- Edge ports use intra-box cables to enable longer physical reach, but staying within proposed VSR budgets
- **Pros:** System does not incur cost or power of additional retimers, commonality with existing "retimerless" designs
- **Cons**: Increases mechanical complexity, may impact airflow, cost of cable and associated mechanicals

Multi-ASIC Linecards (Chassis Systems)



- Each ASIC can connect to fewer, closer module ports, which are supported within VSR proposed budget
- **Pros:** Similar "PHYless" design to current generation systems
- **Cons:** Does not address single ASIC "fixed box" designs forecast to be the dominant volume of the datacenter market

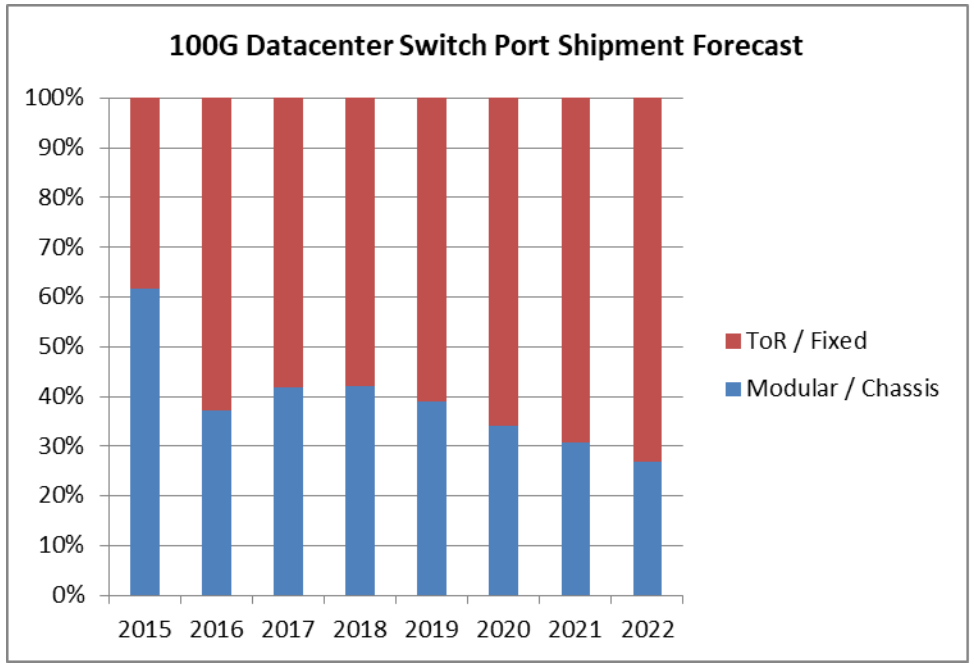
MR Capable Modules

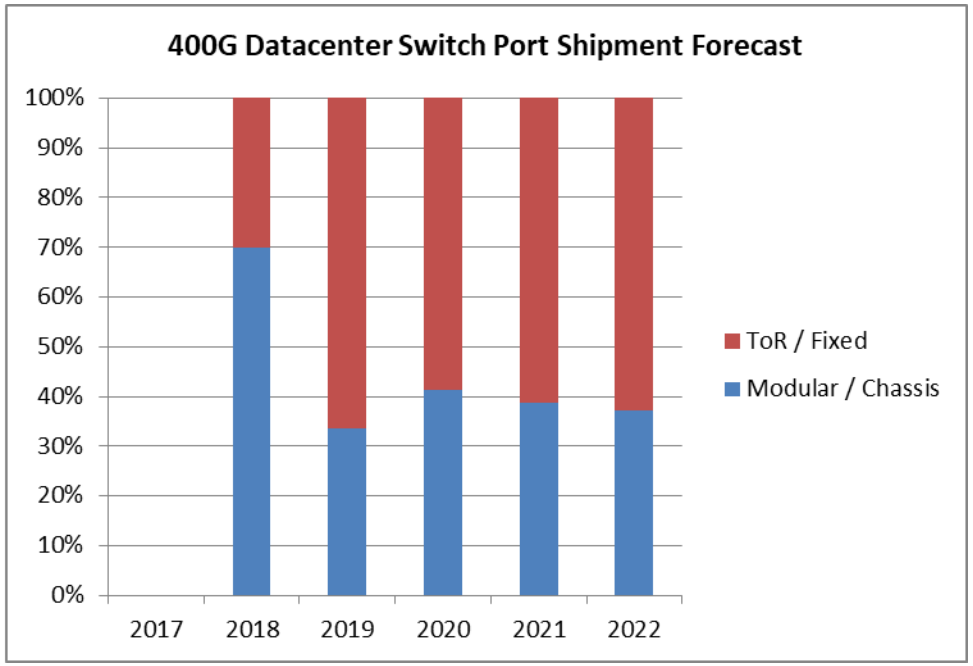


- Enable modules with MR capability
- **Pros:** Similar "PHYless" design to current generation systems
- **Cons:** Requires MR support in modules, potentially increasing module power. Serdes may require training, and appropriate management support. Doesn't work on all ports with DAC – so not a universal port

BROADCOM

Datacenter Switch Market Architectural Forecast





- Aggregate port shipment data presented for 100G and 400G[1]
- Datacenter switches are forecast to migrate from chassis to majority fixed / ToRs
- Important to find a low power, cost effective solution which supports the majority of this market!

[1]Source: *January 2018 CREHAN Long-range Forecast Data Center Switch* (used with kind permission)

BROADCOM

Possible Additional Work Needed for Different Architectures

- Retimers – nothing (VSR budget as proposed appears OK for this approach)
- Intra-box cables – can these enable retimerless systems?
- Asymmetric switch and host losses (see *ghiasi_3ck_01_0518.pdf)*?
  - An option for endpoint connections but not applicable for all single ASIC switch – switch DAC links using a retimerless system
- Multi-ASIC Chassis Linecards – nothing (VSR budget as proposed looks workable)
- MR based modules
  - Training required? In-band or out-of band?
  - Additional management complexity?
  - Increase in module power?

BROADCOM

Summary

- Short host channels as proposed have undesirable implications for single ASIC solutions and will lead to higher system power and cost
  - Such single ASIC designs are forecast to be the dominant part of the datacenter deployments at 100 and 400GE
- The power / complexity savings associated with shorter physical reach host channels needs to be quantified to ensure we are making the right trade-off to align with the market need

BROADCOM

Thank You

BROADCOM®