Thoughts on 50Gb/s ASIC IO and backwards compatibility considerations for 50G, 100G and 200G

Gary Nicholl - Cisco

IEEE Atlanta, Jan 18-22, 2016

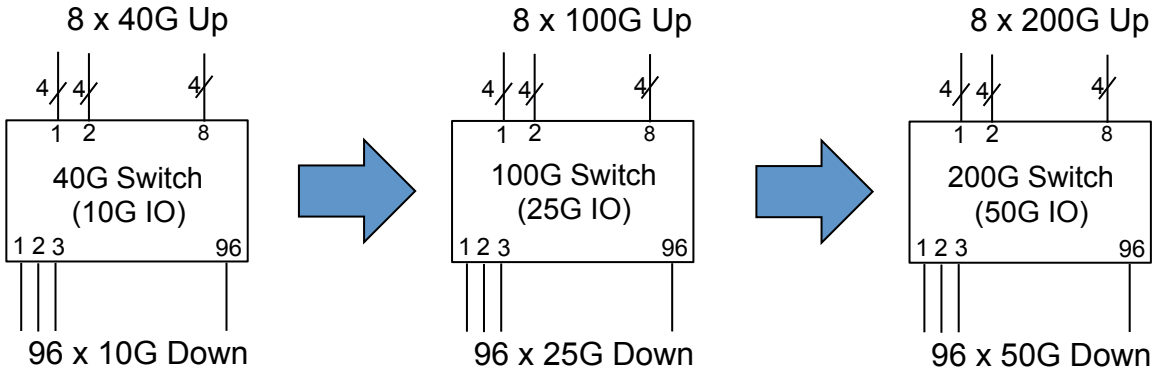
Topics

- Terminology
- Switch Chip evolution
  - 10Gb/s IO > 25Gb/s IO
  - 25Gb/s IO > 50Gb/s IO
- Switch upgrade scenarios
- Higher density (port counts) for legacy (lower rate) interfaces
- Conclusions

Terminology

- Multiple terminology in use:
  - Backwards compatibility
  - Backwards commonality
  - Legacy support
- Important considerations/desires:
  - connecting new equipment to legacy equipment in field (*backwards compatibility*)
    - support new interfaces in legacy equipment
    - support legacy interfaces in new equipment
  - support new interfaces on legacy equipment (cheaper PMDs)
  - upgrading network speeds and feeds (e.g. 10G to 25G, etc)
  - optimizing new ASIC/Chip designs (don’t want too many options)

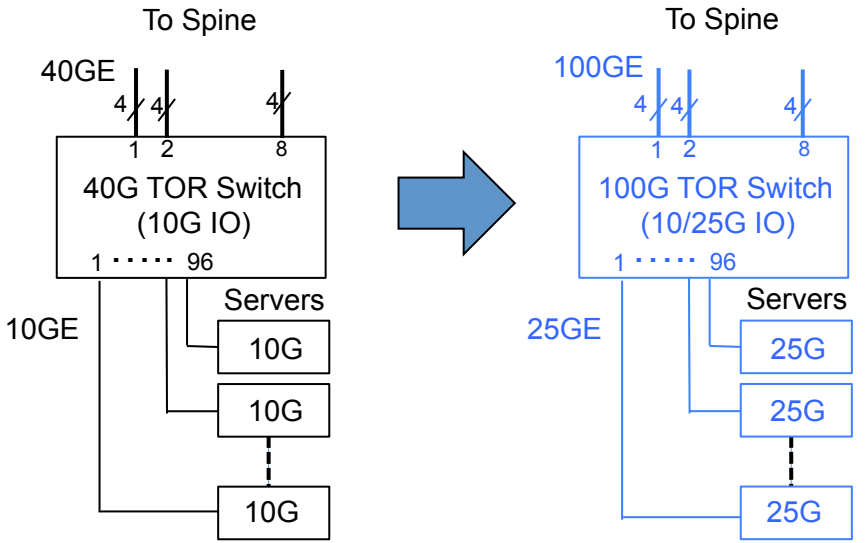
Switch Chip Evolution



Capacity	1.28T	3.2T	6.4T
Serdes rate	10Gb/s	25Gb/s	50Gb/s
Serdes count	128	128	128
Uplink ports	8 x 40GE	8 x 100GE	8 x 200GE
Downlink ports	96 x 10GE	96 x 25GE	96 x 50GE
Oversub ratio	3:1	3:1	3:1
Uplink switch radix	8	8	8

Why should 25G>50G transition be different from 10G>25G transition ?

Switch Upgrade Scenario 1 (Swap out)

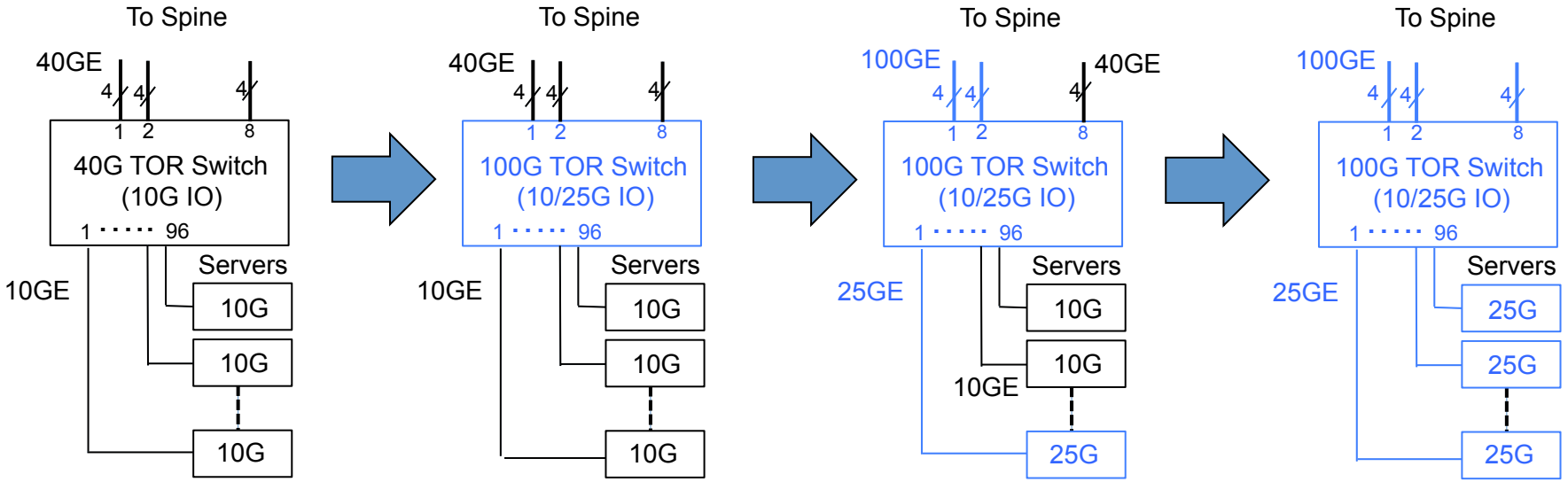


- Day 1 (starting point)
- 40GE/10GE config

- Swap out complete rack
- 100GE/25GE config
- no need to ‘downspeed’ switch IO. Run at 25G from Day 1.

Note: Same # of ports and links after the upgrade as before (8 up and 96 down), just running at double the rate.

Switch Upgrade Scenario 2 (Incremental)



- Day 1 (starting point)
- 40GE/10GE config
- 8x40G up, 96x10G down.

- Install shiny new 100G TOR switch
- Same 40GE/10GE config as Day 1
- Must “Downspeed” all switch IO.

- Incrementally upgrade uplinks to 100GE and downlinks to 25GE
- Hybrid configuration

- Upgrade complete
- 100GE/25GE config

Note: This example illustrates the power of the SFP/QSFP eco-system (new switch accepts legacy SFP/QSFP modules)

Observations

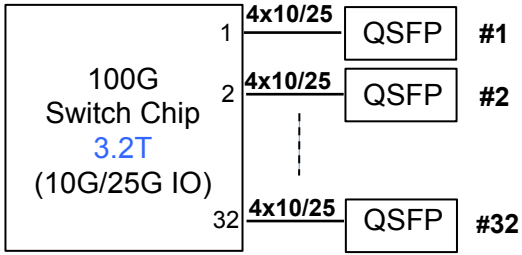
- Some observations on the 10G to 25G serdes transition and associated network upgrade scenarios:
  - New switch chip must support a ‘down speed’ serdes mode (to connect to legacy interfaces)
  - During the upgrade the number of ports and links in the network didn’t change (they just migrated to running 2x faster)
  - Number of serdes on the switch chip didn’t change (stayed at 128)
  - Number of QSFP ports on the TOR switch didn’t change (stayed at 32)
  - No requirement (in these examples) for higher density of lower speed (legacy) ports

Support higher density of lower speed ports?

- Is there ever a need for a new switch chip to support higher density of lower speed (legacy) ports than on the previous generation of switch chip?
- Not in the case of the scenarios shown in the previous slides
- But are there potential other applications where this may be needed/ desired ?
- Let's again look at how this was dealt with during the transition from 10G IO based switch chips to 25G IO based switch chips.

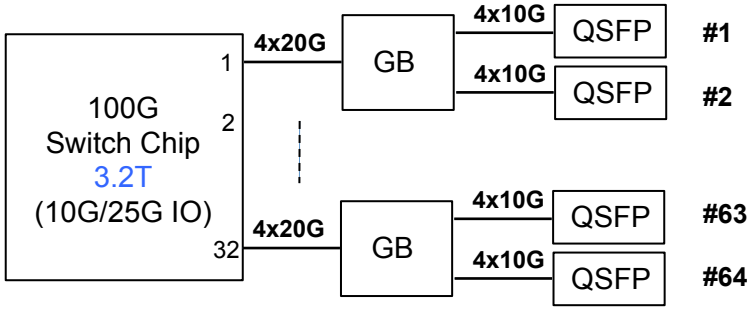
High density 40GE off a 100GE switch chip ?

High Density 100G Application



- "direct connect" to QSFP
- Mainstream application (32 ports in 1RU)
- 32 ports of 100GE/4x25GE b/o (3.2T)
- 32 ports of 40GE/4x10GE b/o (1.28T)
- 128 ports of 10GE (1.28T)
- Legacy ports run in 'downspeed' mode

High Density 40G application



- 1:2 external gearbox breakout
- 40GE optimized application (64 ports in 2RU)
- 64 ports of 40GE w/o breakout (2.56T)
- requires new 40GE AUI (i.e. 2 x 20G)
- 1/2 density of 100GE/10GE (i.e. 32/128 ports in 2RU)

How popular is the product configuration on the right ?

Conclusions

- If the 25G>50G ASIC IO transition mirrors what happened during the 10G>25G ASIC IO transition, then:
  - A ‘downspeed’ mode will be required (to support legacy 100GE, 40GE, and 10GE PMDs)
  - Running at reduced chip capacity , but with the same port density, for legacy interfaces is acceptable (and likely the primary application)
  - Application for a higher density mode for legacy interfaces is unclear
- The above comments appear to have held true for previous Ethernet rate transitions, i.e. from 100M>1G>10G>25G>40G>100G
- The FEC choice should be optimized based on the signaling rate, rather than the Ethernet MAC rate.