Global Foundation Services

Microsoft

Global Networking Services

Representative Cloud Scale Data Center Design

Tom Issenhuth

IEEE 802.3 400Gb/s Ethernet Study Group

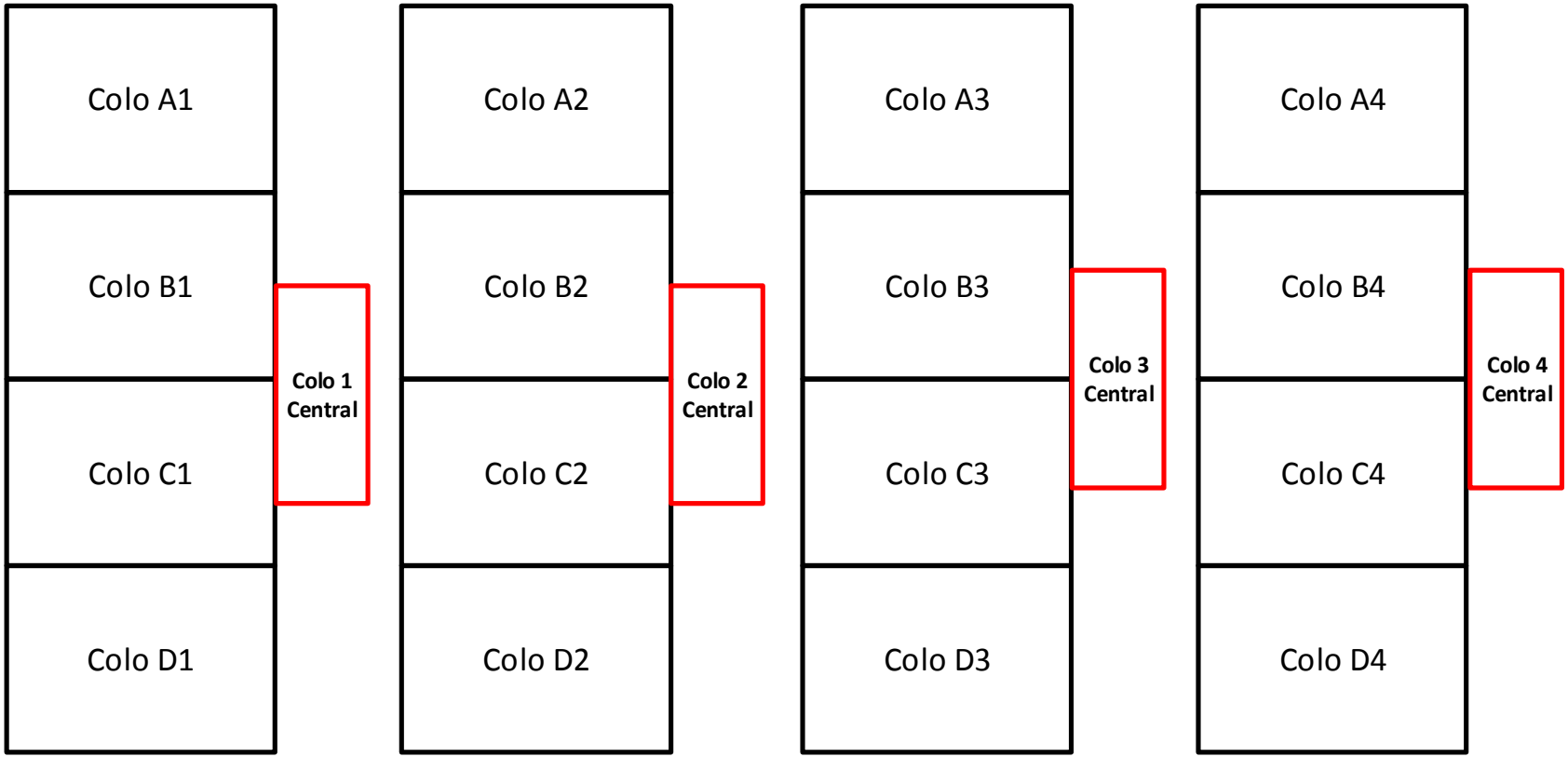
IEEE 802 July 2013 Plenary

Geneva, Switzerland

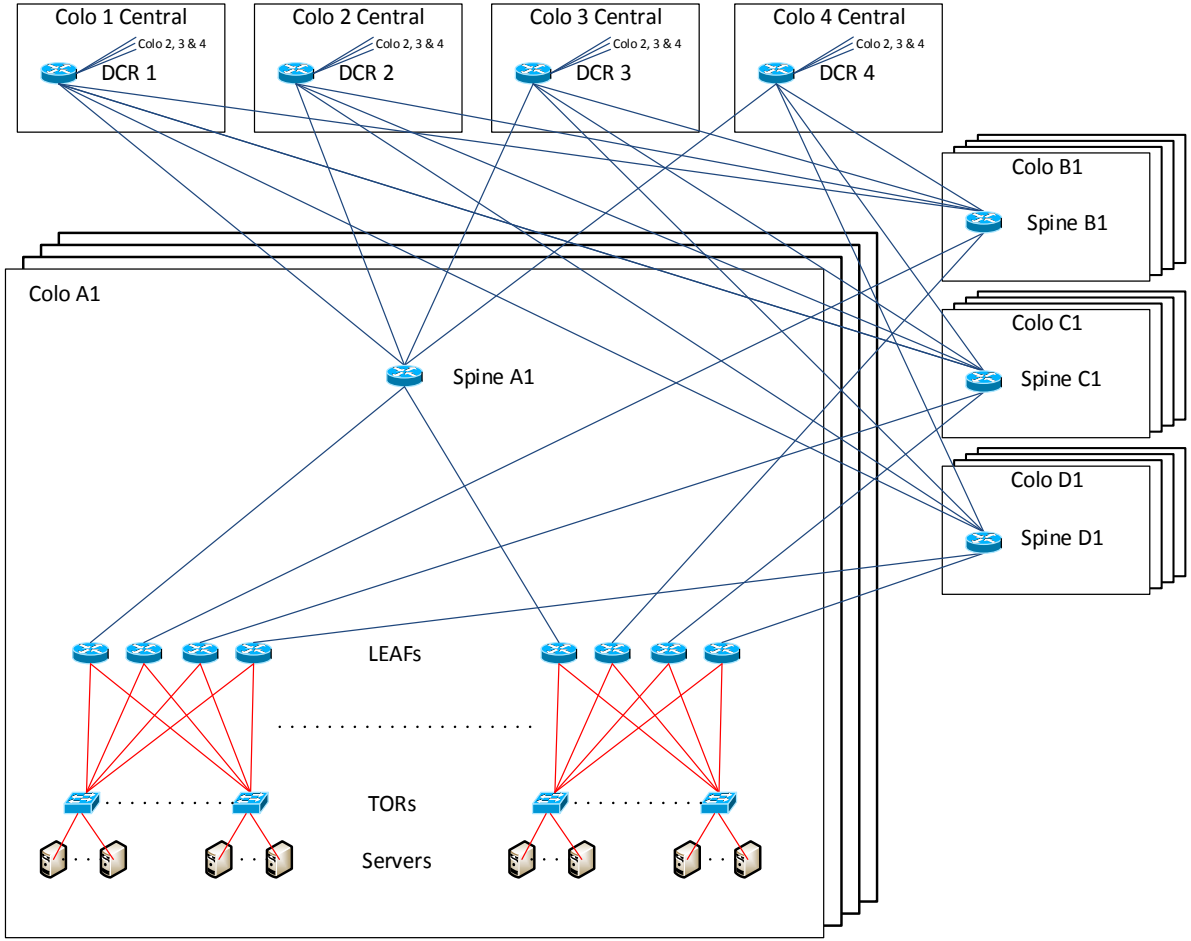
Cloud Scale Data Centers

- There is no single design or size for a data center
- While we attempt to standardize designs, differences are driven by generation of design, location and scale.
- While the overall traffic flow within different data centers is similar the design differences drive different link requirements
- Here is an overview of a typical web scale data center and the interconnections that would be required

Representative Data Center Campus Design



Representative Data Center Campus Interconnections



Interconnections

- Multiple interconnection lengths are required
- There are multiple colo areas per data center so the total number of links will vary
- All link quantities are per colo area

A End	Z End	Link Quanity	Link Length	Type of interconnection
Server	TOR	10,000s	.5-3m	TwinAx
TOR	LEAF	1,000s	1-20m	AOC
LEAF	SPINE - local	100s	20-300m	SM fiber
LEAF	SPINE - inter building	1,000s	100-400m	SM fiber
SPINE	DCR	100s	100-1,000m	SM fiber
INTRA METRO		100s	1,000m+	SM fiber

Interconnection Speed Requirements

- Today server interfaces are 10GE
  - Some 40GE servers are being deployed
- All links from the TOR upward are 40GE
- 1:1 oversubscription would be ideal but due to lack of sufficient ports and port speed in the TORs the uplinks are oversubscribed 3:1
- 40GE servers will start becoming more common and make the oversubscription problem worse
- TOR Uplink requirements
  - Assuming an average of 60 10GE servers per rack the uplink capacity ideally would be 600Gb
  - A transition to 40GE servers takes this capacity requirement towards 2.4Tb
    - At 100GE this would require 24 uplinks for 1:1 oversubscription
    - Moving to 3:1 oversubscription which would be the maximum oversubscription acceptable would require 8 uplinks
- The TOR and supporting network interconnection speeds need to stay one step ahead of the servers to minimize the oversubscription problem

What is coming next

- Server outputs increase every year by double digit %s
- Switch capacities are doubling every 2 years
- Clearly higher speed interfaces will be required to keep up with amount of data to be transported
  - It is not practical to only increase the number of ports, the port speeds must also increase
- When will the step beyond 100GE interconnects be required?
  - Availability of interface needs to match silicon and switch capabilities
  - All-in per-bit costs always need to be decrease with increased speed
  - The all-in first generation costs need to meet the cost reduction requirements
  - We would not pay a premium for higher speeds to be an early technology mover
- 400GE appears to be the next logical step
- We need to start thinking of what comes after 400GE as the amount of data to be supported is ever increasing