Optical switching for link bandwidth adaptation in future data-center networks

Michael Frankel
Vladimir Pelekhaty
John Mateosky
Current deployed state-of-art: Clos, Torus, Dragonfly, etc.

**Electrical Packet Fabric**
- Structured / Hierarchical
  - Good performance
  - Visualizable & manageable
  - Standard routing protocols
  - Constrained scaling
  - Failure sensitivity
  - ASIC port limit issue

**Loosely Structured / Flat, Regular Graph**
- Excellent performance
- Hard to visualize
- Off-standard routing protocols
- Excellent scaling
- Failure tolerance
- Reduced ASIC port limit issue

We explore this direction

**Optical IO on ASIC**
- Power & Cost reduction
- Complex development
- Complex supply chain
- Fixed optical IO technology

**Optical Switch Changing topology**
- Improved scalability
- Buffering, cost, power, switch time, control plane, etc.

**Optical Circuit Switch Fixed topology**
- Reduce $ and W
- Switch speed and control plane
- Coupling to application layer and routing algorithms

We explore this direction
Objective of this work

- Leverage baseline static network performance
- Only **low-radix electrical** switches
- Add optical switch to improve performance
  - Only **slow** optical switching
  - **Standard protocols** at edge (application and server layer)
  - **Avoid centralized** control plane and scheduling
  - **Static routing tables** $\rightarrow$ no update delays
  - **Standard optical transceivers** (no $\lambda$ tuning, no burst mode RX, etc.)

**STRAT** – Structured ReArranged Topology based on flat, regular graph
Optical switching resolves bandwidth hot-spots
Alternative to Clos is possible: STRAT flat mesh topology

(Circular representation)

- Eliminate Leaf / Spine switch layers!
- TOR is the only switch that needs to be deployed and managed
- Fully passive static optical interconnect among TORs

More details in paper W1J.5
STRAT flat mesh is promising … but is it **Scalable**?

- Easy scale with low hop count (<5)
- Low radix TORs are sufficient
  - Still need high per-port bandwidth
  - Aggregate ports into large bandwidth pipes
- Good for high bit rate optical transceivers
- Good for reducing switch ASIC I/O fanout

**TOR Count:**

\[
 n \sim \frac{d^{0.84k + 0.1532}}{0.1467(k - 2)^2 + 0.767}
\]

**MEGA-DC**

- TORs with ~10 to 16 ‘ports’ are sufficient!

**Large Enterprise DC**

- STRAT Computation:
  - N – servers/TOR
  - 2.5N – Net ports/TOR
  - 3.5N = 8 – total TOR ports
  - N = 8/3.5 ~ 2 servers/TOR
  - # of TORs = 128 servers/2 = 64
STRAT based on optical interconnect

ALL-OPTICAL interconnect may be static or augmented with optical switching
Spatially separated for failure decorrelation
Each TOR has 16 near-neighbors (i.e. 16 network ‘ports’)

Typical DC Layout

Racks
(16x16)

Fiber Patch Panels
Small network with Mesh U/ECMP example (i.e. Unequal Multi-Path)

Congestion decisions are purely local

- Let's look at the above network, considering Switch 1 only
- The following ‘ECMP’ table is created at Switch 1 with corresponding port assignments

<table>
<thead>
<tr>
<th>Destination Switch</th>
<th>Primary ECMP port assignments</th>
<th>Alternate ECMP port assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2, 3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4, 5</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4, 5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1, 2, 3</td>
<td>4, 5</td>
</tr>
</tbody>
</table>
Mega DC: Clos vs. STRAT (no optical switch)

65536 Server (~20 MW) data center shown
64 port ASIC assumed

**Clos Baseline**
- Clos 3 Level
- 65536 servers
- 5120 Switches
- 131K links (bidir)

**Baseline Load:** 1

- 25% Random Fail Load: ~0.5
- 25% Upper Fail Load: ~0.6
- 25% Cluster Fail Load: ~0.2

**Rapid degradation of ~50%, far above link failure**

**STRAT (static)**
- 8 Patch Panels
- 65536 servers
- 4096 Switches
- 98.3K links (bidir)

**Baseline Load:** ~1.7

- 25% Random Fail Load: ~1.4

**280% higher throughput**

- 16 network ports
- 3x port granularity (or higher)
- Denser faceplate, cheaper XCVR

**>20% less HW**

- Graceful degradation of ~20%, sub-linear to link failure

- Rapid degradation of ~50%, far above link failure

Uses 1x port granularity
Need high E-W fan-out

25% link failure
Example small STRAT network – with Optical Switch
(X8 illustrative topology)

Skewed loading within ECMP groups

-fully loaded
- - partially loaded
- - - - empty

All-to-All Traffic Before Reassignment

5-port ECMP Grouping

~30% increase in Throughput
Topology stays fixed!!!

All-to-All Traffic After Reassignment
Illustrated operation of optical switch

1. 1\textsuperscript{st} port filled
2. Different port filled
3. 1\textsuperscript{st} port Queue starts getting busy $\rightarrow$ fill 2\textsuperscript{nd} channel
4. 1\textsuperscript{st} and 2\textsuperscript{nd} queues busy $\rightarrow$ fill 3\textsuperscript{rd} channel
5. Packet Switch detects ECMP group near exhaust
   - Packet switch requests more bandwidth from Optical switch
   - Optical switch reallocates free ports to busy ECMP group
   - Push new ECMP port association to affected Packet switches (forwarding table stays the same)
STRAT based on optical switch

Control Plane switch Coordination -- Slow

Optical Switch R1

Optical Switch R2

Optical Switch R8

1 2 3 4 128

1 2 3 4 128

1 2 3 4 128

Loopbacks waiting For upgrade Rows

Loopbacks waiting For upgrade Rows

TOR 1

TOR 2...

TOR 32

TOR 33

TOR ...

TOR 64

TOR

TOR...

TOR 4096

ROW 1

ROW 2

ROW 128

servers

Conduit

*  *  *

Copyright © Ciena Corporation 2016. All rights reserved. Confidential & Proprietary.
Operation and requirement

Operation:
- TOR: detect local congestion + local idle ports → send BW request and idle port list to control plane
- Control plane: reserve end-point idle ports (bi-dir), flip optical switch, update ECMP groups
- Newly “idle” ports do not need explicit broadcast to control plane

Requirements and Limitations:
- TOR port ECMP members must be separable at centralized switch
  - Parallel fibers (PSM transceiver) → may be hard to cable, but avoids wavelength blocking issues
  - WDM → demux/remux at optical switch → some wavelength blocking constraints exist
  - Benefits from more parallelized Transceivers, i.e. 8 x 50Gb
- TOR port ECMP members use weighted fill order → available from commercial ASICs
- As flowlets stop/start → ECMP loading shifts away from high-weight ports (to low-weight)

NOTE: Electrical Xpoint switch is also possible
- Solves optical granularity and wavelength conversion issues
- Doubles transceiver count
Summary

- **STRAT** offers excellent baseline static network performance
- Only low-radix electrical switches (10 to 16 network ports)
- Optical switch improves performance
  - Only slow optical switching
    - Burst traffic absorbed at electronic edge queues
  - Standard protocols at edge (application and server layer)
    - No need to: coordinate schedules, separate mice/elephant, etc.
  - Distributed, localized control plane
  - Static routing tables → no update delays
  - Standard optical transceivers (no λ tuning, no burst mode RX, etc.)
Thank You!