Hybrid Datacenter Networks

George Papen
Department of Electrical and Computer Engineering
University of California at San Diego

Hot Interconnects 2013
Research teams

- **REACToR**
  - *Students*: He Liu, Feng Lu, Rishi Kapoor, Malveeka Tewari, Alex Forencich
  - *Senior Researchers*: Stefan Savage, Geoff Voelker, George Papen, Alex C. Snoeren, George Porter

- **Mordia**
  - *Students*: Nathan Farrington, Alex Forencich, Richard Strong
  - *Senior Researchers*: Joe Ford, Pang Chen-Sun, Tajana Rosing, Yshaiahu Fainman, George Papen, George Porter, Amin Vahdat
Outline

- Motivation and Background
  - Scale-out Datacenters
  - Hybrid Networks
- Research Issues
  - Circuits in a Packet-based World
  - Burstiness of Traffic
  - Scheduling
  - Optical Circuit Switching
- Conclusions and Acknowledgements
Scale-out data centers

- Massive scale; large number of nodes (100k+)
- Applications are different, unpredictable, uncoordinated
- Bi-section bandwidth, low latency, and reliability critical
- Scale-out designs [VL2,FatTree,...]:
  - No oversubscription
  - Cost, Power, Complexity
Scalability $\rightarrow$ High costs

- $S_{N,0}$, $S_{N,1}$, ..., $S_{N,k/2}$ = Core transceiver
- $S_{2,0}$, $S_{2,1}$, $S_{2,2}$, $S_{2,3}$, ..., $S_{2,k}$ = Edge transceiver

Number of components is a strong function of the number of layers.

Scalability à→ High costs
Sources of cost, power, and complexity

<table>
<thead>
<tr>
<th>Network</th>
<th># nodes</th>
<th># levels</th>
<th>Switch radix</th>
<th>Core Transceivers</th>
<th>Core Transceivers per host</th>
</tr>
</thead>
<tbody>
<tr>
<td>10G</td>
<td>27,648</td>
<td>3</td>
<td>48</td>
<td>138,240</td>
<td>5</td>
</tr>
<tr>
<td>10G</td>
<td>65,536</td>
<td>3</td>
<td>64</td>
<td>393,216</td>
<td>5</td>
</tr>
<tr>
<td>40G, redundancy</td>
<td>15,552</td>
<td>5</td>
<td>12 (effective)</td>
<td>139,968</td>
<td>9</td>
</tr>
</tbody>
</table>

Faster links
Fault tolerance \(\rightarrow\) Smaller switch radix \(\rightarrow\) More layers
More cost
Network Demand

More “network-centric”

Random traffic

Datacenter workloads can vary across this space

Many datacenter traffic patterns are a mixture

Correlated traffic

More “processor-centric”
Network Demand

More “network-centric”

Datacenter workloads can vary across this space

More “processor-centric”

Random traffic

Correlated traffic

Measured datacenter traffic
(data from Microsoft - Kandula 2009)
Rank-Ordered Demand

- Rank-Ordered Demand
- Rank-Ordered Connection Number

Traffic Per Connection

1

Rank-Ordered Connection Number

$N^2$

Circuit traffic

Packet traffic

$1 \rightarrow n_c \rightarrow N^2$
Hybrid Electrical/Optical Networks

- Circuit switching
  - Decouple line rate from speed of control plane
  - Used for persistent high-data rate traffic – must be scheduled

- Packet switching
  - Handle ‘tail’ of traffic demand
  - Can correct for errors in circuit schedule

● $S_{0,0}$, $S_{0,1}$, $S_{0,2}$, $S_{0,3}$, ..., $S_{0,k}$

○ $= Edge transceiver$

○ $Pkt$

○ $OCS_{k \times k}$

○ $H_i$
Hybrid Switching Architecture for Data Centers
Hybrid Switching Architecture for Data Centers

Optical Circuit-Switched Ring

Optical circuit switch
- High port count (1k x 1k)
- Fast switch time (ns or us)
- Low loss (multi-stage friendly)

Top-of-Rack Switch
Server Rack

Optical/Electrical Interconnect

N-Layers
Future Hybrid Needs

Optical Circuit-Switched Ring

Optical circuit switch
- High port count (1k x 1k)
- Fast switch time (ns or us)
- Low loss (multi-stage friendly)

Circuit-enabled ToR switches
TOR <-> Host control plane
Push circuits to the server
Future Hybrid Needs

Optical Circuit-Switched Ring
Optical/Electrical Interconnect

Optical circuit switch
- High port count (1k x 1k)
- Fast switch time (ns or us)
- Low loss (multi-stage friendly)

Can be under-provisioned, saving cost/power/complexity

Circuit-enabled ToR switches
TOR <-> Host control plane
Push circuits to the server
Outline

- Motivation and Background
  - Scale-out datacenters
  - Distributed vs. Centralized Network Control

- Research Issues
  - Circuits in a Packet-based World
  - Burstiness of Traffic
  - Scheduling
  - Optical Circuit Switching

- Conclusions and Acknowledgements
High-Level Diagram of REACToR

Standard Switch

Input Ports

Fast Buffers

Circuit Switch

Fast Buffers

Output Ports

REACToR

Circuit-switched fabric

Host A

Slow DRAM

Host B

Slow DRAM
REACToR Architecture

N x N port EPS

k x k port OCS

N x k MUX
Circuit-friendly NIC – More Queues
Testbed Components

- Classifier and Controller (Master) - HXT100G Virtex6
- End-Host
- Electrical Packet Switch - Fulcrum Monaco
- Crosspoint Circuit Switch - MindSpeed M21048
- Classifier and Controller (Slave) - HXT100G Virtex6
- End-Host
- Control Host/Scheduler
- Clock
- Reconfiguration Schedule
- Reconfiguration Schedule/Clock
- Circuit Control
- Data Path
- Control Path
- Intel NIC
SFP+ Adaptor board

Virtex-6

Electrical Circuit Switch (Mindspeed)

Virtex-6

SFP+ Adaptor board
Pausing and synchronization

- Relying on 802.1Qbb
  - aka Priority Flow Control
- Eight endhost queues
  - Maintained by O/S
  - Non-realtime
- Queues “paused” by control packets from Hybrid TOR
  - Maintained by NIC
  - Real-time
Pausing and synchronization

Dash-Dot – Turn on command

Dots – Turn off command


Time (µs)
Outline

• Motivation and Background
  • Scale-out datacenters
  • Distributed vs. Centralized Network Control

• Research Issues
  • Using Circuits in a Packet-based Environment
  • Burstiness of Traffic
  • Scheduling
  • Optical Circuits

• Discussion
Burstiness of Traffic

- How fast does circuit switch need to be?
  - Try to match to the burstiness of traffic

- What determines burstiness?
  - Where the circuit switch is deployed (ToR vs. core)
  - Application dependent
    - All-to-all vs. highly coherent (traffic matrix coherence)
  - Also the O/S, TCP, the NIC (e.g., TCP offloading)
How fast is fast enough?

Burst behavior of Intel 82599 NIC with workloads taken from [Helios, Sigcomm10]
Traffic matrix coherence

- Circuit traffic
- Packet traffic
- Fewer Circuits
- More Circuits
- Different Traffic Pattern
- Higher Data Rate per Port

Rank-Ordered Connection Number

Traffic Per Connection

1 \[ \rightarrow \] \[ n_c \] \[ \rightarrow \] \[ N^2 \]
Traffic conditioning

100us circuit = 75 MTU-sized frames per circuit

- Natural correlation
  - Macro: Multi-packet data objects
  - Micro: TCP segmentation offloading (TSO)

- Induced correlation
  - Example: sort
  - Coordinate shuffle phase to create skew at small timescales

**Bullet Trains: A Study of NIC Burst Behavior at Microsecond Timescales**

*Rishi Kapoor, Alex C. Snoeren, Geoffrey M. Voelker, George Porter*

(In submission to ACM CoNEXT’13)
Outline

- Motivation and Background
  - Scale-out datacenters
  - Distributed vs. Centralized Network Control

- Research Issues
  - Using Circuits in a Packet-based Environment
  - Persistence of Traffic
  - Scheduling
  - Hardware for Optical Circuits

- Discussion
Previous approaches: Hotspot Scheduling

Step 1. Observe network traffic
Step 2. Compute schedule
Step 3. Reconfigure

Time

Assign circuits to elephants
Limitations of Hotspot Scheduling

1. Observe
2. Reconfigure
3. Observe

Goal

Time

TM(t)
Choosing a schedule

1) Maximal matching:

2) Matrix decomposition:

\[ TM = \sum_{i} t_i P_i \]
Traffic Matrix Scheduling

Step 1. Gather traffic matrix TM

Step 2. Scale TM into TM’

Step 3. Decompose TM’ into schedule

Step 4. Execute schedule in hardware
Evaluating Schedules

- Perfect decomposition
  - E.g., Birkhoff-von Neumann: “equal”

\[
P_1 t_1 + P_2 t_2 + P_3 t_3 + P_4 t_4 + P_5 t_5 + P_6 t_6 + P_7 t_7 + P_8 t_8 + \ldots + P_M t_M
\]

- Longest-time slot first: “approximate”

Remainder routed using packet switch

\[O(N^2)\] terms
Scheduling Summary

- The faster the circuit switch, the more of the overall matrix can be scheduled using circuits.
- Schedule for circuit switch does not need to be perfect – rest is routed over packet switch.
Outline

- Motivation and Background
  - Scale-out datacenters
  - Hybrid Networks

- Research Issues
  - Circuits in a Packet-based World
  - Burtiness of Traffic
  - Scheduling
  - Optical Circuit Switches

- Conclusions and Acknowledgements
Circuit switching in data centers

- Speed of optical circuits (~10 of ms) means that can be used at aggregation level or higher
- May be sufficient for some kinds of networks such as Google’s B4 network
Fast Optical Circuit Switch- Mordia

Each station:
1. Inserts signals into ring
2. “Tunes into” subset of signals

Data plane:
Physical ring w/ 24 DWDM signals

Each device transmits on its own frequency

Photonics Technology Letters
Vol. 25(16) 1589-1592, 2013
Mordia Station Architecture

WSS – Wavelength selective switch  VOA – variable optical attenuator  EDFA – Optical amplifier
Measured 10G Data through WSS
End-to-end reconfiguration time

(N, μ, σ) = (705, 11.55, 2.36)
Conclusions

- Hybrid networking has potential to reduce cost, power, and complexity
- Complements current trends towards SDNs
- Where in the network it is deployed depends on the persistence of the traffic matrix and the applications
- Deployment at the ToR level is an open research topic.

Thank you!
Acknowledgements

• This work has been supported by the following sources:
  • NSF Engineering Research Center for Integrated Access Networks (CIAN)
  • Center for Networked Systems (CNS)
  • Google
  • Cisco
  • Corning
  • Mindspeed