Data-center network (DCN) architectures with Reduced Power Consumption

“Flow/Application triggered SDN controlled electrical/optical hybrid switching data-center network: HOLST”

Satoru Okamoto, Keio University
okamoto@ieee.org
Co-Authors and Acknowledgement

• Co-Authors
  – Yukihiro Imakiire, Masayuki Hirono, and Naoaki Yamanaka (Keio University)

• Acknowledgement
  – This work is partly supported by the “HOLST (High-speed Optical Layer 1 Switch system for Time slot switching based optical data center networks) Project” funded by New Energy and Industrial Technology Development Organization (NEDO) of Japan.
Outline

• Data-center Electricity Consumption
• Data-center network architecture
  – Leaf-Spine Electrical Switching
  – Optical data-center network
    • Optical Circuit Switching
    • Optical Slow Switching
• HOLST data-center network
• Summary
Data-Center (DC) Electricity Consumption

- Infrastructure Savings
- Network Savings
- Storage Savings
- Server Savings

70 BkWh@2014 (2 % of US)

4 % increase Per 5 years

Savings: 620 billion kWh

Breakdown of the Power Consumption in DC

- **Cooling** 37%
- **Equipment** 50%
  - Server + Storage 33%
  - Network 17%
- **Converter Loss** 10%
  - (DC/DC)
Basic Data-Center Network (DCN) Architecture

- **Leaf-Spine Architecture (Layer 2 or Layer 3)**
  - DCN capacity can be adjustable by changing # of Spine Switches.
Power consumption: Optical vs. Electrical


MEMS-based Optical Circuit Switching!!
1st Generation: Helios (2010 UC San Diego)

- MEMS-based Optical Circuit Switching (OCS) is introduced to the Leaf-Spine architecture

How to accommodate “big flows” into Optical Circuit Switching Network

• First, all flows are accommodated into Electrical Switching Network.

• If “Elephant Flow” is observed, then the flow is rearranged to Optical Circuit Switching Network.
  – On-line Flow Classification “Elephant Trap”
  – Observation-based flow assignment
    • Maximum weight matching problem

2\textsuperscript{nd} Generation: Optical Slot Switching (OSS)

- **Fixed Length $\mu$s-order Slot Switching** + SDN control
  - ICTON 2017 Mo.B3.4 "NEPHELE" (National Technical Univ. of Athens)
    - High-speed (10 ns) 2x2 Optical Switch
      - All Optical, Ring Topology
  - ECOC 2017 We.2.A.3 "Cloud BOSS" (Nokia Bell Labs)
    - High-speed (100 ns) tunable Tx for making a slot
      - All Optical, Ring Topology
  - ECOC 2017 We.2.A.4 "COSIGN" (Univ. of Bristol)
    - High-speed (25 ns) 4x4 Optical Switch
      - OCS (MEMS) + OSS
HOLST: High-speed optical layer 1 switch system for time slot switching based optical data center networks

- Slot Switching-based DCN developing project
  - Keio University, OA Laboratory, and Epi Photonics
  - Electrical and Optical (Circuit and Slot) hybrid switching network
  - High-speed (10 ns) 8x8 and 16x16 Optical Switch is developing
  - Application triggered SDN-based DCN control
    - ECOC 2017 We.2.A.2 “Hadoop-based Application Triggered Automatic Flow Switching in Electrical/Optical Hybrid Data-Center Network” (Keio Univ.)
HOLST System Architecture

- Spine_Switch
- TOR_Switch
- Mice Flow
- Doggy Flow
- Elephant Flow
- Ultra_High_Speed_Optical_L1_Switch
- PLZT_Switch
- MEMS_Switch
- OSS network
- OCS network
- SDN Controller
- hadoop
- Controlplane
Power Reduction by OSS + OCS

- # of ToRs = 256
  - 30 servers/rack, NIC 10 GE, mixed traffic (Web search and Data mining)
How to accommodate “Elephant and Doggy flows” into OCS and OSS Network in HOLST

• First, all flows are accommodated into Electrical Switching Network.

• If “Elephant Flow” is observed, then the flow is rearranged to OCS Network.

• If “Doggy Flow” is observed, then the flow is rearranged to OSS Network.
  – Observation-based flow assignment
  – On-line Flow Classification
  – Application (Hadoop) triggered flow assignment
Observation-based Doggy Flow assignment

• 8x8 Optical Switch is assumed
  – 1 ToR can connect to 7 other ToRs
    • ToR Groups should be found in 256 ToRs’ Traffic Matrix
    • Optimum Grouping problem is NP-hard.
      – Heuristic grouping algorithm is developed.
On-line Flow Classification

- Flow-ID management queue will be set in ToR
  - Hierarchical Least Recently Used (LRU) queue
    - Flow-ID and reference # of the Flow-ID (counter) are stored.
      - If counter exceeds the threshold, the Flow-ID is moved into higher queue
  - Thresholds and queue size are adaptively changed.

MF: Mice Flow
DF: Doggy Flow
EF: Elephant Flow
Hadoop Triggered Flow Assignment

• “Hadoop Cluster” is monitored.
  – Newly defined “Shuffle Ratio” is used for classification.

Cluster Manager detects job start
→ Instruct flow monitoring to Traffic Monitor

Calculate “Shuffle-Ratio” from traffic monitor and job information

Set circuit through the SDN Controller
  Shuffle-Ratio is large → Optical
  Shuffle-Ratio is small → Electrical
HOLST PoC experiment

• Small HOLST PoC is constructed.
  – 10 GE L2/L3 Switches, 16x16 MEMS Switch, 4x4 PLZT Switch
  – Software-based OSS adapter, Software-based On-line Flow Classifier
  • Throughput is limited due to the software-based
Optical Slot Switching in HOLST PoC

- Slot Size 200 ms (software) → µs order (developing FPGA)
On-line Flow Classification in HOLST PoC

• ~ 100 Mbps throughput is realized by software emulation.
Hadoop triggered Flow Assignment in HOLST PoC

- In the shuffle phase, the flow is rearranged to OCS
Summary

• “Optical Slot Switching” becomes the hot topic technology in the optical data-center network.

• In case of the Hybrid DCN, flow classification is required to efficiently utilize the optical network.

• In the HOLST project, three flow classification methods are developing.