Optical Packet Switching: Myth, Fact, and Promise

Ken-ichi Kitayama

kitayama@gpi.ac.jp

Graduate School for the Creation of New Photonics Industries,
formerly with Osaka University, Japan
Acknowledgements

This work has been partly supported by the R&D program, “High-Performance Hybrid Optoelectronic Packet Router -Towards Green High-Capacity Data Center Networks-” (2011~2016), funded by NICT.

- Y.Huang, P.J.Argibay-Losada, Y.Yoshida (presently with NICT), and A.Maruta
- I’d also like to thank my collaborators of ;
  - R.Takahashi, NTT Laboratories
  - NEC Corporation
  - Kyushu University
Packet switching: Underlying concept

Figure 1. Three forms of networks: (a) centralized; (b) decentralized; (c) distributed.

Circuit switching vs. packet switching

Arrival

- If not busy
- Busy

Call loss
Queuing delay = 0

Circuit switching

POTS
- QoS guaranteed
- Stream data

Packet switching

Packet loss
Queuing delay

Cost

H. Miyahara, Osaka Univ.-NTT Meeting (2001.11.05).

Internet
- Best effort service: Only connection guaranteed by TCP/IP
- Burst data
Outline

- **Introduction**
  - Niche application of OPS to DC networks
- **Myths**
  - Immature architecture
  - Lack in optical RAM buffer
- **Promise: Enabling technologies at hand**
  - Prototyping energy-efficient optoelectronic packet router
  - OPS/OCS/VOCS DC network performs well against mice & elephant
- **Summary**

OPS: Optical packet switching
OCS: Optical circuit switching
VOCS: Virtual OCS
Outline

- Introduction
  - Niche application of OPS to DC networks

- Myths
  - Immature architecture
  - Lack in optical RAM buffer

- Promise: Enabling technologies at hand
  - Prototyping energy-efficient optoelectronic packet router
  - OPS/OCS/VOCS DC network performs well against mice & elephant

- Summary

OPS: Optical packet switching
OCS: Optical circuit switching
VOCS: Virtual OCS
Transforming DC from hardware-centric to software-centric

Modified after A. Fuetsch, Plenary talk, OFC2016.
Transforming DC fabric design

**Fat-tree network**
- Core
- Aggregation
- ToR

**Leaf & spine network**
- Spine
- Leaf
- ToR

- **Scale-up** approach (10G⇒40G⇒100G)
  - The path and bw are fixed end to end
  - Traffic concentrates at the core switch
  - Blocking the ports by STP* is a waste

- **Scale-out** approach (10G⇒10GxN)
  - Path & bw are changeable end to end
  - Load-balancing is possible
  - STP is replaced with IS-IS* or BGP*

* STP: Spanning tree protocol
* IS-IS: Intermediate system-to-Intermediate system
* BGP: Border gateway protocol

July 4, 2016
OECC / PS2016
Transforming DC fabric design

- Bimodal traffic distribution of short and long packet flows.
- East-to-west long-lived data flows, so-called *elephant* flows.
  and south-to-north short-lived data flow, so-called *mice* flows.
- Single switching protocol cannot perform well for the co-existence.
# Hybrid switching schemes

<table>
<thead>
<tr>
<th>Category</th>
<th>EPS</th>
<th>OCS</th>
<th>OPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current DC</td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helios, C-Through</td>
<td>✔</td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>LION</td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Lightness</td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>COSIGN³</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>-</td>
<td>✔</td>
<td></td>
<td>✔</td>
</tr>
</tbody>
</table>

- AWGR-based wavelength routing.
- MCF switch which can switch individual cores.
- SDN controlled

R&D efforts for photonics and optics in DC

After C. Kachris et al., IEEE Commun. Mag., Sept., 2013

July 4, 2016
Optical circuits break the network topology!

6-D torus network with 4,096 (4^6)-node.

Note: 885 ExPs traverse 22% of 4,096 routers.

Packet Loss Probability

Load from servers [Gbps]

Typical operation range
30-50 %

Pure OPS

3D-torus topology
## Power consumption of DCs

<table>
<thead>
<tr>
<th></th>
<th>2007 [Billion kWh]</th>
<th>2020 [Billion kWh]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data centers</td>
<td>330</td>
<td>1,012</td>
</tr>
<tr>
<td>Telecoms</td>
<td>293</td>
<td>951</td>
</tr>
<tr>
<td>Cloud</td>
<td>623</td>
<td>1,963</td>
</tr>
</tbody>
</table>

- Power & heat management: ~280 NC power plants
- 35%
Power consumption of DCs

- Power & heat management: 35%
- IP look-up & forwarding: 32%
- I/O: 7%
- Control plane: 11%
- Switch fabric: 10%
- Buffers: 5%

Source: G. Epps, Cisco, 2007

The higher the power consumption, the more you waste the power!
Outline

- Introduction
  - Niche application of OPS to DC networks

- Myths
  - Immature architecture
  - Lack in optical RAM buffer

- Promise: Enabling technologies at hand
  - Prototyping energy-efficient optoelectronic packet router
  - OPS/OCS/VOCS DC network performs well against mice & elephant

- Summary

  OPS: Optical packet switching
  OCS: Optical circuit switching
  VOCS: Virtual OCS
Conventional e-router architecture

Segmented into 8 short fixed-length packets and switch at low-speed in parallel

Large buffer required to store packets for the re-assembly

Fill Interframe gap (IFG)

Input buffering & Label processing

Output buffering

Large buffer required to store all the input packets for segmentation
Huge bw of optical components can be fully exploited. Packet payload is transparently forwarded w/o segmentation and w/o o-e-o conversion.
No optical RAM won’t be available in the near future.
- e-RAM is available but it causes substantial delay.
- Small size fiber-delay-line (FDL) is practically implementable.
Outline

- Introduction
  - Niche application of OPS to DC networks

- Myths
  - Immature architecture
  - Lack in optical RAM buffer

- Promise: Enabling technologies at hand
  - Prototyping energy-efficient optoelectronic packet router
  - OPS/OCS/VOCS DC network performs well against mice & elephant

- Summary

OPS: Optical packet switching
OCS: Optical circuit switching
VOCS: Virtual OCS
Small size FDL work well !!!

Fixed-length packet @40~100Gbps in 44km, 12-node MAN

TCP Stop&wait (SAW)

☐ A single FDL increases the application throughput from 160Gb/s to 240Gb/s

Small size FDL work well !!!

Fixed-length packet @40~100Gbps in 12-node MAN

No appreciable improvement between 16~256 FDLs.

Outline

- Introduction
  - Niche application of OPS to DC networks

- Myths
  - Immature architecture
  - Lack in optical RAM buffer

- Promise: Enabling technologies at hand
  - Prototyping energy-efficient optoelectronic packet router
  - OPS/OCS/VOCS DC network performs well against mice & elephant

- Summary

OPS: Optical packet switching
OCS: Optical circuit switching
VOCS: Virtual OCS
Hybrid optoelectronic packet router (HOPR)

To be continued

Targets:
- Power < 100mW/Gbps (=nJ/bit)
- Operation at > 100~400 Gbps
- Latency < 100 ns


100Gbps optical packet

Header    Payload

25Gbps x 4\(\lambda\)

100-Gbps optical packet

CMOS RAM buffer

Fiber delay line buffer

Controller

Label processor

8 x 8 optical switch

100-Gbps BM Tx

CMOS RAM

100-Gbps BM Rx

10GbE

TOR

TOR

TOR

TOR

TOR
Enabling device technologies at hand

OCTA: Optically clocked transistor array
- Self-clocked

Broadcast & select optical switch
- 1x8 EAM gate array
- 8x8 optical switch module

Parallel-to-serial converter
1.6G, 16ch → 25G, 1ch
- NRZ output
- Tolerance to optical trigger

Tunable transmitter
- New type (PRR) TLD
- Integrated with 25G EAM
  (25-G Burst-mode APD-TIA)
- High, flat output power

Optical clock pulse train generator
- Self stabilization
- Compact low power module

Serial-to-parallel converter
25G, 2ch → 1.6G, 32ch
Dually-operating SPC

Burst-mode APD-TIA
- High sensitivity
- Quick response from the 1st bit

Low-jitter latch clock generator

Fiber delay line
Shared buffer
100G Rx
100G Tx
CMOS

Controller
8x8 Optical switch

Label processing
Controller

Label separating circuit

Burst-mode EDFA
8x8 broadcast & select switch

SOA
- Pattern effect
- Poor SNR
- High power
- High WDL/TDL
- Current control
- Very compact
- Extinction, Speed

SOA vs. EAM+EDFA

EAM+EDFA
- No pattern effect
- Small NF
- Low power
- Low WDL/TDL
- Voltage control
- Compact
- Extinction, Speed

Switching speed < 10 ns, Extinction ratio < -40 dB, Power consumption < 4 W
Target specification of HOPR prototype

- **100 Gbps, 8x8 Optical packet switch**
  - **40 W**
  - **5 W / 100G-port**

**Constructed in 2009**

**Next HOPR (3-D torus)**

<table>
<thead>
<tr>
<th>Current</th>
<th>Previous in 2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data rate</td>
<td>100 Gbps</td>
</tr>
<tr>
<td>Switch</td>
<td>8 x 8 (B&amp;S)</td>
</tr>
<tr>
<td>Throughput</td>
<td>1.28 Tbps</td>
</tr>
<tr>
<td>Power consumption</td>
<td>110 W [86 mW/Gbps]</td>
</tr>
<tr>
<td>Latency</td>
<td>100 ns</td>
</tr>
</tbody>
</table>

- **Fiber Delay Line**
- **Shared buffer**
- **Aggregation Switch**
  - **70 W**
  - **10 GbE**
- **Label Processors**
- **ToR**
- **CMOS**
- **100G Rx**
- **100G Tx**
- **ToR**
- **40 W**

July 4, 2016

OECC / PS2016
Energy consumption: EPS vs. HOPR

Electronic router

- A. Power & heat management: 35%
- B. IP look-up & forwarding engine: 32%
- C. I/O: 7%
- D. Control plane: 11%
- E. Switch fabric: 10%
- F. Buffers: 5%

Source: G. Epps, Cisco, 2007

HOPR

- A: 16%
- B: 22%
- C: 44%
- D: 13%
- E: 3%
- F: 2%

Source: R. Takahashi, NTT, 2016
Outline

- **Introduction**
  - Niche application of OPS to DC networks

- **Myths**
  - Immature architecture
  - Lack in optical RAM buffer

- **Promise**: Enabling technologies at hand
  - Prototyping energy-efficient optoelectronic packet router
  - OPS/OCS/VOCS DC network performs well against mice & elephant

- **Summary**

OPS: Optical packet switching
OCS: Optical circuit switching
VOCS: Virtual OCS
Best mixture of protocol: OPS and OCS

**Optical packet switching**
- Best effort service
- Relatively small data
- High resource utilization due to statistical multiplexing
- Packet loss

**Optical circuit switching**
- QoS required
- Set of large/stream data
- Dedicated optical path
- QoS guaranteed
- Optical path destroys the NW topology

6-D torus network

Open Flow controller

Internet

Gateway

100Gbps = 25Gbps x 4λ

HOPR

10GbE

ToR

July 4, 2016

OECC / PS2016
Contention resolution strategy

**Deflection routing**

- S1 → S2 → S3 → (6, 6, 6)
- S4 → S5 → S6

**Buffering strategy**

- HOPR Contention!
- FDL
- RAM
- Destined node

**Breaking path loop**

- Destined node
- FDL
- HOPR
- Buffering strategy

**Ordinary packet path**

- Express path
- Ordinary packet path

July 4, 2016

OECC / PS2016
## Parameters in numerical simulations

**HOPR**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of FDLs ($D$)</td>
<td>4</td>
</tr>
<tr>
<td>Number of 10 Gbit/s links ($A$)</td>
<td>20</td>
</tr>
<tr>
<td>Number of ports between buffer and switch ($M$)</td>
<td>2</td>
</tr>
<tr>
<td>Size of each input queue in the shared buffer ($Q$)</td>
<td>10</td>
</tr>
</tbody>
</table>

**Network**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimension ($N$)</td>
<td>6</td>
</tr>
<tr>
<td>Number of nodes ($4^6 = 4096$)</td>
<td></td>
</tr>
<tr>
<td>Additional TTL ($\alpha$)</td>
<td>50</td>
</tr>
</tbody>
</table>

**Traffic injected by each server & Simulation**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic pattern</td>
<td>Uniform$^b$</td>
</tr>
<tr>
<td>Injection process</td>
<td>Bernoulli$^b$</td>
</tr>
<tr>
<td>Packet length</td>
<td>1500B (Deterministic)</td>
</tr>
<tr>
<td>Ordinary packet load ($\rho$)</td>
<td>0.1~1.0</td>
</tr>
<tr>
<td>Express flow length ($L$)</td>
<td>10 packets</td>
</tr>
<tr>
<td>Express packet load ($\rho'$)</td>
<td>0.01, 0.001</td>
</tr>
<tr>
<td>Cycle period (Time-slot)</td>
<td>120 ns (=1500B @100Gbps)</td>
</tr>
</tbody>
</table>

---

Drastic reduction of packet dropping

**W/o contention resolution**

**With contention resolution**

Extended work to leaf & spine fabric, TuF2.

**Note**: 885 ExPs traverse 22% of 4,096 routers.

Further improvement by virtual OCS (VOCS)

OPS + OCS

OPS + VOCS

Packet loss probability

Packet

Pure OPS & VOCS

Further improvement by virtual OCS (VOCS)

OPS + OCS

OPS + VOCS

Summary

- Niche application of OPS would be DC networks.

- Optical RAM buffer is a long shot in the near future, but FDL can improve the performance.

- Enabling device technologies for extremely energy-efficient optoelectronic packet router are at hand. But it’s a challenge to gain their cost-competitiveness.

- OPS/OCS/VOCS performs best against mice & elephant data flows.
Thank you!

E-mail: kitayama@gpi.ac.jp
http://www.kitayama.tech (personal)