Cross-layer Hybrid and Optical Packet Switching

Artur Minakhmetov

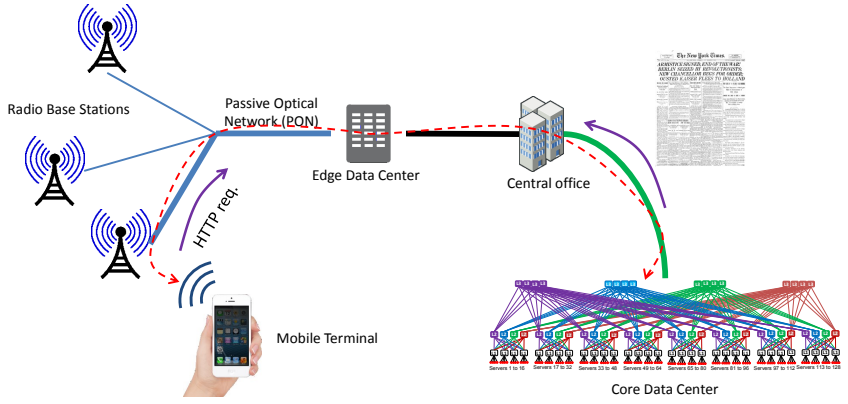
LTCI, Télécom Paris, Institut Polytechnique de Paris

Thesis Defense: Télécom Paris, December 4, 2019

Outline

Introduction into Optical Networks

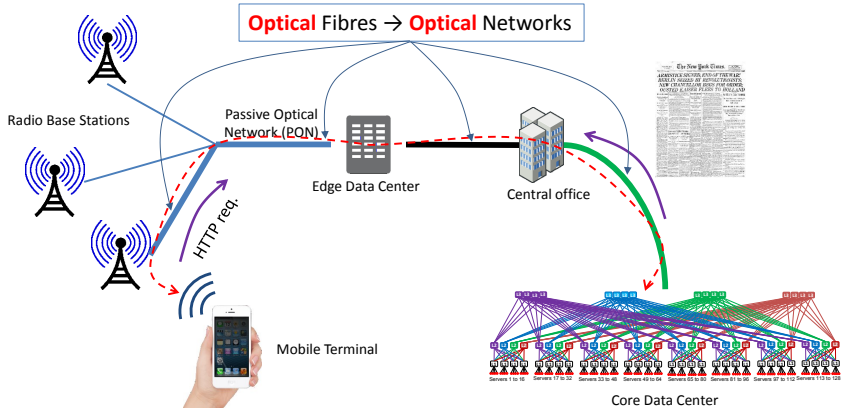
Optical Network in Our Daily Lives



- Simple news reading involves **connections through** all kinds of networks.

Introduction into Optical Networks

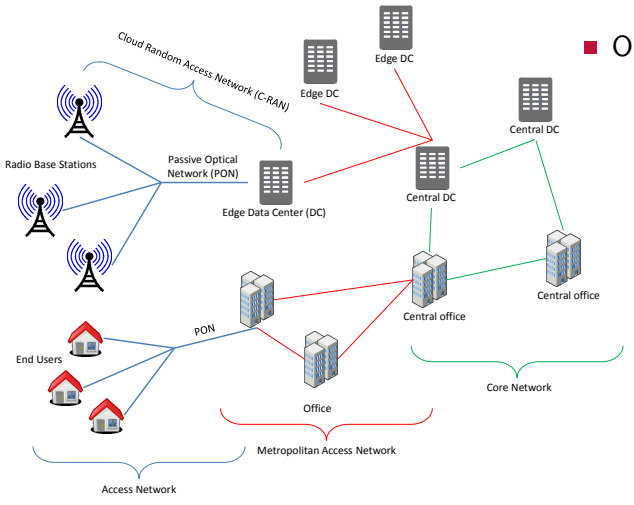
Optical Network in Our Daily Lives



- Simple news reading involves **connections through** all kinds of networks.
- These are **optical** networks.

Introduction into Optical Networks

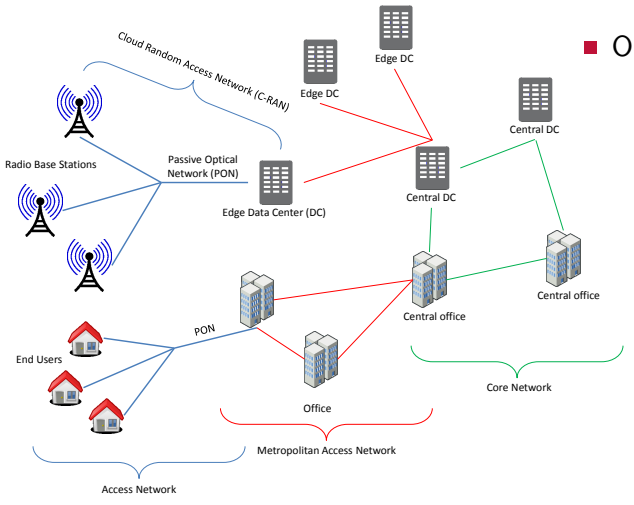
Types of Network



- Optics enables **all** network segments:
  - Access Networks:
    - 4G, 5G enabled by C-RAN
    - Broadband Access by PONs

Introduction into Optical Networks

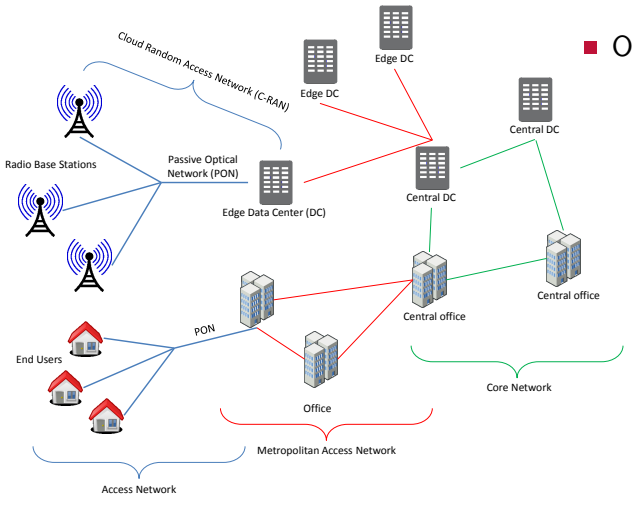
Types of Network



- Optics enables **all** network segments:
  - Access Networks:
    - 4G, 5G enabled by C-RAN
    - Broadband Access by PONs
  - Metropolitan and Core Networks:
    - Enabled by Dense Wavelength Division Multiplexing (DWDM)

Introduction into Optical Networks

Types of Network

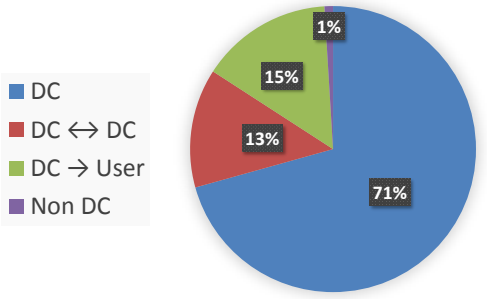


- Optics enables **all** network segments:
  - Access Networks:
    - 4G, 5G enabled by C-RAN
    - Broadband Access by PONs
  - Metropolitan and Core Networks:
    - Enabled by Dense Wavelength Division Multiplexing (DWDM)
  - Data Center (DC) Networks (DCN):
    - Need **low power**, **high bandwidth** communications $\Rightarrow$ use optical transceivers and fibers

Introduction into Optical Networks

Traffic Distribution

2021 Yearly Global Traffic: 20.8 ZBs



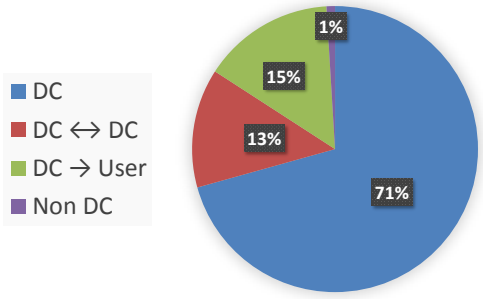
- "Non DC"+"DC⇒User" traffic:
  - will grow **3 times** from 2017 till 2022
- Intra-DC traffic:
  - will grow **3 times** from 2016 till 2021
  - bigger than Inter-DC by **factor of 5**
  - bigger than "Non DC" by **factor of 70**

Cisco Global Cloud Index: Forecast and Methodology, 2016–2021
Cisco Visual Networking Index: Forecast and Trends, 2017–2022

Introduction into Optical Networks

Traffic Distribution

2021 Yearly Global Traffic: 20.8 ZBs

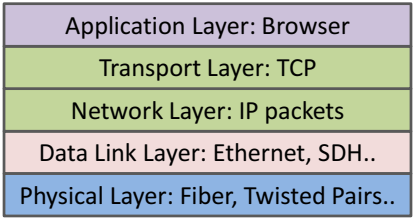


- "Non DC"+"DC⇒User" traffic:
  - will grow **3 times** from 2017 till 2022
- Intra-DC traffic:
  - will grow **3 times** from 2016 till 2021
  - bigger than Inter-DC by **factor of 5**
  - bigger than "Non DC" by **factor of 70**

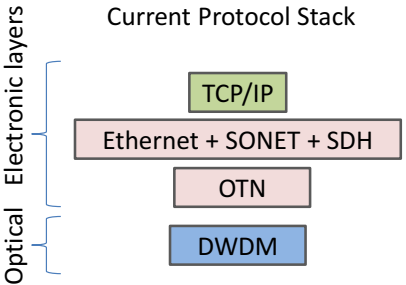
- Because of fast traffic growth and dominance of Intra-DC traffic:
  - Data Center Networks is the **future bottleneck** and will require **new solutions**

Cisco Global Cloud Index: Forecast and Methodology, 2016–2021
Cisco Visual Networking Index: Forecast and Trends, 2017–2022

Layered Structure of Telecommunications



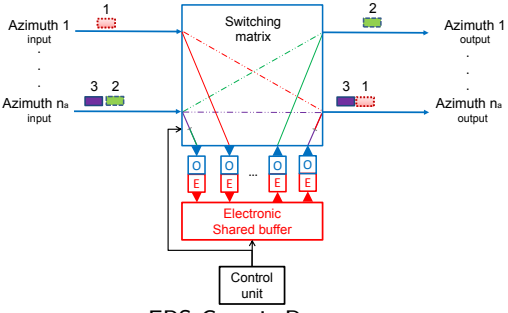
- Application Lv. manages data transmission as a whole.
- Transport Lv. cuts file on Data Units (DU), sends DUs.
- Network Lv. encapsulates transport DU in IP packets.
- Data Link Lv. puts IP pckt. in frames to send over a link.
- Physical Lv. transmission media in a link.



- Packets/frames a routed **electronically** in network.
- DWDM can route *streams* of packets/frames **optically**, but still relaying on point-to-point links, now λ-specific .

Electronic Packet Switching (EPS)

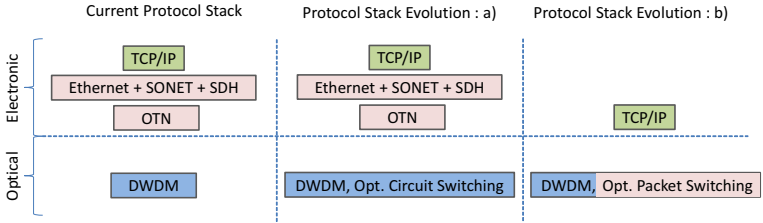
Electronic Packet Switching – core packet routing technology, currently applied in networks.



EPS Generic Router

- Numerous Optical-Electronic-Optical (OEO) conversions entail **high power consumption.**
- Store-and-forward mode of switching contributes to **high network latency**.

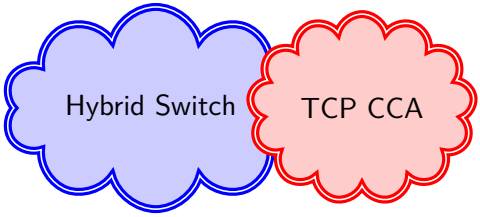
Towards All-Optical Networks through Cross-Layer



Possible evolution of protocol stack towards all-optical networks

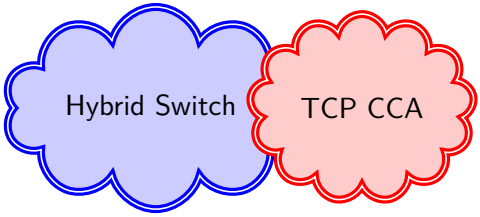
- Energy, latency and traffic needs require going away from EPS:
  - By evolving protocol stack, going cross-layer.
  - By using existing technologies:
    - Optical Circuit Switching (OCS).
    - Optical Packet Switching (OPS).
  - By proposing new solutions:
    - Using TCP Congestion Control Algorithm (CCA) to manage OPS.
    - Hybrid Optical Packet Switching (HOPS).

Thesis Contributions



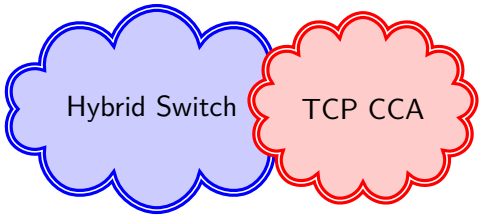
- This Ph.D. research project:
  - Considering OCS and OPS vulnerability.

Thesis Contributions



- This Ph.D. research project:
  - Considering OCS and OPS vulnerability.
  - Considering OPS enablers:
    - device level: Hybrid Optical Packet Switching (HOPS).
    - network level: specific TCP CCA for OPS.

Thesis Contributions

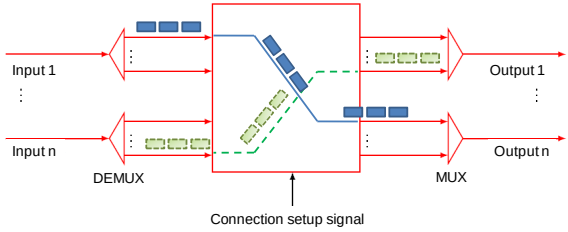


- This Ph.D. research project:
  - Considering OCS and OPS vulnerability.
  - Considering OPS enablers:
    - device level: Hybrid Optical Packet Switching (HOPS).
    - network level: specific TCP CCA for OPS.
  - Investigates combination of HOPS with TCP CCA in Data Centers:
    - Are there protocols that are adapted to be used for HOPS?
    - What would be overall gain in throughput, energy saving and latency?
    - Can we apply class-specific witching rules for HOPS?

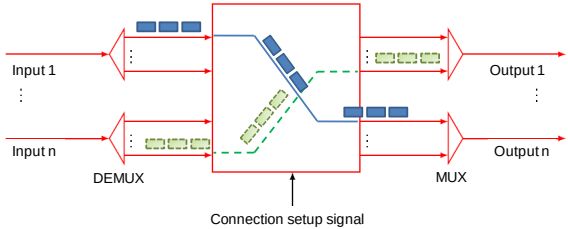
Outline

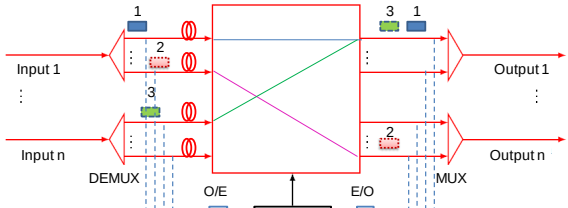
- Motivation
- All Optical Data Centers Networks (AO-DCNs) Solutions
  - Optical Circuit and Packet Switching (OCS and OPS)
  - Device Level OPS Enabler: Hybrid Switch (HOPS)
  - Network Level OPS Solution: Use TCP Stop-And-Wait (SAW)
  - New TCP: SAW -> SAWL for HOPS
  - TCP SACK, adaptation for HOPS
- Switching and Data Center Network Model
- AO-DCN: General Network Performance
- AO-DCN: Energy Consumption
- AO-DCN: Latency
- AO-DCN: Classes of Service
- Conclusion

Optical Circuit and Packet Switching (OCS and OPS)



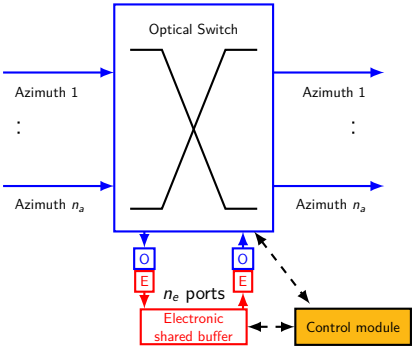
Optical Circuit and Packet Switching (OCS and OPS)





Device Level OPS Enabler: Hybrid Switch (HOPS)

Hybrid Optical Packet Switching



Hybrid Packet Switch Concept

- Hybrid switch = cut-through **all-optical** switch + shared **electronic** buffer.
  - Switch has n_e Input/Output (I/O) ports of buffer.
  - If packet is blocked: put it into shared electronic buffer.
  - If output port is release, packet is re-emitted FIFO.

Network Level OPS Solution: Use TCP Stop-And-Wait (SAW)

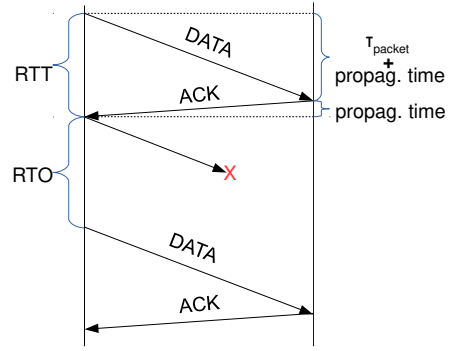
TCP Congestion Control Algorithms (CCA) application

- TCP Congestion Control Algorithms:
  - Embrace contention and high PLR.
  - ACK of packet required to send next.
  - No ACK after Retransmission Time Out (RTO) $\longrightarrow$ retransmit.

Network Level OPS Solution: Use TCP Stop-And-Wait (SAW)

TCP Congestion Control Algorithms (CCA) application

- TCP Congestion Control Algorithms:
  - Embrace contention and high PLR.
  - ACK of packet required to send next.
  - No ACK after Retransmission Time Out (RTO) → retransmit.
- TCP Stop-And-Wait (SAW) in data centers (DC):
  - One packet in flight
  - If ACK: $RTO_i = \beta \cdot RTT + (1 - \beta) \cdot RTO_{i-1}$
  - Else: $RTO_i = \alpha \cdot RTO_{i-1}$



TCP SAW Working principle

Conditions: $RTO_1 = 1\ ms, \max(RTO) = 60\ s, \alpha > 1, \beta \in (0, 1)$, RTT =Round Trip Time

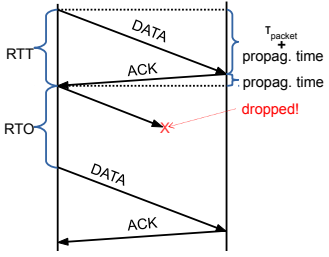
P.J. Argibay-Losada et al, Using Stop-and-Wait to Improve TCP Throughput in Fast Optical Switching (FOS) Networks over Short Physical Distances, INFOCOM 2014

New TCP: SAW → SAWL for HOPS

- Introducing TCP Stop-And-Wait-Longer (SAWL):
  - Adapted for a data center network with hybrid switches.
  - Verified in simulations.

New TCP: SAW → SAWL for HOPS

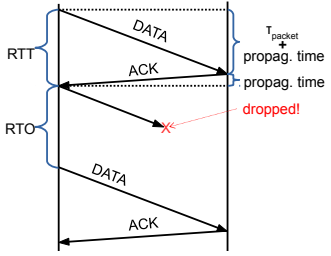
- Introducing TCP Stop-And-Wait-Longer (SAWL):
  - Adapted for a data center network with hybrid switches.
  - Verified in simulations.



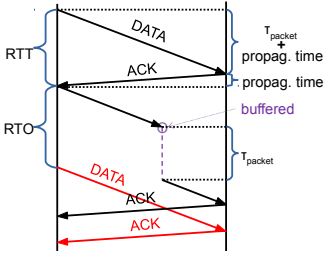
SAW for OPS

New TCP: SAW → SAWL for HOPS

- Introducing TCP Stop-And-Wait-Longer (SAWL):
  - Adapted for a data center network with hybrid switches.
  - Verified in simulations.



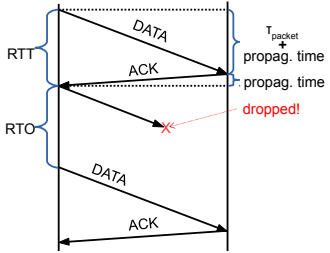
SAW for OPS



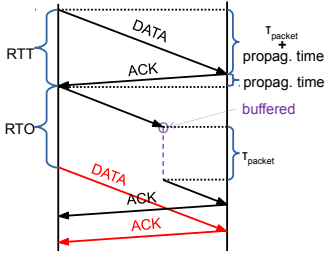
SAW for HOPS

New TCP: SAW → SAWL for HOPS

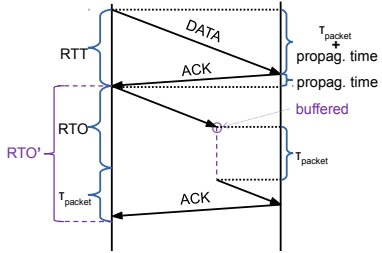
- Introducing TCP Stop-And-Wait-Longer (SAWL):
  - Adapted for a data center network with hybrid switches.
  - Verified in simulations.



SAW for OPS



SAW for HOPS



SAWL for HOPS

New TCP: SAW → SAWL for HOPS

- Introducing TCP Stop-And-Wait-Longer (SAWL):
  - Adapted for a data center network with hybrid switches.
  - Verified in simulations.
- SAW and SAWL differences:
  - SAW considers buffered packet as lost and retransmits prematurely.
  - SAWL increase RTO so packet buffered p times, wouldn't be considered lost.

Event	TCP SAW	TCP SAWL
If ACK:	$RTO_i = \beta \cdot RTT + (1-\beta) \cdot RTO_{i-1}$	$\mathbf{RTO'_i = RTO_i + p \cdot \tau}$
Else:	$RTO_i = \alpha \cdot RTO_{i-1}$	$RTO'_i = \alpha \cdot RTO'_{i-1}$

RTO definition for SAW and SAWL

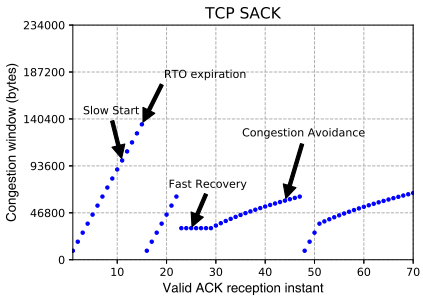
SAWL conditions: τ = data packet duration, depending on emitter bit-rate, $p = 4$ for simulations of SAWL

TCP SACK, adaptation for HOPS

- TCP SACK:
  - Based on conventional CCA – TCP Reno.
  - Use conventional RTO update rules, but $RTO_{init} = 1\ ms$ contrary to $RTO_{init} = 1\ s$.
  - May have several packets in flight, regulated by Congestion WiNDow (CWND) (Bytes).
  - Use Selective ACK (SACK), i.e. acknowledge data range received, not just a packet.

TCP SACK, adaptation for HOPS

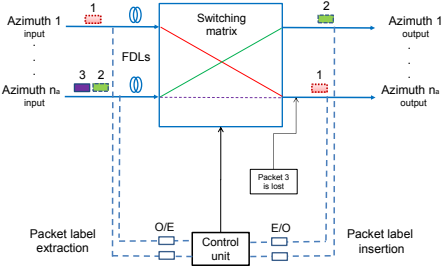
- TCP SACK:
  - Based on conventional CCA – TCP Reno.
  - Use conventional RTO update rules, but $RTO_{init} = 1\ ms$ contrary to $RTO_{init} = 1\ s$.
  - May have several packets in flight, regulated by Congestion WiNDow (CWND) (Bytes).
  - Use Selective ACK (SACK), i.e. acknowledge data range received, not just a packet.
- Phases of CWND evolution:
  - Exponential growth during "Slow Start".
  - Constant level during "Fast Recovery".
  - Linear growth during "Congestion Avoidance".



CWND evolution for SACK

Outline

Optical Packet Switching Model

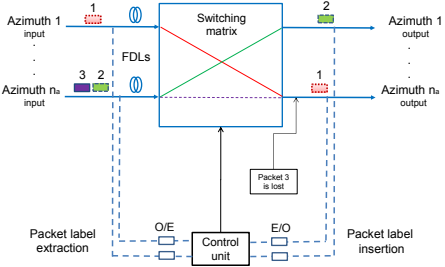


General architecture of all-optical packet switch

- Packet 1 is switched optically.
- Packet 2 is switched optically.
- Packet 3 is blocked by 1st and dropped.

- Data Packets:
  - travel along with labels, containing routing information.
  - labels are read by Control Unit (CU) passing through OEO conversions.
  - are delayed by Fiber Delay Lines (FDL) so CU can configure switching matrix.
  - are switched optically without OEO conversion and labels are regenerated.

Optical Packet Switching Model

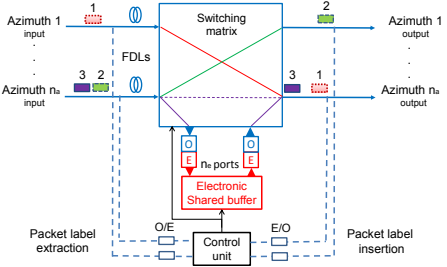


General architecture of all-optical packet switch

- Packet 1 is switched optically.
- Packet 2 is switched optically.
- Packet 3 is blocked by 1st and dropped.

- Switching blocks are generic:
  - Switching matrix can be realized by Broadcast & Select scheme with Semiconductor Optical Amplifiers (B&S+SOA), or Mach–Zehnder Interferometers (MZIs) array.
  - Label management, labels can be extracted by 90:10 splitter, or carried out of band.
  - Control Unit (CU) can be realized by Field-Programmable Gate Array (FPGA).
  - Switching time is on order of ns (we consider 0 for simulations).

Hybrid Optical Packet Switching Model

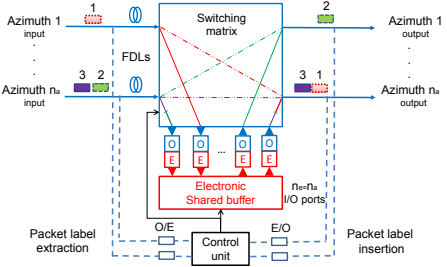


- Packet 1 is switched optically.
- Packet 2 is switched optically.
- Packet 3 is blocked by 1st and switched electronically through buffer.

General architecture of hybrid optical packet switch

- Hybrid Switch has *shared* electronic buffer with n_e buffer I/O, realized by Burst Transceivers.
- Buffer accepts blocked packets if free to re-emit them FIFO, otherwise they are dropped.
- $n_e = 0$ corresponds to Optical Packet Switching case.
- Packets switched in dual mode:
  - *cut-through* mode by optical switching matrix.
  - *store-and-forward* mode by buffer $\Rightarrow$ **P.3** is delayed with respect to **P.1**.

Electronic Packet Switching Model

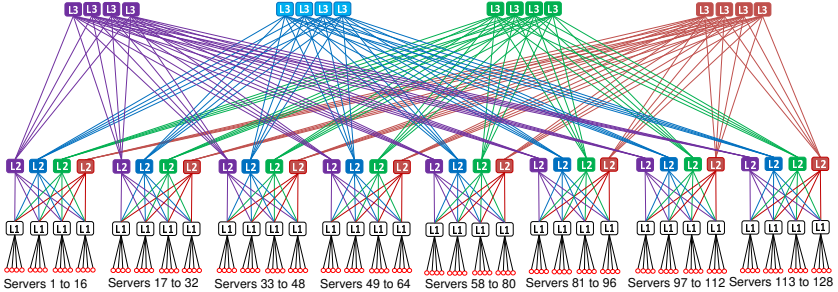


- Packet 1 is switched electronically.
- Packet 2 is switched electronically.
- Packet 3 is switched electronically.

General architecture of all-electronic packet switch

- Electronic Switch can be considered as special case of hybrid switch with $n_a = n_e$, but with all packets passing buffer, none blocked.
- Each packet corresponds to OEO conversion $\Rightarrow$ high energy consumption.
- All packets are switched in *store-and-forward* mode $\Rightarrow$ high latency.

Data Center Network Topology



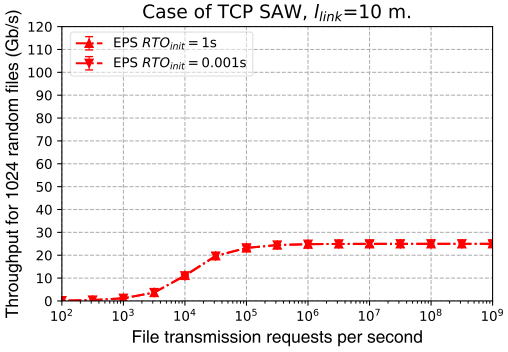
8-ary fat-tree DC network (related to **Facebook DC** network)

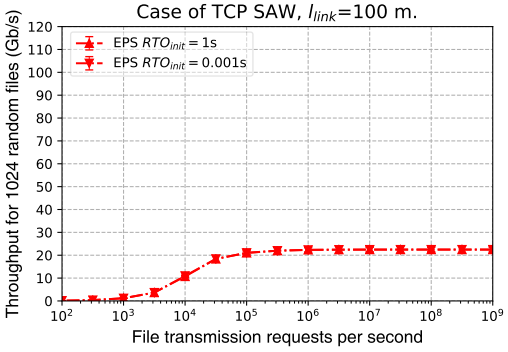
- 3 level of switches with each $n_a = 8$ I/O ports interconnected by $l_{link} = \{10, 100\} m$.
- All switches are of the same type: OPS($n_e = 0$), HOPS ($n_e = var$) or EPS.
- File transmission through TCP connection simulated, with packet size = 9kB on 10 Gbit/s.
- Log-normal distribution of files.
- Load – mean number of file transmission requests/s (req/s) in Poissonian process.
- Network *throughput, energy consumption, latency* are studied as function of load.

Outline

TCP SAW Throughput Analysis

TCP SAW for OPS, HOPS and EPS

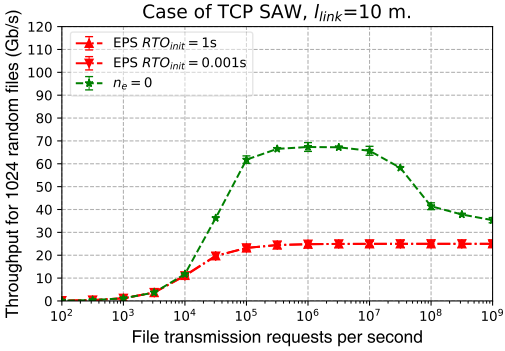


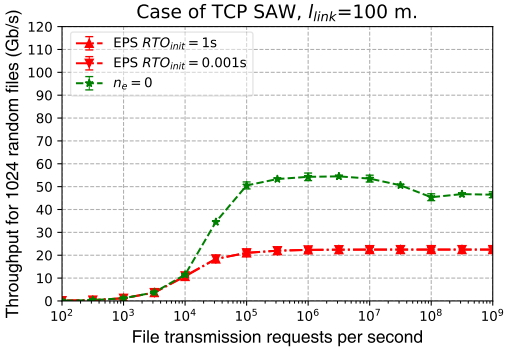


- EPS performs poorly due to high latency, invoked by store-and-forward mode.

TCP SAW Throughput Analysis

TCP SAW for OPS, HOPS and EPS

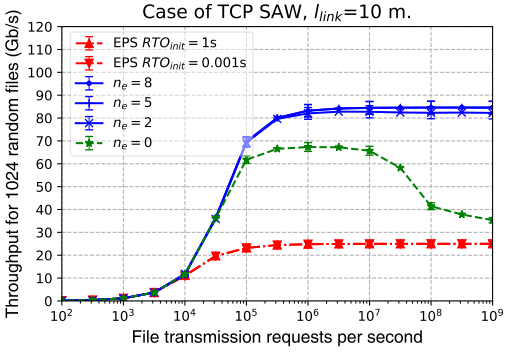


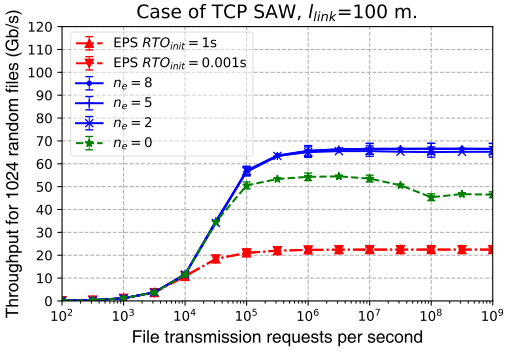


- EPS performs poorly due to high latency, invoked by store-and-forward mode.
- Performance of OPS drops on high load.
- SAW is sensible to l_{link} changes.

TCP SAW Throughput Analysis

TCP SAW for OPS, HOPS and EPS

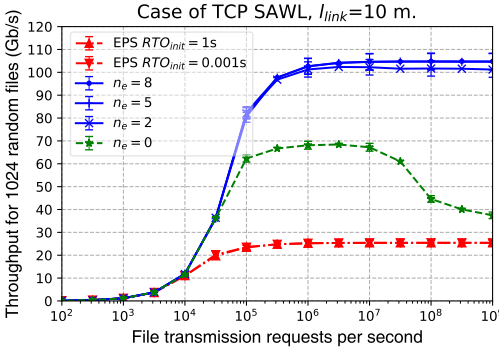


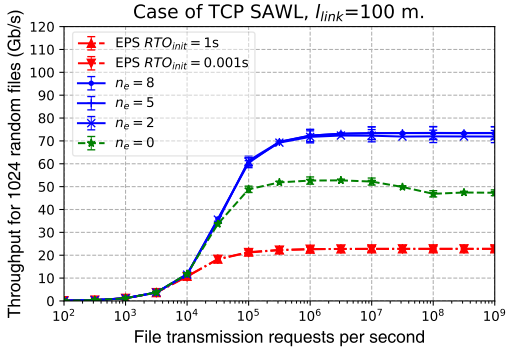


- EPS performs poorly due to high latency, invoked by store-and-forward mode.
- Performance of OPS drops on high load.
- SAW is sensible to l_{link} changes.
- HOPS **outperforms** OPS with few n_e even with SAW, without drop on high load.

TCP SAWL Throughput Analysis

TCP SAWL for OPS, HOPS and EPS

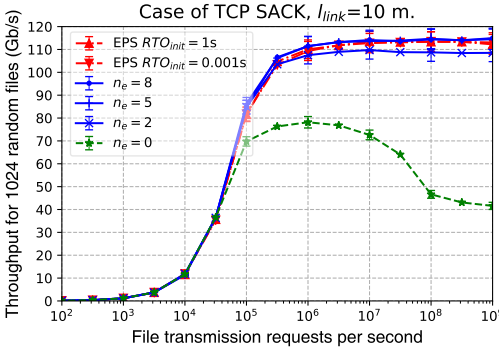


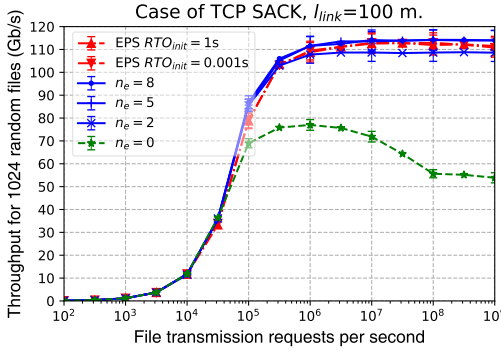


- SAWL outperforms SAW by 25% for $l_{link} = 10$ m and 10% for $l_{link} = 100$ m on HOPS.
- SAWL on HOPS ($n_e = 2$) outperforms OPS by **50%** at least.
- SAWL on HOPS ($n_e = 2$) outperforms SAW on OPS more than by **200%** on 10^9 req/s.
- SAWL is still sensible to l_{link} changes.

TCP SACK Throughput Analysis

TCP SACK for OPS, HOPS and EPS





- SACK outperforms SAWL by 10% for $l_{link} = 10$ m and 50% for $l_{link} = 100$ m on HOPS.
- SACK on HOPS ($n_e = 2$) outperforms OPS by $\approx 40\%$, when OPS performs at its best.
- SACK on HOPS ($n_e = 2$) is very close to EPS, and on HOPS ($n_e = 5$) outperforms EPS.
- SACK is unaffected by l_{link} change.

Discussion on Solutions for Best Throughput

- HOPS **outperforms** OPS with few n_e even with SAW.
- SAWL+HOPS ($n_e = 2$) outperforms SAW+OPS more than by **200%** on 10^9 req/s.
- SACK outperforms SAWL by **only 10%** for $l_{link} = 10m$ and 50% for $l_{link} = 100m$ on HOPS.
- SAWL is close to SACK for $l_{link} = 10m$ in throughput, but who is better in terms of:
  - **OEO reduction** & **Latencies** (RTTs)?

Discussion on Solutions for Best Throughput

- HOPS **outperforms** OPS with few n_e even with SAW.
- SAWL+HOPS ($n_e = 2$) outperforms SAW+OPS more than by **200%** on 10^9 req/s.
- SACK outperforms SAWL by **only 10%** for $l_{link} = 10m$ and 50% for $l_{link} = 100m$ on HOPS.
- SAWL is close to SACK for $l_{link} = 10m$ in throughput, but who is better in terms of:
  - **OEO reduction** & **Latencies** (RTTs)?

Now let us consider energy consumption aspect.

Outline

Motivation for Energy Efficient Data Center Networks

- IT sector energy consumption **growing 9%/year**, currently 4% carbon emissions

Motivation for Energy Efficient Data Center Networks

- IT sector energy consumption **growing 9%/year**, currently 4% carbon emissions
  - Up to **60 %** of energy consumption is for switching and transport in DCN.
  - Currently: Electronic Packet Switching (EPS) over optical fiber network
  - Packets need Optical-Electrical-Optical conversion at every switch!
  - Packet Loss Ratio (PLR) $\simeq 0$

Motivation for Energy Efficient Data Center Networks

- IT sector energy consumption **growing 9%/year**, currently 4% carbon emissions
  - Up to **60 %** of energy consumption is for switching and transport in DCN.
  - Currently: Electronic Packet Switching (EPS) over optical fiber network
  - Packets need Optical-Electrical-Optical conversion at every switch!
  - Packet Loss Ratio (PLR) $\simeq 0$
- Optical or Hybrid Packet Switching (OPS/HOPS):
  - More efficient capacity use (packet mode)
  - Packets pass switches without OEO conversion
  - Reduces number of transceivers, can use burst mode $\Rightarrow$ less light emission (80% of transceiver power)

Motivation for Energy Efficient Data Center Networks

- IT sector energy consumption **growing 9%/year**, currently 4% carbon emissions
  - Up to **60 %** of energy consumption is for switching and transport in DCN.
  - Currently: Electronic Packet Switching (EPS) over optical fiber network
  - Packets need Optical-Electrical-Optical conversion at every switch!
  - Packet Loss Ratio (PLR) $\simeq 0$
- Optical or Hybrid Packet Switching (OPS/HOPS):
  - More efficient capacity use (packet mode)
  - Packets pass switches without OEO conversion
  - Reduces number of transceivers, can use burst mode $\Rightarrow$ less light emission (80% of transceiver power)
- HOPS is the best *practical* candidate for energy consumption reduction over OPS & EPS:
  - higher throughput than OPS and OEO conversions only for buffering w.r.t EPS.

Energy Savings for Data Transport

Metric for Electronic-Optical Conversions Reduction

- Burst Transceiver can spend $>$**80 %** of power on Tx $\rightarrow$ EO conversions predominant

Energy Savings for Data Transport

Metric for Electronic-Optical Conversions Reduction

- Burst Transceiver can spend $>$**80 %** of power on Tx $\rightarrow$ EO conversions predominant
- Metric to measure EO conversions:

$$\text{Bit transport energy factor} = \frac{Data_{pckt}[B] \times EO_{data} + Ack_{pckt}[B] \times EO_{ack}}{Payload[B]}$$

  - Defined as how many bits should be physically emitted to ensure delivery of one bit

Energy Savings for Data Transport

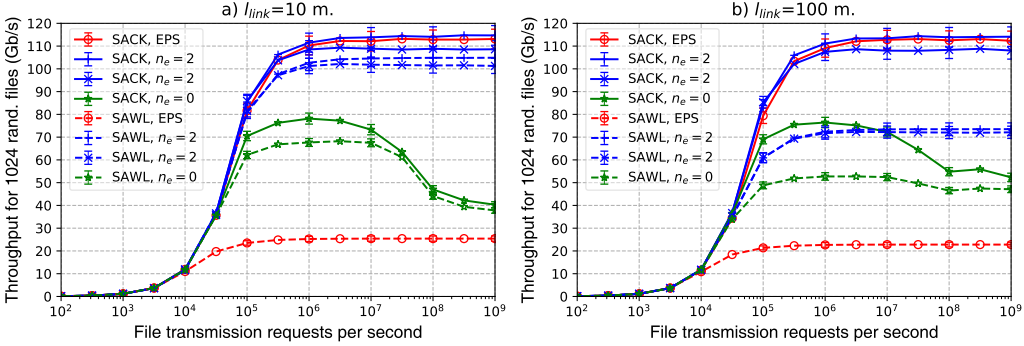
Metric for Electronic-Optical Conversions Reduction

- Burst Transceiver can spend $>$**80 %** of power on Tx $\rightarrow$ EO conversions predominant
- Metric to measure EO conversions:

$$\text{Bit transport energy factor} = \frac{Data_{pckt}[B] \times EO_{data} + Ack_{pckt}[B] \times EO_{ack}}{Payload[B]}$$

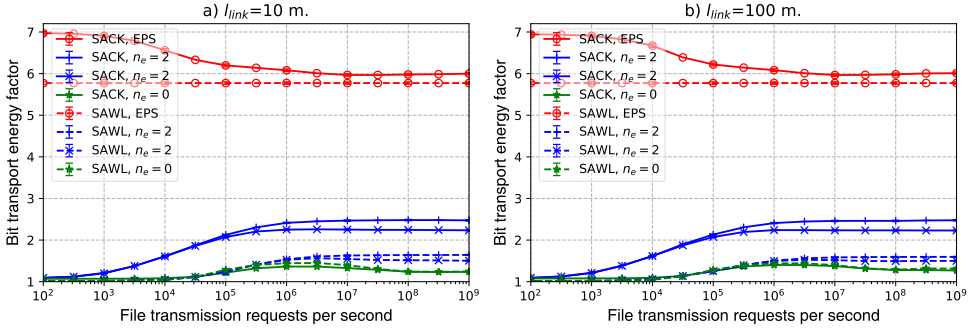
  - Defined as how many bits should be physically emitted to ensure delivery of one bit
  - Takes into account RTO re-transmissions induced by TCP CCA
  - Takes into account EO conversions induced by buffer of a Hybrid Switch
  - Estimates energy consumption by multiplying with [J/b] of a transmitters used
- "Transmission energy cost" measures BTEF under varying network load

Network performance: Throughput (recap for SACK and SAWL)



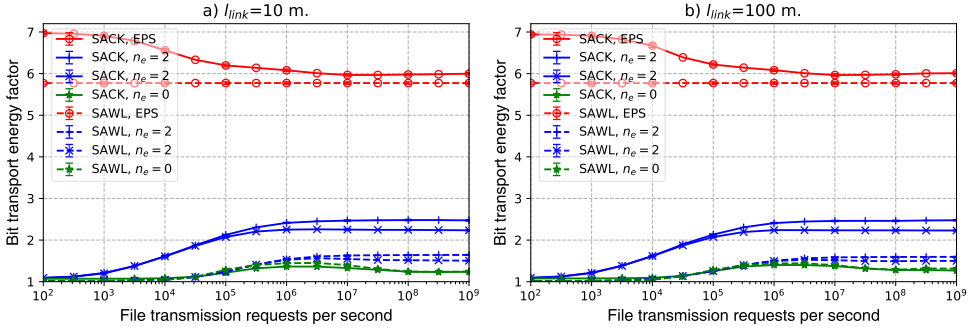
- Performance on high load of OPS drops, HOPS holds with **few n_e** and **outperforms** OPS.
- SACK outperforms SAWL **by only 10%** for $l_{link} = 10$ m and 50% for 100 m on HOPS.
- SACK on HOPS, $n_e = 2$ is very close to EPS, and with $n_e = 8$ **outperforms** EPS.

Network performance: Transmission energy cost (1/2)



- EPS performs **the worst** in terms of energy consumption (all packets OEO).
- Worst case of HOPS outperforms best case of EPS by **factor more than 2**.
- OPS performs **the best** energy-wise (but not throughput-wise).
- No change for different l_{link}.

Network performance: Transmission energy cost (2/2)



- SACK + HOPS consumes $\approx \times 1.5$ more than SAWL + HOPS.
- SAWL + OPS consumes least energy, but as well has lowest throughput.
- SAWL + HOPS, $n_e = 2$ is a trade-off solution for $l_{link} = 10m$ DCN:
  - Throughput: SAWL + HOPS, $n_e = 2$ outperformed **by only 10%** by SACK+EPS.
  - Energy: SAWL + HOPS, $n_e = 2$ saves **4 times** than SACK+EPS.

Discussion on Solutions for Best Energy Savings

- **HOPS** = **robust solution** in OPS data center network with **few n_e**.
- HOPS + SACK delivers **best throughput**, better than EPS +SACK, and energy consumption **reduced by factor of 2** at least.
- HOPS + SAWL delivers **only 10% lower** throughput than best, but help reduce energy consumption energy consumption **by factor of 4**.

Outline

Latency in Data Center Networks

- Current state of things:
  - Low latency,i.e. RTT, in Data Center Network (DCN) is a must.
  - DCN function on Electronic Packet Switching (EPS), and use solutions adapted to it:
    - Data Center TCP (DCTCP), based on existing TCP CCA (i.e. SACK)

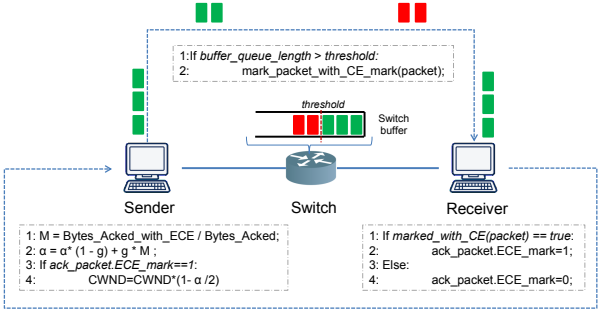
Latency in Data Center Networks

- Current state of things:
  - Low latency,i.e. RTT, in Data Center Network (DCN) is a must.
  - DCN function on Electronic Packet Switching (EPS), and use solutions adapted to it:
    - Data Center TCP (DCTCP), based on existing TCP CCA (i.e. SACK)
- Question to answer:
  - How OPS and HOPS perform latency-wise compared to EPS?
  - What is performance of DCTCP on HOPS?
  - What TCP is the best to use for HOPS, OPS and EPS?

Latency in Data Center Networks

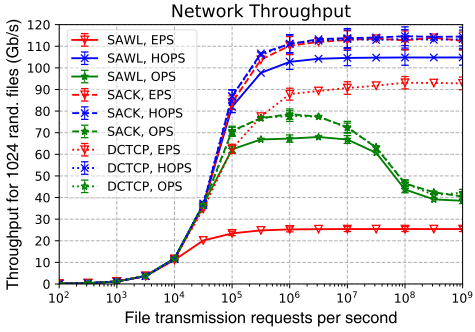
- Current state of things:
  - Low latency,i.e. RTT, in Data Center Network (DCN) is a must.
  - DCN function on Electronic Packet Switching (EPS), and use solutions adapted to it:
    - Data Center TCP (DCTCP), based on existing TCP CCA (i.e. SACK)
- Question to answer:
  - How OPS and HOPS perform latency-wise compared to EPS?
  - What is performance of DCTCP on HOPS?
  - What TCP is the best to use for HOPS, OPS and EPS?
- What metrics to use:
  - Average:
    - Round Trip Time – gives idea about latency.
    - Flow Completion Time (FCT) – time required to finish transmission of flow/file.
  - 99th percentile RTT, FCT – the "worst" case scenario, required by operators.

Data Center TCP (DCTCP)



- Explicit Congestion Notification (ECN) mechanism with indirect sender notification is used.
- Packets gets marked by switch if buffer threshold k passed.
- The Sender calculates a ratio M of data ack-ed with CE to generally ack-ed, over CWND.
- The Sender estimates a parameter α based on M and weight g.
- Upon reception of marked ACK: $CWND_i = \frac{CWND_{i-1}}{1-\alpha/2}$ once per CWND.

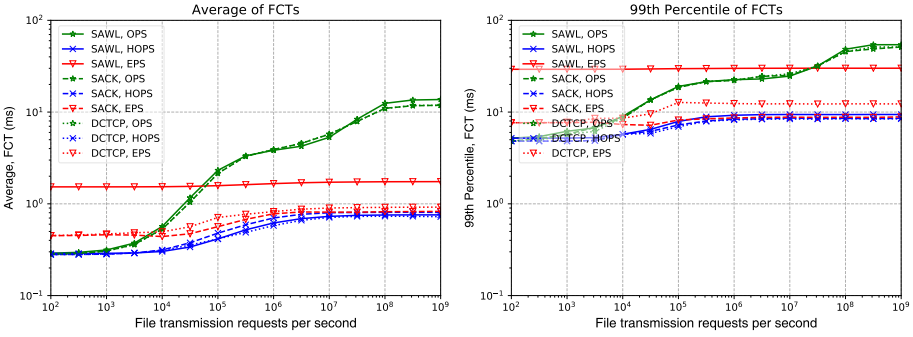
Throughput for DCTCP and other TCP CCAs



Throughput dependence on CCA and load for $l_{link} = 10\ m$

- Throughput – **best** performance:
  - HOPS with DCTCP or SACK
  - EPS with SACK
- Throughput – *acceptable* performance:
  - HOPS with SAWL
  - EPS with DCTCP

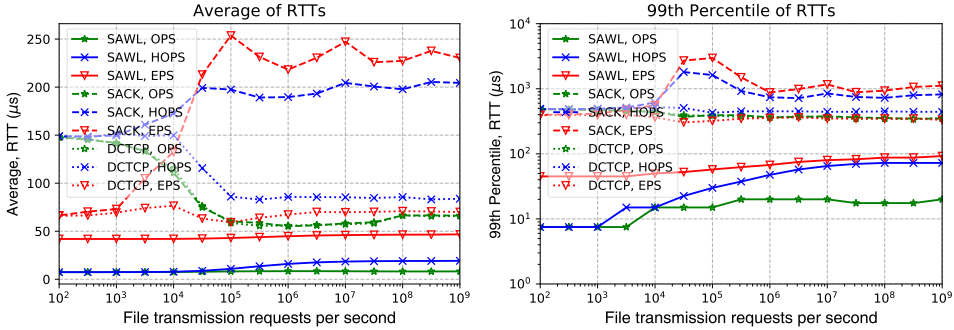
Average FCT for DCTCP and other TCP CCAs



Average and 99th percentile FCT dependence on CCA and load for $l_{link} = 10\ m$

- HOPS outperforms OPS and EPS: < 1 ms for average and < 10 ms for 99th percentile.
- HOPS+SAWL edged by HOPS+DCTCP by 0.1 ms for average and 1 ms for 99th percentile.
- HOPS+DCTCP delivers the **best** performance.
- OPS performs poorly.

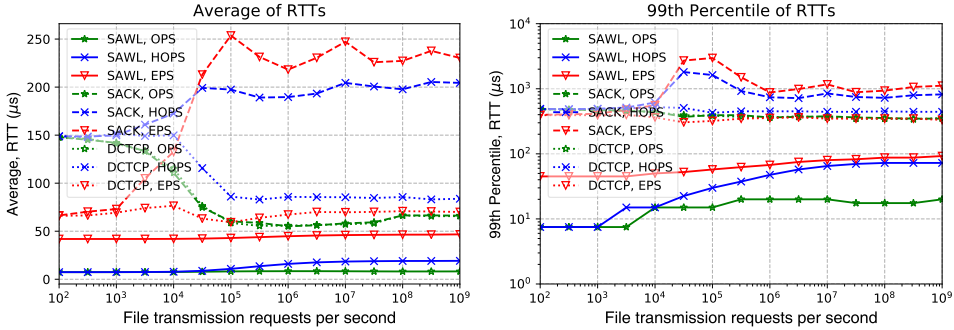
RTT for DCTCP and other TCP CCAs



Average and 99th percentile FCT dependence on CCA and load for $l_{link} = 10\ m$

- SAWL >> DCTCP; DCTCP $\geq$ SACK, but DCTCP+EPS > DCTCP+HOPS.
- OPS+SAWL is better than HOPS+SAWL, which is better than EPS+SAWL.

RTT for DCTCP and other TCP CCAs



Average and 99th percentile FCT dependence on CCA and load for $l_{link} = 10$ m

- SAWL >> DCTCP; DCTCP $\geq$ SACK, but DCTCP+EPS > DCTCP+HOPS.
- OPS+SAWL is better than HOPS+SAWL, which is better than EPS+SAWL.
- HOPS+SAWL is close to best: average $<$ **20 µs**; 99th percentile: $<$ **75 µs**
- HOPS+DCTCP: average $<$ 150 µs; 99th percentile: $<$ 500 µs

Discussion on Solutions for Best Latency

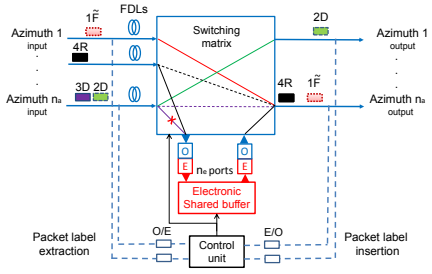
- 1st choice in DC network – HOPS+SAWL
  - Achievable Throughput – **100 Gbit/s**
  - Close to best average RTT: $<$ **20 µs**
  - Close to best 99th RTT percentile: $<$ **75 µs**

Discussion on Solutions for Best Latency

- 1st choice in DC network – HOPS+SAWL
  - Achievable Throughput – **100 Gbit/s**
  - Close to best average RTT: $<$ **20 µs**
  - Close to best 99th RTT percentile: $<$ **75 µs**
- 2nd choice in DC network – HOPS+DCTCP
  - Achievable Throughput – **115 Gbit/s**
  - Average RTT: $<$ **150 µs**
  - 99th RTT percentile: $<$ **500 µs**

Outline

Hybrid Switch with Class Specific Switching Rules



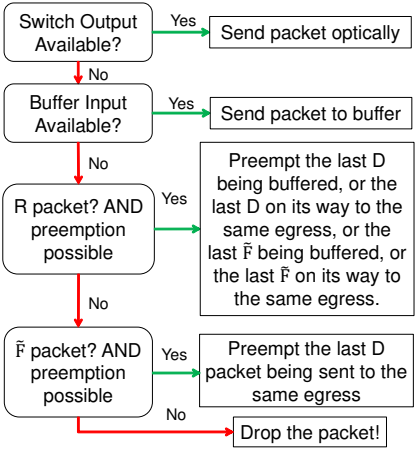
Class Specific Switching Rules on Hybrid Switch

- Classes with priorities: Reliable(R)>Not-So-Fast($\tilde{F}$)>Default(D).
- $1\tilde{F}$ is switched optically.
- 2D is switched optically.
- 3D is blocked by $1\tilde{F} \Rightarrow$ starts buffering.
- 4R is blocked by $1\tilde{F}$ & 3D $\Rightarrow$ 3D is preempted, 4R switched through buffer.

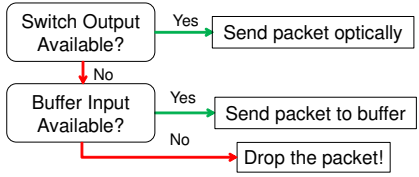
- Why: need to manage the Packet Loss Ratio (PLR) $\Rightarrow$ lose one, but save another.
- How: introduce logic into switching rules.
- Particularity:
  - Needed when $n_a < n_e$, i.e. for OPS or HOPS.
  - Would not work for EPS or $n_a = n_e$ HOPS.

Switching Rules

Preemption strategy switching



Class agnostic switching



Service Classes and DC connections distribution:

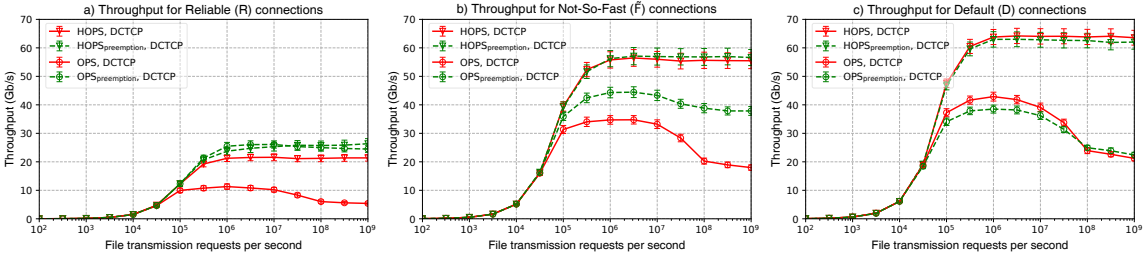
- Reliable (R) – 10%
- Not-So-Fast ($\tilde{F}$) – 40%
- Default (D) – 50%

Source: W. Samoud et. al.,"Performance Analysis of a Hybrid Optical–Electronic Packet Switch Supporting Different Service Classes," J. Opt. Commun. Netw. 7, 952-959 (2015)

Preemption strategy gains

Throughput

HOPS ($n_e = 2$) and OPS ($n_e = 0$)

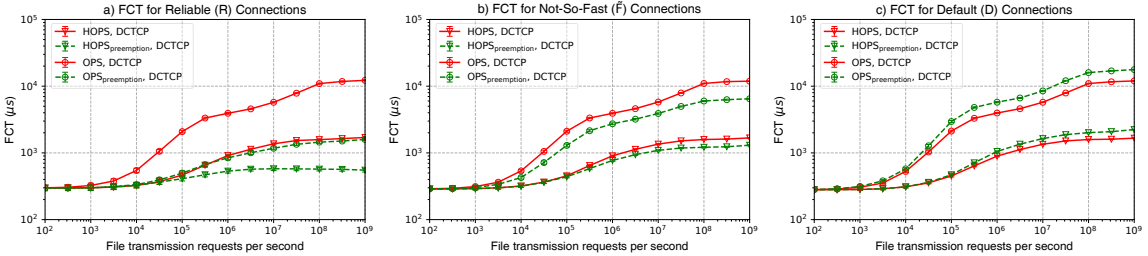


- Reliable Class:
  - HOPS: increase by 25%
  - OPS: increase by 150%
- Not-So-Fast Class:
  - HOPS: almost no change
  - OPS: increase by 30-100%
- Default Class:
  - HOPS: almost no change
  - OPS: decrease by 10%

Preemption strategy gains

Average FCT

HOPS ($n_e = 2$) and OPS ($n_e = 0$)



- Reliable Class:
  - HOPS: reduce by $\times 2$
  - OPS: reduce by $\times 8$
- Not-So-Fast Class:
  - HOPS: reduce by 25%
  - OPS: reduce by $\times 2$
- Default Class:
  - HOPS: slight increase
  - OPS: slight increase

Results Discussion

- Class specific switching rules are for light-weight HOPS solutions: $n_e = 0, 2$
- HOPS and preemption strategies let us to:
  - increase Throughput and decrease FCT in DCN for R, $\tilde{F}$ connections
  - without losing a lot of performance for D connections

Outline

- Motivation
- All Optical Data Centers Networks (AO-DCNs) Solutions
- Switching and Data Center Network Model
- AO-DCN: General Network Performance
- AO-DCN: Energy Consumption
- AO-DCN: Latency
- AO-DCN: Classes of Service
- Conclusion

Conclusion: Research Result Highlights

- HOPS + TCP CCA:
  - Delivers the same throughput as EPS.
  - Saves up to **4 times** of energy compared to EPS.
  - Brings down latency by **factor of 3** compared to EPS.

Conclusion: Research Result Highlights

- HOPS + TCP CCA:
  - Delivers the same throughput as EPS.
  - Saves up to **4 times** of energy compared to EPS.
  - Brings down latency by **factor of 3** compared to EPS.
- **HOPS** = **robust solution** in AO-DCN with **few n_e**.
- TCP **CCAs** discoves potential of hybrid switches and **boosts** network **performance**.
- TCP CCA + hybrid switches = solution for making AO-DCN a reality.

Conclusion: Research Result Highlights

- HOPS + TCP CCA:
  - Delivers the same throughput as EPS.
  - Saves up to **4 times** of energy compared to EPS.
  - Brings down latency by **factor of 3** compared to EPS.
- **HOPS** = **robust solution** in AO-DCN with **few n_e**.
- TCP **CCAs** discoves potential of hybrid switches and **boosts** network **performance**.
- TCP CCA + hybrid switches = solution for making AO-DCN a reality.
- What's next?
  - Study on learning of **p** parameter in SAWL during transmission.
  - Consideration of heterogeneity of networks (EPS+HOPS+OPS).
  - Application of DWDM and study of Wide Area Network (WAN) topologies.
  - Consideration of All-Optical Wavelength Converters (AO-WC).
  - Validation of simulation results in the laboratory.

Thank You!

- Jury Members: Nicola Calabretta, Stefano Secci, Hind Castel, Mounia Lourdiane, Emmanuel Lochin, Daniel Kilper, Nihel Djoher Benzaoui
- Thesis Advisers: Cedric Ware, Luigi Iannone.
- Family: my mother Roza, and my partner Anastasia.
- Friends: Dima, Sérioja, Anton, Marina B.,
- Friends from Télécom: Samet, Akram, Abby, Julien, Alaa, Vincent .
- Friends from Moscow: Anton Ch., Anotn Sch., Volodia Ts., Jaeyeol R., Grisha Ch.