

Hybrid Networking for Cloud Resource Management

T. S. Eugene Ng
Rice University



RICE

Guohui Wang, David Andersen, Michael Kaminsky, Konstantina Papagiannaki, Eugene Ng, Michael Kozuch, Michael Ryan,
"c-Through: Part-time Optics in Data Centers", SIGCOMM'10

Hamid Bazzaz, Malveeka Tewari, Guohui Wang, George Porter, Eugene Ng, David Andersen, Michael Kaminsky, Michael Kozuch, Amin Vahdat,
"Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks", SOCC'11

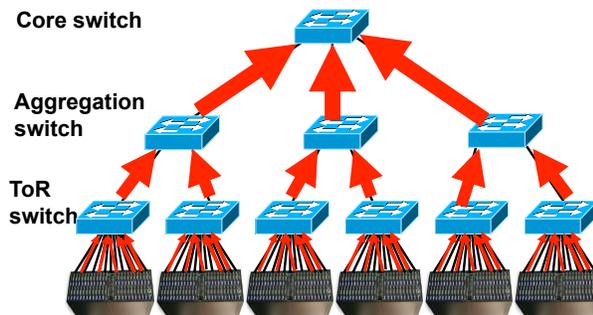
1



RICE

Bandwidth bottleneck in data center networks

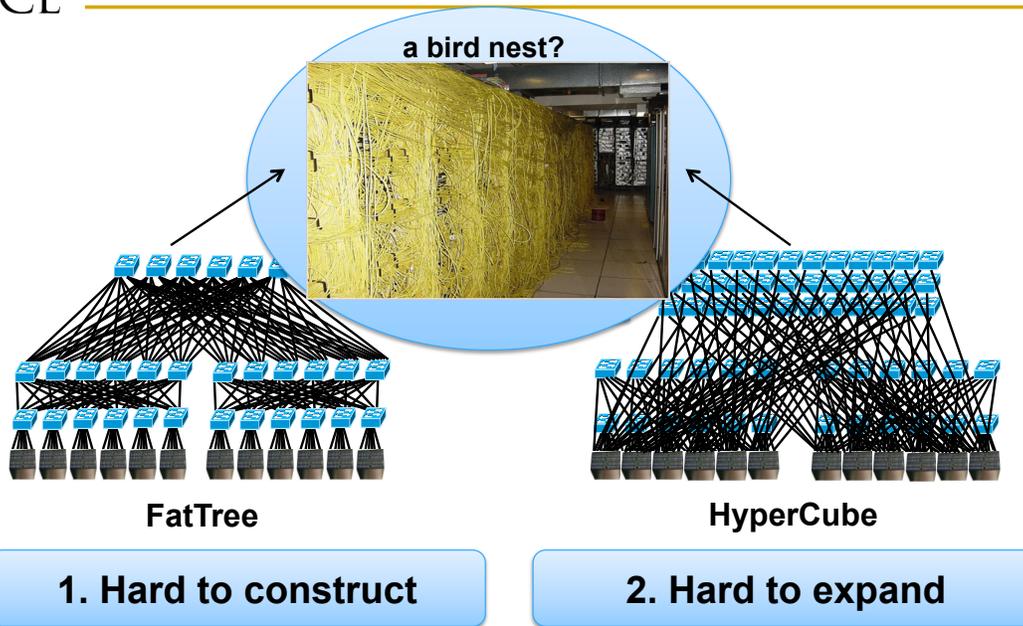
- Traditional data center network:
 - tree-structure Ethernet



Severe bandwidth bottleneck in aggregation layers.

2

Previous packet switching solutions for increasing data center network bandwidth



Optical Channels

- Ultra-high bandwidth



40G, 100Gbps technology has been developed.



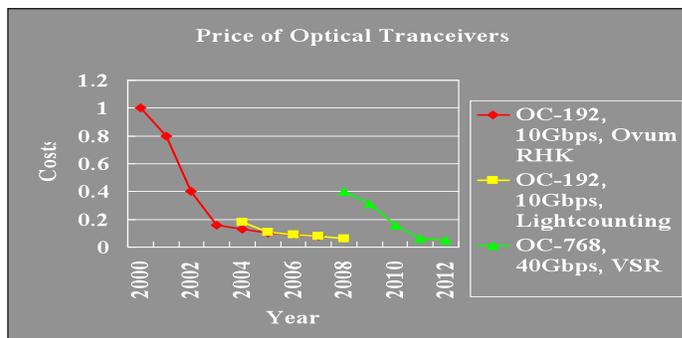
Bell Labs uses 155 lasers to beam ridiculous amounts of data over 7,000 kilometers

engadget.com Oct 1, 2009

Let's say you have a monumental collection of data at your place. Like, say, everything ever posted to the Pirate Bay. And let's say the Feds are beating down your door and you need to dump that data to a secure off-site...

15.5Tbps over a single fiber!

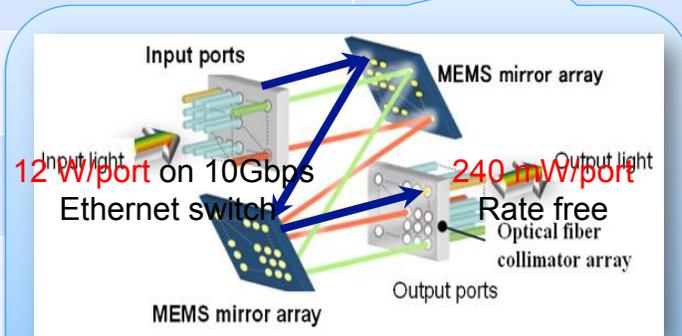
- Dropping prices



Price data from: Joe Berthold, Hot Interconnects '09

Optical circuit switching v.s. Electrical packet switching

	Electrical packet switching	Optical circuit switching
Switching technology	Store and forward	Circuit switching
Switching capacity		
Energy efficiency		
Switching time		

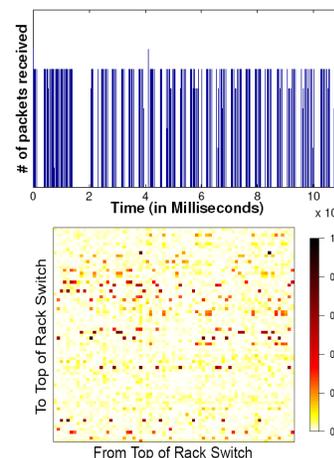


e.g. MEMS optical switch

5

Why is optical circuit switching worth considering?

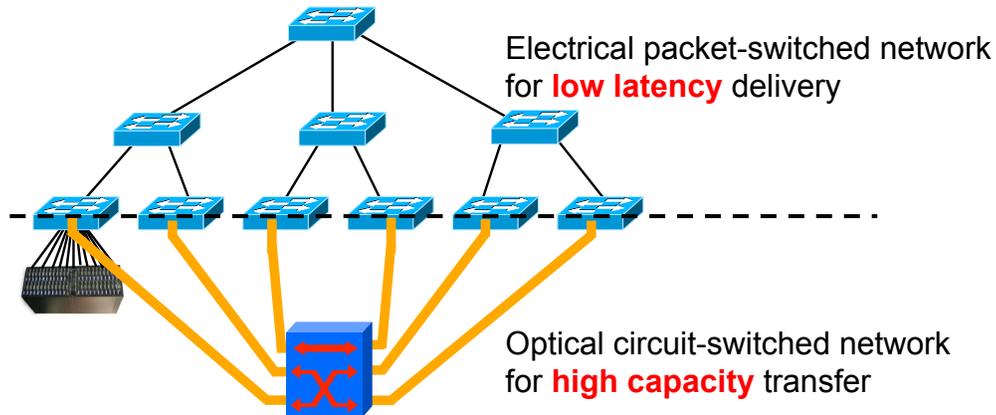
- Many measurement studies have suggested evidence of traffic concentration.
 - [SC05]: “... the bulk of inter-processor communication is bounded in degree and changes very slowly or never. ...”
 - [WREN09]: “... We study packet traces collected at a small number of switches in one data center and find evidence of ON-OFF traffic behavior...”
 - [IMC09][HotNets09]: “Only a few ToRs are hot and most their traffic goes to a few other ToRs. ...”



Full bisection bandwidth at packet level may not be necessary.

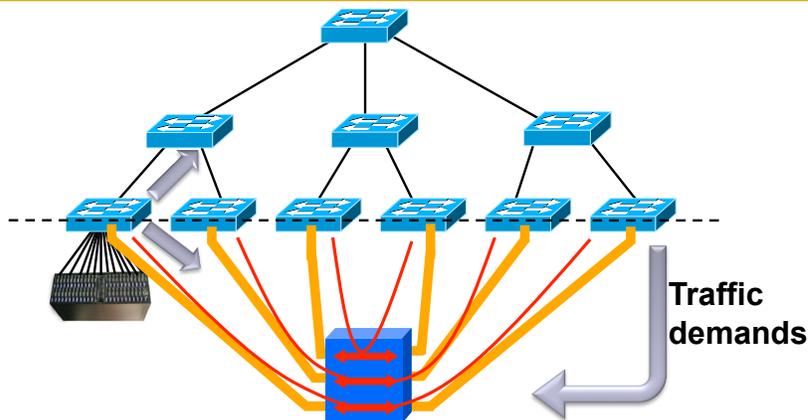
6

The HyPaC architecture (Hybrid Packet and Circuit)



- Optical paths are provisioned rack-to-rack
 - A simple and cost-effective choice
 - Aggregate traffic on per-rack basis to better utilize optical circuits

Design requirements

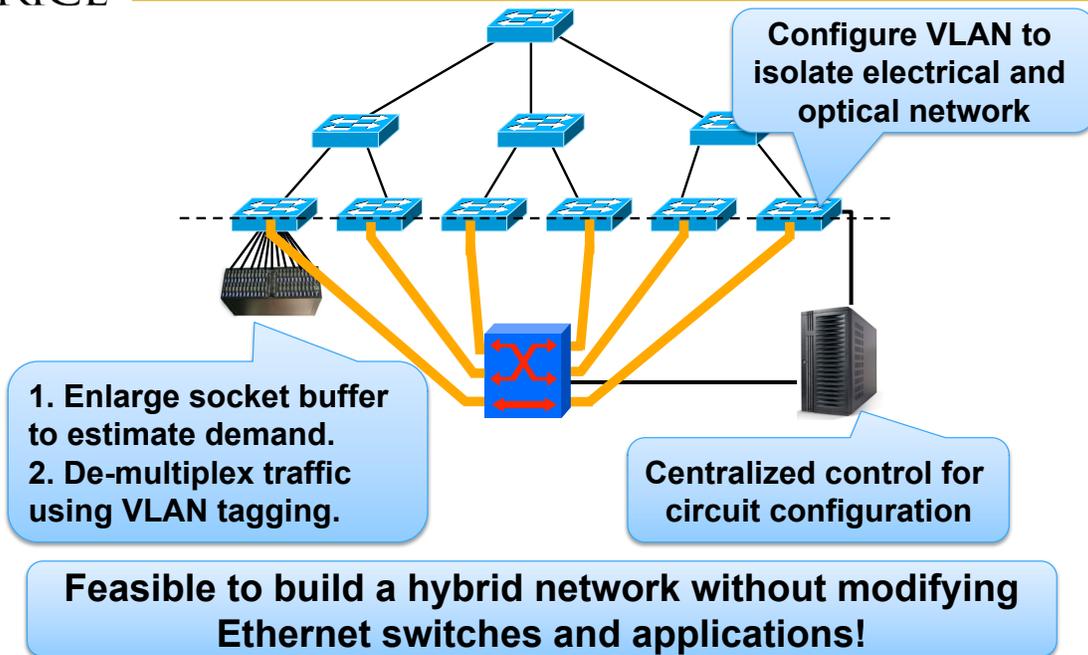


- Control plane:
 - Traffic demand estimation
 - Optical circuit configuration
- Data plane:
 - Dynamic traffic de-multiplexing
 - Optimizing circuit utilization (optional)



RICE

c-Through design



9

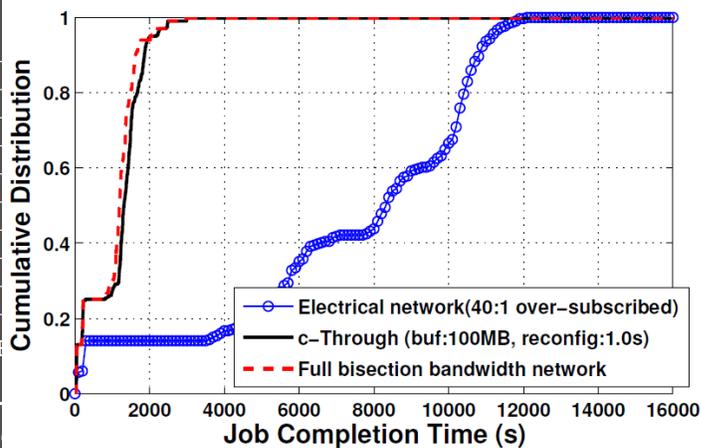
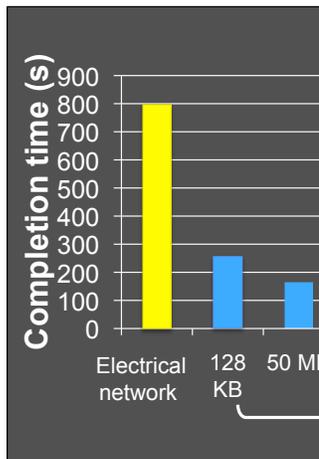


RICE

Performance of hybrid network

MapReduce performance

Gridmix performance



Close-to-optimal performance even for applications with all-to-all traffic patterns.

10



RICE

Related work

c-Through

[HotNets'09, SIGCOMM'10]

- Rack level optical paths
- Estimating demand from server socket buffer
- Traffic control in server kernel

Helios

[SIGCOMM'10]

- Pod level optical paths
- Estimating demand from switch flow counters
- Traffic control by modifying switches

Others

- **Proteus** [HotNets'10]: all optical data center network using WSS
- **DOS** [ANCS'10]: all optical data center network using AWGR

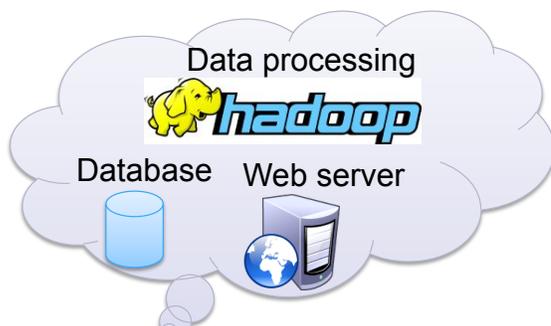
11



RICE

Circuit control in the wild

- Sharing is the key of cloud data centers



- Share at fine grain

- Complicated data dependencies

- Heterogeneous applications



12



RICE

Problems in c-Through and Helios

1. Treating all traffic as independent flows
 - Suboptimal performance for correlated applications
2. Inaccurate information about traffic demand
 - Vulnerable to ill-behaved applications
3. Restricted sharing policies
 - Limited by the control platform of Ethernet switches

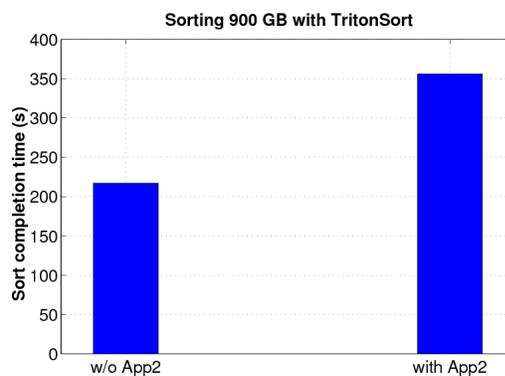
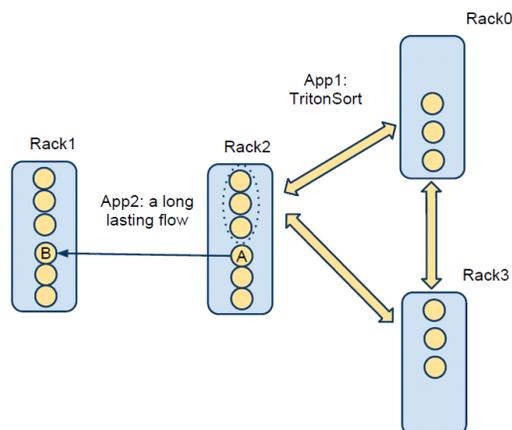
13



RICE

Problem 1: traffic dependencies

- Effect of correlated flows



14



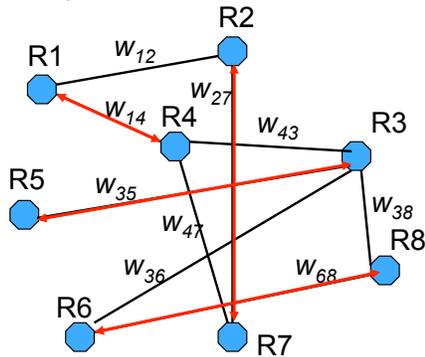
Circuit configuration with correlated traffic

● Problem formulation

Basic configuration: a matching problem

Graph $G: (V, E)$

$$w_{xy} = \text{vol}(Rx, Ry) + \text{vol}(Ry, Rx)$$



Modeling correlated traffic:

Definition of correlated edge groups:

$EG = \{e_1, e_2, \dots, e_n\}$, so that

$$w(e_i) += \Delta(e_i), i = 1, \dots, n$$

when EG is part of the matching.

Conflicting edge groups:

Two edge groups are conflict if they have edges sharing one end vertex.

Maximum weight matching with correlated edges

15



Algorithm design (1)

- If there is only one edge group
 - Intuition: test if including the edge group in the match will improve the overall weight.
 - Equation:

$$\text{benefit}(EG, G) = \text{Weight}(EG + \text{Edmonds}(G - EG)) - \text{Weight}(\text{Edmonds}(G))$$

Accept
Not accept

- If no conflict among edge groups:
 - A greedy algorithm
 - Iteratively accept all the edge groups with positive benefits;
 - Proven to achieve maximum overall weight;

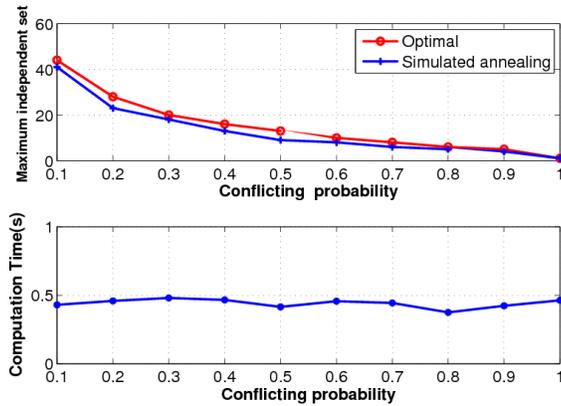
16



Algorithm design (2)

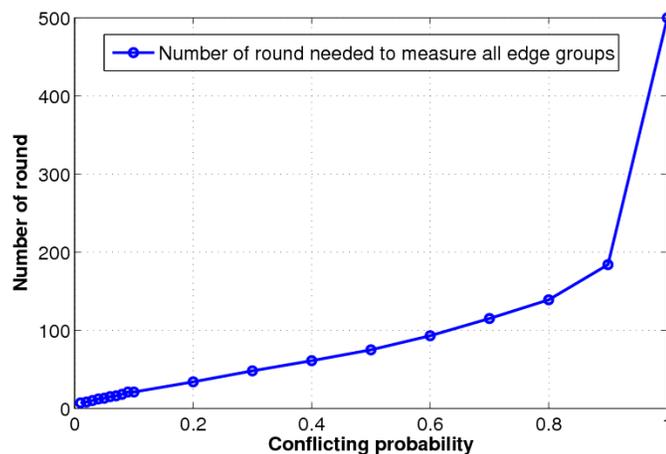
- If there are conflicts among edge groups
 - Finding the best non-conflict edge groups is NP-hard.
 - Equivalent to maximum independent set problem.

– An approximation algorithm based on simulated annealing works well.



Inferring correlated edge groups (1)

- Locations known, demand unknown:
 - Measuring maximal number of non-conflicting edge groups in each round.





RICE

Inferring correlated edge groups (2)

- Location unknown, demand unknown:
 - Hard problem

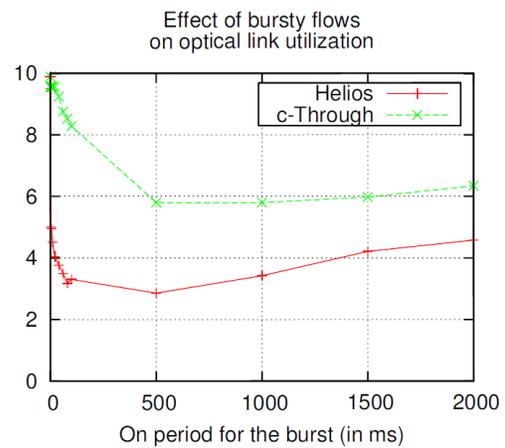
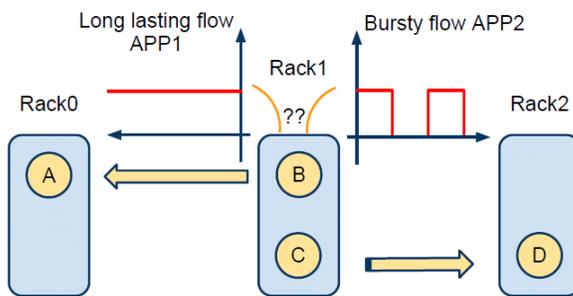
19



RICE

Problem 2: Inaccurate demand

- Effect of bursty flow



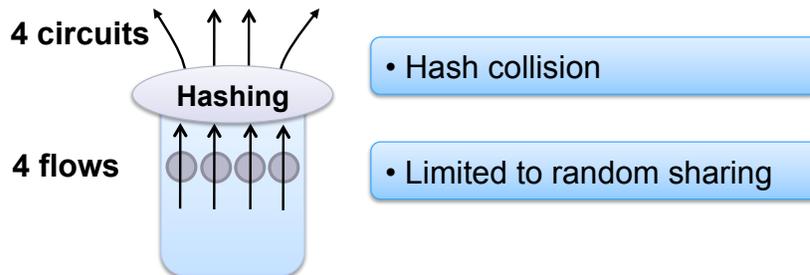
20



RICE

Problem 3: Restricted sharing policy

- An example problem:
 - Random hashing over multiple circuits.



- Potential solution:
 - Flexible control using programmable OpenFlow switches.

21



RICE

Summary

- HyPaC architecture has lots of potentials by marrying the strengths of packet and circuit switching
- Lots of open problems in the HyPaC control plane
- New physical layer capabilities (e.g. optical multicast) bring additional benefits and challenges

22