Machine-Learning-Based Flow scheduling in OTSS-enabled Datacenters

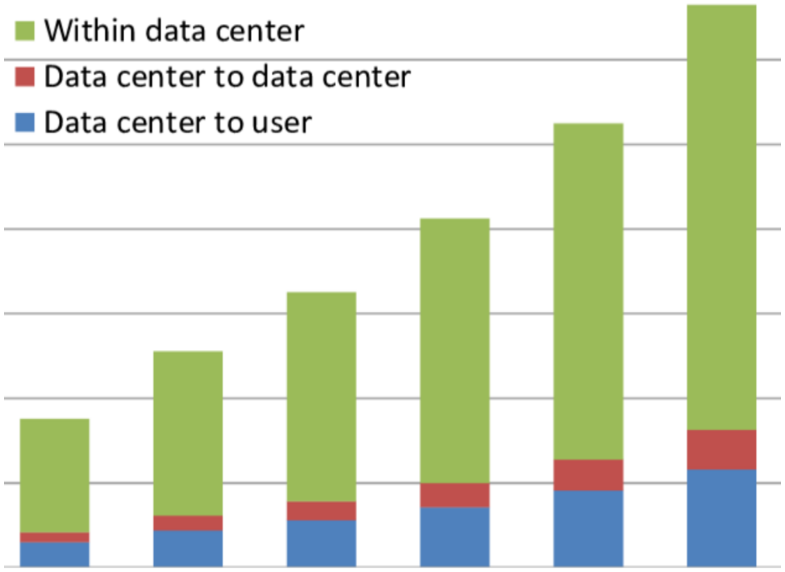
Speaker: Lin Wang

Research Advisor: Biswanath Mukherjee

UCDAVIS

Motivation

- **Traffic demand increasing in datacenter networks**
  - Cloud-service, parallel-computing, etc., lead to huge amount of intra datacenter traffic growth.
  - Cisco forecasts 31% increase per year of datacenter traffic by 2021



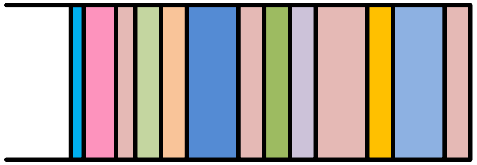
Datacenter traffic loads is growing

Introduction

- **Datacenter traffic measurements**
- A large fraction of datacenter traffic is carried in a small fraction of flows.
- 90% of flows carry less than 1MB of data, called " ant flows".
- More than 90% of bytes are transferred in flows greater than 100MB, called "elephant flows"

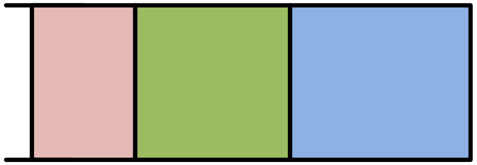
A. Greenberg, J, et al., "A scalable and flexible data center network," SIGCOMM, 2009.
S. Kandula, S., et al., The nature of datacenter traffic: Measurements & analysis," IMC, 2009.

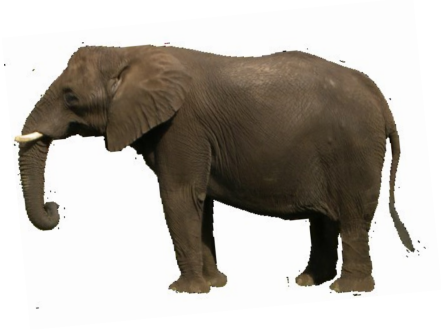
Mice VS. Elephant Flow



- Small size packet
- Large number
- Short flow
- Short-lived

transactional traffic, web browsing, search queries (≈ 90% traffic)





- Large size packet
- Small number
- Large volume flow
- Long-lasting

bulk data transfer, data backup, virtual machine migration (≈ 10% traffic)

Elephant flow detection

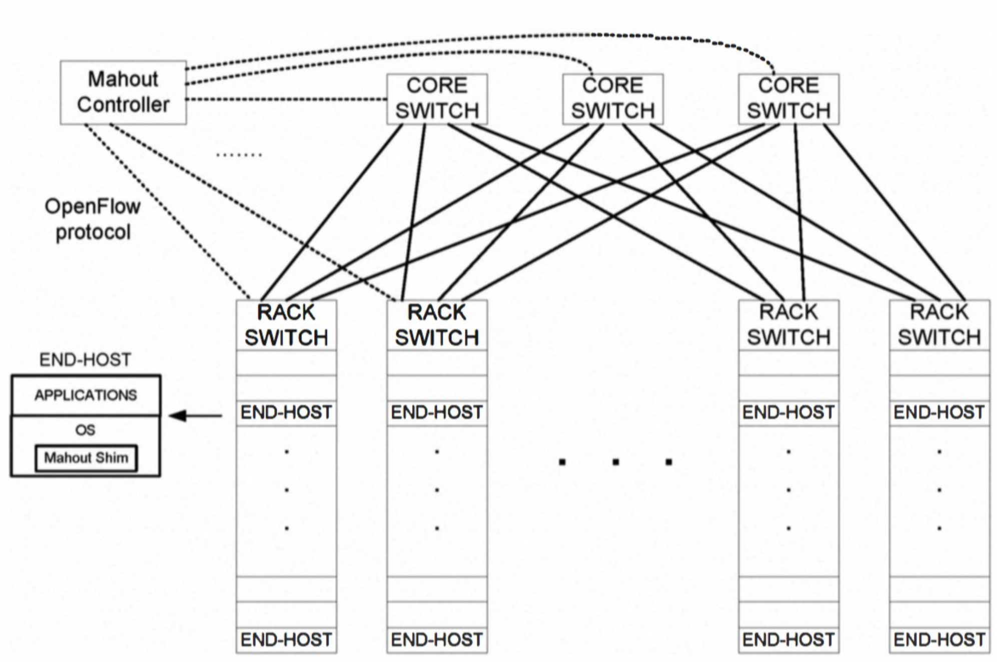
- **Application identify elephant flows**
- Impractical for traffic management in datacenter, as each application needs to be modified to support.

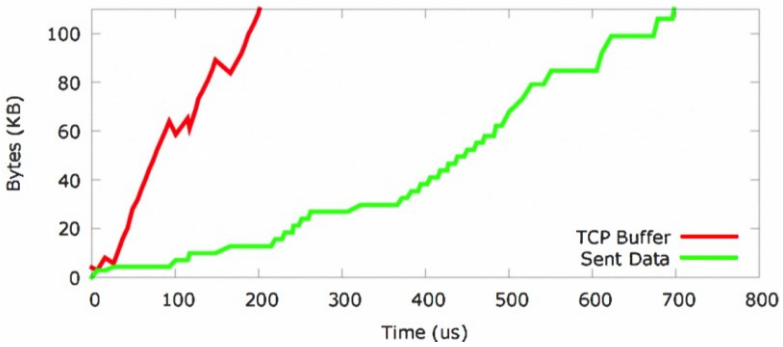
- **Maintain per-flow statistics**
- Not scale to large datacenter networks

- **Sampling**
- Is not reliable to detect an elephant flow before it has carried more than 10K packets, roughly 15MB.

Elephant flow detection

- **End host monitor**
- Monitor flows at origin end hosts. When detect an elephant flow, it marks subsequent packets of that flow using in-band signaling mechanism.





Amount of data observed in the TCP buffers vs. data observed at the network layer for a flow.

Curtis, Andrew R., Wonho Kim, and Praveen Yalagandula. "Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection." In INFOCOM, Proceedings IEEE, pp. 1629-1637. IEEE, 2011.

Needs for a transparent fine-grained optical network

- **Optical networks: enormous transmission bandwidth.**
  - High-order modulation (PAM4): increase per-channel capacity.
  - space-division-multiplexing: increase spatial channels.

- **Mismatch between application demands and optical channel capacity.**
  - Traffic grooming is the first proposal.
  - Drawback of grooming: energy, latency, security, etc.

Build a transparent, bufferless, fine-grained, WDM-like network

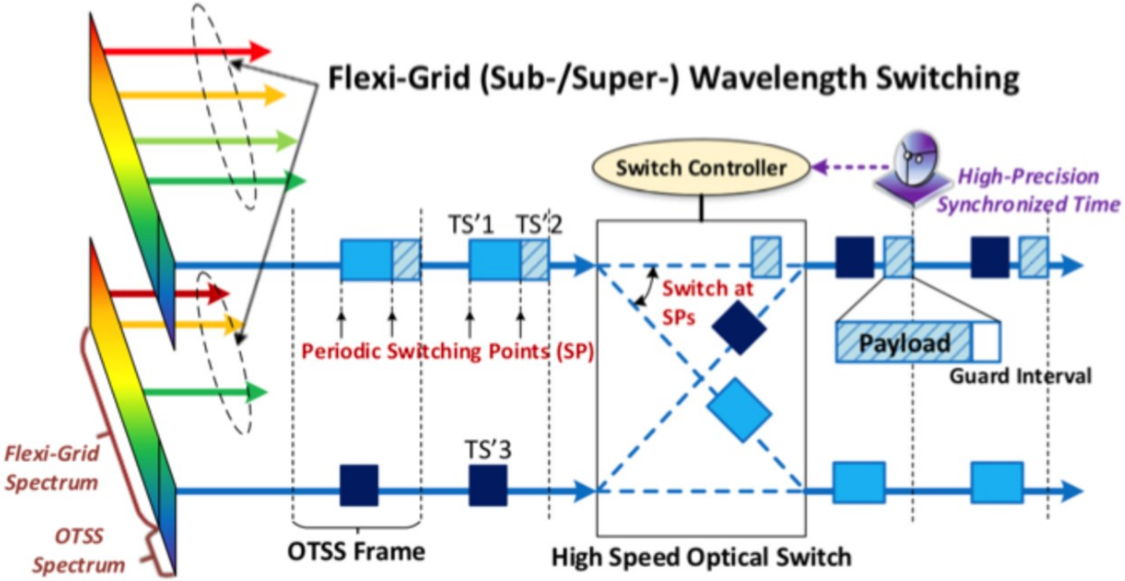
- Why WDM can avoid collision?
  - Wavelength channels are separated by a global coordinate.
  - (frequency! All the same in different nodes)

- Time synchronization: a global coordinate in temporal domain
  - Definite time, all nodes are synchronized for a global coordinate.

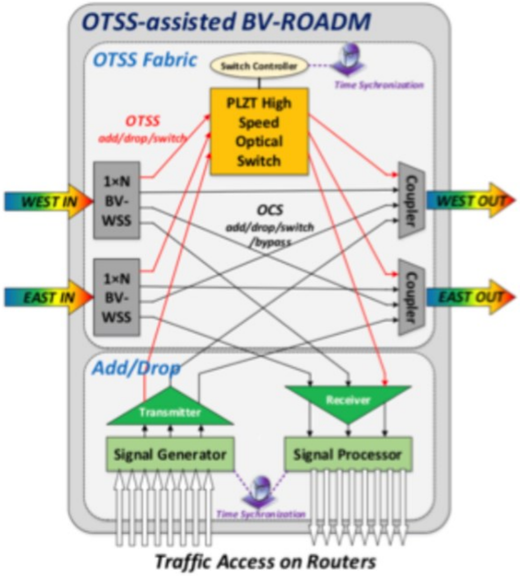
- Temporally-statistical multiplexing for asynchronous transmission based on synchronized global time.
  - We call it: Optical Time Slice Switching (OTSS).

OTSS Principle

- Designing a WDM-like TDM switching paradigm
  - WDM: all nodes have same frequency coordinate.
  - OTSS: all node should have same time coordinate!


Flexi-Grid (Sub-/Super-) Wavelength Switching
Switch Controller
High-Precision Synchronized Time
TS'1
TS'2
TS'3
Switch at SPs
Periodic Switching Points (SP)
Payload
Guard Interval
Flexi-Grid Spectrum
OTSS Spectrum
OTSS Frame
High Speed Optical Switch

OTSS switching principle



OTSS node architecture

Flow scheduling

- **Flow scheduling analysis**
  - Ant flows should be transmitted via OTSS switching paradigm.
  - Elephant flows should be transmitted via WDM switching paradigm.

- **Flow classification**
  - Apply C4.5 Decision Tree (C4.5), Naïve-Bayes Discretisation (NBD) to detect whether a flow belongs to “elephant flows”.

Machine-learning-based Coflow scheduling on OTSS

- **First step**
  - Detect whether a flow belongs to elephant flows.
- **Second step**
  - Transmit ant flows using OTSS switching paradigm.
  - Transmit elephant flows using WDM switching paradigm.
- **Third step**
  - Using switch controller to schedule flow transmitting.

Mathematical Formulations

Parameters

G(N, E): network topology in a unidirectional graph, where N and E denotes the set of nodes and fiber links, respectively.

R: set of traffic requests.

C: maximum link capacity.

s_r, d_r, b_r: source, destination, and bandwidth of traffic request r, r $\in$ R. Here, bandwidth is calculated in terms of the number of time slots.

η_1, η_2: parameters for optimization sequence, $\eta_1 \gg \eta_2$.

d(i, j): length of fiber link (i, j).

Max: a maximum number.

Variables

$\lambda_{(i,j)}^{r,t}$: binary variable, which equals 1 if request r occupies time slot t on fiber link (i, j).

ρ_r: binary, which equals 1 if request r is accepted.

Mathematical Formulations

Objective:

$$\mathrm{A}_{\mathrm{throughput}} = \sum_{r \in R} \rho_r * b_r$$

$$\mathrm{A}_{\mathrm{resource}} = \sum_{r \in R} \sum_{t \in T} \sum_{(i,j) \in E} \lambda_{(i,j)}^{r,t}$$

Maximize: $\eta_1 * A_{throughput} - \eta_2 * A_{resource}$

Mathematical Formulations

Constraints:

$$\sum_{j\in N}\sum_{t\in T}\lambda_{(i,j)}^{r,t}-\sum_{j\in N}\sum_{t\in T}\lambda_{(j,i)}^{r,t}=\begin{cases}\rho_r * b_r, i = s_r\\ -\rho_r * b_r, i = d_r\ , \forall r\in R\\ 0, otherwise\end{cases}$$

$$\sum_{t\in T}\lambda_{(i,j)}^{r,t}\left(\lambda_{(i,j)}^{r,t}-\lambda_{(j,k)}^{r,t}\right)\geq 0,\quad \forall r\in R, t\epsilon T, (i,j),\quad (j,k)\epsilon E$$

$$\sum_{r\in R}\lambda_{(i,j)}^{r,t}\leq 1, \forall t\epsilon T, (i,j)\ \epsilon E$$

amlwang@ucdavis.edu