Scalable Multi-Class Traffic Management in Data Center Backbone Networks

Amitabha Ghosh, Edward Crabbe, Jennifer Rexford

What we're doing

Joint optimization of rate control and routing taking into account application performance constraints and business priority

Some Selected Motivations

- Make efficient use of network resources
- Globally optimize throughput taking into account relative traffic characteristics and priority
- Datacenters may be run by single operator
  - we control the horizontal and the vertical
  - if optimization control loop is not too flabby, there may not as much need to encode priorities in packets
- protocols running inter-datacenter may not implement fairness objectives
  - that play well with TCP
    - at all, either implicitly or explicitly
- there may be many demand priority levels (> 8)

Investigation

- computation is distributed across multiple tiers of optimization machinery via optimization decomposition
- optimization machinery itself may be geographically distributed, although this clearly increases the length of time for each optimization cycle
- result is provably optimal

1. TRUMP
2. DaVinci

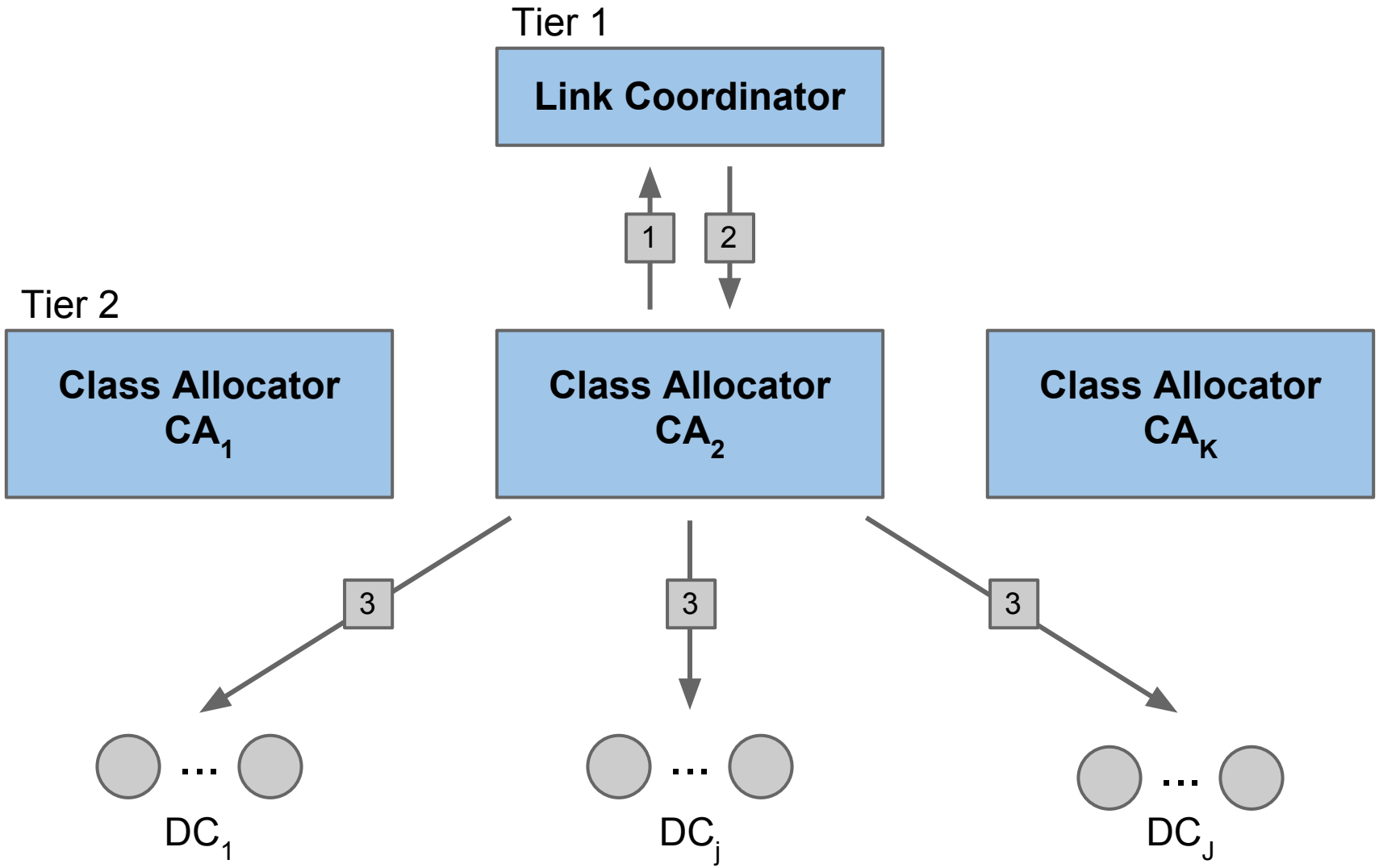
Flow Utility

$$U_s^k = w_s^k \left[a^k f^k(\cdot) - b^k g^k(\cdot) \right]$$

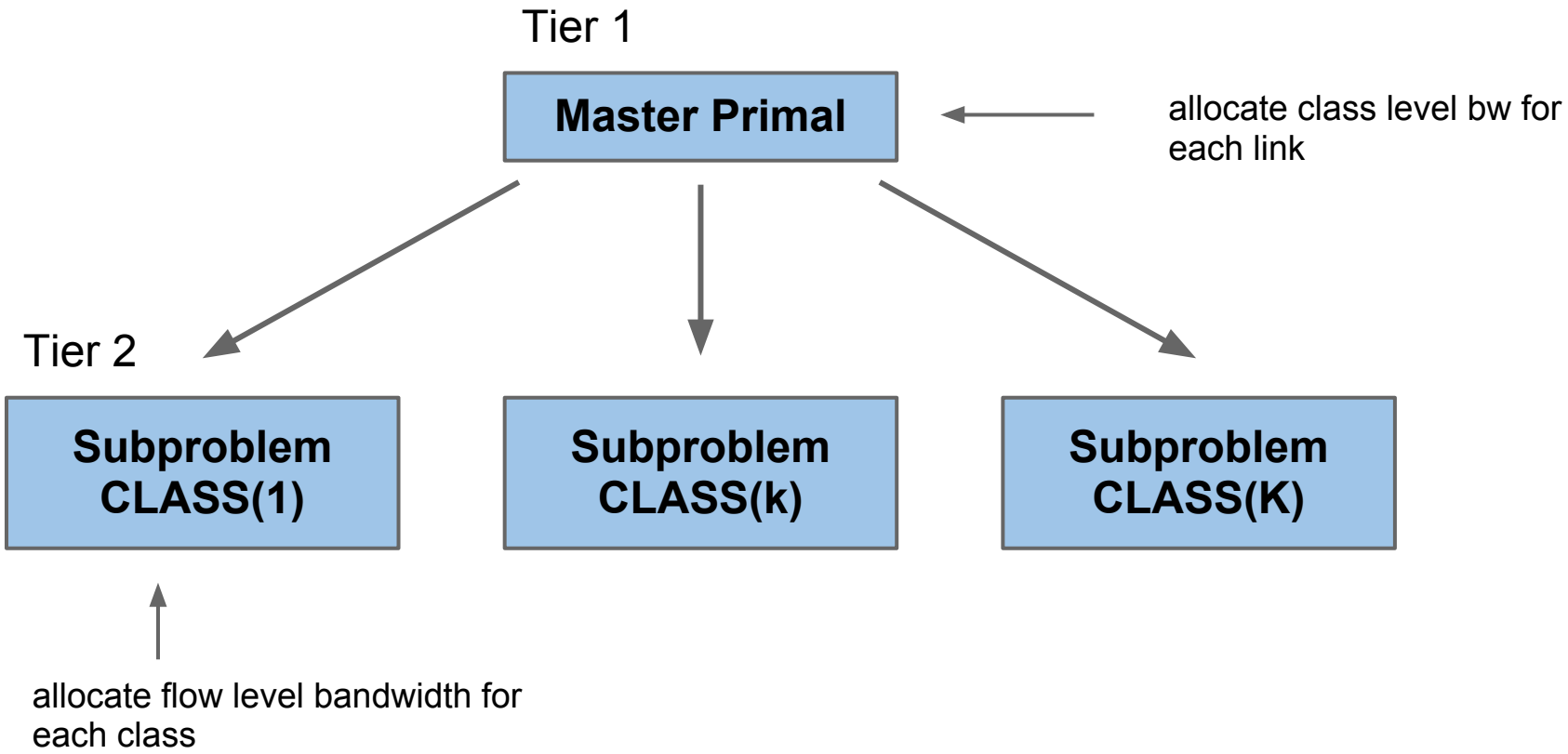
Weighted Utility of flow s of traffic class k

U^k_s	Utilitiy of flow s of class k
k	traffic class
s	flow
f^k	throughput / loss sensitivity
g^k	delay sensitivity
w^k_s	weight of flow s of class k
a^k	weight coefficient for throughput / loss sensitivity of class k
b^k	weight coefficient for latency sensitivity of class k

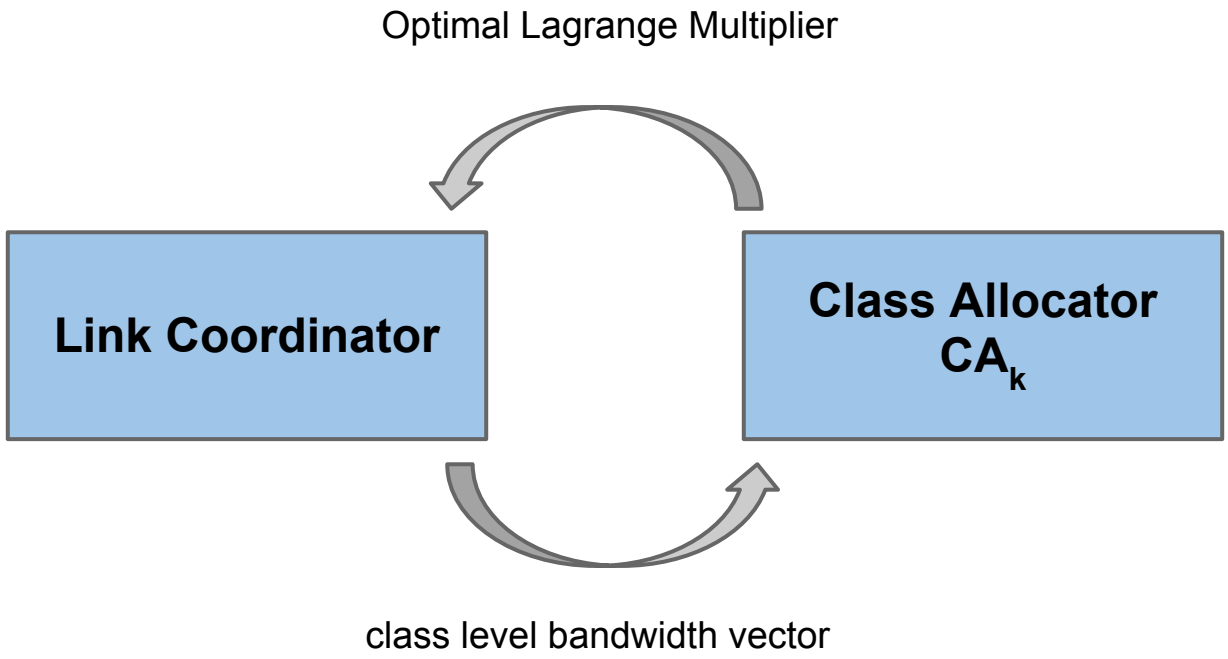
Two Layer Architecture



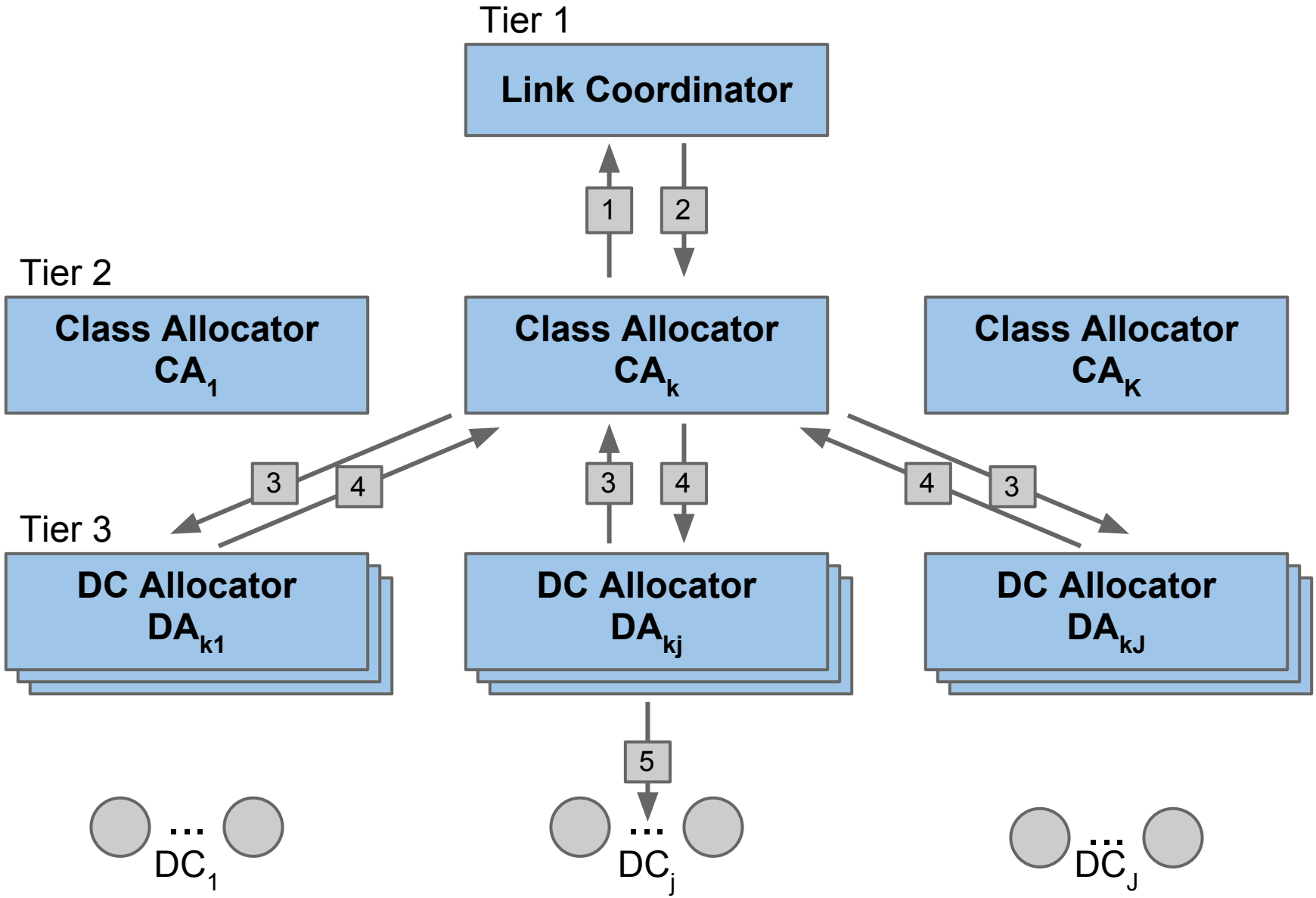
Two Layer Decomposition



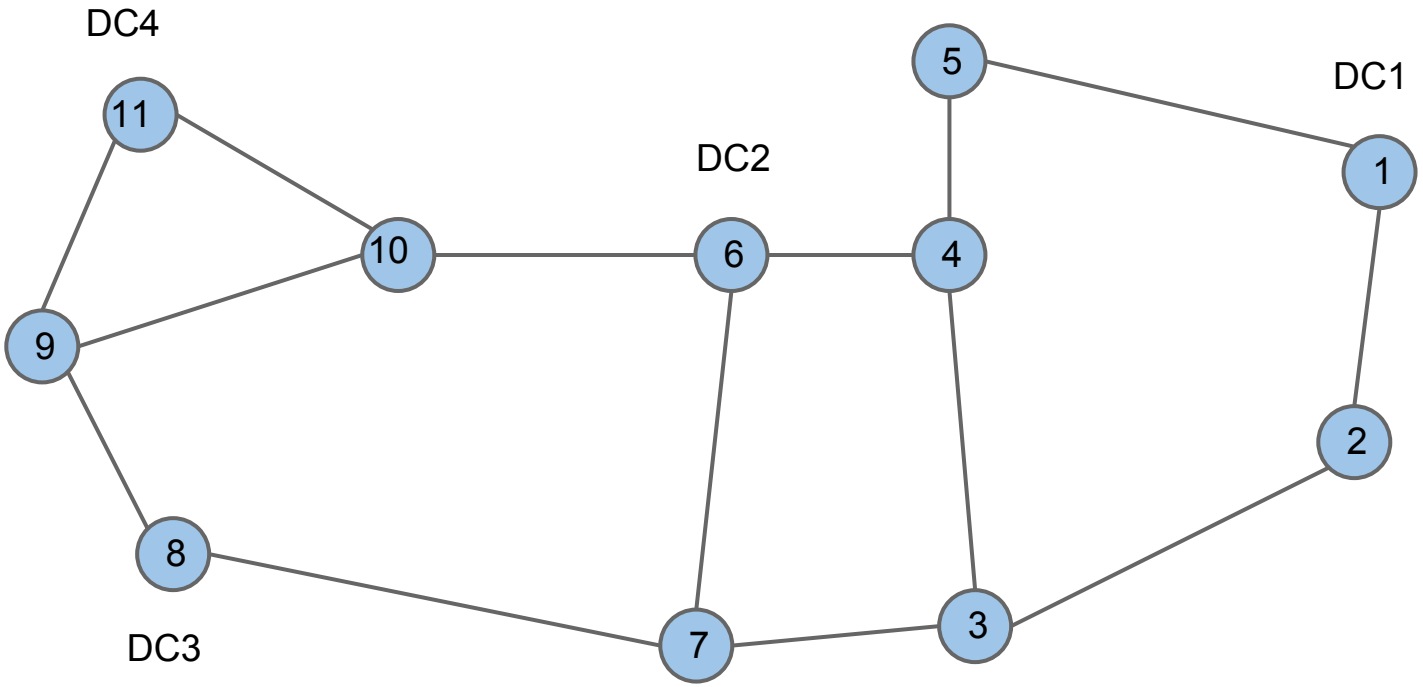
Message Passing



Three Layer Architecture



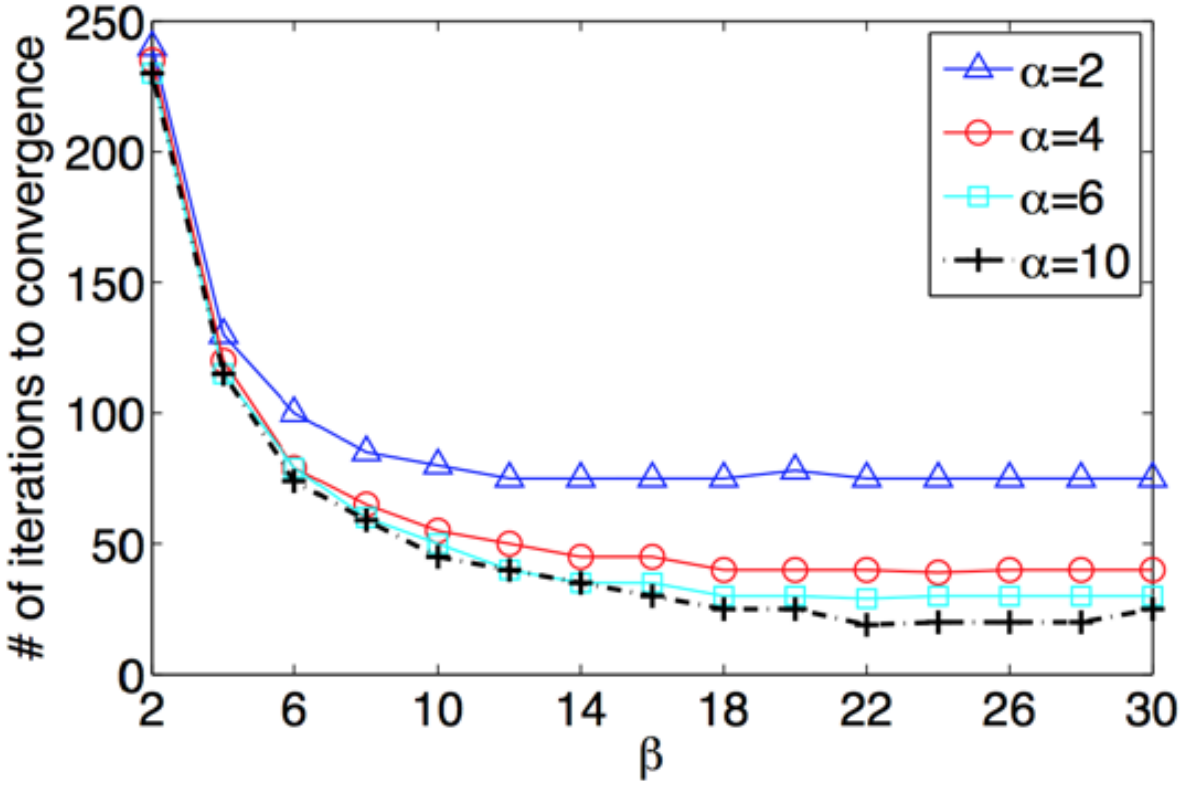
Simulation



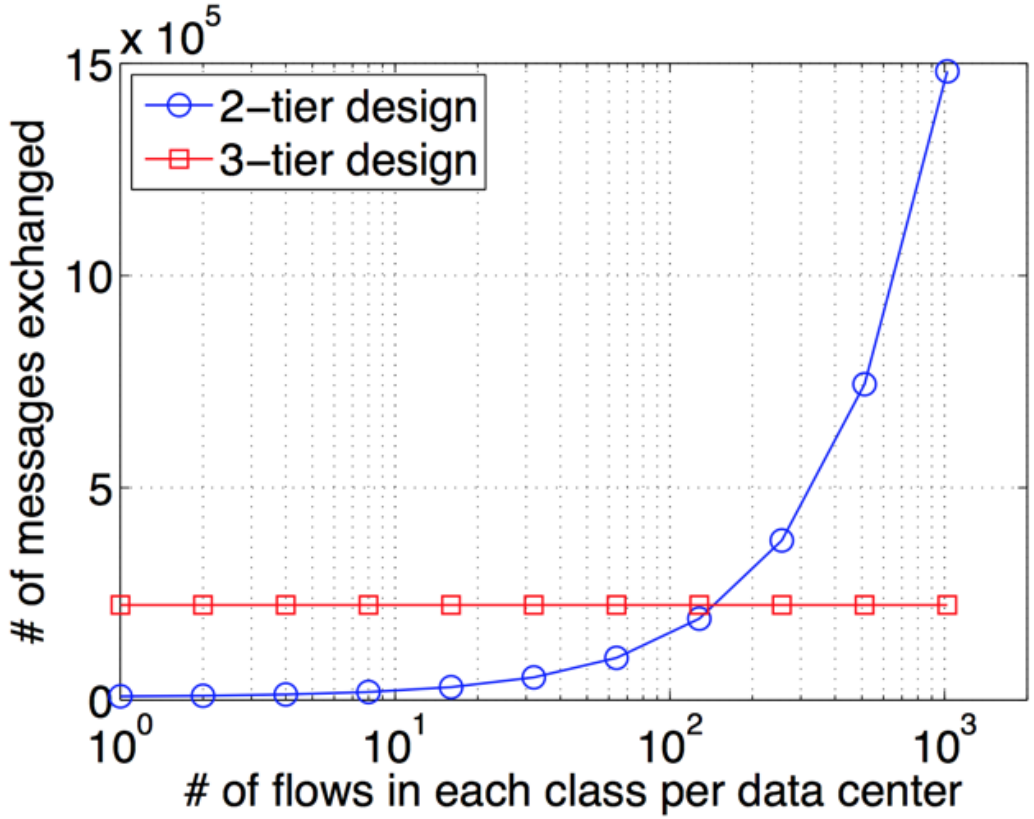
Abilene Topology

Two Classes of traffic, randomized flows

Results: Convergence



Results: Message Passing



Backup

GLOBAL optimization used for experiments

$$\begin{aligned}
\text{maximize} \quad & \mathcal{U} = \sum_{k} \sum_{s \in \mathcal{F}^k} w_s^k \left[a^k f^k \left(x_s^k \right) - b^k g^k \left(u_l^k \right) \right] \\
\text{subject to} \quad & \sum_{s \in \mathcal{F}^k} \sum_{p} A_{lp} R_{sp}^k \, z_{sp}^k \leq y_l^k, \quad \forall k, l \\
& \sum_{k} y_l^k \leq c_l, \quad \forall l \\
\text{variables} \quad & z_{sp}^k \geq 0, \quad \forall k, s, p \\
& y_l^k \geq 0, \quad \forall k, l
\end{aligned}$$

$$A_{lp} = \begin{cases} 1, & \text{if link } l \text{ lies on path } p \\ 0, & \text{otherwise.} \end{cases}$$

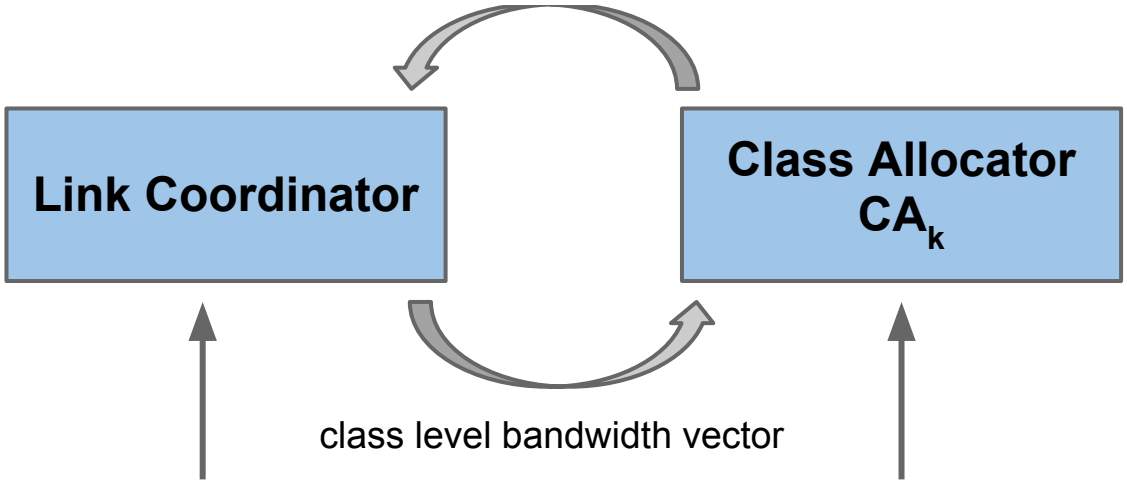
$$R_{sp}^k = \begin{cases} 1, & \text{if flow } s \text{ of class } k \text{ uses path } p \\ 0, & \text{otherwise.} \end{cases}$$

$\mathcal{F}$	Set of all flows across all classes.
$\mathcal{F}^k$	Set of flows in class k.
c_l	Capacity of link l.
w_s^k	Weight of flow s of class k.
z_{sp}^k	Rate of flow s of class k on its p^{th} path.
y_l^k	Bandwidth allocated for class k on link l.

Two Tier Message Passing

optimal lagrange multiplier subject to constraints of Class(k)

$$\mathbf{y}^k(t+1) = \left[\mathbf{y}^k(t) + \beta(t) \cdot \boxed{\boldsymbol{\lambda}^{*k}}(\mathbf{y}^k(t))\right]_{\mathcal{X}}$$



MASTER-PRIMAL:

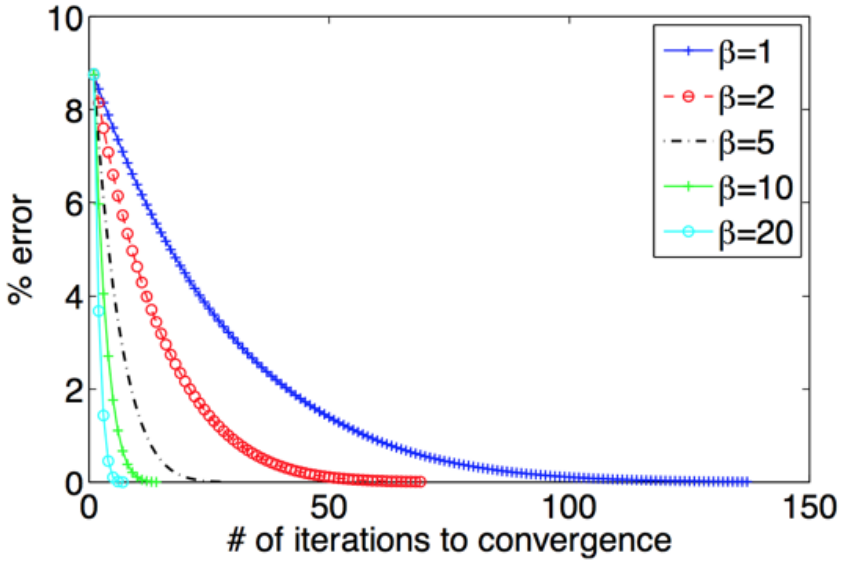
$$\begin{aligned} \text{maximize} \quad & \mathcal{U} = \sum_k \mathcal{U}^{*k}\left(\mathbf{y}^k\right) \\ \text{subject to} \quad & \sum_k y_l^k \le c_l, \quad \forall l \\ \text{variables} \quad & y_l^k \ge 0, \quad \forall k, l \end{aligned}$$

CLASS(k):

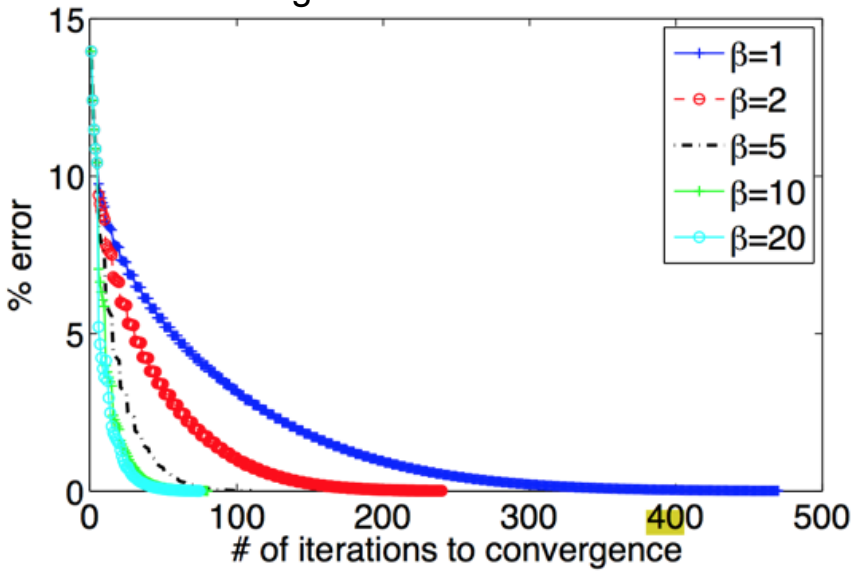
$$\begin{aligned} \text{maximize} \quad & \mathcal{U}^k = \sum_{s \in \mathcal{F}^k} w_s^k \left[a^k f^k\left(x_s^k\right) - b^k g^k\left(u_l^k\right)\right] \\ \text{subject to} \quad & \sum_{s \in \mathcal{F}^k} \sum_p A_{lp} R_{sp}^k \, z_{sp}^k \le y_l^k, \quad \forall l \\ \text{variables} \quad & z_{sp}^k \ge 0, \quad \forall s \in \mathcal{F}^k, p \end{aligned}$$

Rate of Convergence vs Class Level Step Size

2 Tier Covergence

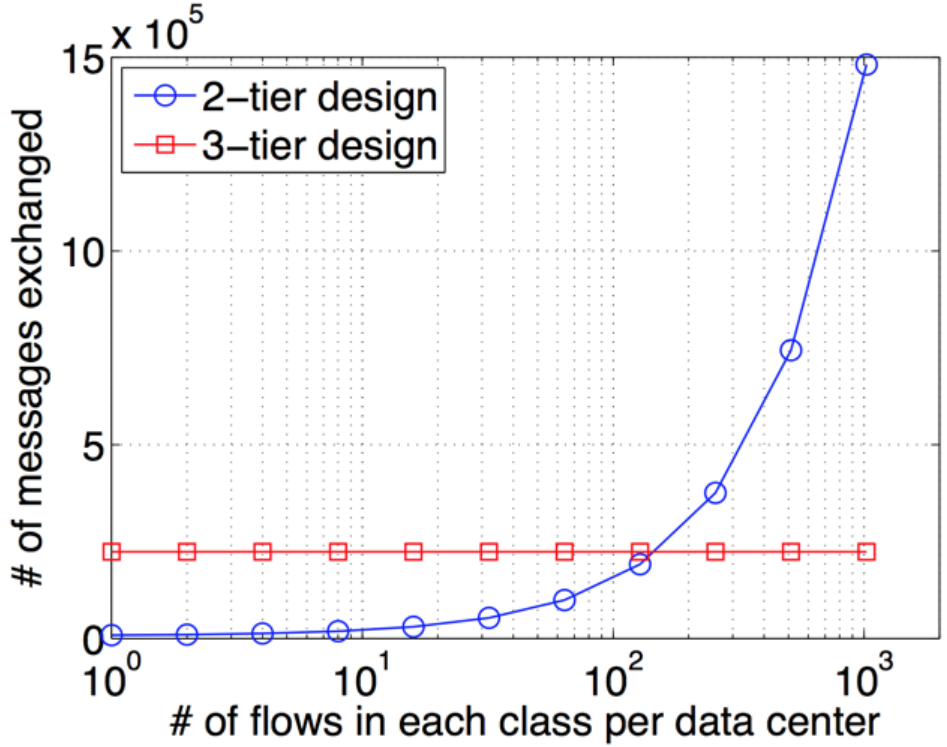


3 Tier Covergence



Class-level stepsize β	**2-tier design**	**3-tier design**
small $\beta = 1, 2$	slow	very slow, all α
medium $\beta = 5, 10$	moderate	slow, all α
large $\beta = 20, 30$	fast	moderate, all α
very large $30 < \beta < 40$	fast	moderate, $\alpha \leq 16$
extremely large $40 \geq \beta < 50$	fast	does not converge
$\beta \geq 50$	does not converge	does not converge

Message Passing



2 tier:

$$\#\text{ of messages } = N\left(2KL + \sum_k \sum_j \sum_{s \in \mathcal{F}^{kj}} \sum_p R^k_{sp}\right)$$

3 tier:

$$\#\text{ of messages } = N'(2KL + 2JKLM).$$