BwE: Flexible, Hierarchical Bandwidth Allocation for WAN Distributed Computing

ALOK KUMAR
NIKHIL KASINADHUNI
BJORN CARLIN
SUSHANT JAIN
ENRIQUE ZERMENO
MIHAI AMARANDEI-STAVILLA
STEPHEN STUART
UDAY NAIK
C. STEPHEN GUNN
MATHIEU ROBIN
AMIN VAHDAT
ANAND RAGHURAMAN
JING AI
ASPI SIGANPORIA

Google Inc.
(Sigcomm'15)

25.05.2018

ETH Zurich

Presented by :
Ritu Sriram

Main idea

- Service-level bandwidth allocation
  - Prioritized bandwidth functions
  - Isolation between services
- Suitable for distributed computing applications running across dedicated private WANs
- Mapping of bandwidth allocation policies onto packets at the host machines

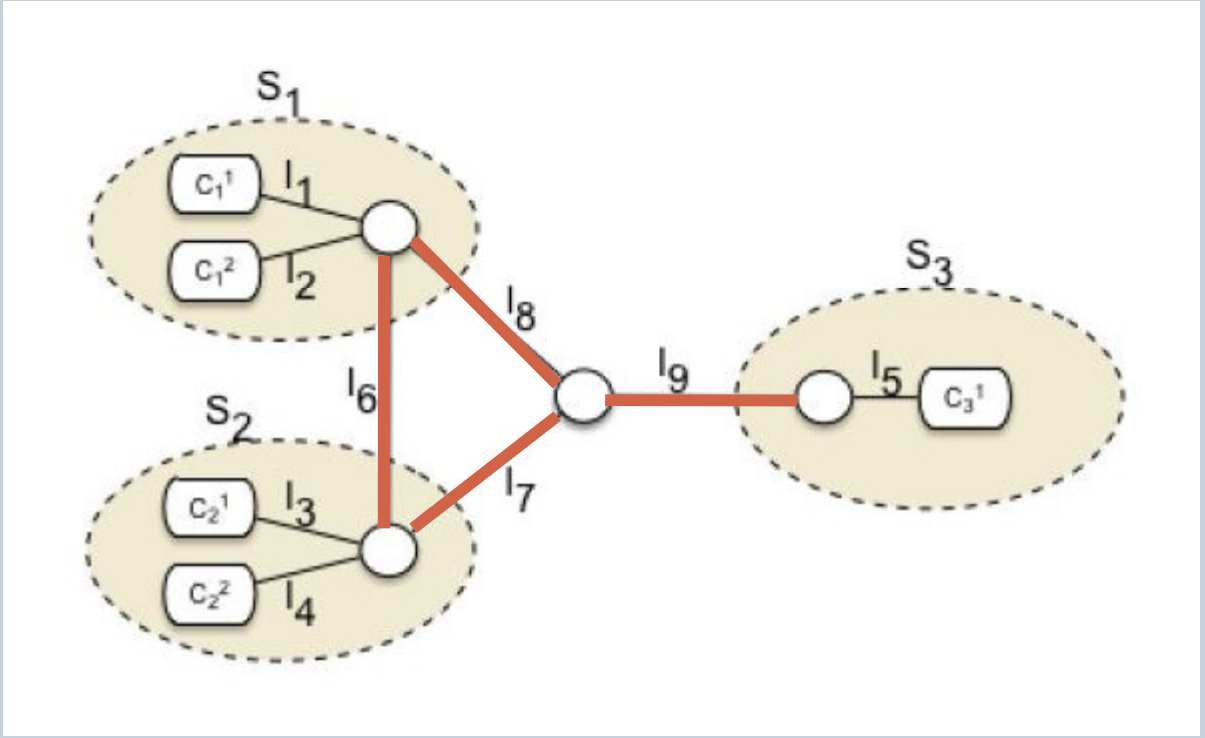
Existing Approaches

- TCP-based bandwidth allocation
  - All flows are of equal priority
  - Allocation is proportional to the # of active flows
- QoS and MPLS tunnels
  - Fixed the equal priority assumption
  - Scaling issues
  - Not enough flexibility in allocation policy

Contributions

- Unified, hierarchical control plane for bandwidth management at the end-hosts
- Integrating BwE with existing TE solutions
- Hierarchical max-min fair bandwidth allocation algorithm

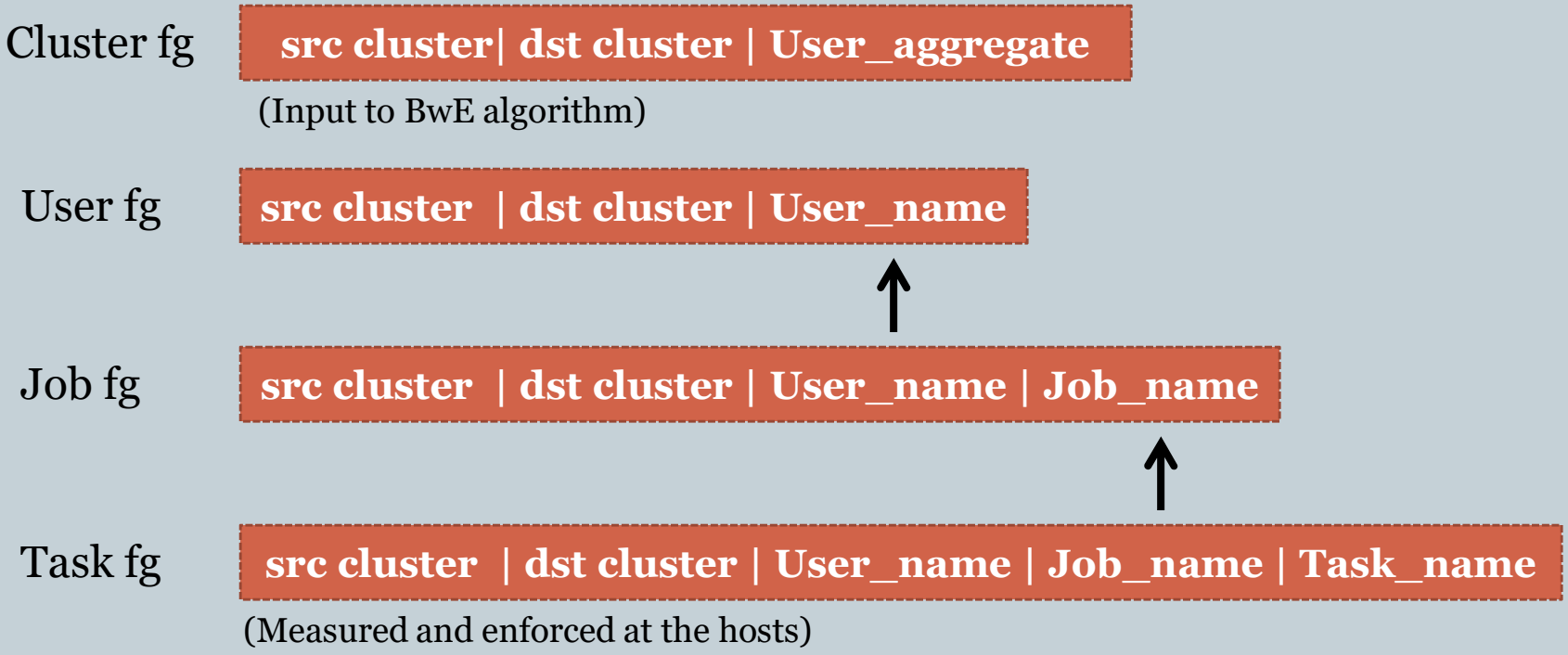
Concepts : WAN Network Model



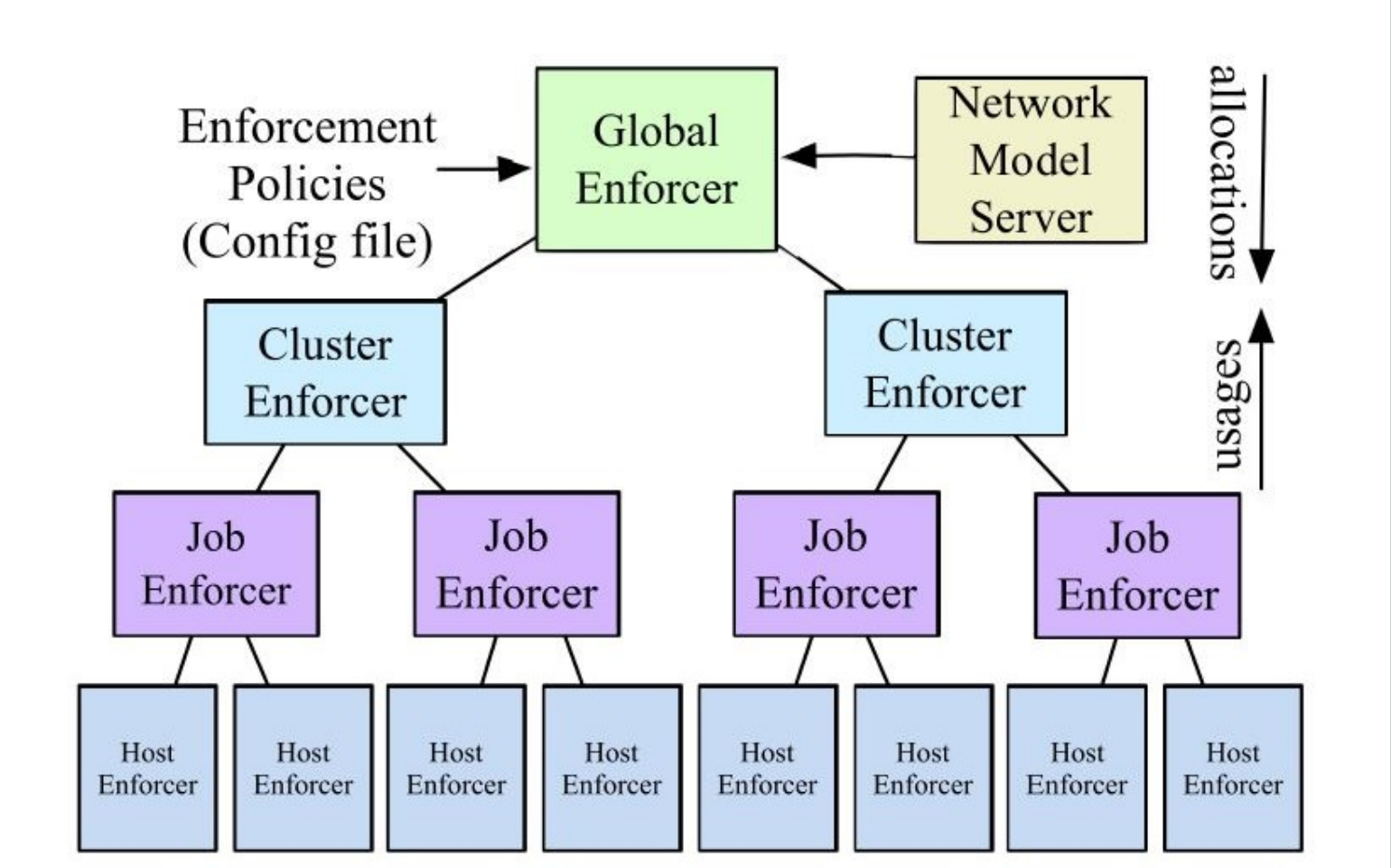
S – sites
C – clusters
l – links

Concepts : FlowGroups or Traffic Aggregates

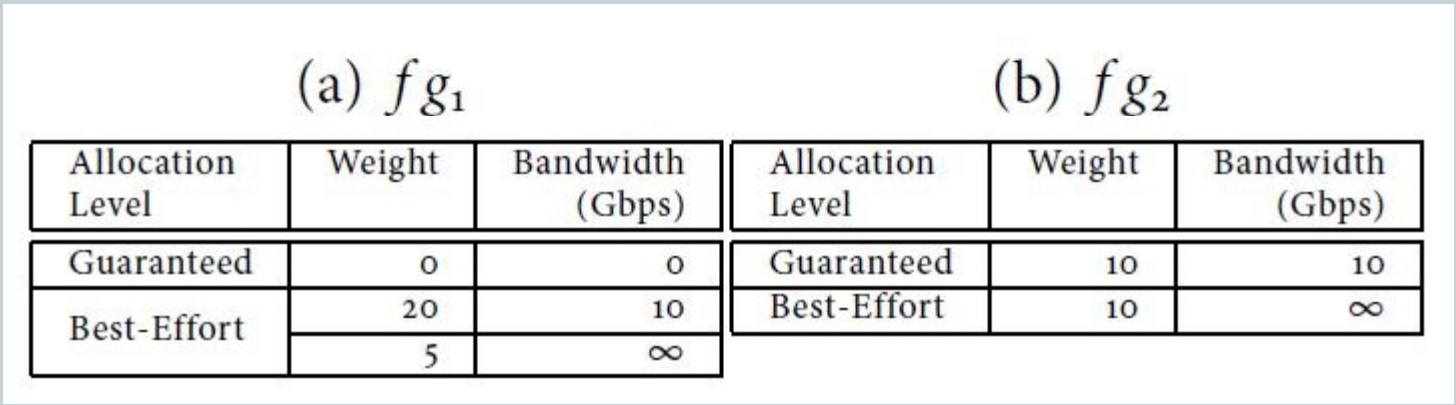
- Different granularities



BwE Architecture



Concepts : BwE configuration



(a) fg_1

Allocation Level	Weight	Bandwidth (Gbps)
Guaranteed	0	0
Best-Effort	20	10
	5	∞

(b) fg_2

Allocation Level	Weight	Bandwidth (Gbps)
Guaranteed	10	10
Best-Effort	10	∞

Concepts : Bandwidth functions

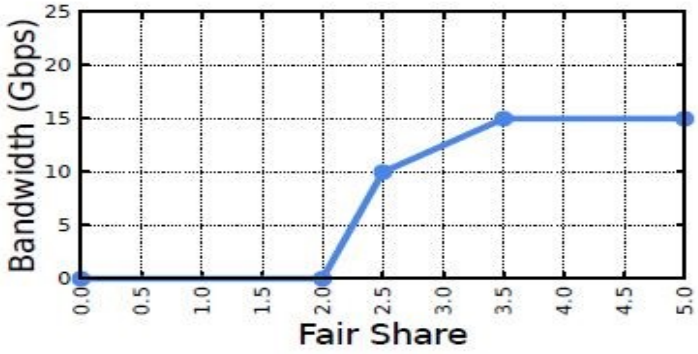
- Specifies the bandwidth allocation to a FlowGroup as a function of its relative priority and 'fair-share'

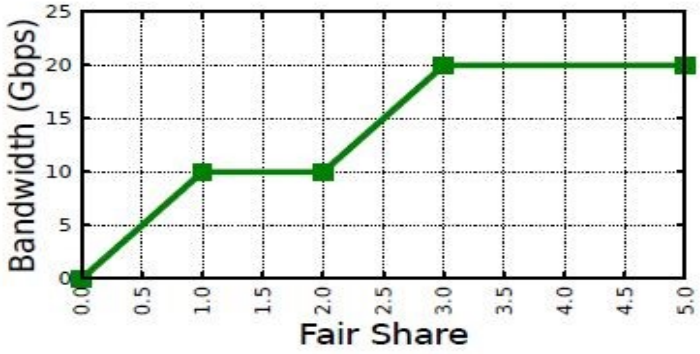
(a) fg_1

Allocation Level	Weight	Bandwidth (Gbps)
Guaranteed	0	0
Best-Effort	20	10
	5	∞

(b) fg_2

Allocation Level	Weight	Bandwidth (Gbps)
Guaranteed	10	10
Best-Effort	10	∞

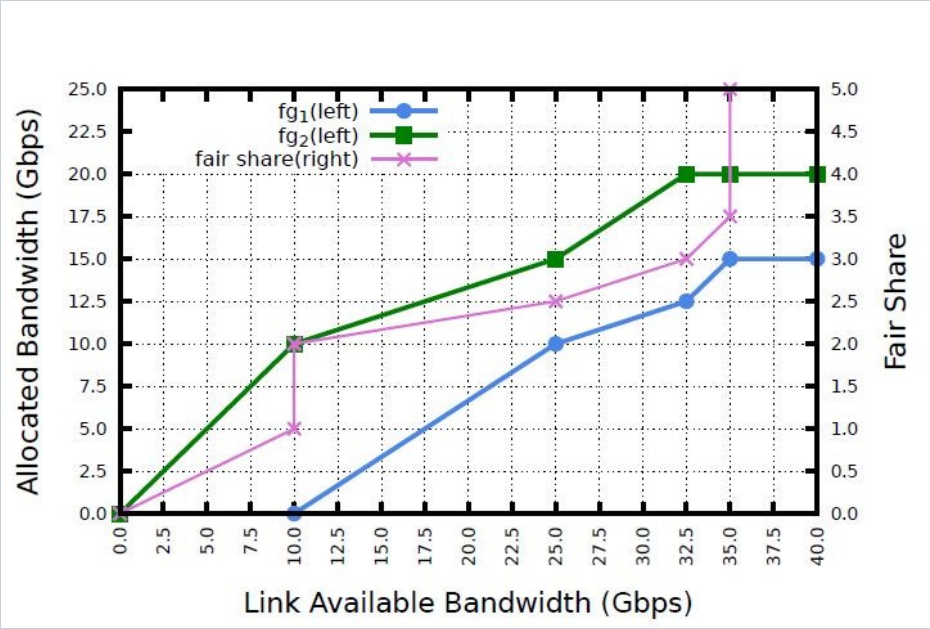




Fair-share: measure of available capacity to be shared fairly

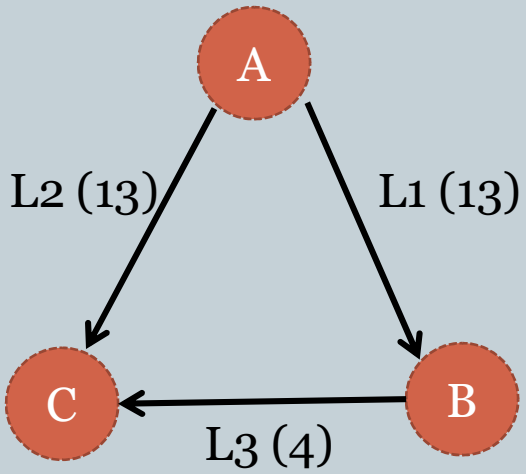
BwE algorithm : Global MPFA

- Inputs
  - Task fg's demands aggregated to the top level
  - Network topology , optimal paths and weights from the underlying TE system
- Output



MPFA Problem : Setup

Topology



Inputs

Flow Group	Src	Dst	Paths
f_1	A	C	l_2: 0.75 $l_1 \rightarrow l_3$: 0.25
f_2	A	C	l_2: 1.00
f_3	A	B	l_1: 1.00

- All flows have demand – 18
- Weights of flows (priorities) :
  - f1 – 1
  - f2 – 2
  - f3 – 3

* Link capacities in Gbps

MPFA : Bandwidth functions (FlowGroups)

- $B_{f1}(s) = \min(18 , s)$
- $B_{f2}(s) = \min(18 , 2s)$
- $B_{f3}(s) = \min(18 , 3s)$

Flow Group	Src	Dst	Paths
f_1	A	C	l_2: 0.75 $l_1 \rightarrow l_3$: 0.25
f_2	A	C	l_2: 1.00
f_3	A	B	l_1: 1.00

- All flows have demand – 18
- Weights of flows:
  - f1 – 1
  - f2 – 2
  - f3 – 3

s – fair share measure

MPFA : Fraction of traverse

- Formula: Sum of weights for all paths of a fg contaning the given link

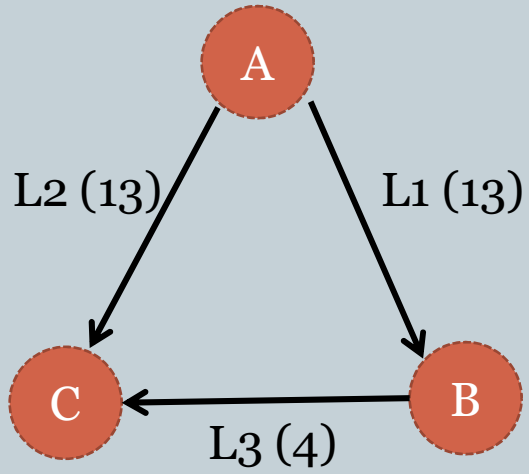
$$FR(f_i, l_k) = \sum_{1 \le j \le n_{f_i} | l_k \in p_j^{f_i}} w_j^{f_i}$$

- FR(f1 , l1) = 0.25
  FR(f1 , l2) = 0.75
  FR(f1 , l3) = 0.25
- FR(f2, l2) = 1
- FR(f3, l1) = 1

Flow Group	Src	Dst	Paths
f_1	A	C	l_2: 0.75 $l_1 \rightarrow l_3$: 0.25
f_2	A	C	l_2: 1.00
f_3	A	B	l_1: 1.00

- All flows have demand – 18
- Weights of flows:
  - f1 – 1
  - f2 – 2
  - f3 – 3

MPFA : Bandwidth functions (Links)



Flow Group	Src	Dst	Paths
f_1	A	C	l_2: 0.75 $l_1 \rightarrow l_3$: 0.25
f_2	A	C	l_2: 1.00
f_3	A	B	l_1: 1.00

- Aim : Allocate bandwidth fairly without exceeding link capacity

$B_{f1}(s) = \min(18 , s)$; $B_{f3}(s) = \min(18 , 3s)$;

$FR(f1 , l1) = 0.25$; $FR(f3, l1) = 1$;

$$B_{l_1}(s) = \begin{pmatrix} 0.25(\min(18, s)) \\ +\min(18, 3s) \end{pmatrix} = \begin{cases} 3.25s & : 0 \le s < 6 \\ 0.25s + 18 & : 6 \le s < 18 \\ 22.5 & : s \ge 18 \end{cases}$$

MPFA : Bandwidth functions (Links)

$$B_{l_1}(s) = \begin{pmatrix} 0.25(\min(18, s)) \\ +\min(18, 3s) \end{pmatrix} = \begin{cases} 3.25s & : 0 \le s < 6 \\ 0.25s + 18 & : 6 \le s < 18 \\ 22.5 & : s \ge 18 \end{cases}$$

$$B_{l_2}(s) = \begin{pmatrix} 0.75(\min(18, s)) \\ +\min(18, 2s) \end{pmatrix} = \begin{cases} 2.75s & : 0 \le s < 9 \\ 0.75s + 18 & : 9 \le s < 18 \\ 31.5 & : s \ge 18 \end{cases}$$

$$B_{l_3}(s) = \quad 0.25(\min(18, s)) = \begin{cases} 0.25s & : 0 \le s < 18 \\ 4.5 & : s \ge 18 \end{cases}$$

MPFA : Finding bottleneck

- We find a bottleneck fair share for each link such that its bandwidth reaches link capacity.

$$B_{l_1}(s) = \begin{pmatrix} 0.25(\min(18, s)) \\ +\min(18, 3s) \end{pmatrix} = \begin{cases} 3.25s & : 0 \le s < 6 \\ 0.25s + 18 & : 6 \le s < 18 \\ 22.5 & : s \ge 18 \end{cases}$$

Link		Bottleneck fair share
1	(3.25 * s) = 13	s = 4;
2	(2.75 * s) = 13	s = 4.72 ;
3	(0.25 * s) = 4	s = 16 ;

L1 has reached **full capacity first**, it no longer participates in MPFA !

L1 is the bottleneck link

MPFA : Update functions

Flow Group	Src	Dst	Paths
f_1	A	C	l_2: 0.75 $l_1 \rightarrow l_3$: 0.25
f_2	A	C	l_2: 1.00
f_3	A	B	l_1: 1.00

Frozen FlowGroups

- Need to update Bandwidth functions for links 2 and 3 beyond fair share of 4

Accounted for f1
We found s = 4 when it has max fair allocation

$$B_{l_2}(s) = \begin{pmatrix} 0.75(\cancel{\min(18, s)}) \\ + \min(18, 2s) \end{pmatrix}$$

MPFA : Update functions

$$B_{l_2}(s) = \begin{cases} 2.75s & : 0 \le s < 4 \\ 2s + 3 & : 4 \le s < 9 \\ 21 & : s \ge 9 \end{cases}$$

$$B_{l_3}(s) = \begin{cases} 0.25s & : 0 \le s < 4 \\ 1 & : s \ge 4 \end{cases}$$

Link		Bottleneck fair share
1	-	s = 4;
2	2s + 3 = 13	s = 5 ;
3	Never hits full capacity	s = ∞

L2 has reached **full capacity** next, it no longer participates in MPFA !

L2 is the bottleneck link

MPFA : Final output

- Terminate when all FlowGroups are frozen
- OUTPUT : allocation in fairshare = (f1:4, f2:5, f3:4)

$B_{f1}(s) = \min(18 , 4) = 4$ Gbps

$B_{f2}(s) = \min(18 , 2*5) = 10$ Gbps

$B_{f3}(s) = \min(18 , 3*4) = 12$ Gbps

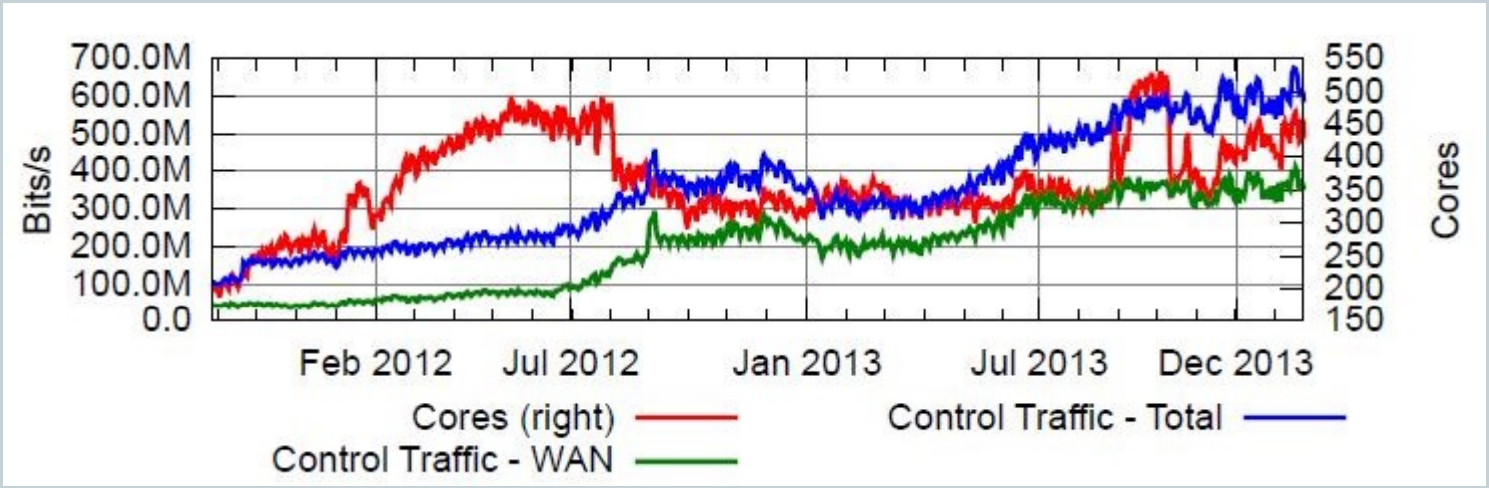
Best effort

and

As per priority

Evaluation

- BwE has run in production for many years on Google WAN



- Algorithm run time

	Algo Run-time		Algo Interval(s)	Reporting Interval(s)
	Max(s)	Mean(s)		
Global Enforcer	3	-	10	10
Cluster Enforcer	.16	.15	4	10
Job Enforcer	<0.01	<0.01	4	5

Conclusion

- Self-contained paper
- Well structured
- Mentions drawbacks and future optimization scope

Discussion ?