Flowtune

Jonathan Perry

Joint work with Hari Balakrishnan and Devavrat Shah.

PETASCAPE

Flowtune is..

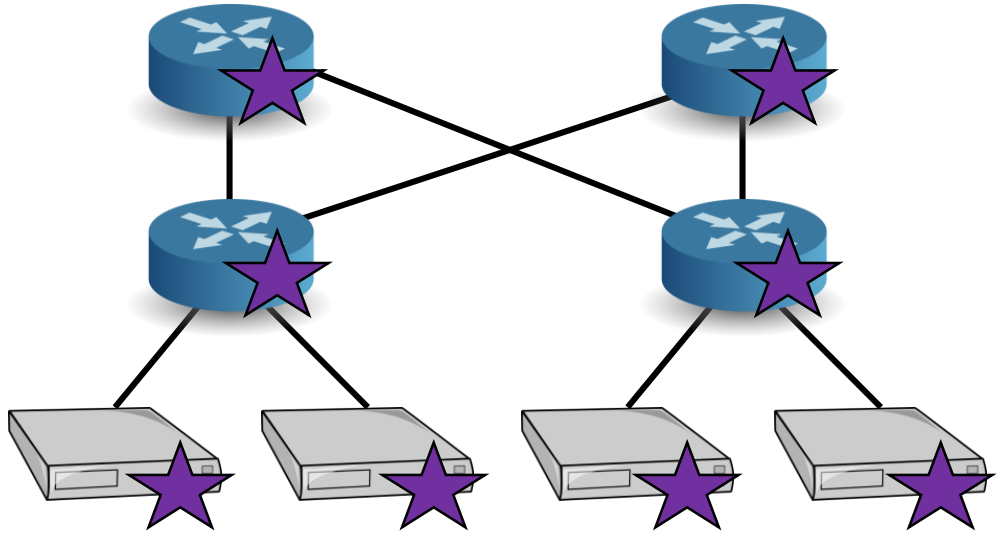
Allocate network resources

- Quickly
- Explicitly (maximize utility)
- Flexibly (in software)

Traditional approach is packet-centric

Switch Algorithms

Server Algorithms



Implicit Allocation

Several RTT to converge

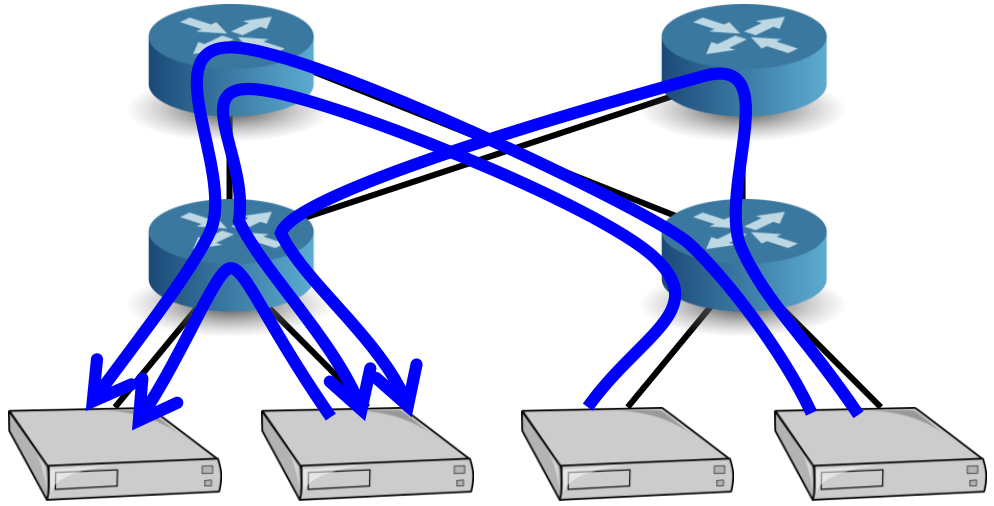
Changes many components

Flowtune's approach

1. *Flowlet control*

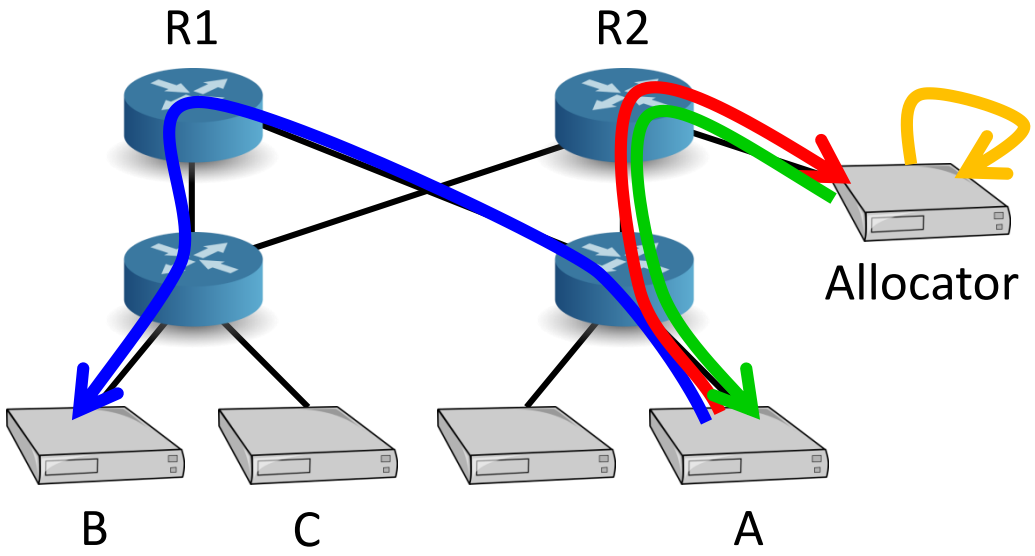
   Allocation changes *only* when:
   - Flowlets arrive
   - Flowlets terminate

2. Logically centralized
   - Reduce RTT dependence



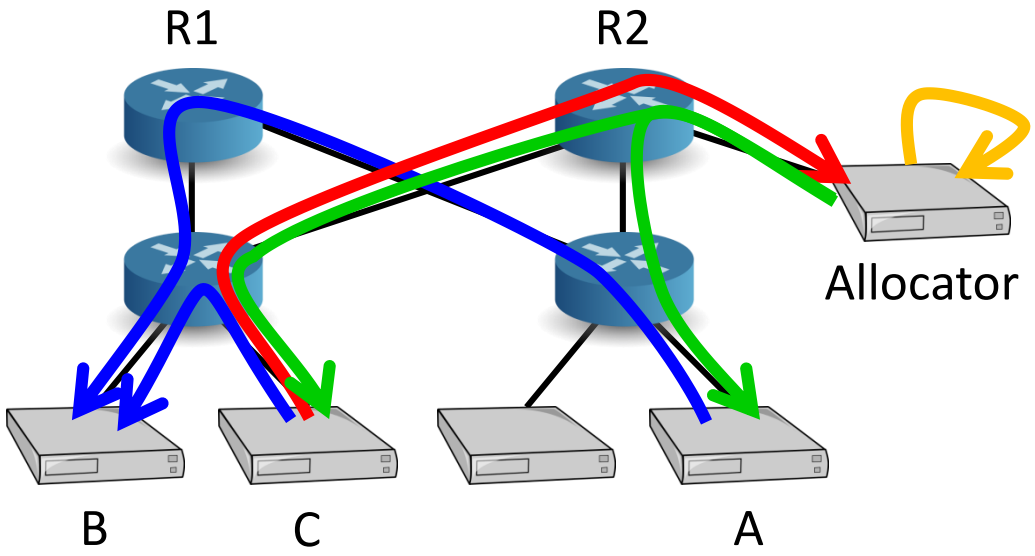
Example

A → Allocator	“Hadoop on A has data for B”
Allocator	Assign rates
Allocator → A	“Send at 10Gbps”



Example

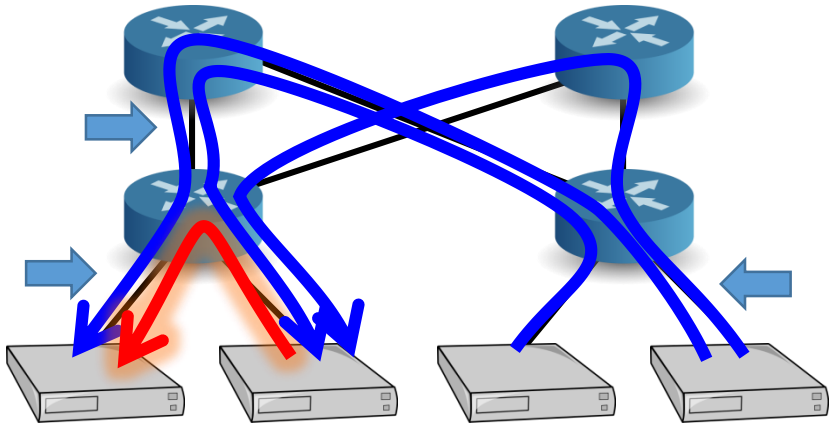
C → Allocator	“ads_update on C has data for B”
Allocator	Assign rates
Allocator → A	“Send at 1Gbps”
Allocator → C	“Send at 9Gbps”



Why is this hard?

Need to choose rates given active flowlets

1. Updates cascade!
2. What is the goal? To act like TCP?



NUM Iterative Optimizer

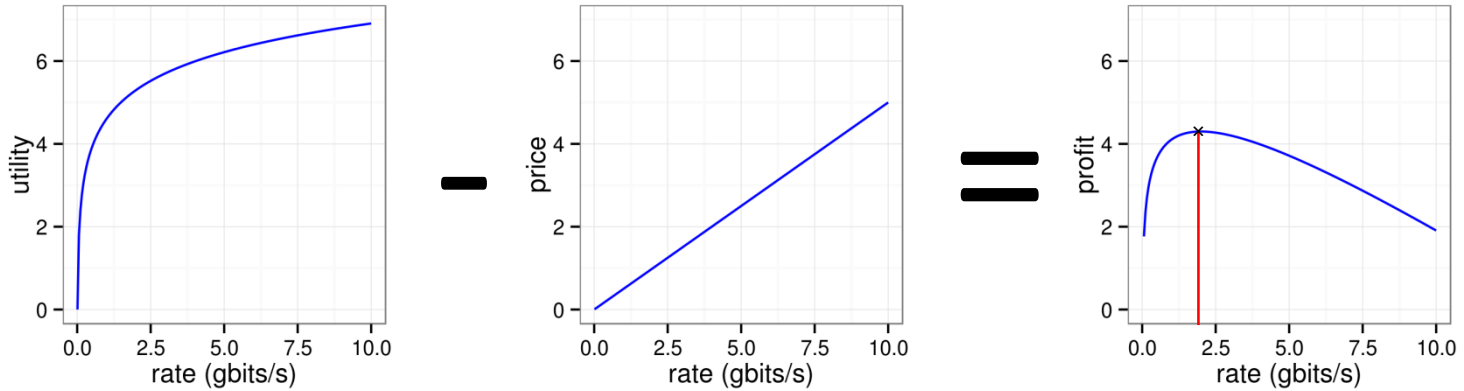
1. Each link ℓ chooses price p_ℓ

$$\sum_{s \in S(\ell)} x_s - c_\ell$$

Supply

Demand

2. Each flow s chooses rate x_s



3. Goto 1

Kelly et al., Journal of the Operational Research Society, 1998

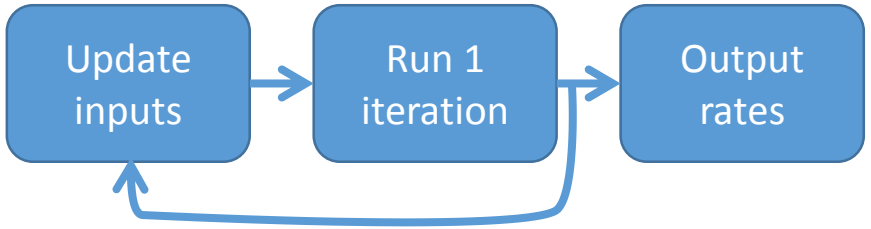
How to reduce latency?

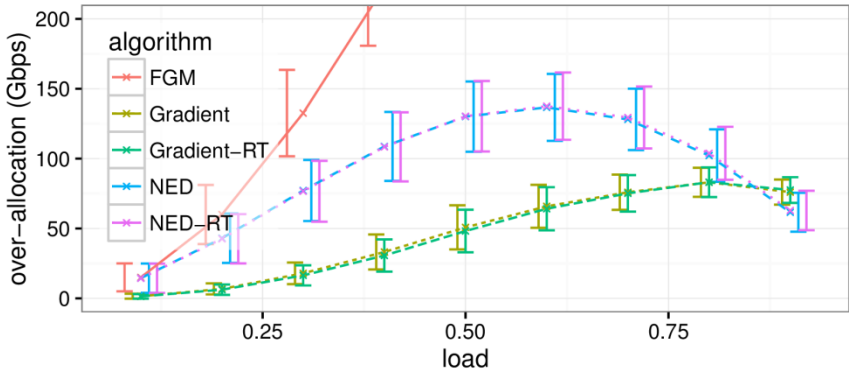
Solution 1:

But: too slow!

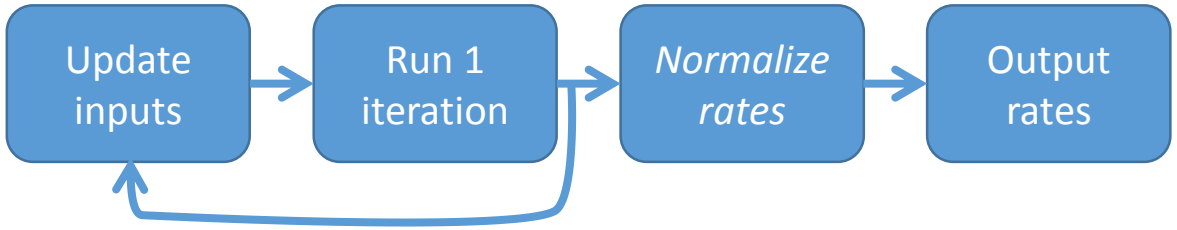
But: links are over-allocated!

Solution 2:



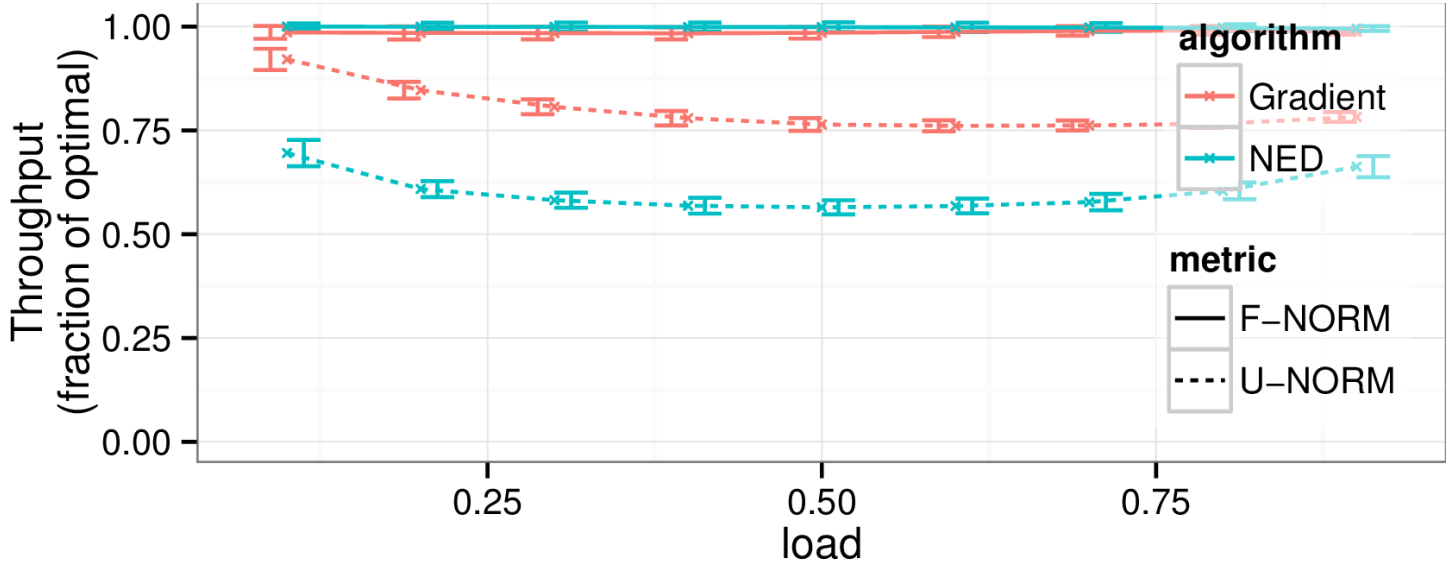


Solution 3:

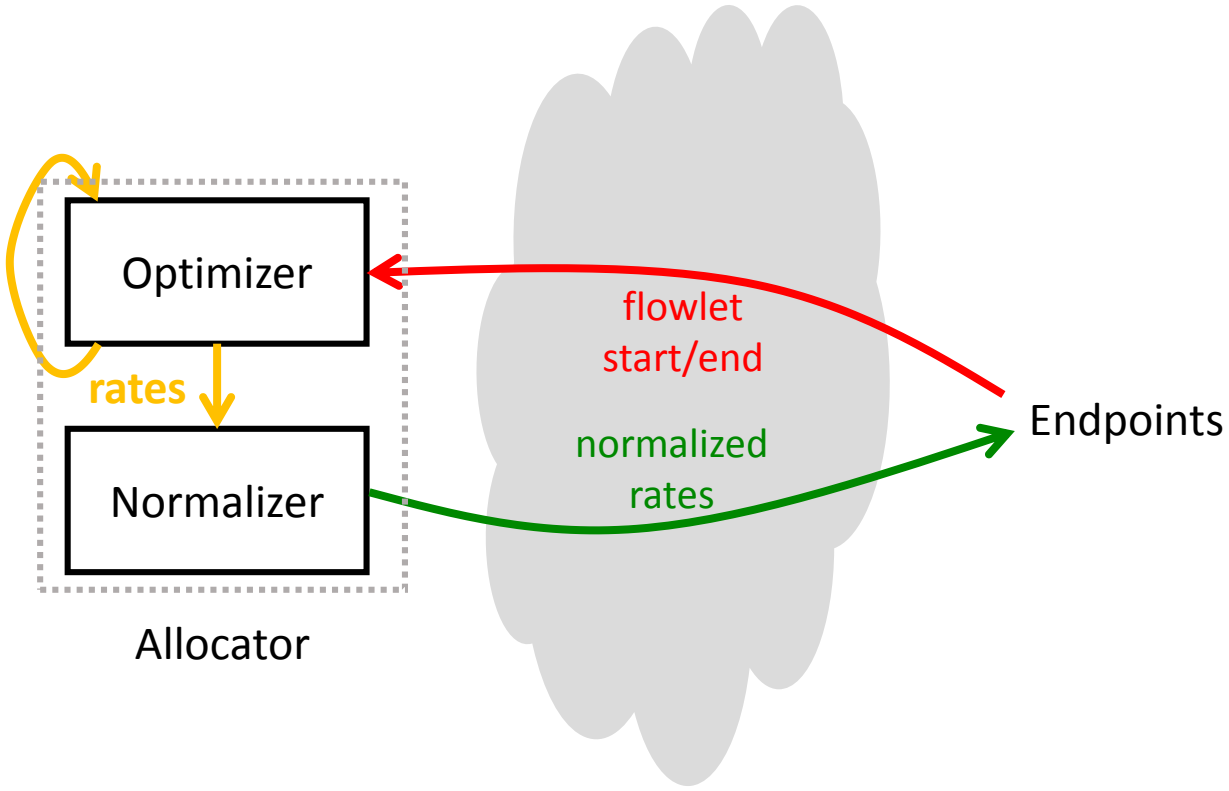


Flowtune normalizes rates between iterations

- For each flow:
  - Find link ℓ on path with largest $\mathrm{r}_\ell = \frac{\sum flow\ rates}{link\ capacity}$
  - Normalize: $x_s \leftarrow {}^{x_s}/_{r_\ell}$

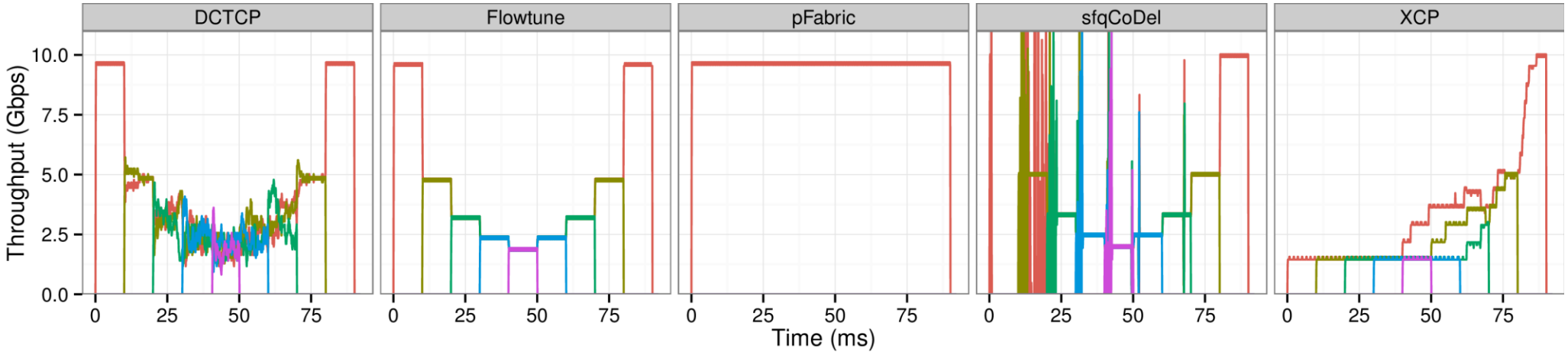


Architecture



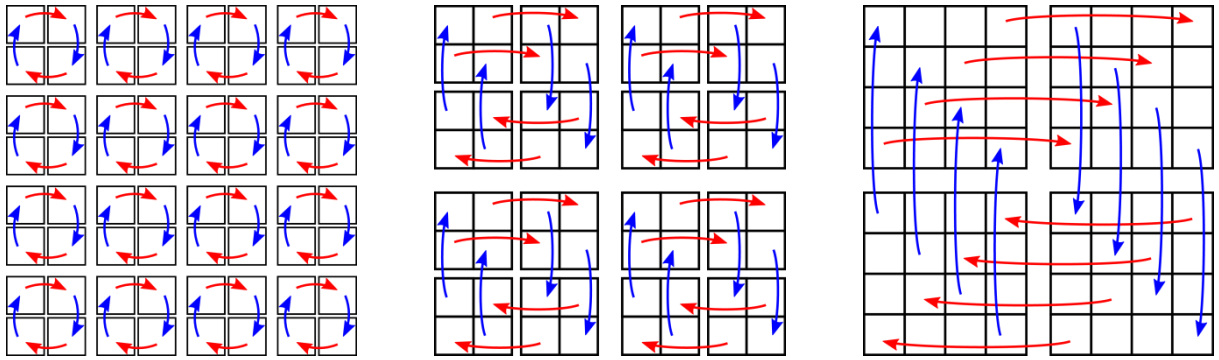
Flowtune converges quickly to fair allocation

1. Every 10 milliseconds add sender, up to 5 senders
2. Every 10 milliseconds remove sender



Flowtune scales to 49K flows on 64 cores

Cores	Nodes	Flows	Cycles	Time
4	384	3072	19896.6	8.29 µs
16	768	6144	21267.8	8.86 µs
64	1536	12288	30317.6	12.63 µs
64	1536	24576	33576.2	13.99 µs
64	1536	49152	40628.5	16.93 µs
64	3072	49152	57035.9	23.76 µs
64	4608	49152	73703.2	30.71 µs



Flowtune

Allocate network resources

- Quickly (centralized)
- Explicitly (maximize utility)
- Flexibly (in software)