FairCloud: Sharing the Network in Cloud Computing

Lucian Popa (HP Labs)
Gautam Kumar (UC Berkeley)
Mosharaf Chowdhury (UC Berkeley)
Arvind Krishnamurthy (Univ Washington)
Sylvia Ratnasamy (UC Berkeley)
Ion Stoica (UC Berkeley)
Motivation
Context

Networks are more difficult to share than other resources
Context

• Several proposals that share network differently, e.g.:
  – proportional to # source VMs (Seawall [NSDI11])
  – statically reserve bandwidth (Oktopus [Sigcomm12])
  – ...

• Provide specific types of sharing policies

• Characterize solution space and relate policies to each other?
This Talk

1. Framework for understanding network sharing in cloud computing
   – Goals, tradeoffs, properties

2. Solutions for sharing the network
   – Existing policies in this framework
   – New policies representing different points in the design space
Goals

1. Minimum Bandwidth Guarantees
   - Provides predictable performance
   - Example: file transfer finishes within time limit

\[
\text{Time}_{\text{max}} = \frac{\text{Size}}{B_{\text{min}}}
\]
Goals

1. Minimum Bandwidth Guarantees

2. High Utilization
   - Do not leave useful resources unutilized
   - Requires both work-conservation and proper incentives

![Diagram with three sections: Both tenants active, Non work-conserving, Work-conserving]
Goals

1. Minimum Bandwidth Guarantees
2. High Utilization
3. Network Proportionality

– As with other services, network should be shared proportional to payment

– Currently, tenants pay a flat rate per VM ➔ network share should be proportional to #VMs (assuming identical VMs)
Goals

1. Minimum Bandwidth Guarantees
2. High Utilization
3. Network Proportionality
   - Example: A has 2 VMs, B has 3 VMs

\[
\frac{Bw_A}{Bw_B} = \frac{2}{3}
\]

When exact sharing is not possible use max-min
Goals

1. Minimum Bandwidth Guarantees
2. High Utilization
3. Network Proportionality

Not all goals are achievable simultaneously!
Tradeoffs

Min Guarantee  ❯  Network Proportionality

High Utilization

Not all goals are achievable simultaneously!
Tradeoffs

Min Guarantee  →  Network Proportionality
Tradeoffs

Min Guarantee

Network Proportionality

Access Link L
Capacity C

Min Guarantee

Network Proportionality

Minimum Guarantee

\(Bw_A = \frac{1}{2} C\)

\(Bw_B = \frac{1}{2} C\)

Network Proportionality

\(Bw_A = \frac{2}{13} C\)

\(Bw_B = \frac{11}{13} C\)

\(Bw_A \approx \frac{C}{N_T} \to 0\)

#VMs in the network

10 VMs
Tradeoffs

Network Proportionality

High Utilization
Tradeoffs

Network Proportionality

High Utilization

A_1 → A_2
A_3 → A_4
B_1 → B_2
B_3 → B_4
Tradeoffs

Network Proportionality

High Utilization

Network Proportionality

\[
\text{Bw}_A = \frac{1}{2} C
\]

\[
\text{Bw}_B = \frac{1}{2} C
\]
Tradeoffs

Uncongested path

Network Proportionality

High Utilization
Tradeoffs

Network Proportionality

Network Proportionality

\[ Bw_A^P + Bw_A^L = Bw_B^L \]

\[ Bw_A^L < Bw_B^L \]

Tenants can be disincentivized to use free resources

If A values \( A_1 \rightarrow A_2 \) or \( A_3 \rightarrow A_4 \) more than \( A_1 \rightarrow A_3 \)
Tradeoffs

Network proportionality applied only for flows traversing *congested* links *shared* by multiple tenants.
Tradeoffs

Uncongested path

Network Proportionality
Congestion Proportionality

High Utilization

Congestion Proportionality

\[ Bw_A^L = Bw_B^L \]
Tradeoffs

Network Proportionality
Congestion Proportionality

High Utilization

Still conflicts with high utilization
Tradeoffs

Network Proportionality

Congestion Proportionality

High Utilization

\[ C_1 = C_2 = C \]
Tradeoffs

Network Proportionality
Congestion Proportionality

High Utilization

\[ C_1 = C_2 = C \]

![Diagram showing network and congestion proportionality]

Congestion Proportionality

\[ B_{wA}^{L1} = B_{wB}^{L1} \]

\[ B_{wA}^{L2} = B_{wB}^{L2} \]
Tradeoffs

Network Proportionality

Congestion Proportionality

High Utilization

$C_1 = C_2 = C$

Demand drops to $\varepsilon$
Tradeoffs

Network Proportionality

Congestion Proportionality

High Utilization

Tenants incentivized to not fully utilize resources

\[ C_1 = C_2 = C \]
Tradeoffs

Network Proportionality
Congestion Proportionality

High Utilization

Tenants incentivized to not fully utilize resources

\[ C_1 = C_2 = C \]

\[ \text{Tenants incentivized to not fully utilize resources} \]
Tradeoffs

Network Proportionality

Congestion Proportionality

High Utilization

Tenants incentivized to not fully utilize resources

C_1 = C_2 = C

L_1

A_1

B_1

L_2

A_3

B_3

L_1

A_2

B_2

L_2

A_4

B_4

C/2

C/2

C - 2\epsilon

\epsilon

Uncongested
Tradeoffs

Network Proportionality

Congestion Proportionality

Link Proportionality

High Utilization

Proportionality applied to each link \textit{independently}
Tradeoffs

Network Proportionality

Congestion Proportionality

Link Proportionality

High Utilization

Full incentives for high utilization
Goals and Tradeoffs

Min Guarantee

Network Proportionality
- Congestion Proportionality
- Link Proportionality

High Utilization
Guiding Properties

Min Guarantee

Network Proportionality

Congestion Proportionality

Link Proportionality

High Utilization

Break down goals into lower-level necessary properties
Work Conservation

• Bottleneck links are fully utilized
• Static reservations do not have this property
Properties

- Min Guarantee
- Network Proportionality
  - Congestion Proportionality
  - Link Proportionality
- High Utilization
  - Work Conservation
  - Utilization Incentives
Utilization Incentives

- Tenants are not incentivized to lie about demand to leave links underutilized
- Network and congestion proportionality do not have this property
- Allocating links independently provides this property
Communication-pattern Independence

- Allocation does not depend on communication pattern
- Per flow allocation does not have this property
  - (per flow = give equal shares to each flow)
Properties

- Min Guarantee
- Network Proportionality
  - Congestion Proportionality
  - Link Proportionality
- High Utilization
  - Work Conservation
  - Utilization Incentives
- Symmetry
- Comm-Pattern Independence

[Diagram showing relationships between properties]
Symmetry

- Swapping demand directions preserves allocation
- Per source allocation lacks this property
  - (per source = give equal shares to each source)
Goals, Tradeoffs, Properties

Min Guarantee

Network Proportionality
- Congestion Proportionality
- Link Proportionality

Symmetry
Comm-Pattern Independence

High Utilization

Work Conservation
Utilization Incentives
Outline

1. Framework for understanding network sharing in cloud computing
   – Goals, tradeoffs, properties

2. Solutions for sharing the network
   – Existing policies in this framework
   – New policies representing different points in the design space
Per Flow (e.g. today)

Min Guarantee

Network Proportionality
- Congestion Proportionality
- Link Proportionality

Symmetry
Comm-Pattern Independence

High Utilization
- Work Conservation
- Utilization Incentives
Per Source (e.g., Seawall [NSDI’11])

- Min Guarantee
- Network Proportionality
  - Congestion Proportionality
  - Link Proportionality
- Symmetry
- Comm-Pattern Independence
- Work Conservation
- Utilization Incentives
- High Utilization
Static Reservation (e.g., Oktopus [Sigcomm’11])

- Min Guarantee
- Network Proportionality
  - Congestion Proportionality
  - Link Proportionality
- Symmetry
- Comm-Pattern Independence
- High Utilization
- Work Conservation
- Utilization Incentives
New Allocation Policies

3 new allocation policies that take different stands on tradeoffs
Proportional Sharing at Link-level (PS-L)

Min Guarantee

Network Proportionality

Congestion Proportionality

Link Proportionality

Symmetry

Comm-Pattern Independence

Work Conservation

Utilization Incentives

High Utilization
Proportional Sharing at Link-level (PS-L)

- Per tenant WFQ where weight = # tenant’s VMs on link

\[ WQ_A = \# \text{VMs A on L} \]

\[ \frac{Bw_A}{Bw_B} = \frac{\# \text{VMs A on L}}{\# \text{VMs B on L}} \]

Can easily be extended to use heterogeneous VMs (by using VM weights)
Proportional Sharing at Network-level (PS-N)
Proportional Sharing at Network-level (PS-N)

- Congestion proportionality in severely restricted context
- Per source-destination WFQ, total tenant weight = # VMs
Proportional Sharing at Network-level (PS-N)

- Congestion proportionality in severely restricted context
- Per source-destination WFQ, total tenant weight = # VMs
Proportional Sharing at Network-level (PS-N)

- Congestion proportionality in severely restricted context
- Per source-destination WFQ, total tenant weight = \# VMs

\[
\frac{WQ_A}{WQ_B} = \frac{\# VMs A}{\# VMs B}
\]
Proportional Sharing on Proximate Links (PS-P)

- Min Guarantee
- Network Proportionality
  - Congestion Proportionality
  - Link Proportionality
- Symmetry
- Comm-Pattern Independence
- High Utilization
  - Work Conservation
  - Utilization Incentives
Proportional Sharing on Proximate Links (PS-P)

- Assumes a tree-based topology: traditional, fat-tree, VL2 (currently working on removing this assumption)
Proportional Sharing on Proximate Links (PS-P)

• Assumess a tree-based topology: traditional, fat-tree, VL2
  (currently working on removing this assumption)

• Min guarantees
  – Hose model
  – Admission control
Proportional Sharing on Proximate Links (PS-P)

• Assumptions a tree-based topology: traditional, fat-tree, VL2 (currently working on removing this assumption)

• Min guarantees
  – Hose model
  – Admission control

• High Utilization
  – Per source fair sharing towards tree root
Proportional Sharing on Proximate Links (PS-P)

• Assumes a tree-based topology: traditional, fat-tree, VL2 (currently working on removing this assumption)

• Min guarantees
  – Hose model
  – Admission control

• High Utilization
  – Per source fair sharing towards tree root
  – Per destination fair sharing from tree root
Deploying PS-L, PS-N and PS-P

• **Full Switch Support**
  – All allocations can use hardware queues (per tenant, per VM or per source-destination)

• **Partial Switch Support**
  – PS-N and PS-P can be deployed using CSFQ [Sigcomm’98]

• **No Switch Support**
  – PS-N can be deployed using only hypervisors
  – PS-P could be deployed using only hypervisors, we are currently working on it
Evaluation

• Small Testbed + Click Modular Router
  – 15 servers, 1Gbps links

• Simulation + Real Traces
  – 3200 nodes, flow level simulator, Facebook MapReduce traces
Many to one

One link, testbed

PS-P offers guarantees

Bw_A

Bw_B

N

N

Per-Flow
Per-Source
PS-L
PS-P
MapReduce

One link, testbed

$Bw_A$  

$Bw_B$  

$5$  

$M$  

$R$  

$M+R = 10$

$PS-L$ offers link proportionality

Graph showing $Bw_B$ (Mbps) vs. $M$ with different allocation methods: Per-Flow, Per-Source, PS-L, and PS-P.
MapReduce

Network, simulation, Facebook trace
MapReduce

Network, simulation, Facebook trace

PS-N is close to network proportionality
MapReduce

Network, simulation, Facebook trace

PS-N and PS-P reduce shuffle time of small jobs by 10-15X
Conclusion

• Sharing cloud networks is not trivial

• First step towards a framework to analyze network sharing in cloud computing
  – Key goals (min guarantees, high utilization and proportionality), tradeoffs and properties

• New allocation policies, superset properties from past work
  – PS-L: link proportionality + high utilization
  – PS-N: restricted network proportional
  – PS-P: min guarantees + high utilization