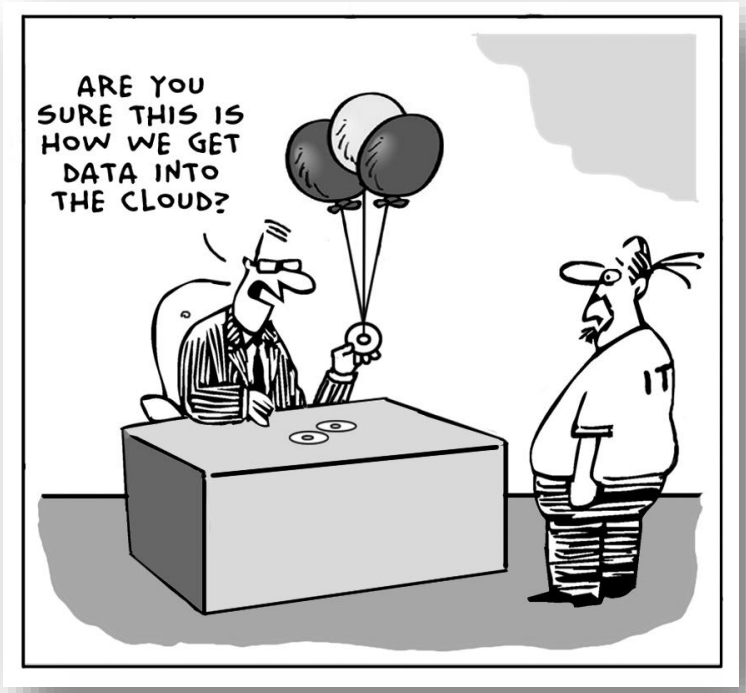
Microsoft Research

Microsoft



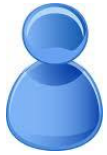
Bridging the Gap Between Applications and Networks in Data Centers

Paolo Costa
Microsoft Research Cambridge

joint work with
Hitesh Ballani, Thomas Karagiannis, Ant Rowstron

Motivation

Tenant



Request VMs

Small Instance – default*

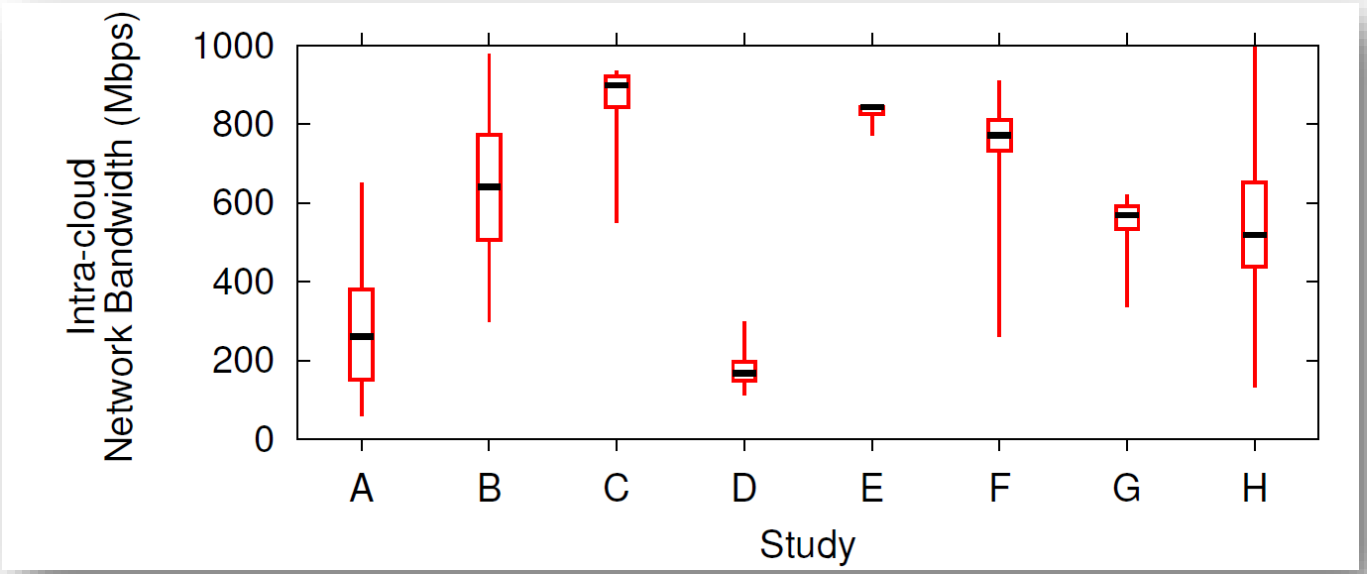
1.7 GB memory
1 EC2 Compute Unit (1 virtual core with 1 EC2 Compute Unit)
160 GB instance storage
32-bit platform
I/O Performance: Moderate
API name: m1.small

Large Instance

7.5 GB memory
4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each)
850 GB instance storage
64-bit platform
I/O Performance: High
API name: m1.large

✘ Network performance is not guaranteed!

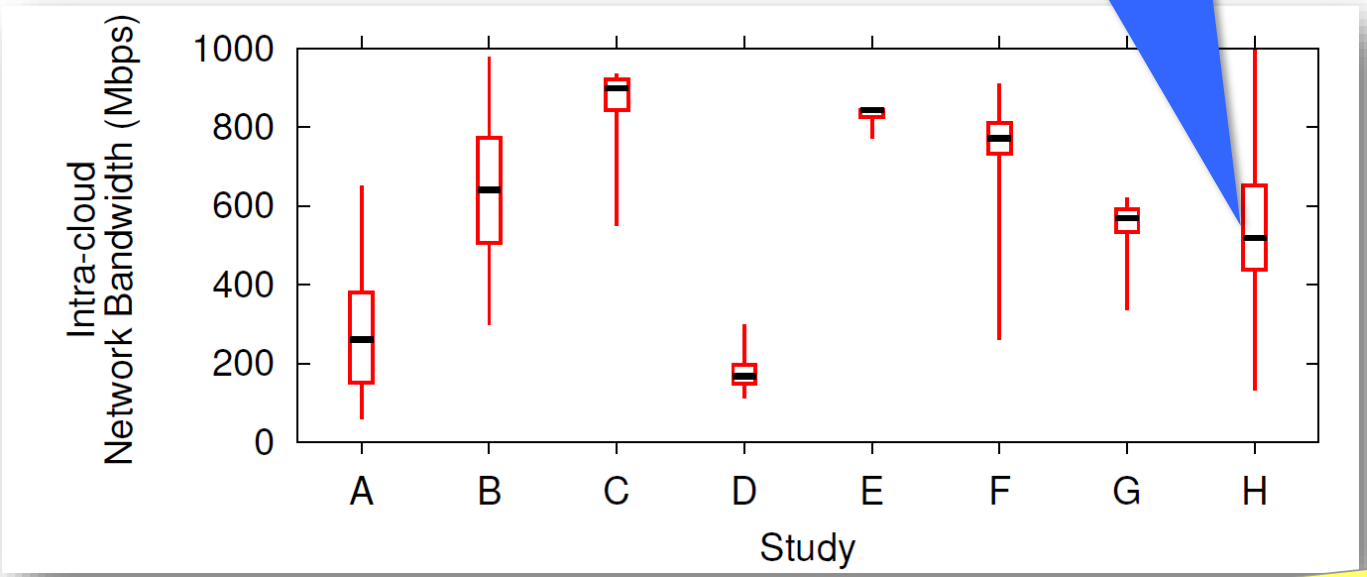
Is this really the case?



Study	Study	Provider	Duration
A	[Giurgui'10]	Amazon EC2	n/a
B	[Schad'10]	Amazon EC2	31 days
C/D/E	[Li'10]	(Azure, EC2, Rackspace)	1 day
F/G	[Yu'10]	Amazon EC2	1 day
H	[Mangot'09]	Amazon EC2	1 day

Is this really the case?

Up to 5x variability



Study	Study	Provider	
A	[Giurgui'10]	Amazon EC2	
B	[Schad'10]	Amazon EC2	
C/D/E	[Li'10]	(Azure, EC2, Rackspa	
F/G	[Yu'10]	Amazon EC2	
H	[Mangot'09]	Amazon EC2	1 day

Several causes
- VMs placement
- Network load
- Protocols used
- …

Fair enough…it's free so who cares?

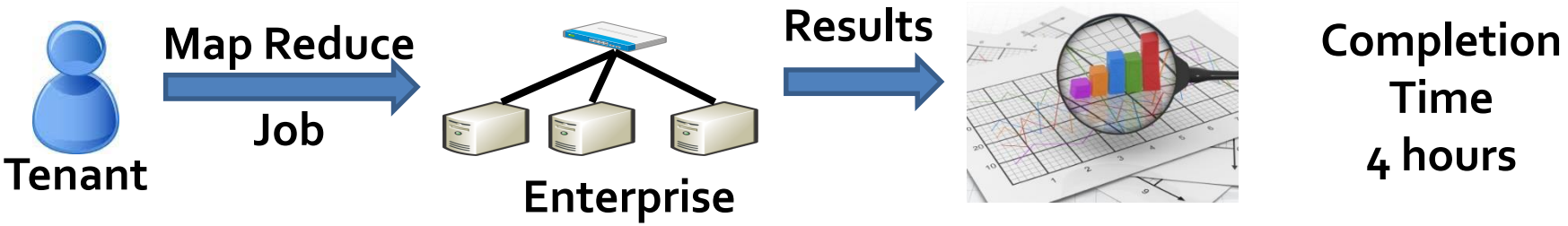
Region: US East (Virginia)

	Pricing
Data Transfer IN	
All data transfer in	$0.000 per GB

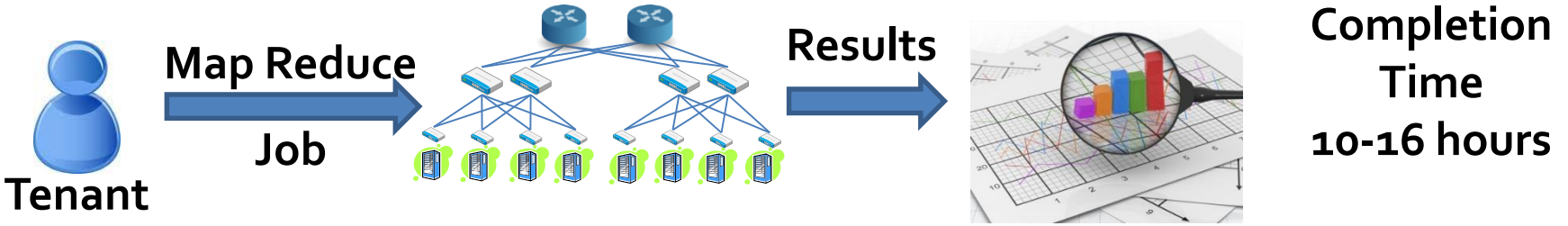
Implicit rather than explicit cost

Fair enough…it's free so who cares?

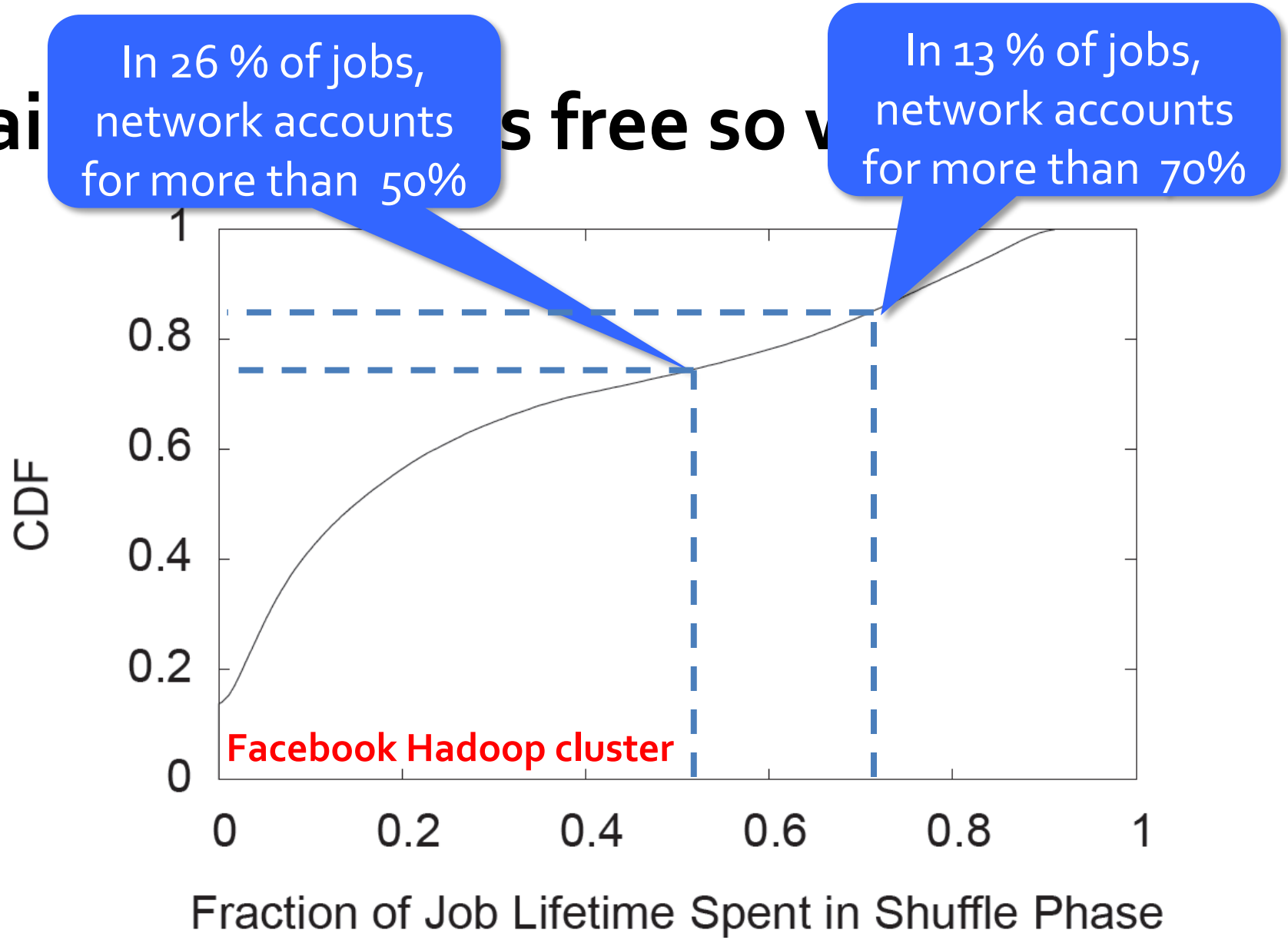
Data analytics on an isolated cluster



Data analytics in a multi-tenant data center



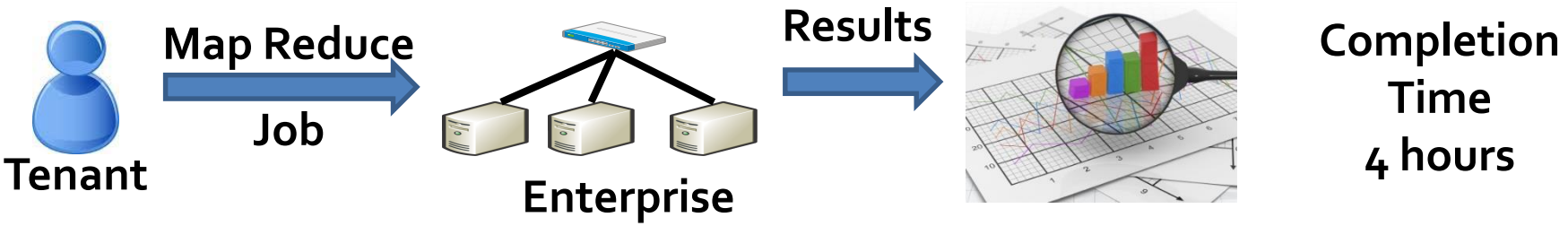
Fai s free so v



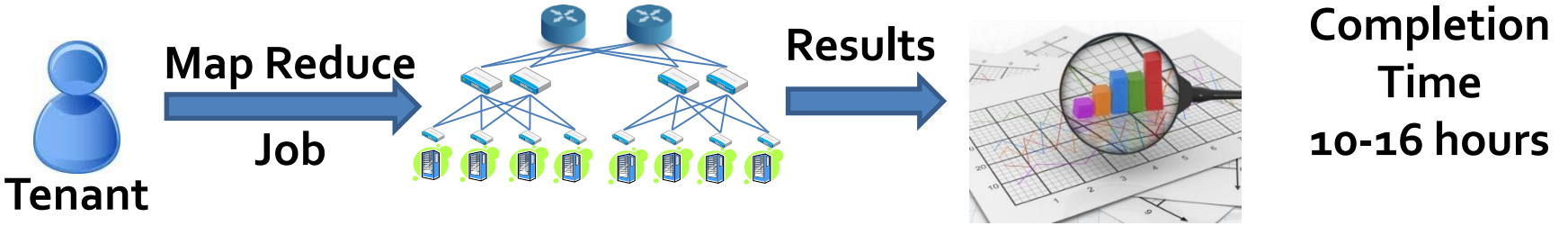
from Mosharaf Chowdhury et al. "Managing data transfers in computer clusters with orchestra", SIGCOMM'11

Fair enough…it's free so who cares?

Data analytics on an isolated cluster



Data analytics in a multi-tenant data center



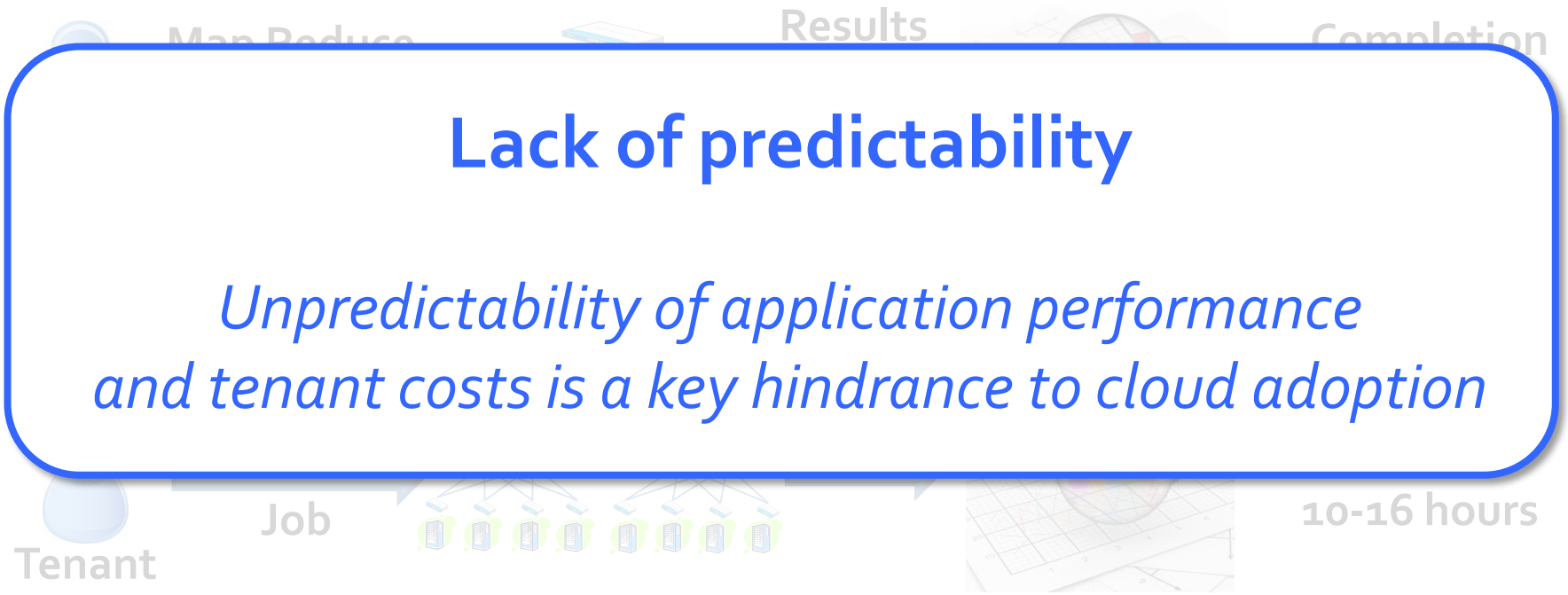
Variable tenant costs

Expected cost (based on 4 hour completion time) = $100

Actual cost = $250-400

Fair enough…it's free so who cares?

Data analytics on an isolated cluster



Lack of predictability

Unpredictability of application performance and tenant costs is a key hindrance to cloud adoption

Variable tenant costs

Expected cost (based on 4 hour completion time) = $100

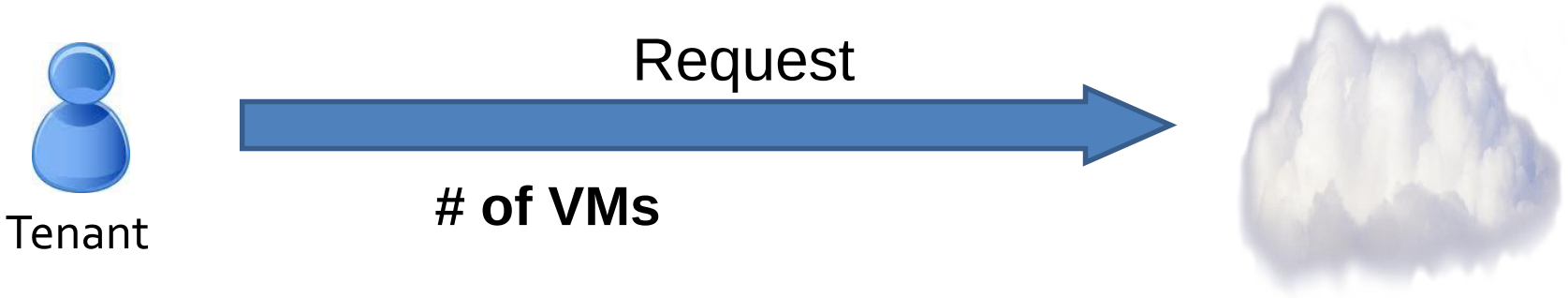
Actual cost = $250-400

Addressing the Problem

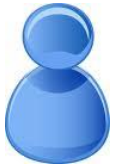
- Amazon EC2 cluster compute
  - ✓ Guaranteed 10 Gbps bandwidth
  - ✘ Very expensive for provider / tenants
    - Requires full-bisection bandwidth and 1 VM/server
    - $ 1.68 / hour ≅ 20 small instances
  - ✘ Very limited tenant flexibility
    - either 10Gbps or nothing

23 GB of memory
33.5 EC2 Compute Units (2 x Intel Xeon X5570, quad-core "Nehalem" architecture)
1690 GB of instance storage
64-bit platform
I/O Performance: Very High (10 Gigabit Ethernet)
API name: cc1.4xlarge

Predictable Data Center Networks



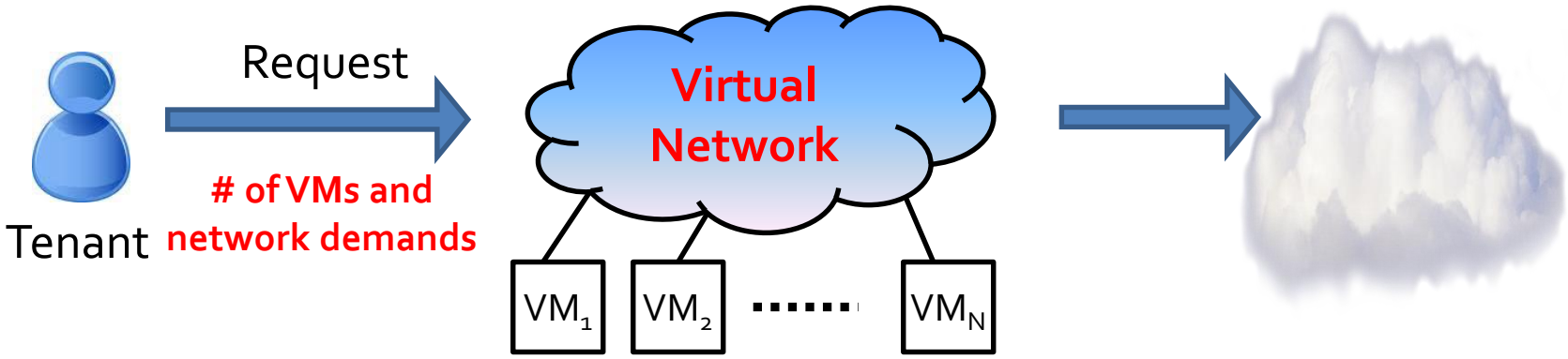
Predictable Data Center Networks



Tenant

# of VMs and network demands

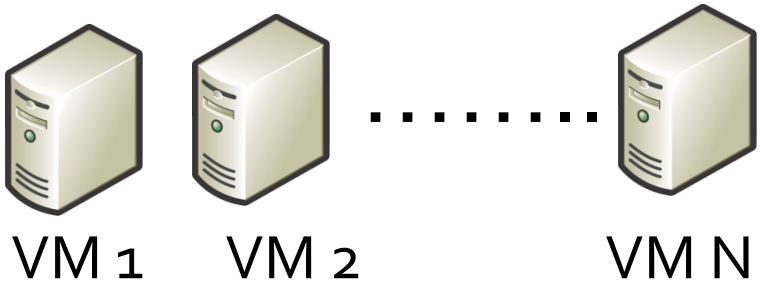
Predictable Data Center Networks



- Extend the tenant-provider interface to account for the network
- Easier transition for tenants
  - Tenants should be able to predict the performance of applications running atop the virtual network
- Provider flexibility
  - Providers should be able to multiplex many virtual networks on the physical network

These are competing design goals
Our abstractions strive to strike a balance between them

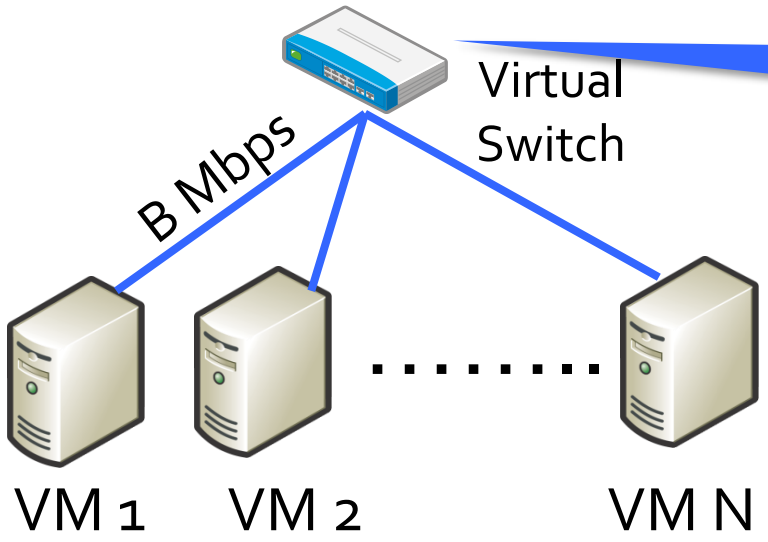
Abstraction #1: Virtual Cluster



Request <N>

N VMs

Abstraction #1: Virtual Cluster



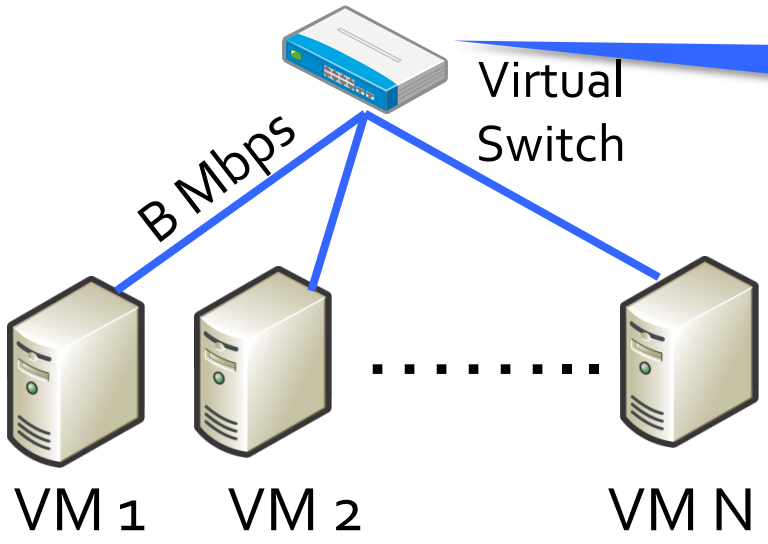
Total bandwidth
= N * B

Request <N, B>

N VMs. Each VM can send and receive at B Mbps

Abstraction #1: Virtual Cluster

- **Motivation**: In enterprises, tenants run applications on dedicated Ethernet clusters



Total bandwidth = N * B

Request <N, B>

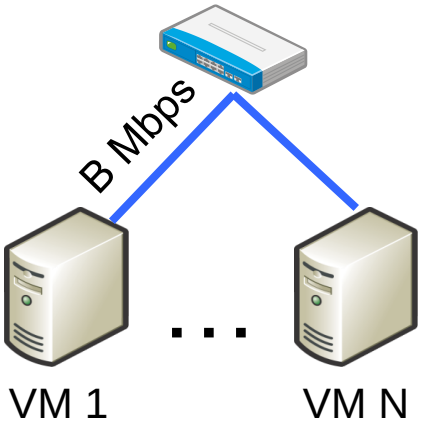
N VMs. Each VM can send and receive at B Mbps

Tenants get a network with no oversubscription

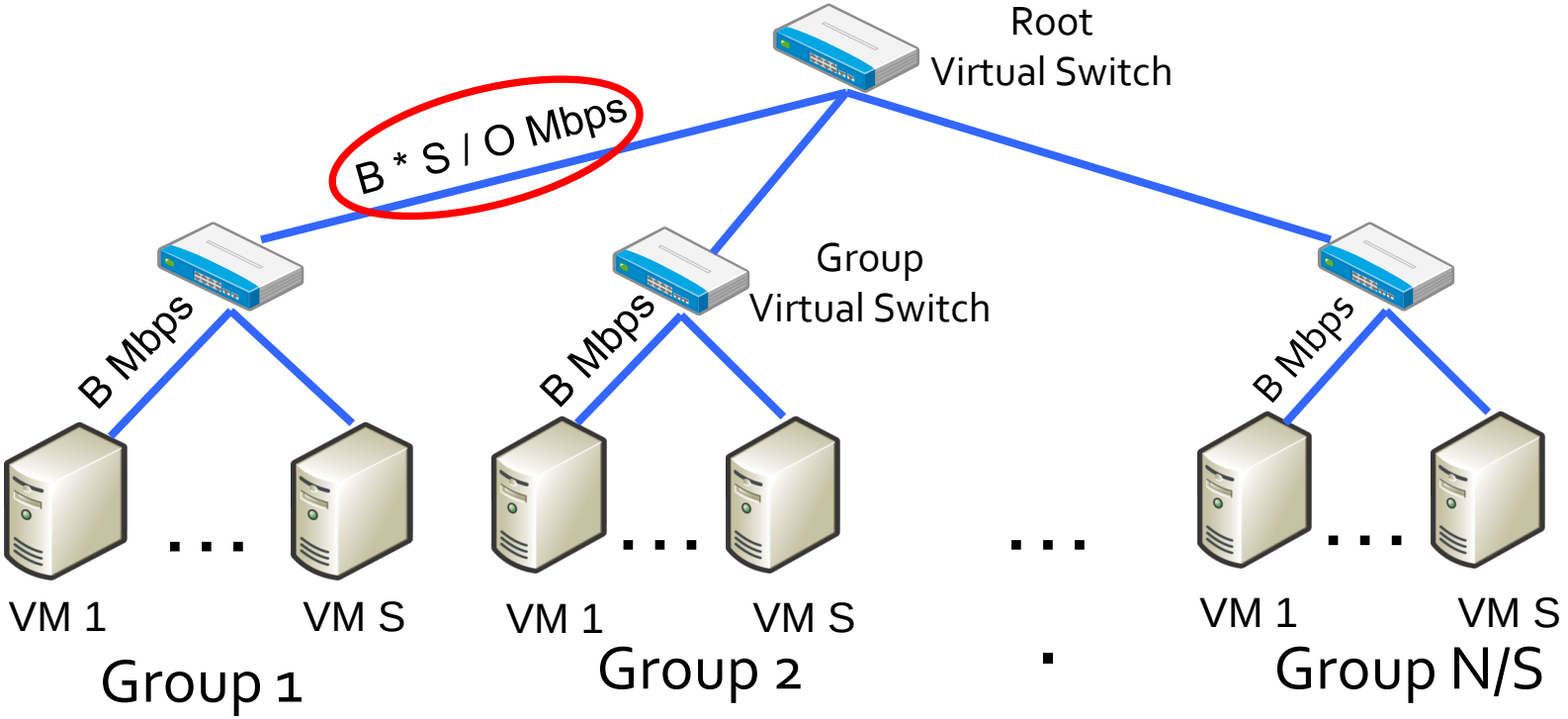
✓ Suitable for data-intensive apps. (MapReduce, BLAST)

✘ Moderate provider flexibility

Abstraction #2: Virtual Oversubscribed Cluster



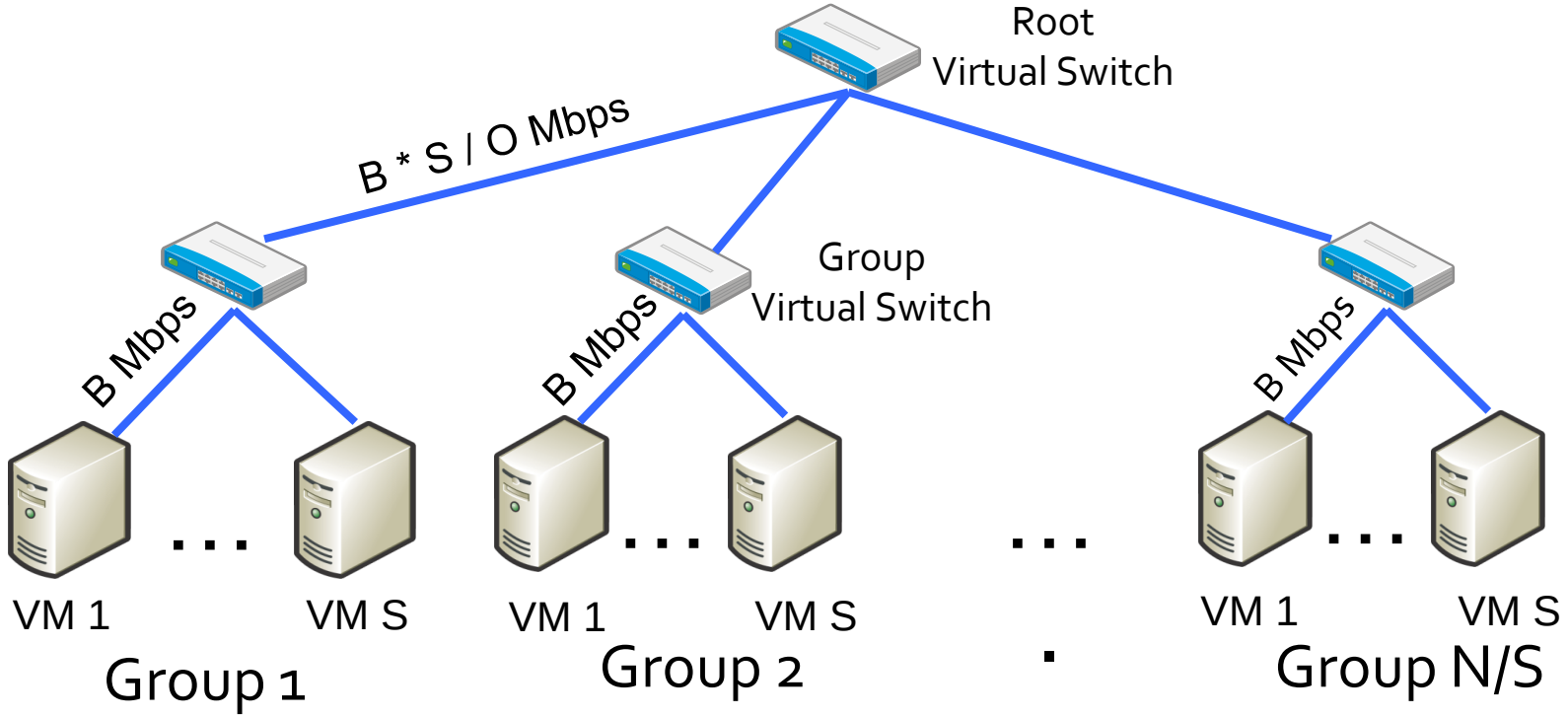
Abstraction #2: Virtual Oversubscribed Cluster



Request <N, B, S, O>

N VMs in groups of size S. Oversubscription factor O

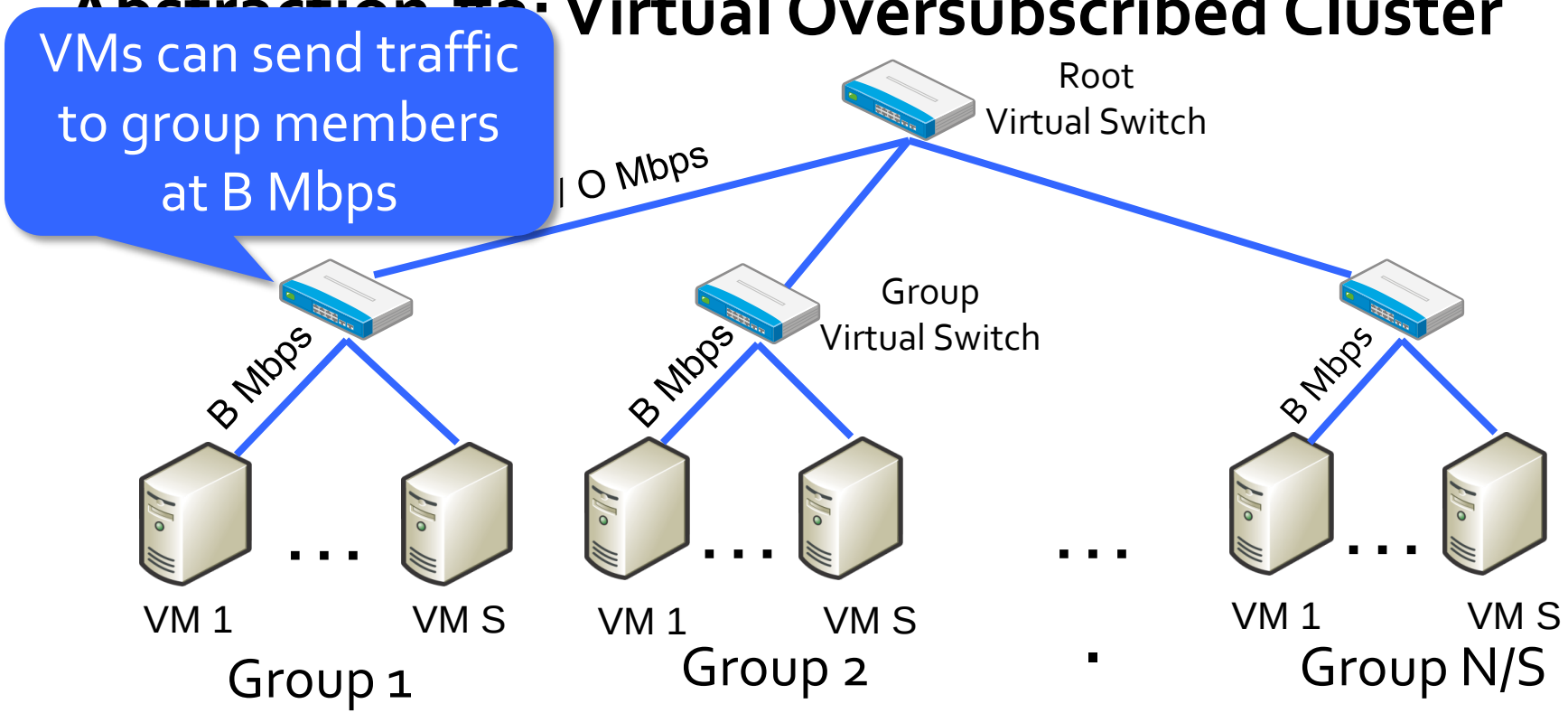
Abstraction #2: Virtual Oversubscribed Cluster



Motivation: Many applications moving to the cloud have localized communication patterns

Applications are composed of groups with more traffic within groups than across groups

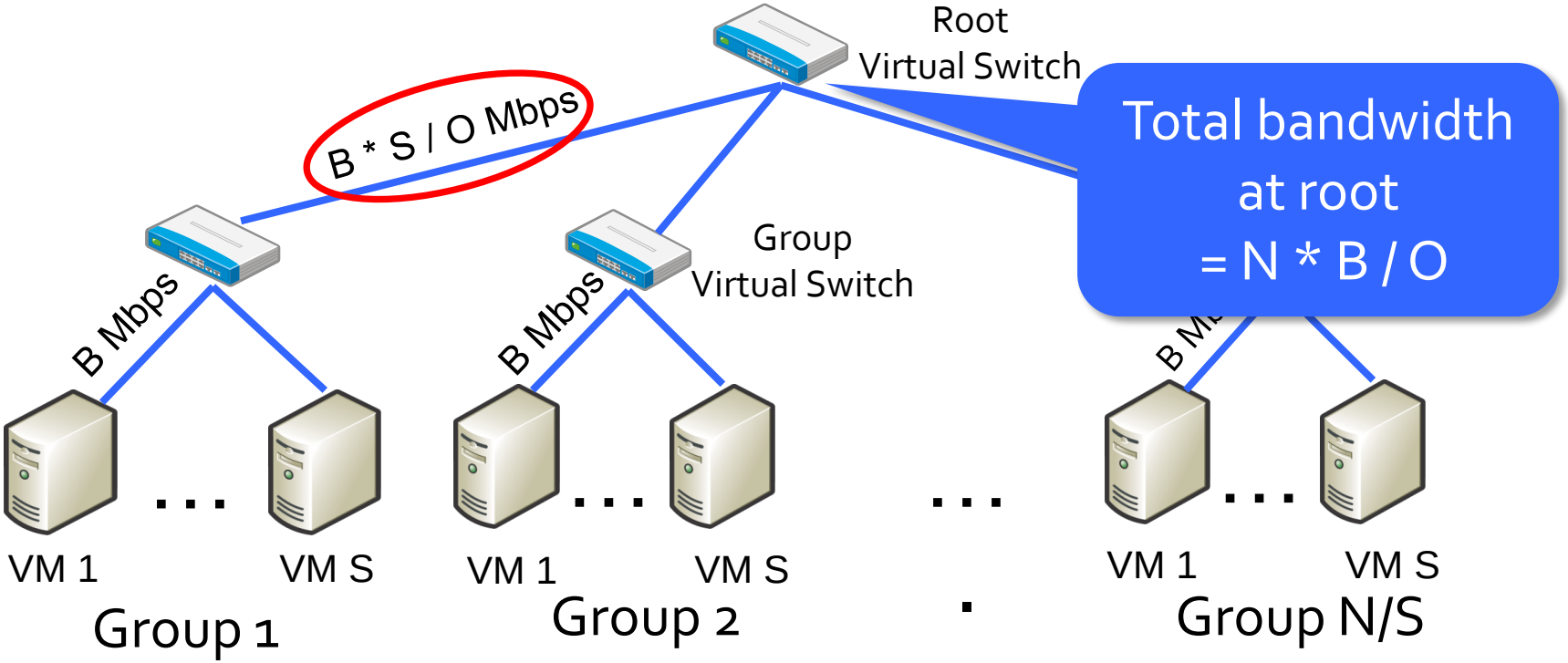
Abstraction #2: Virtual Oversubscribed Cluster



No oversubscription for intra-group communication

Intra-group communication is the common case!

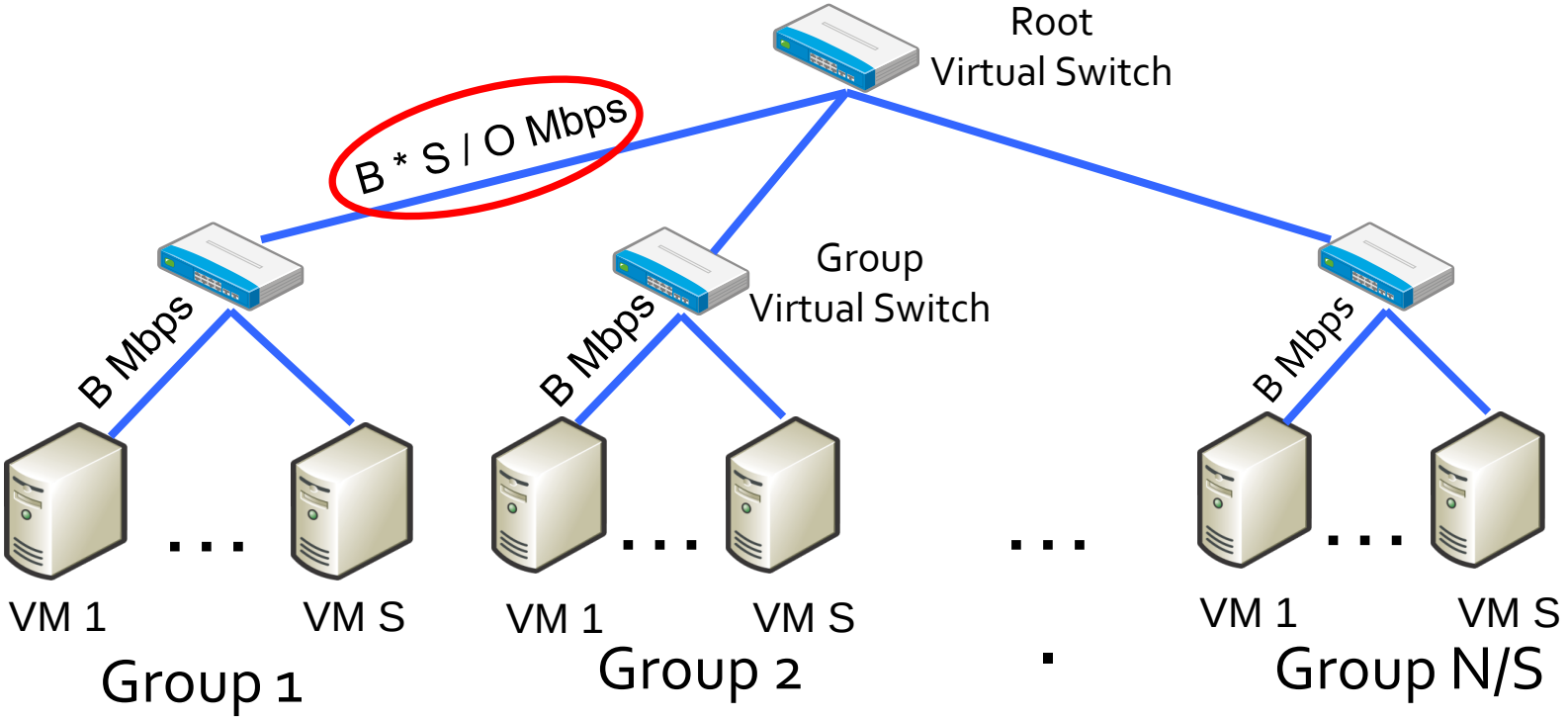
Abstraction #2: Virtual Oversubscribed Cluster



Oversubscription factor O for inter-group communication

(captures the sparseness of inter-group communication)

Abstraction #2: Virtual Oversubscribed Cluster



VOC capitalizes on tenant communication patterns

✓ Suitable for typical applications (*though not all*)

✓ Improved provider flexibility

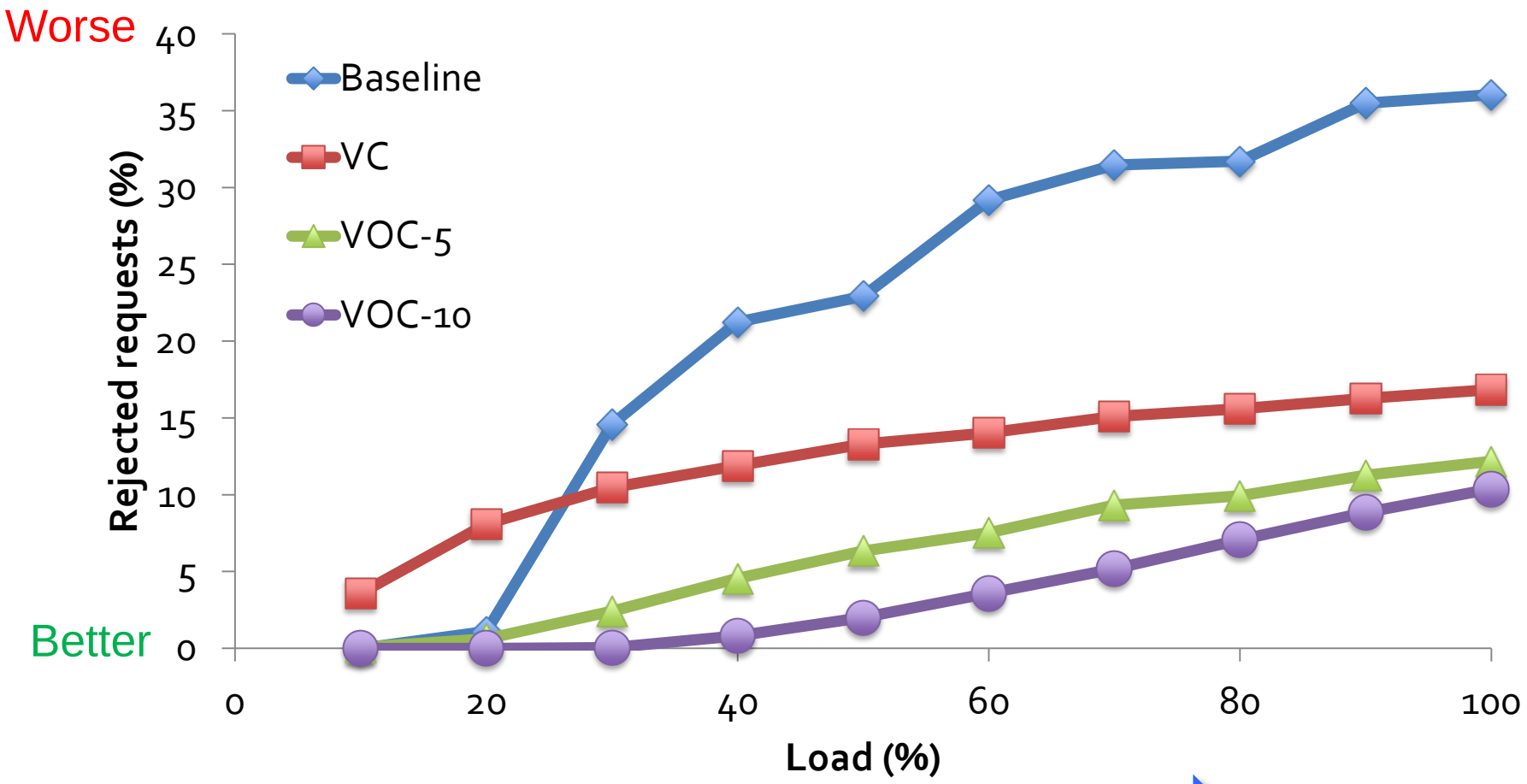
Oktopus [SIGCOMM'11]

- Offers virtual networks to tenants in data centers

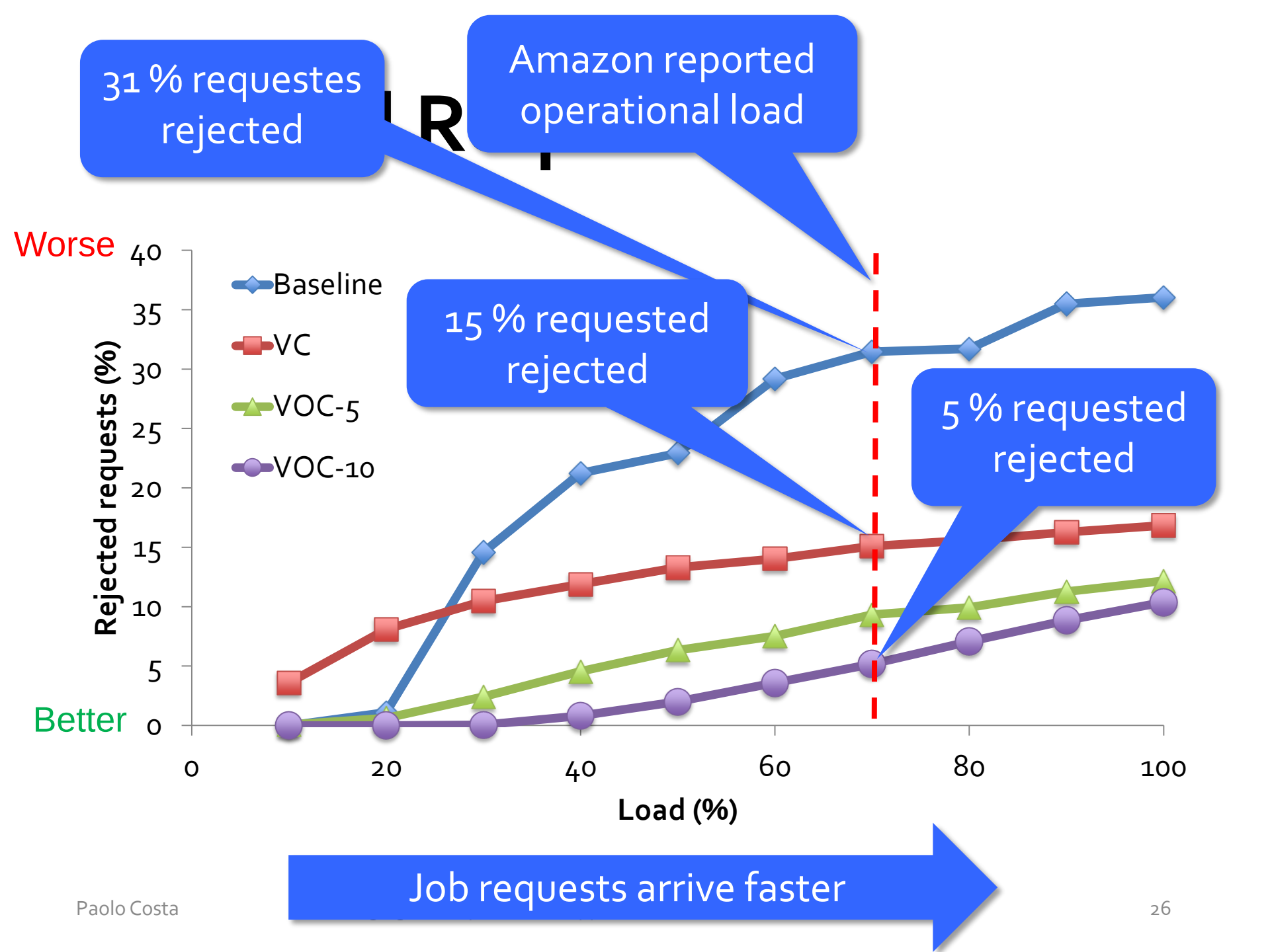
Oktopus [SIGCOMM'11]

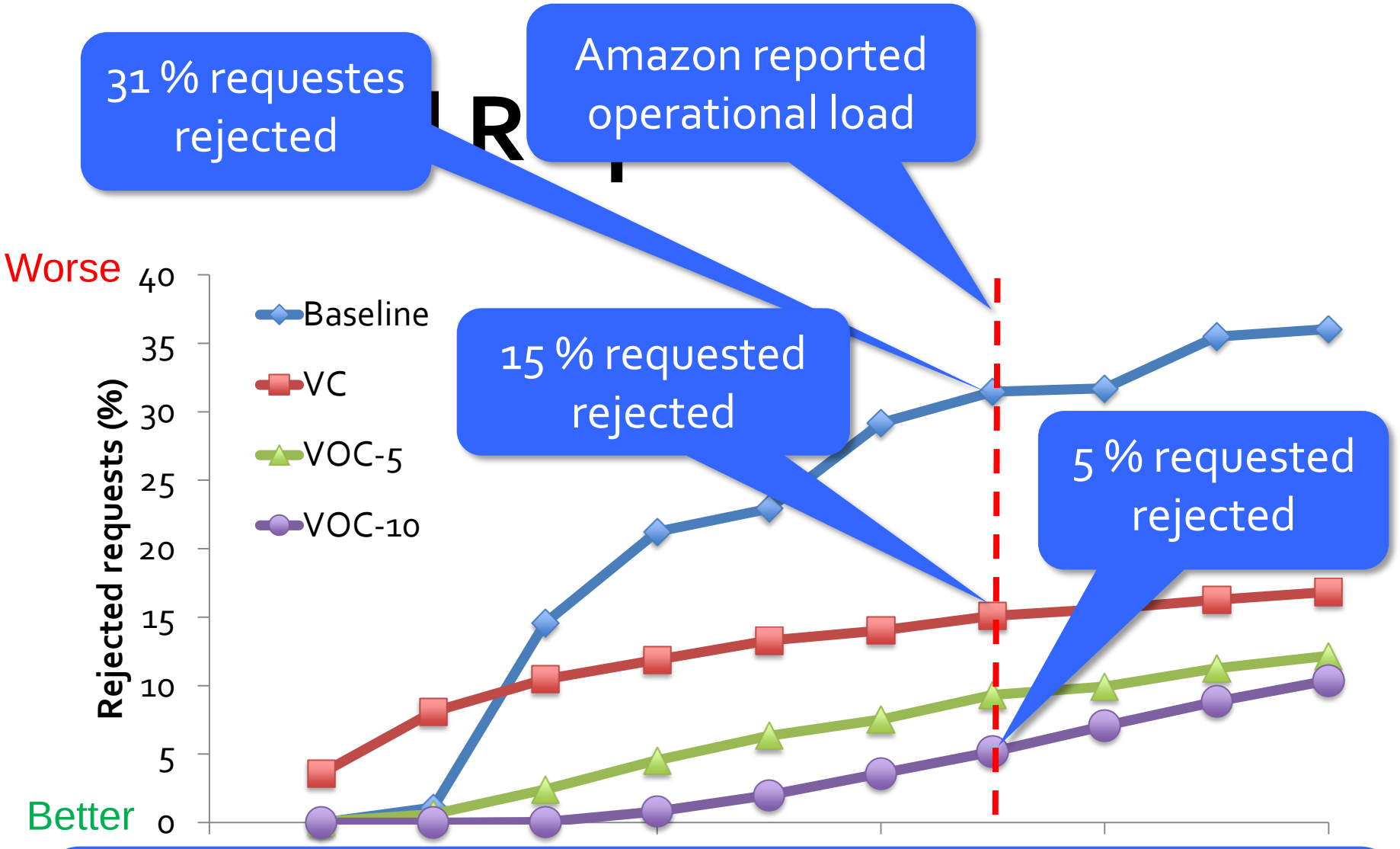
- Offers virtual networks to tenants in data centers
- Management plane: *Allocation of tenant requests*
  - *Goal:* Maximize the ability to allocate future requests
    - Variant of network embedding problem (NP-hard)
    - Existing heuristics don't scale beyond $O(10^2)$ nodes
  - *Key idea*: Restrict the set of virtual and physical topologies
    - Allocation is fast (median time is 0.35ms)
- Data plane: *Enforcement of virtual networks*
  - Enforcement of virtual networks is cheap
    - Implemented on hypervisor only (no network support)
    - It can be deployed today

Rejected Requests



Job requests arrive faster


31 % requestes rejected
Amazon reported operational load
15 % requested rejected
5 % requested rejected
Worse
Better
Baseline
VC
VOC-5
VOC-10
Rejected requests (%)
0
5
10
15
20
25
30
35
40
0
20
40
60
80
100
Load (%)
Job requests arrive faster



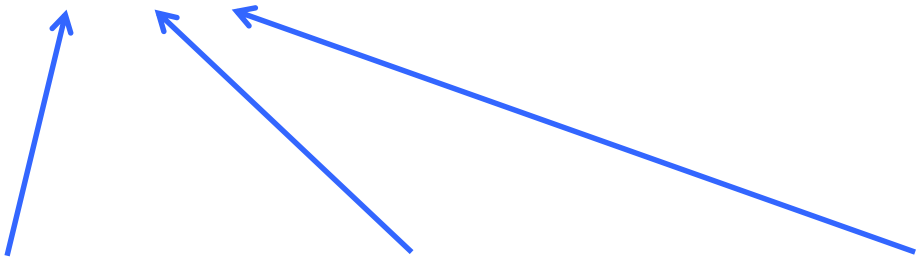
Virtual networks benefit both tenants and provider

Tenants: Job completion is faster and predictable

Provider: Higher data center throughput

How about the tenant cost?

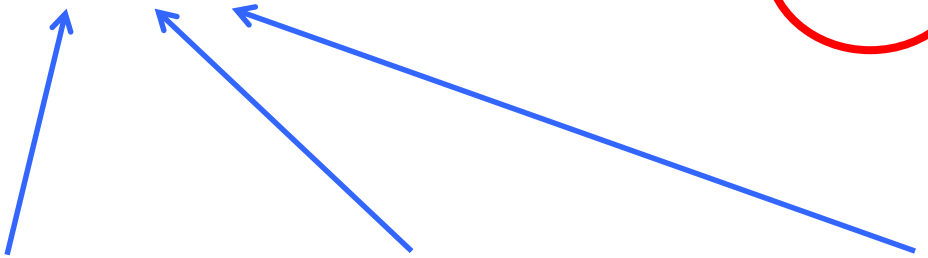
- What should tenants pay to ensure *provider revenue neutrality*?
  - kN_rT_r



k= $0.085 /hour N= # of VMs T = completion time

How about the tenant cost?

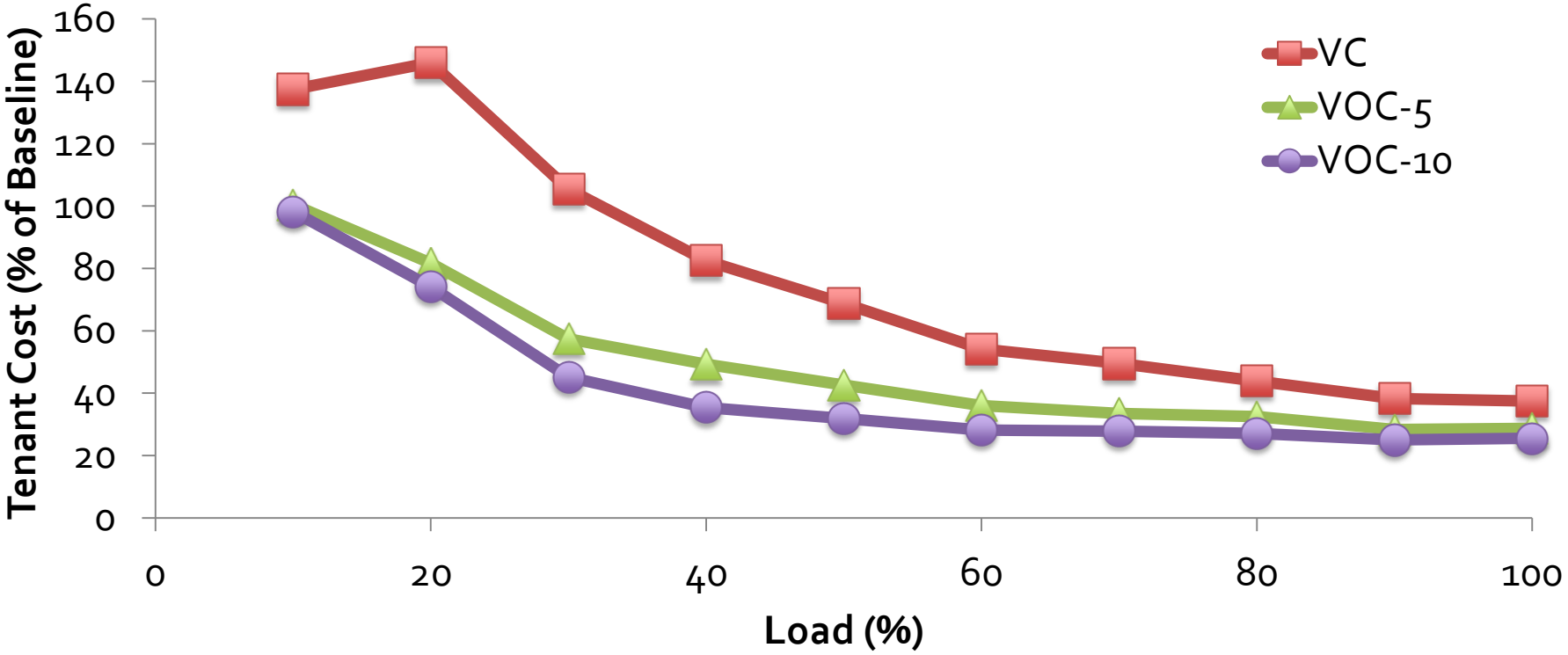
- What should tenants pay to ensure *provider revenue neutrality*?
  - $\sum_r(kN_rT_r) = \sum_r(N_rT_r(k_n + k_bB))$



k= $0.085 /hour N= # of VMs T = completion time

How about the tenant cost?

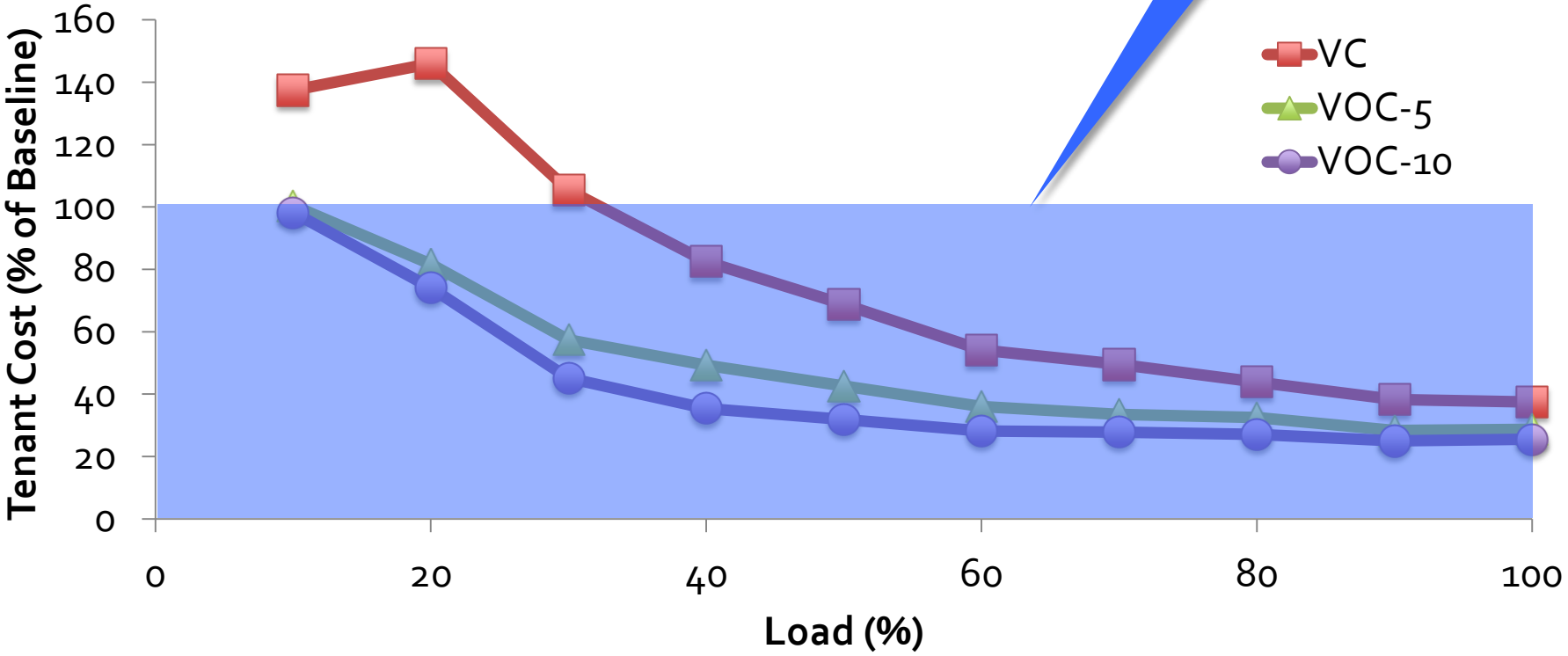
- What should tenants pay to ensure *provider revenue neutrality*?
  - $\sum_r(kN_rT_r) = \sum_r(N_rT_r(k_n + k_bB))$



How about the tenant co

- What should tenants pay to ensure *provider revenue neutrality*?
  - $\sum_r(kN_rT_r) = \sum_r(N_rT_r(k_n + k_bB))$

Tenants are paying less

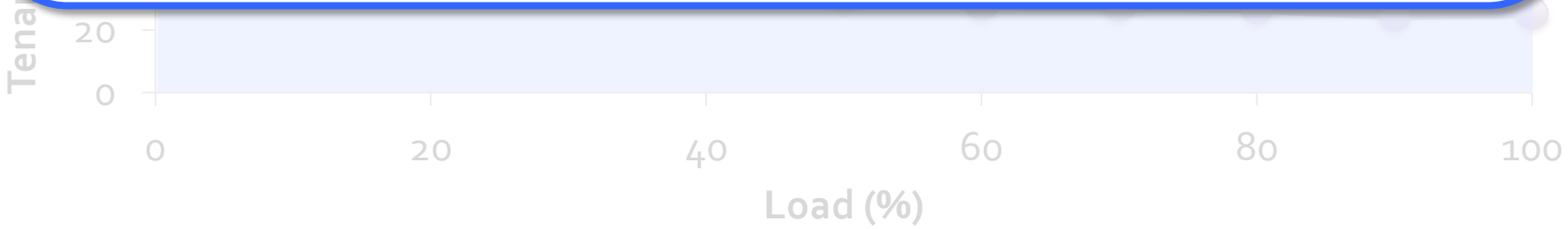


How about the tenant cost?

- What should tenants pay to ensure *provider revenue neutrality*?
  - $\sum_r(kN_rT_r) = \sum_r(N_rT_r(k_n + k_bB))$

Other pricing models are possible, e.g.,
Provider revenue increases while tenants pay less

Provider revenue increases by 20%
and median tenant cost reduces by 42%"



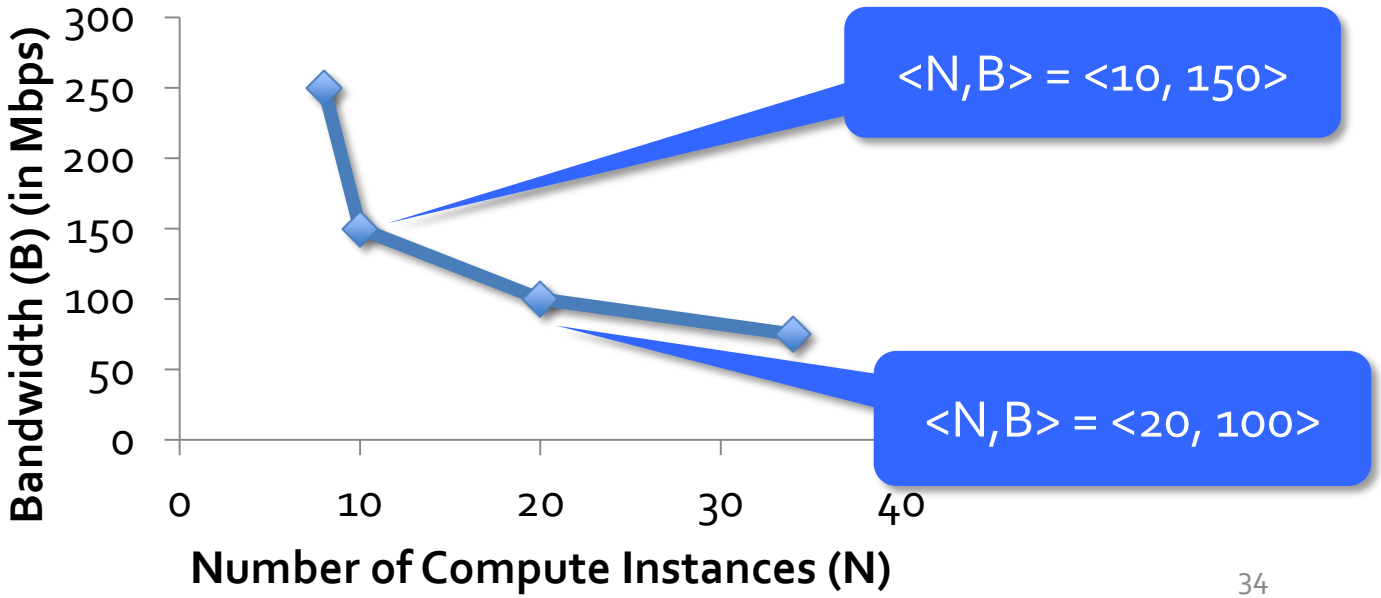
Beyond Oktopus

- Issues with Oktopus model
  - Hard for tenants to derive B (and even N)…
  - Unfortunate <N,B> choices can hurt provider's throughput
- Vision: Job-centric interface
  - *Tenants specify what they want*
    - Performance: complete the job in 1 hour
    - Cost: complete the job for less than $ 100
  - *Providers decide how to achieve this*
    - E.g., job needs 10 VMs and 200 Mbps of network

Why bother?

- Tenants
  - Simpler and more intuitive interface
  - Especially for non-IT experts
- Provider
  - Exploit multi-resource tradeoff

LinkGraph in 300s

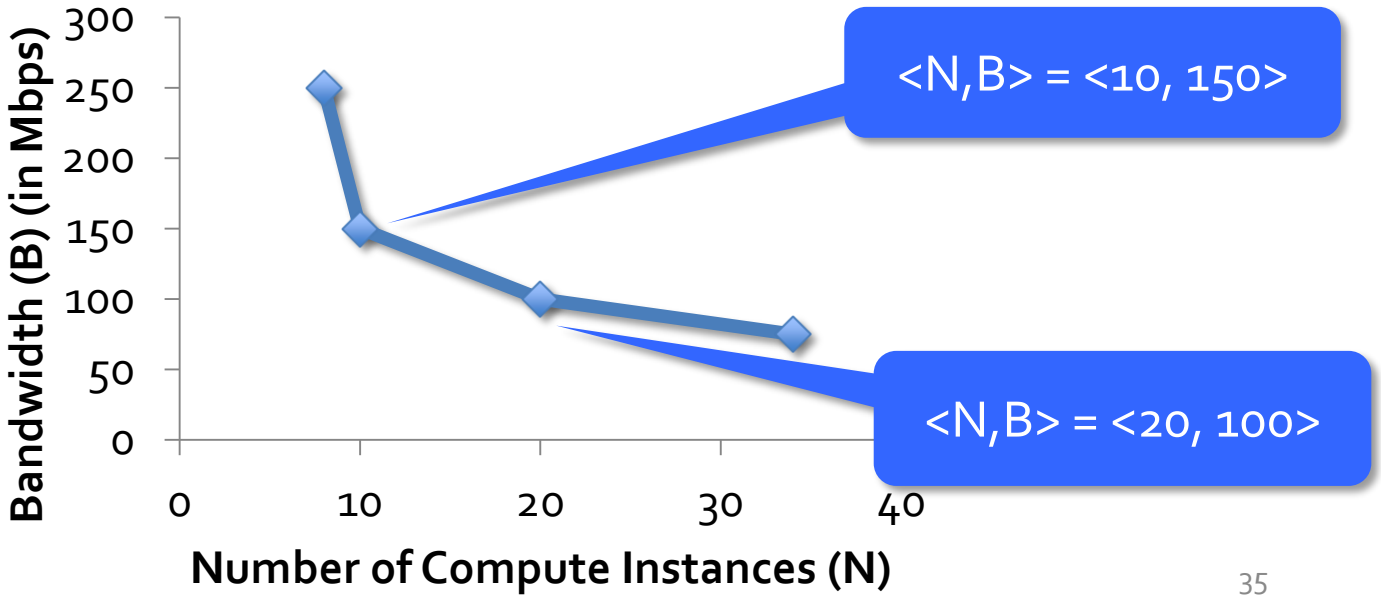


Why bother?

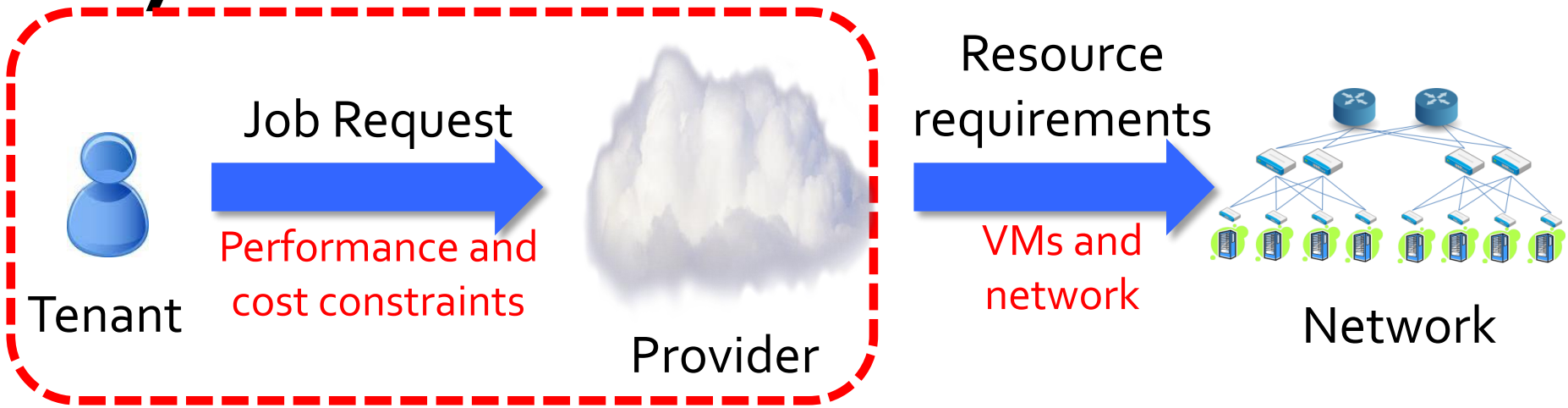
Intuition

Provider can choose the *most convenient* combination of <N,B> that satisfies tenants' goals

LinkGraph in 300s



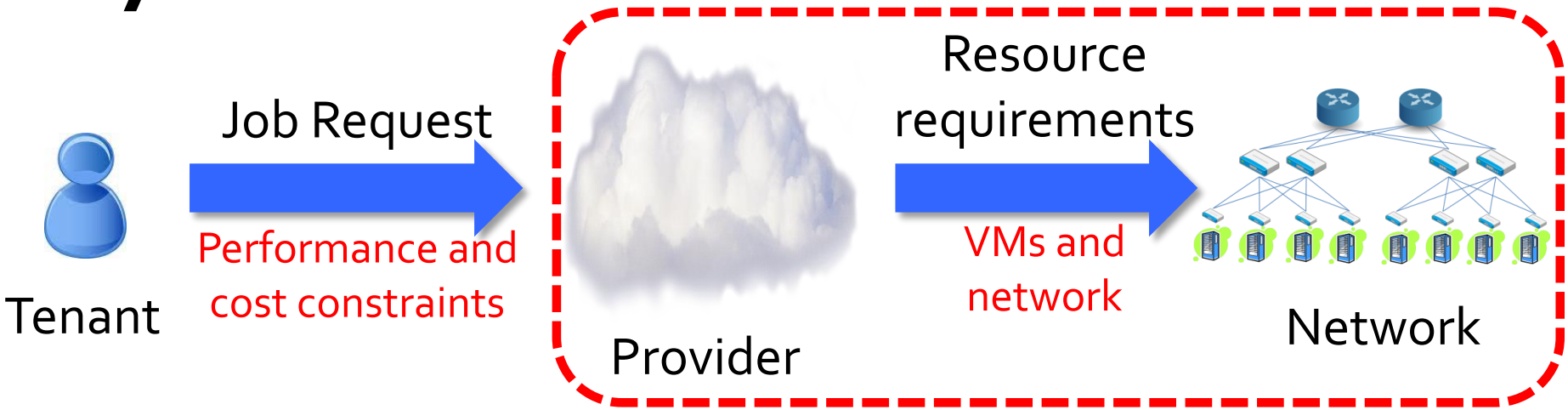
System Model



Tenants expose constraints

Performance: Complete job in less than 3 hours
Cost: Complete job for less than $400

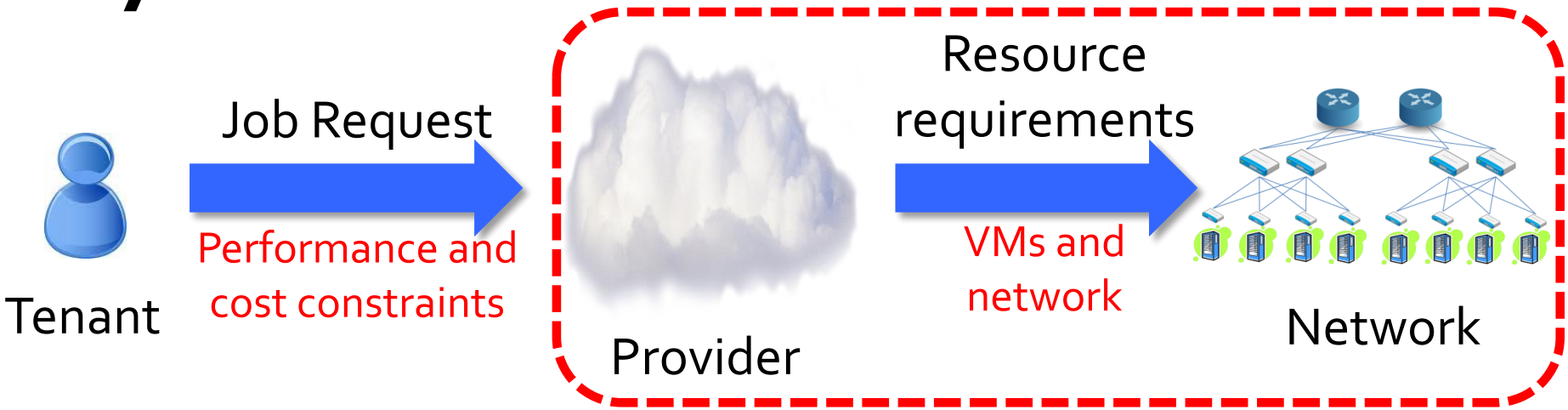
System Model



Bazaar determines resources required

Example, 10 VMs with guaranteed 200Mbps network between them will ensure the job finishes in less than 3 hours

System Model



Multiple resource combinations may satisfy tenant goal

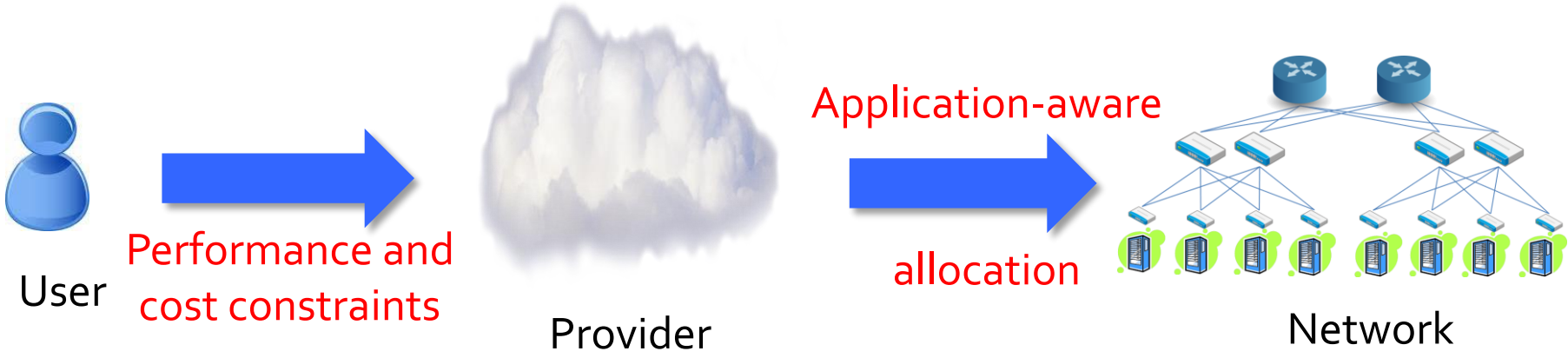
Completion time = 3 hrs ⇨ 1). 25 VMs with 500 Mbps, or
2). 40 VMs with 400 Mbps

System Model

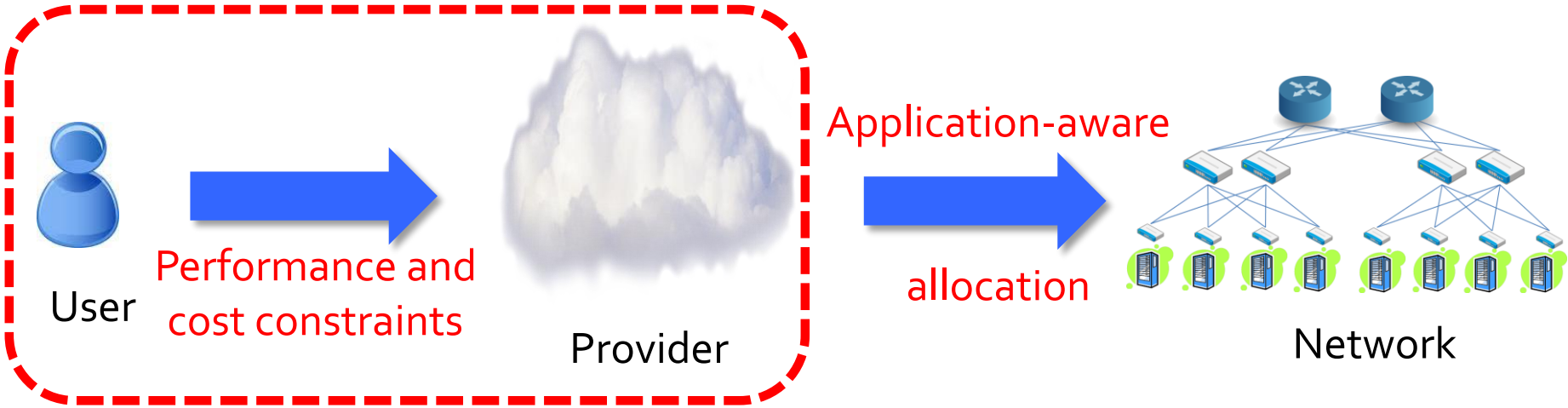
Provider can use this flexibility

{25 VMs with 500 Mbps} vs. {40 VMs with 400 Mbps}
Choose the resource combination to minimize internal impact

Bazaar [SoCC'12]



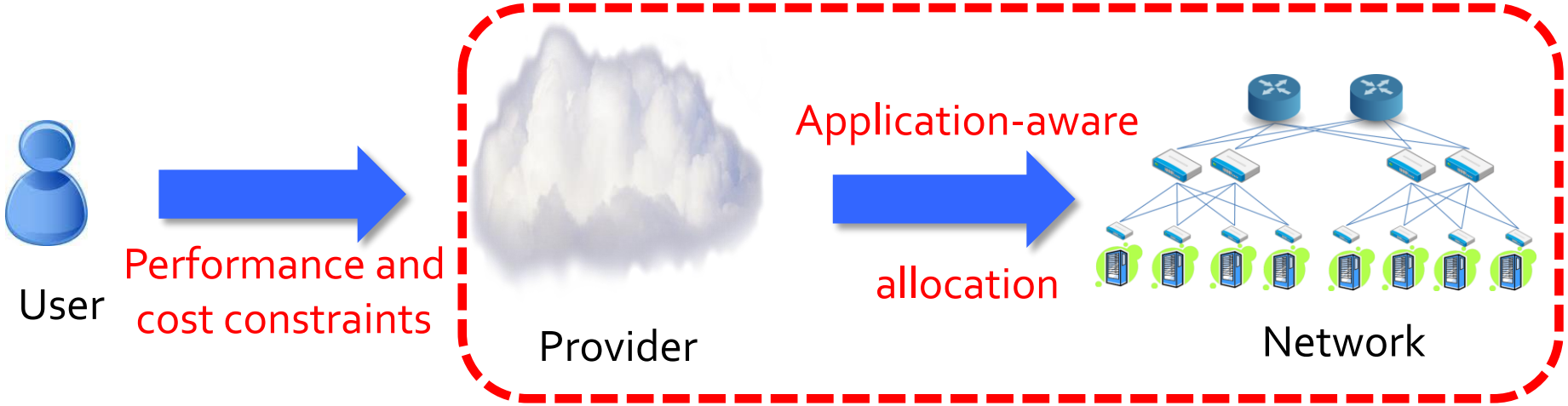
Bazaar [SoCC'12]



Performance prediction

- Translating performance goal is hard
  - For general applications…possibly impossible
  - For popular cloud-based applications…maybe!
- Early evidence
  - Amazon Dynamo DB (e.g., 100K reqs/s)
  - We focus on MapReduce: (simple) analytic model

Bazaar [SoCC'12]

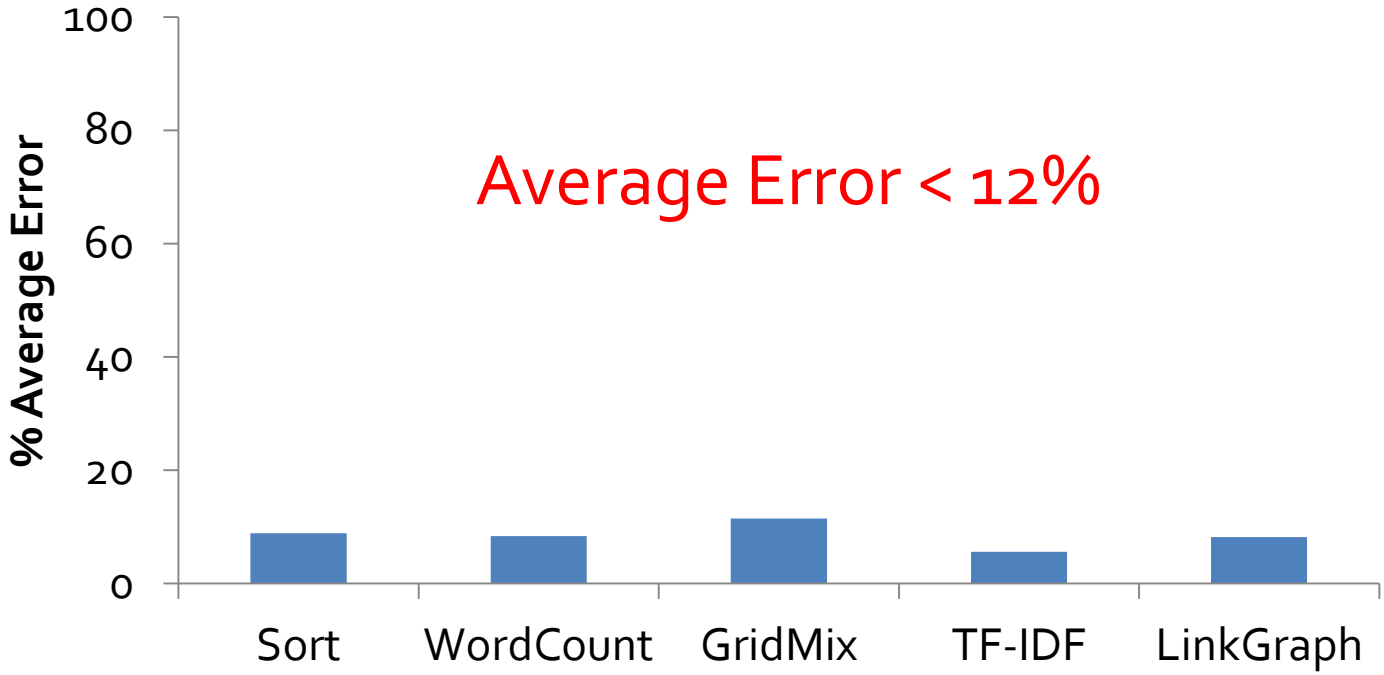


Resource selection

- Multiple resource combination may satisfy user goals
  - E.g., 25 VMs with 500 Mbps or 40 VMs with 400 Mbps
- Similar to Multi-dimensional Bin Packing
- Heuristic: Minimize Resource Imbalance Metric
  - Choose the combination that minimizes resource imbalance

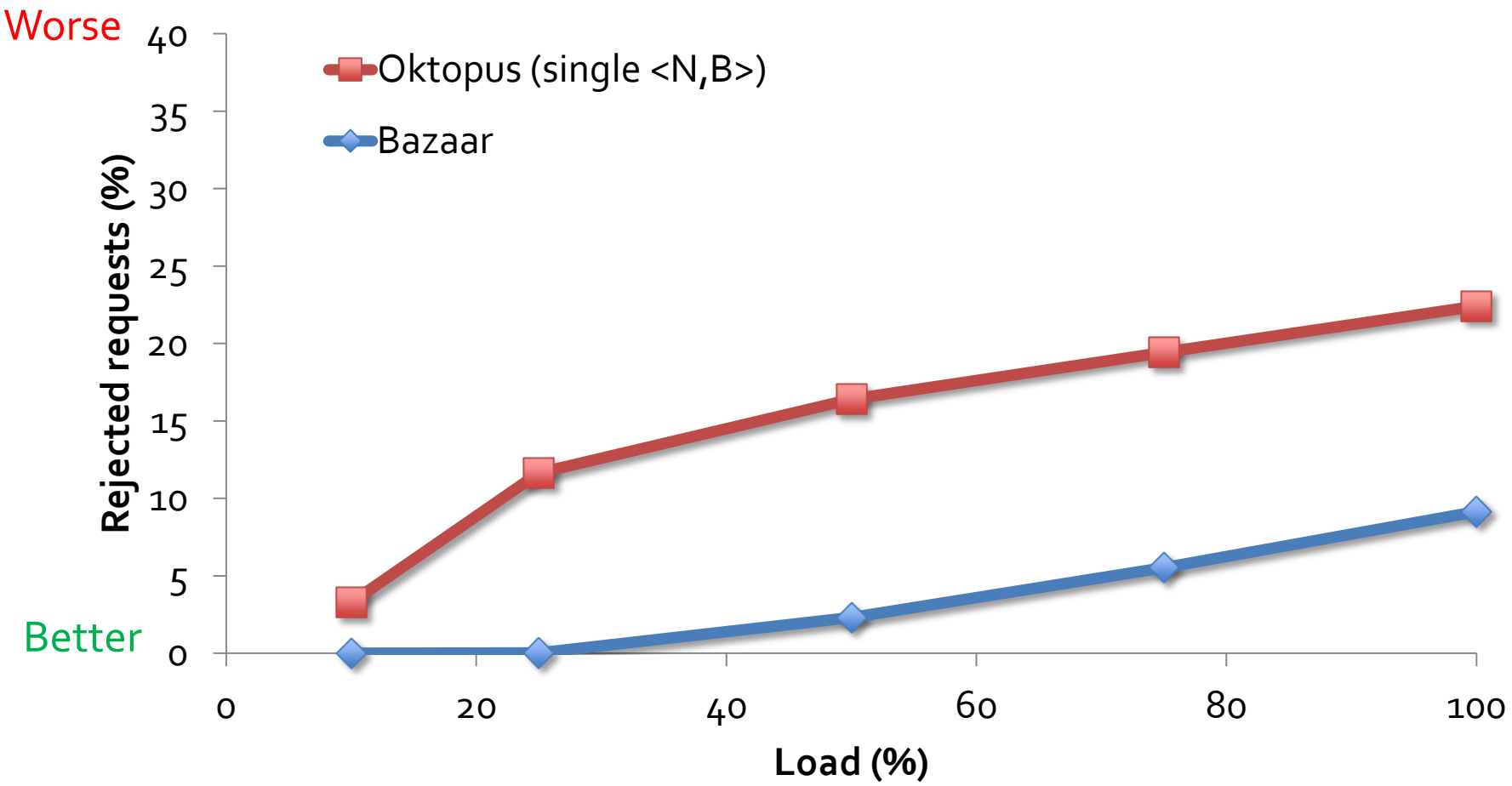
Prediction Accuracy

- Setup: Hadoop on 35-node Emulab cluster
  - 5 MapReduce jobs



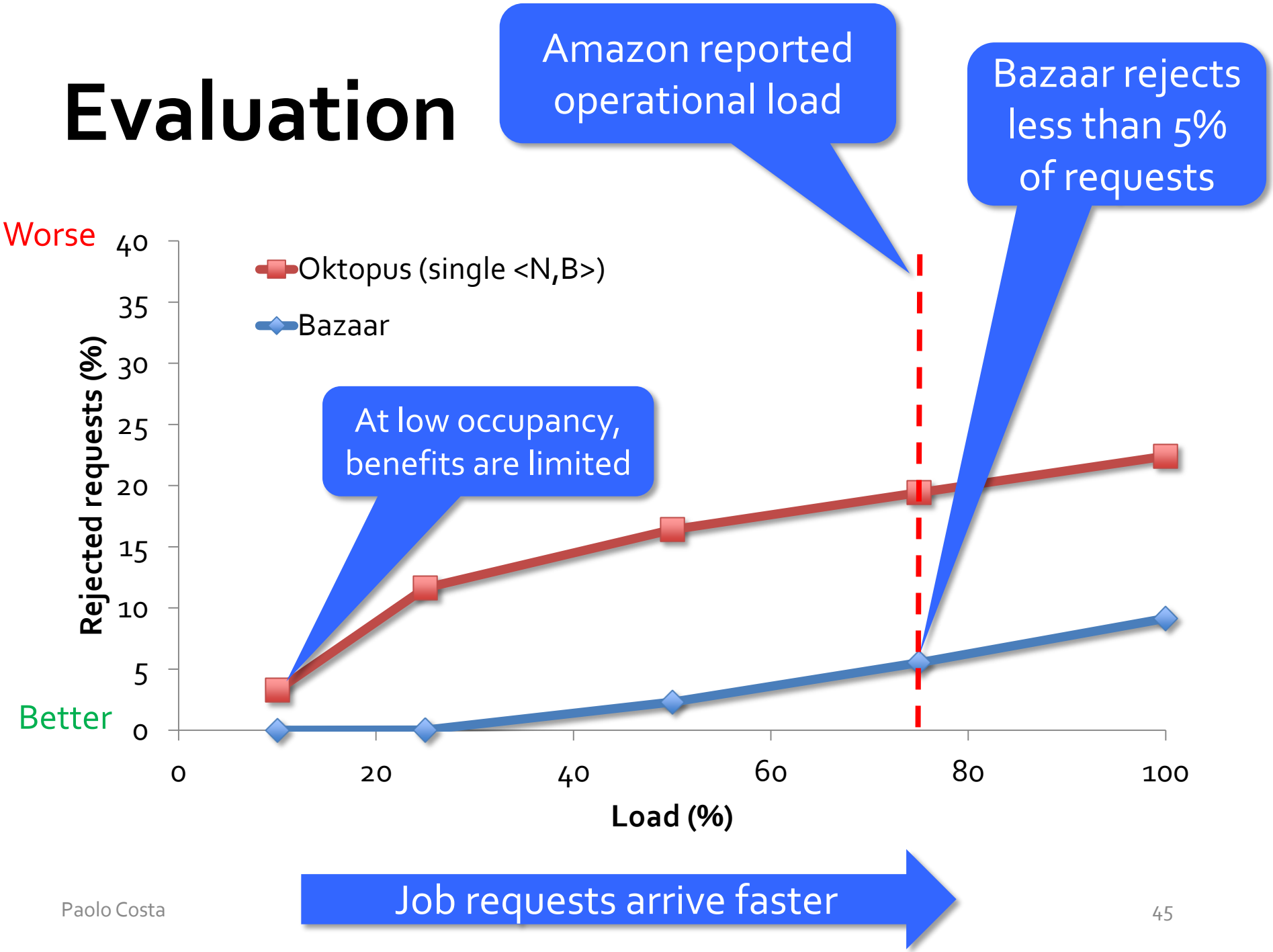
Overcome prediction inaccuracies using slack

Evaluation

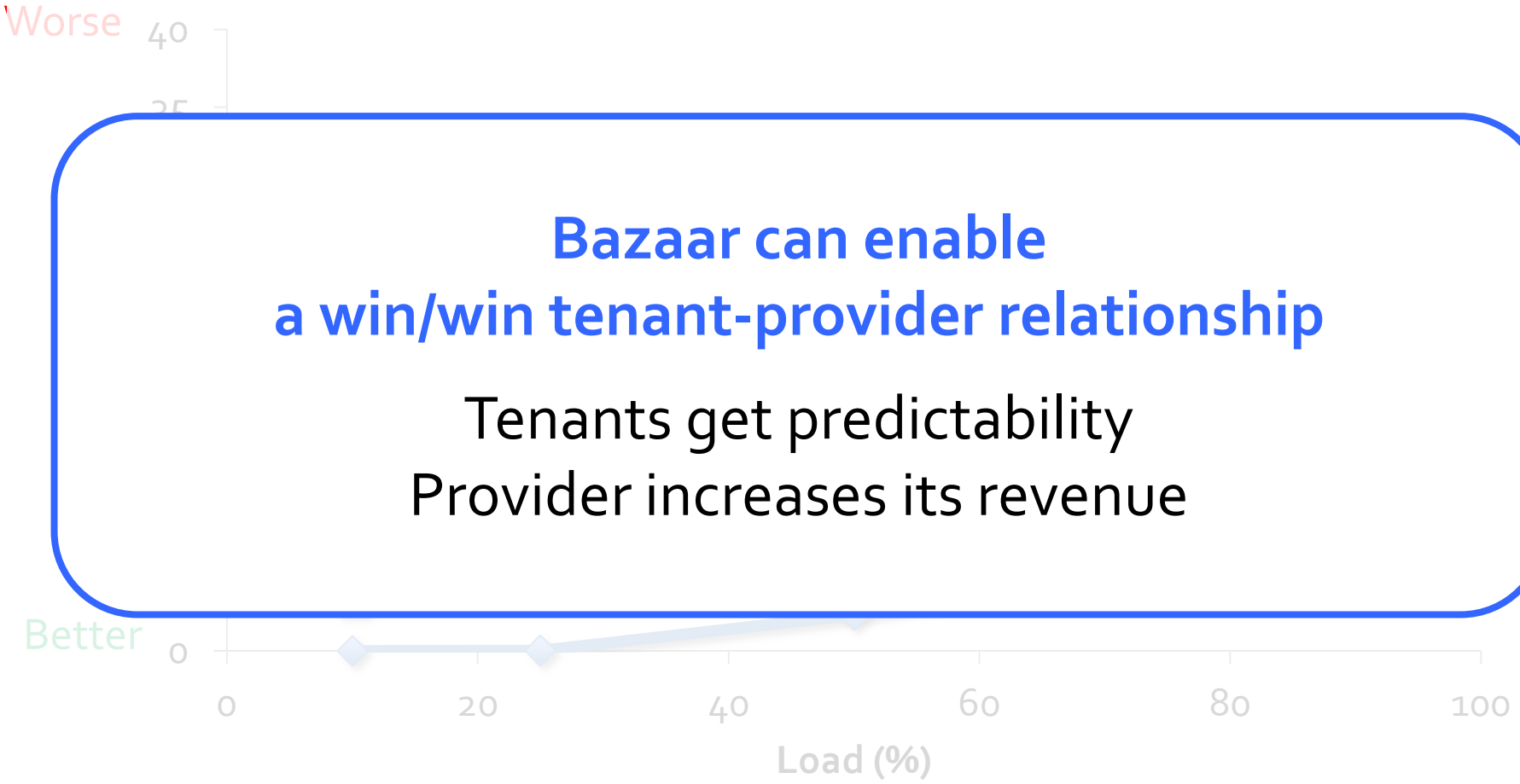


Job requests arrive faster

Evaluation


Amazon reported operational load
Bazaar rejects less than 5% of requests
Worse
Better
Oktopus (single <N,B>)
Bazaar
At low occupancy, benefits are limited
Rejected requests (%)
0
5
10
15
20
25
30
35
40
0
20
40
60
80
100
Load (%)
Job requests arrive faster

Evaluation



Bazaar can enable
a win/win tenant-provider relationship

Tenants get predictability
Provider increases its revenue

Pricing Implications

- Today's resource-based pricing
  - Results in mismatched tenant-provider interests
    - Tenants want cheapest resource tuple <N, B>
    - Provider wants resource tuple with least impact
  - No incentive to reduce completion time!
- Bazaar enables job-based pricing
  - E.g.: Finish Sort over 200GB in 4hrs costs $100
  - Price based on job characteristics and completion time
- Finish jobs "before time"
  - *Intuition*: if spare resources, finish the job earlier so more resources will be available in future

Better alignment of tenant-provider interests

Data Center Network Research

Many proposals:

- **Oversubscription**: Fat-tree[SIGCOMM'08], VL2[SIGCOMM'09], …
- **Path collision**: Hedera[NSDI'10], MPTCP[SIGCOMM'11], SPAIN[NSDI'10], …
- **TCP Incast**: DCTCP [SIGCOMM'10], ICTCP[CoNEXT'10], FDS[OSDI'12],…
- **Traffic prioritization**: Orchestra [SIGCOMM'11], D2TCP[SIGCOMM'11], …
- **Fair sharing**: Seawall [NSDI'11], FairCloud [SIGCOMM'12], …
- **…**

Data Center Network Research

Many proposals:

- **Oversubscription**: Fat-tree[SIGCOMM'08], VL2[SIGCOMM'09], …

The network is a black box
for applications (and vice versa)
Applications need to reverse-engineer the network
The network has to infer application patterns

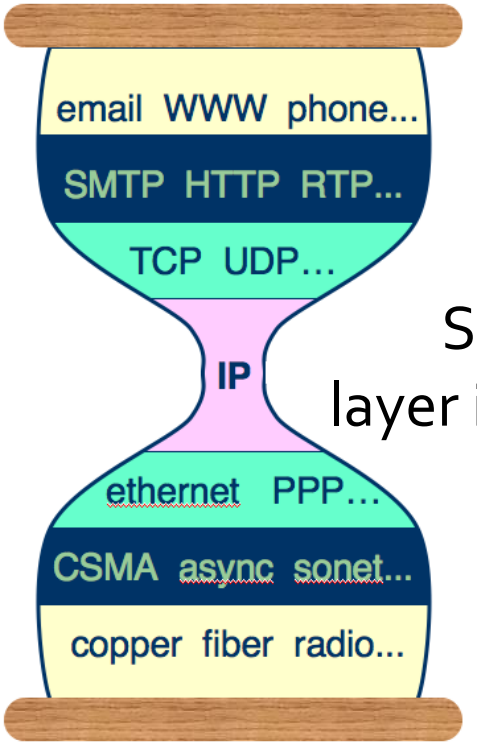
- …

Internet & Data Centers

- This is due to how the Internet was designed…

Internet

- Multiple administration domains
- Heterogeneous HW and network
- Topology not known
- Malicious software



Strict layer isolation

Internet & Data Centers

- This is due to how the Internet was designed…
  - …but data centers are not mini-Internets

Internet

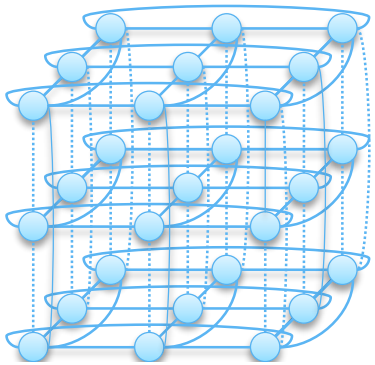
- Multiple administration domains
- Heterogeneous HW and network
- Topology not known
- Malicious software

Data Centers

- Single administration domain
- Homogenous HW and network
  - x86 and Ethernet
- Topology known
  - and can be customised
- Trusted components
  - e.g., using virtualization

Takeaway

- Data centers present both unique challenges and opportunities to network designers
  - Good time to revisit previous assumptions and rethink application and protocol design
- *Bridging the gap*
  - Enable applications to "control" the network
  - Expose application logic to the network



CamCube

[SIGCOMM'10,NSDI'12, HPDC'13]

Predictable Data Centers

[SIGCOMM'11,HotNets'11, SoCC'12]

Takeaway

- Data centers present both unique challenges and opportunities to network designers
  - Good time to revisit previous assumptions and rethink application and protocol design
- *Bridging the gap*
  - Enable applications to "control" the network
  - Expose application logic to the network
- Next steps: Network-as-a-Service (NaaS) [HotIce'12]
  - Fully programmable cloud
  - Remove the distinction between network and computation