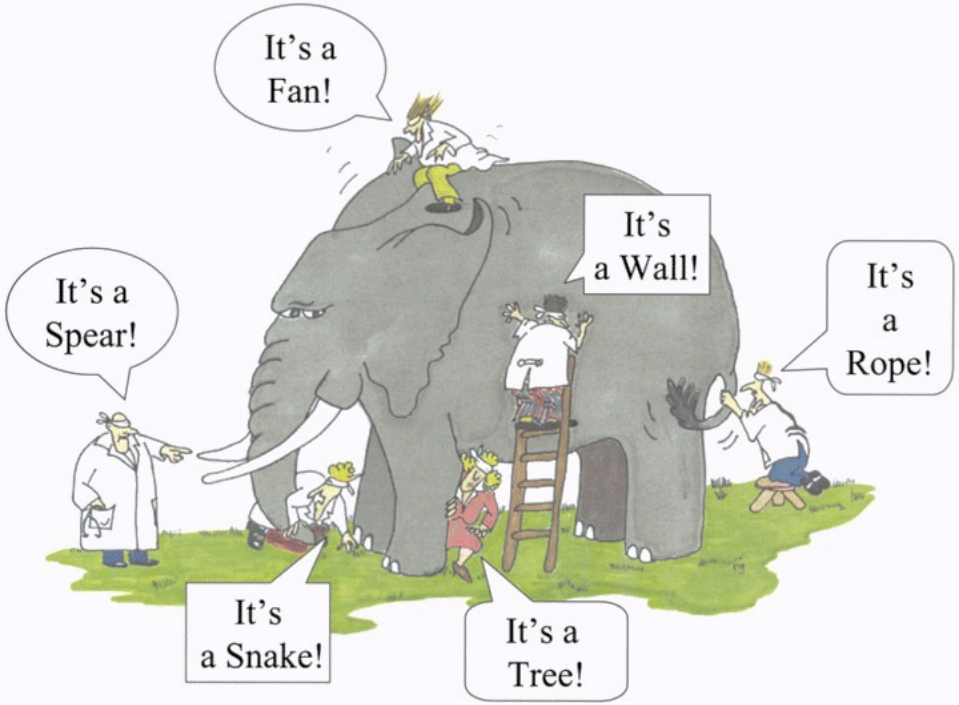
On Blind Mice and the Elephant

Understanding the Network Impact of a Large Distributed System



John S. Otto, Mario A. Sanchez, David R. Choffnes*, Fabián E. Bustamante, Georgos Siganos**

Northwestern, EECS

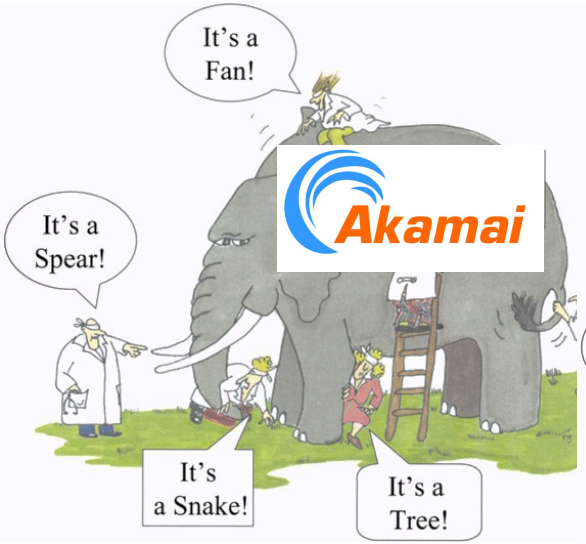
** U. Wash, CSE*

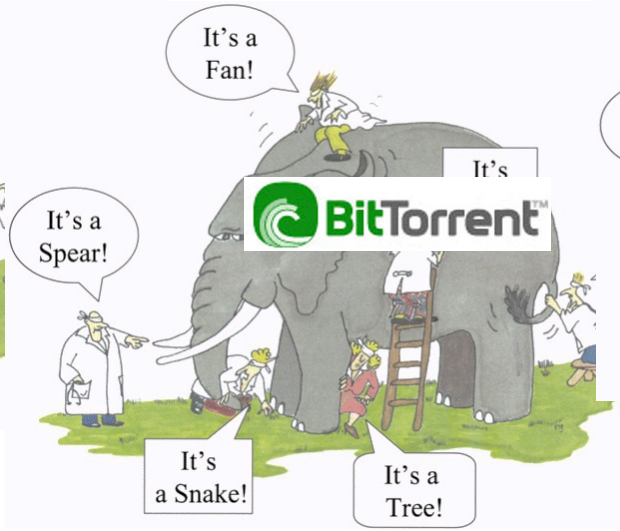
*** Telefónica Research*

AquaLab

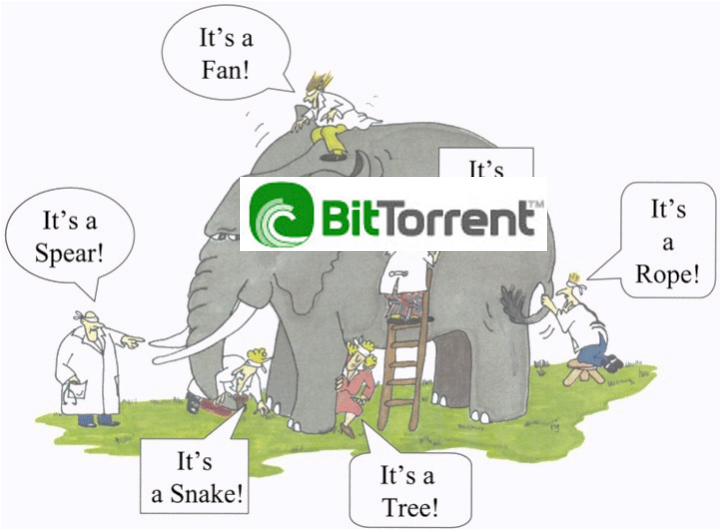
Telefónica

Several elephants of the Internet





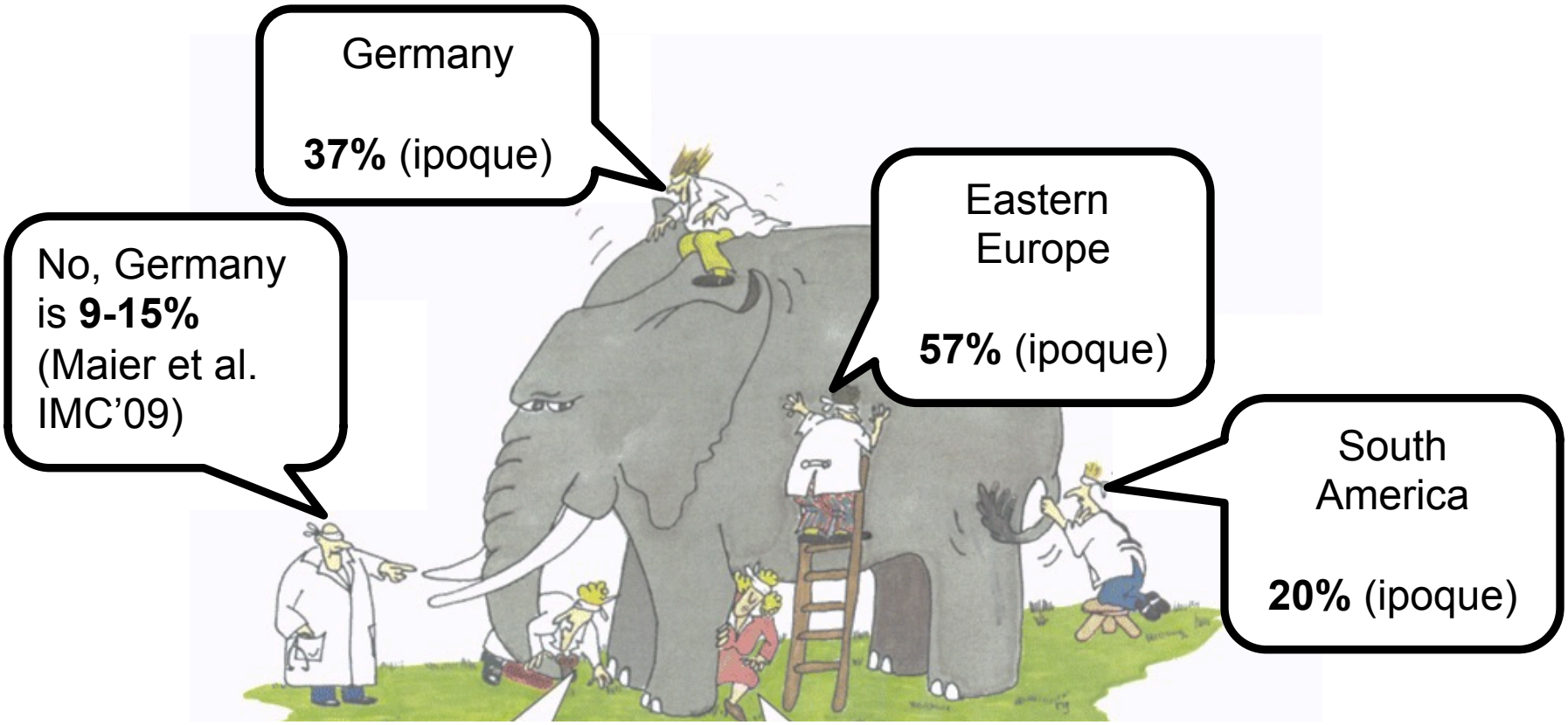
Several elephants of the Internet



- A large, global peer-to-peer system
- Millions of users exchanging content
- Virtually every country in the world

Perspectives on distributed system measurement

- System's *measured* network impact depends on measurement vantage point
  - How much of network traffic is from BitTorrent?



Approach to evaluating system impact

- A view from a broad set of end users
  - To sample its overall network traffic
  - Understand where it flows
  - Who pays for it (and how expensive it is)

- This work
  - Relies on end users as vantage points
    - Captures a sample of *all* BitTorrent traffic
    - Reveals traffic's path through the network
      - Public view is not sufficient to map most BitTorrent traffic
  - ISP data provides *context* to understand cost of BitTorrent traffic

Our diverse end-user perspective

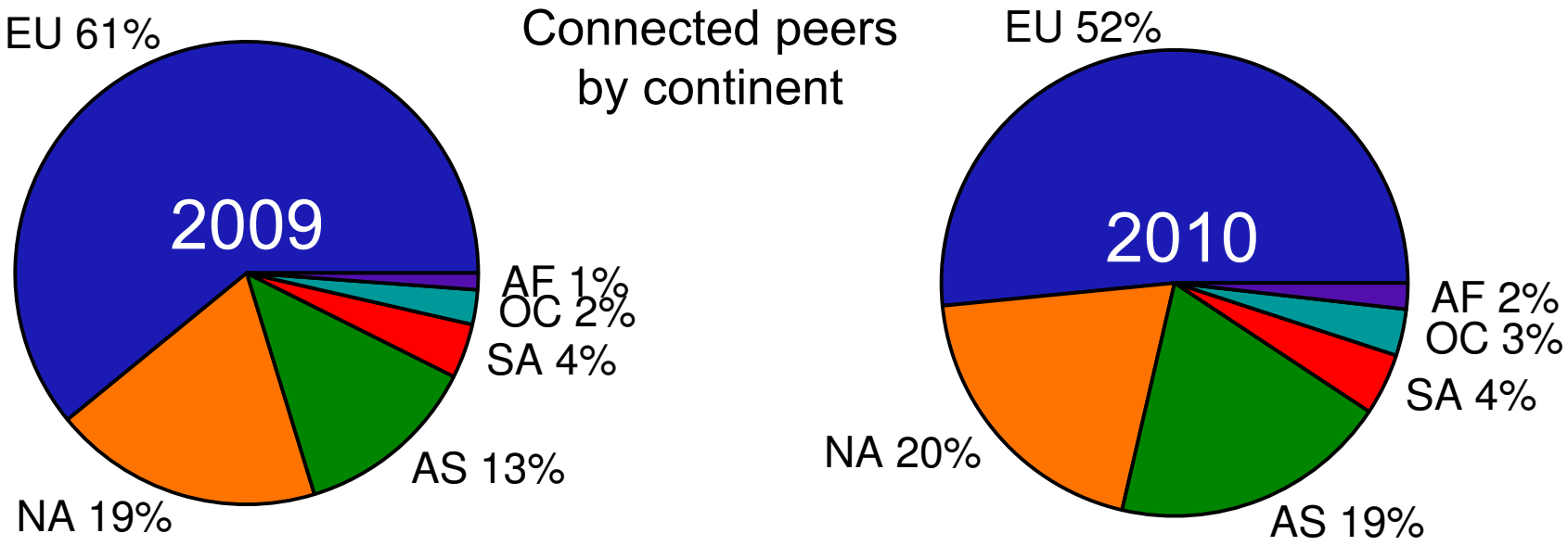
- Representative sample of users
  - 500,000 users, 3,300 networks, 169 countries
- Running extensions (Ono & NEWS) for Vuze BitTorrent client
  - Anonymously report statistics
  - Provide application-level data
    - e.g. session length, per-connection transfer volumes
    - Log 13 TB of traffic per day
  - Conduct active measurements to reveal traffic paths
    - With public view alone, we can map 25% of traffic
    - Supplemented with traceroutes, we can map 89%

Roadmap

- How BitTorrent is being used
  - Who is using BitTorrent?
  - When do people run BitTorrent?
  - How much traffic does it generate?
  - Study data from Nov. 2008 to Nov. 2010
- Where the generated traffic flows
- Who pays for it and how much

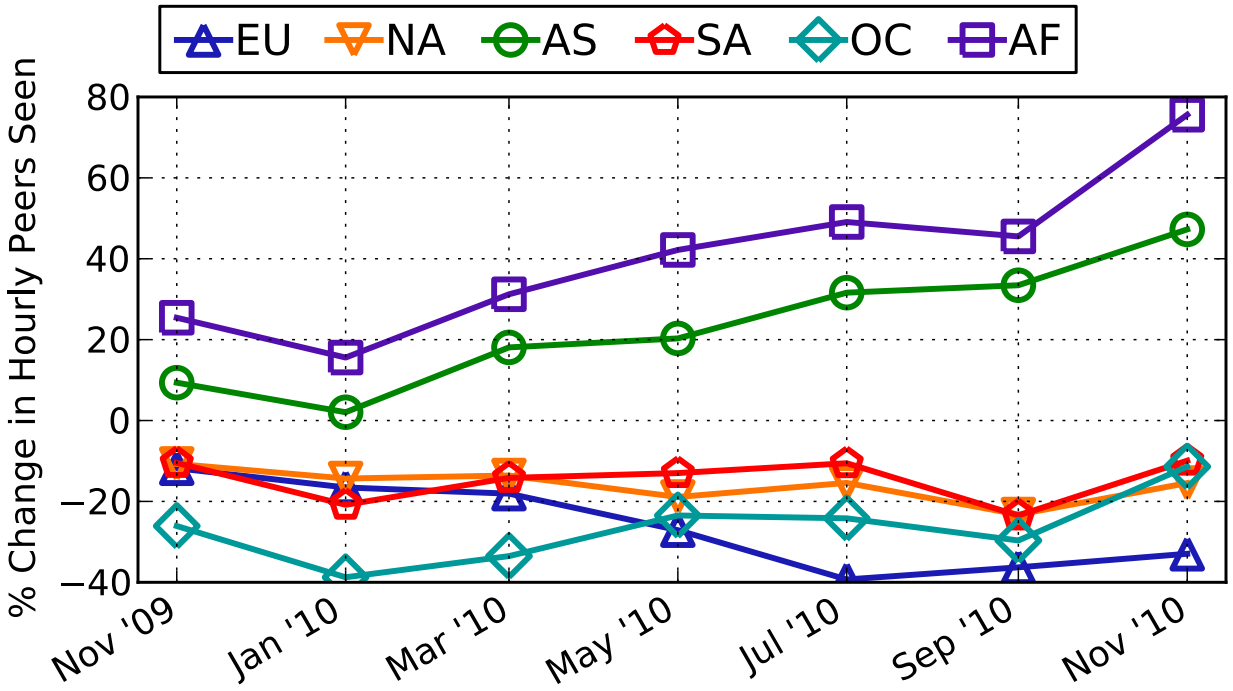
BitTorrent trends: user population

- Overall population reduced by 10%
- Locations of users change over time



- Decrease in Europe
- Increase in Asia, Africa and Oceania

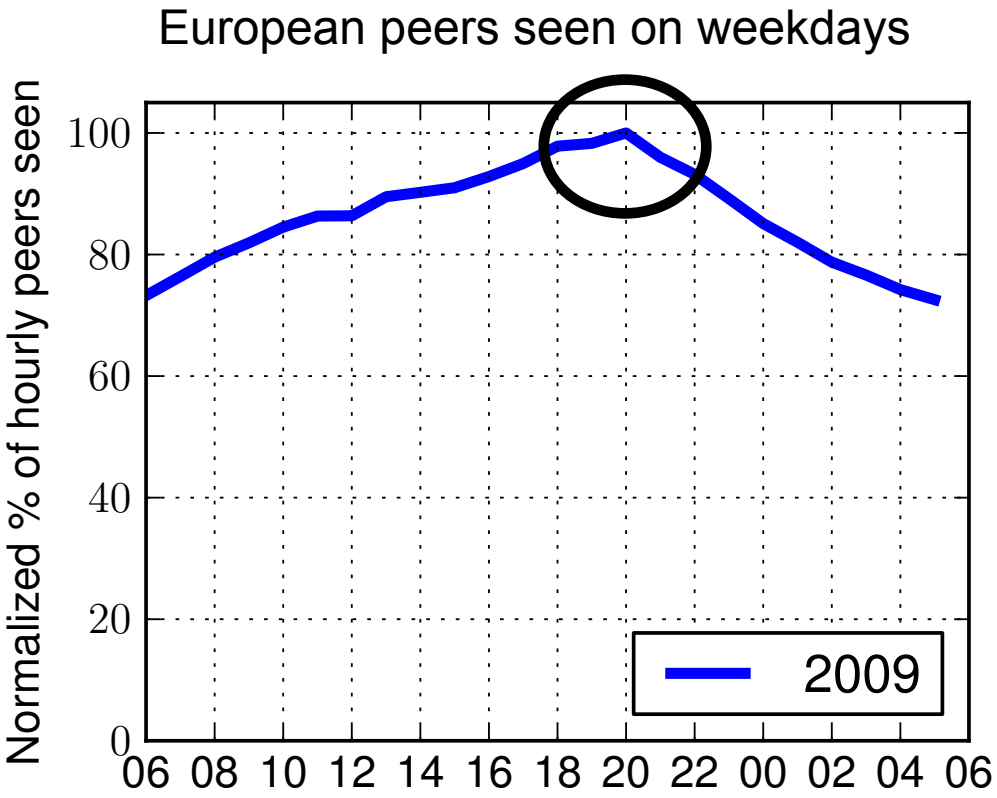
BitTorrent trends: user population



Rate of growth of connected users per continent relative to Nov. 2008

- Europe continues to drop
- N. America, S. America remain stable since 2009
- 76% growth in Africa and 47% in Asia

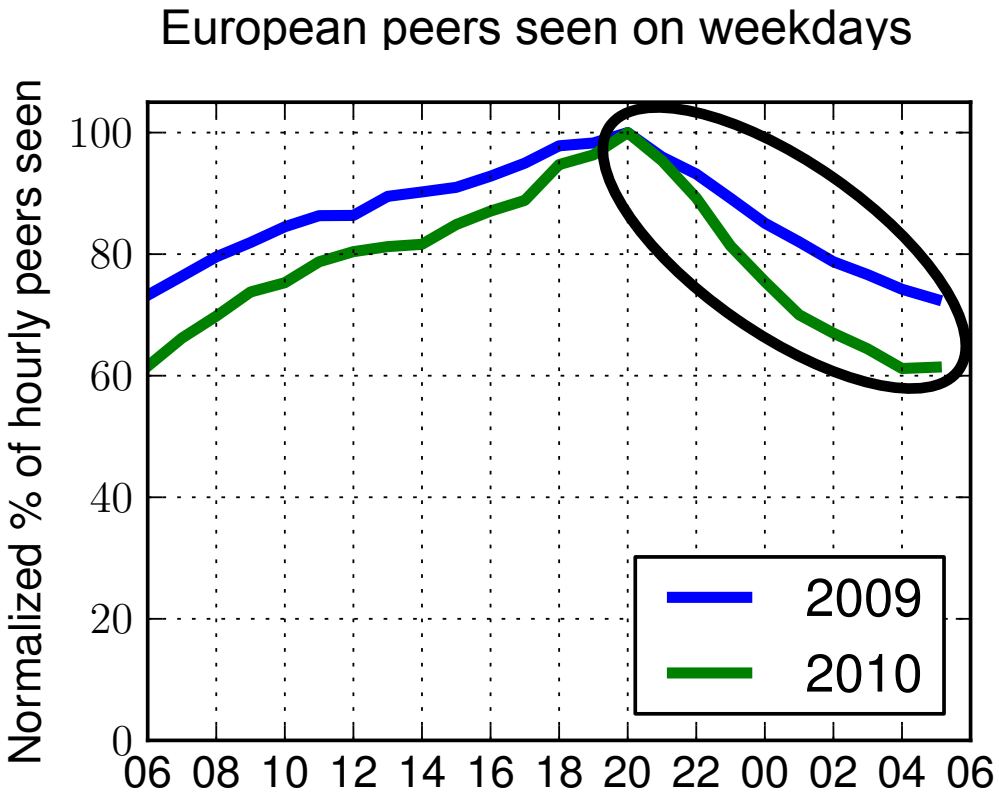
BitTorrent trends: stronger diurnal patterns



Normalized number of peers seen per hour in Europe, depending on time of day

- Shift away from overnight use
- Peak usage aligns with evening hours, local time
  - Potential impact on ISPs' costs under burstable billing

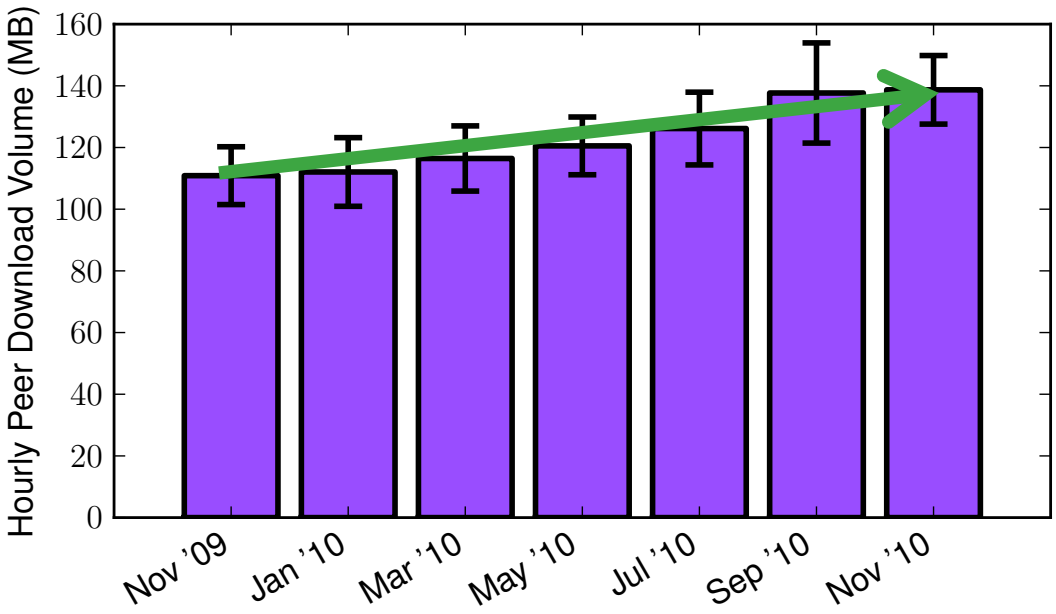
BitTorrent trends: stronger diurnal patterns



Normalized number of peers seen per hour in Europe, depending on time of day

- Shift away from overnight use
- Peak usage aligns with evening hours, local time
  - Potential impact on ISPs' costs under burstable billing

BitTorrent trends: increased traffic volumes



Per-peer hourly download volume (in MB) over the last year

- 25% increase in per-peer hourly download volume
- Despite a 20% drop in total connections, a **12%** increase in overall system traffic

Trends in BitTorrent usage and traffic

- Overall population reduced by 10%
  - But large increase in Africa and Asia
- Peak usage aligns with evening hours
- 12% increase in overall system traffic
  - 25% increase in per-peer hourly download volume

So where's the traffic?

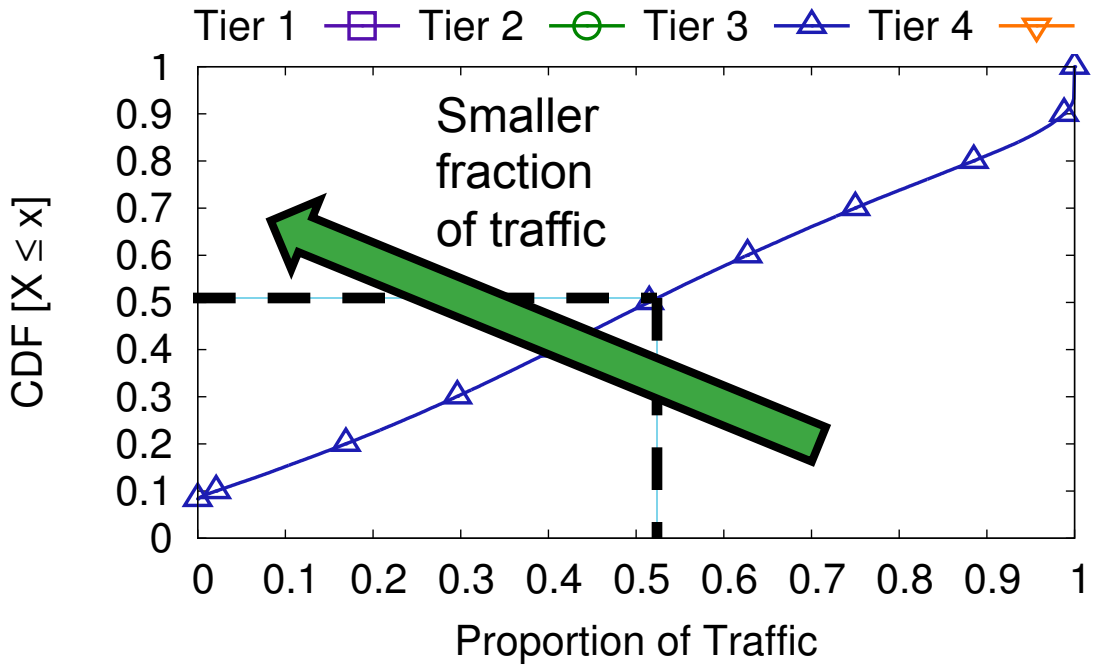
Where BitTorrent traffic flows

- How “deep” does traffic go in the network?
- Who is paying for it?

- Traffic path analysis to see which networks carry most BitTorrent traffic
  - Tier 1: Well-known networks
  - Tier 2: Large transit providers
  - Tier 3: Small transit providers
  - Tier 4: Content/access/hosting providers
    Enterprise customers

Tiers based on Dhamdhere and Dovrolis, IMC 2008

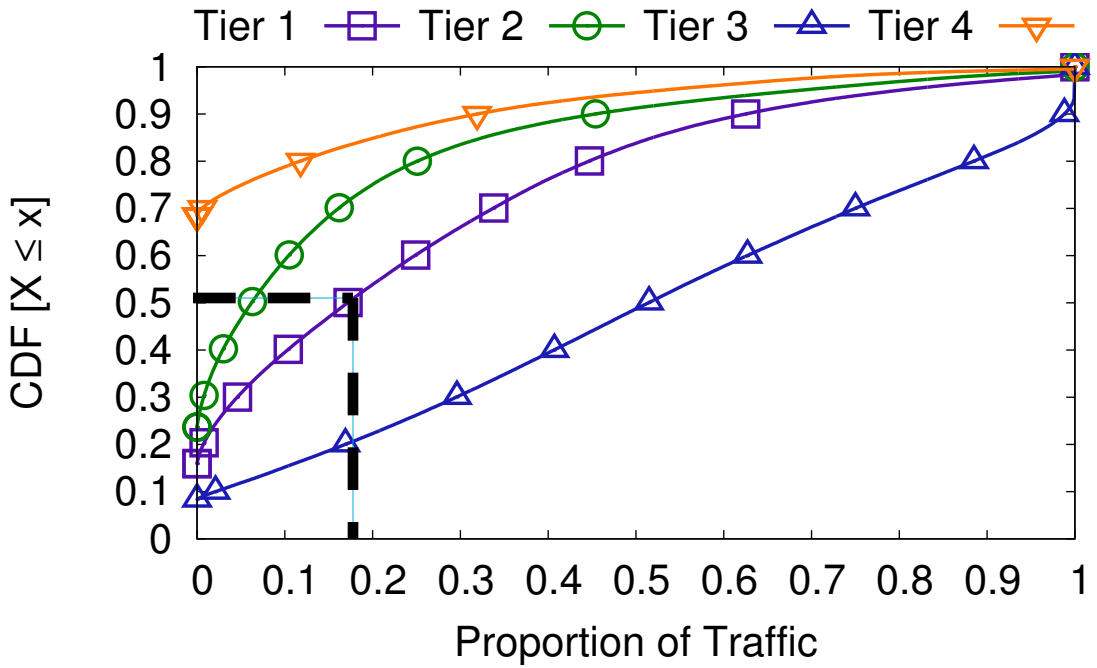
Traffic's "depth" in the network



Fraction of each peer's traffic that reaches Tier X

- Most traffic stays at or below Tier 3
- Significant fraction of traffic never reaches Tiers 1 or 2
  - Typically missed by in-network monitoring studies from the core

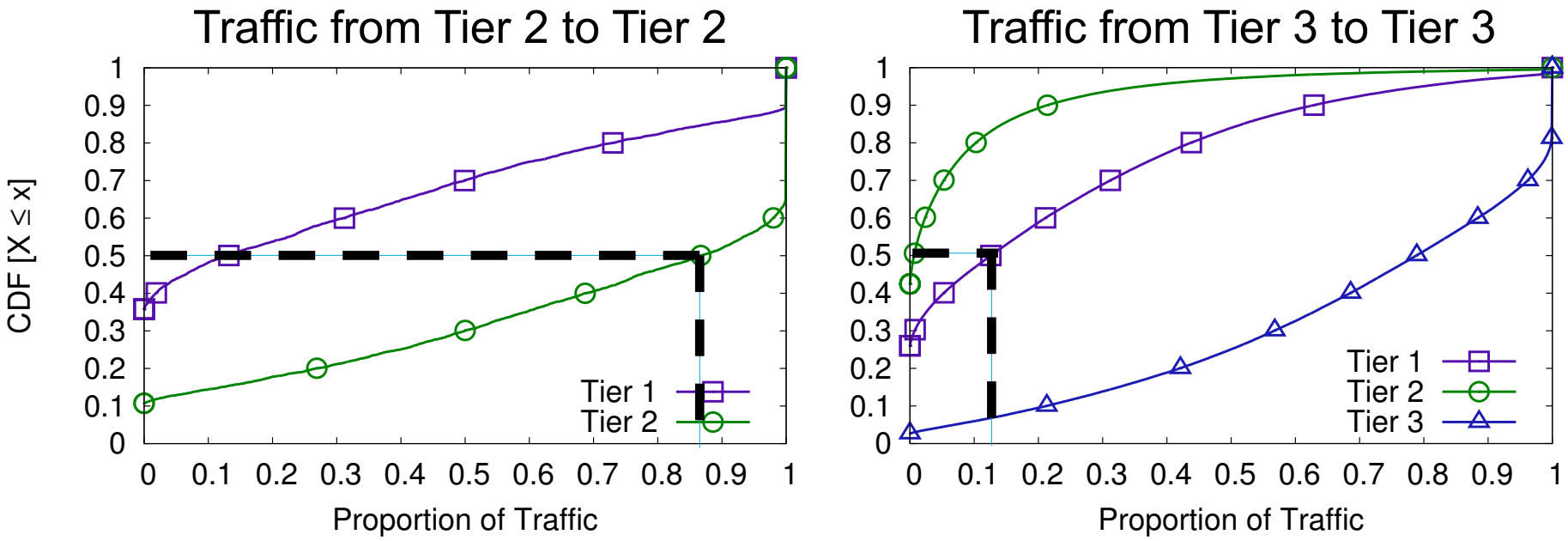
Traffic's "depth" in the network



Fraction of each peer's traffic that reaches Tier X

- Most traffic stays at or below Tier 3
- Significant fraction of traffic never reaches Tiers 1 or 2
  - Typically missed by in-network monitoring studies from the core

Endpoints' tiers determine "depth"



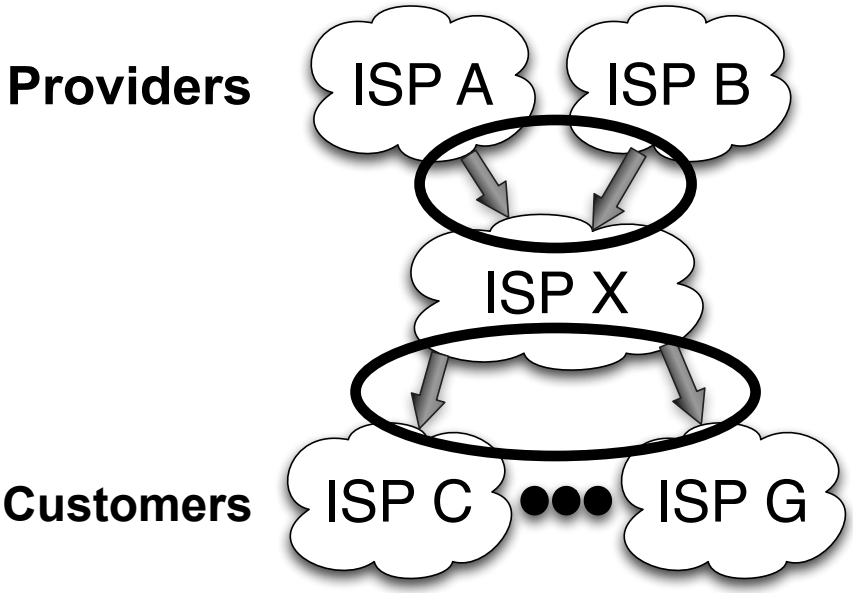
- Traffic generally stays in the originating tier
- Tier 2 networks do not provide "intermediate" level of connectivity between Tiers 1 & 3

Roadmap

- How BitTorrent is being used
- Where the generated traffic flows
  - Most traffic is handled at or below Tier 3
- Who pays for it and how much

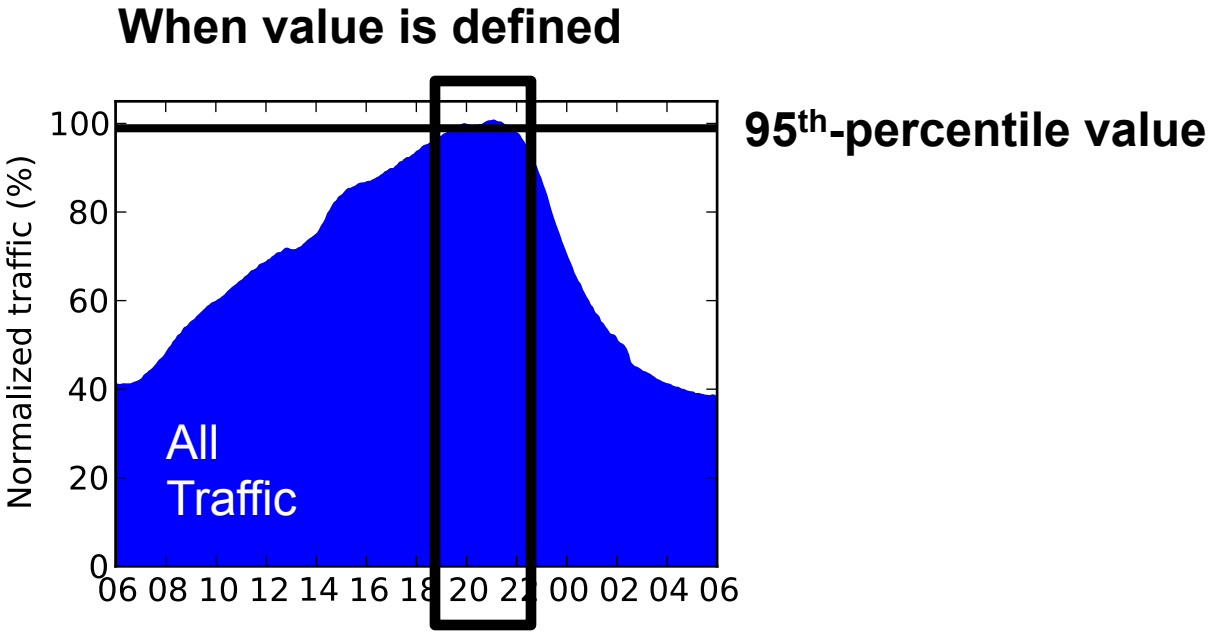
Economic implications for ISPs

- Determine BitTorrent cost relative to other traffic
  - ISP X's data provides context to interpret traffic sample
- Study at granularity of individual network links
- Consider common burstable billing model
  - e.g. 95th-percentile billing
- Data for several
  - of ISP X's links
  - over 1 week

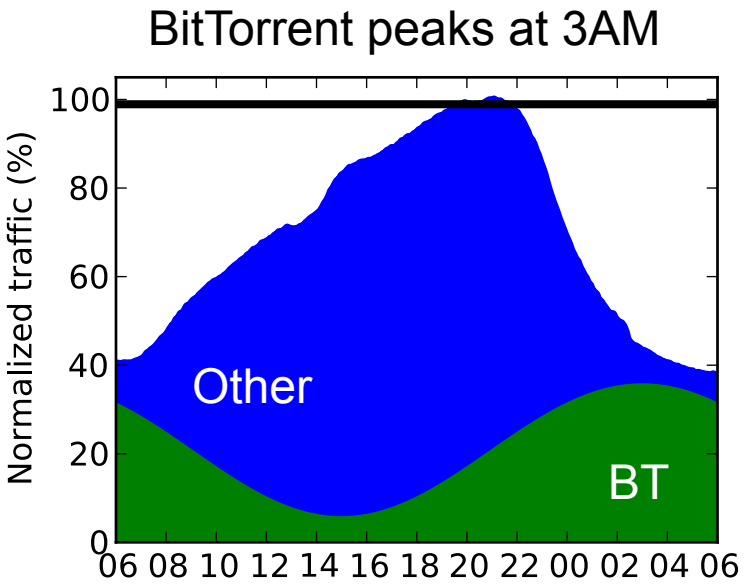


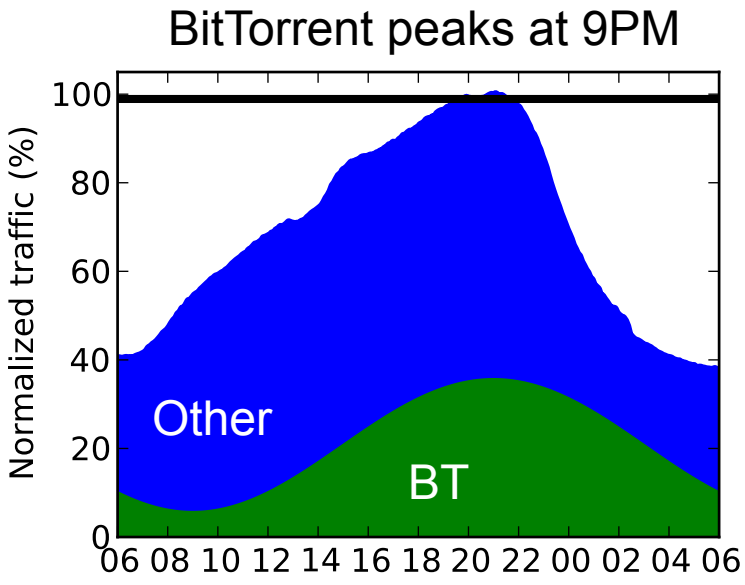
95th-percentile billing

- Aggregate link volume for each 5 minute bin
- Cost is based on 95th-percentile bin's value
- *Under burstable billing model, not all bytes may have the same cost*
  - Peak-hour bytes are more expensive than off-peak



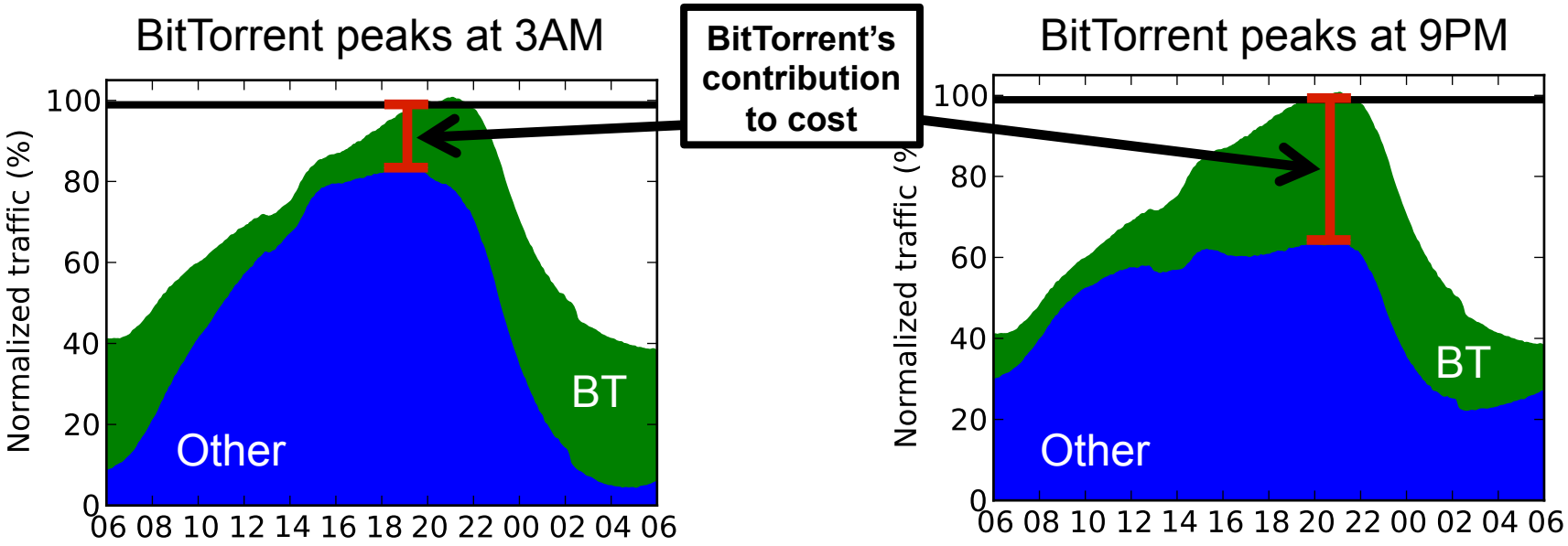
95th-percentile and Shapley value





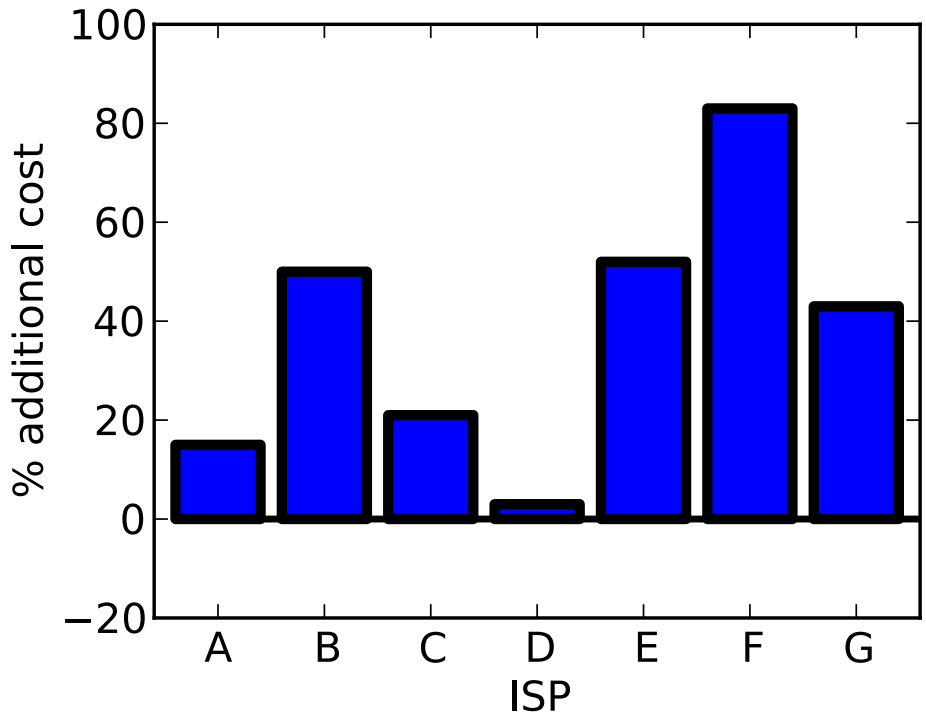
- BitTorrent at peak hour is more expensive
- Use Shapley value to determine *relative cost* of BitTorrent
  - Shapley value gives the cost contribution of BitTorrent traffic
  - Compare to other traffic on the network
  - Is BitTorrent's cost more than its "fair share" by volume?

95th-percentile and Shapley value

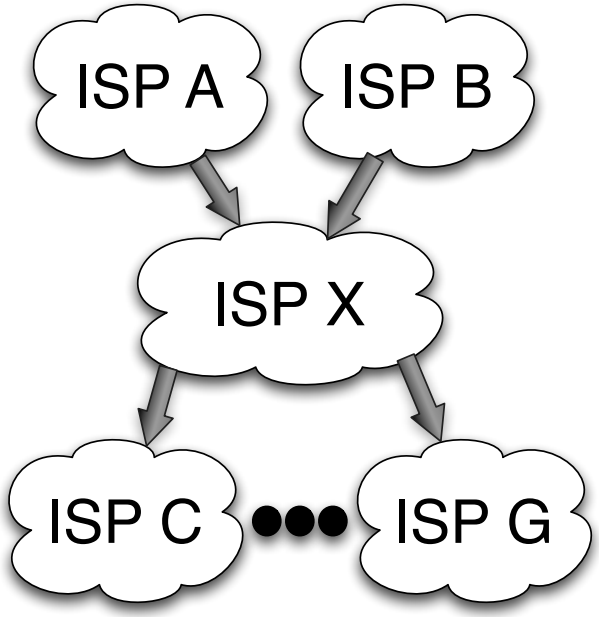


- BitTorrent at peak hour is more expensive
- Use Shapley value to determine *relative cost* of BitTorrent
  - Shapley value gives the cost contribution of BitTorrent traffic
  - Compare to other traffic on the network
  - Is BitTorrent's cost more than its "fair share" by volume?

Relative cost of BitTorrent traffic

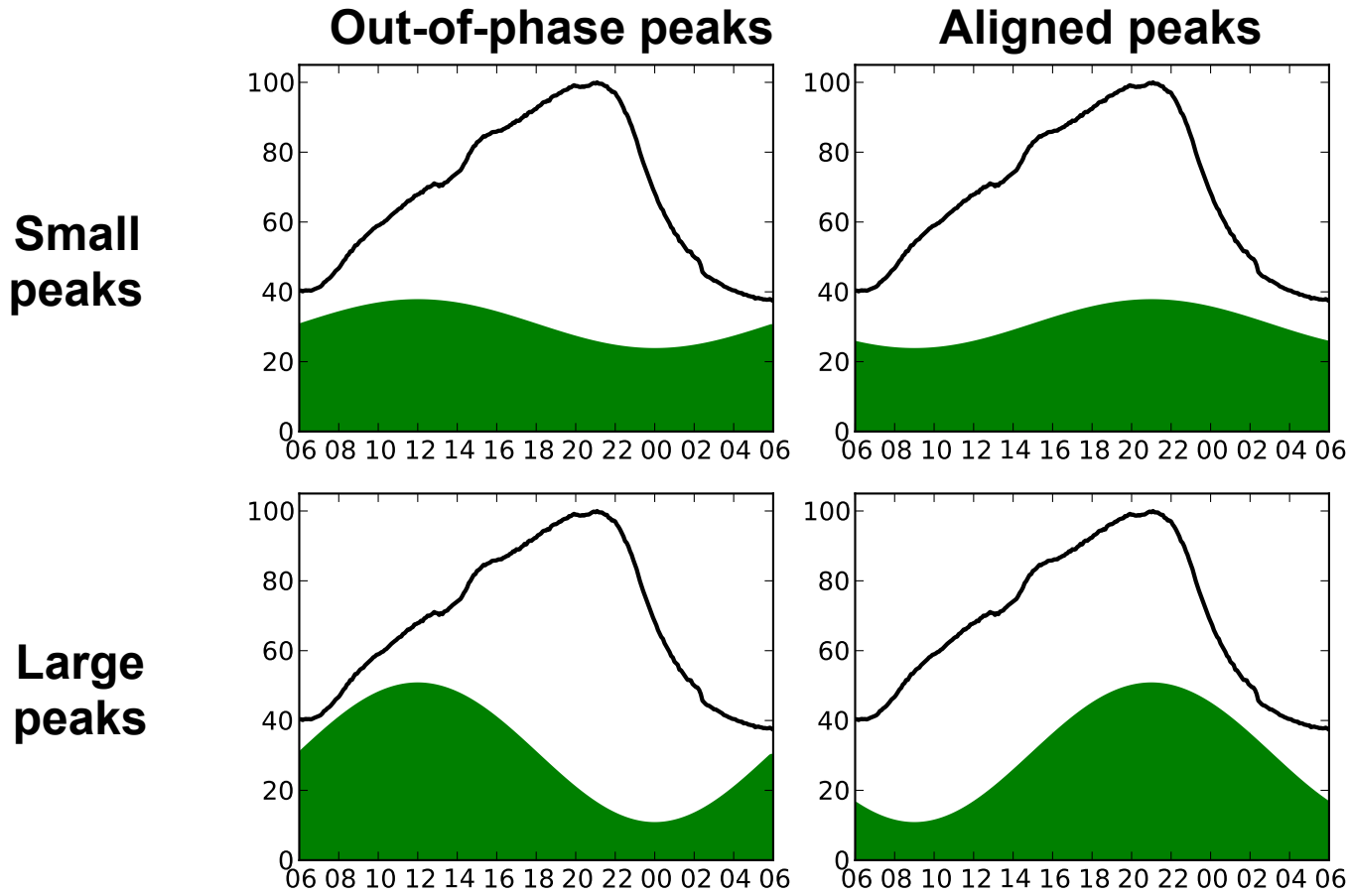


Additional cost of BitTorrent traffic, percent above relative cost of 1



- BitTorrent traffic is generally *more* expensive than other traffic
- What traffic characteristics result in high relative cost?

Traffic characteristics and relative cost



- High relative cost of BitTorrent
  - Large coefficient of variation (“C.V.”, size of peaks in BitTorrent traffic)
  - Small cross-correlation offset (“X-corr”, alignment with overall traffic)

Traffic characteristics and relative cost

	Out-of-phase peaks	Aligned peaks
Small peaks	**ISP A** X-corr: -7.1 hours C.V.: 130% Relative cost: 13%	**ISP E** **X-corr: 1.6** C.V.: 158% **Relative cost: 83%**
Large peaks	**ISP F** X-corr: 7.4 **C.V.: 325%** **Relative cost: 52%**	**ISP B** **X-corr: 3.2 hours** **C.V.: 188%** **Relative cost: 50%**

- High relative cost of BitTorrent
  - Large coefficient of variation (“C.V.”, size of peaks in BitTorrent traffic)
  - Small cross-correlation offset (“X-corr”, alignment with overall traffic)

Conclusions

- BitTorrent is still alive and costly
  - Most traffic stays at the edge of the network
  - It is moving into prime-time
  - Logically, it is relatively *more expensive*
- A broad view from the edge of the network is required to see the system's full usage spectrum
- Approach is general to understanding other distributed systems
  - Video streaming
  - Peer-to-peer CDNs