INTERNET AND THE ECONOMICS OF INTERCONNECTION

Based on: «*Interconnection in the Internet: peering, interoperability and content delivery*» by David D. Clark, William H. Lehr and Steven Bauer

PHISICAL AND ECONOMIC INTERCONNECTION IN THE INTERNET

- The Internet is a network of networks that realizes its global reach by being able to route data from source nodes on one network to destination nodes that, in many cases, are on networks that are owned and operated by different Internet service providers (ISPs)
- Along the end-to-end path, the data may need to cross the networks of still other ISPs
- Supporting the end-to-end, global connectivity, requires that the ISPs be interconnected both physically (i.e., there exists an electronic pathway for transporting packets) and via business relationships
- These business relationships impact both the flow of traffic and the flow of money across the Internet value chain.

NETWORKS AND INTERCONNECTION: A GENERAL VIEW (1)

- Networks are a key component of much of our basic infrastructure, including the networks of roads, water, electricity, and telecommunications that are used ubiquitously in our daily live
- These networks connect:
  - supply (source) and demand (destination) nodes;
  - via a network of interconnected links (transport paths) and switching nodes
- The links that connect the source and destination nodes may be referred to as the access connections

NETWORKS AND INTERCONNECTION: A GENERAL VIEW (2)

- To reduce total costs, it is usually desirable for access nodes to be connected via a hierarchy of network switching nodes, each of which is connected to multiple access or network nodes
- Telecommunications networks are comprised of local switches – or in packet networks, routers – that collect the traffic from access nodes and route it to tandem switches, which are in turn connected to other local or tandem switches in a hierarchy that allows traffic to be routed across town, across the state, or across the globe
- Common features of such networks are that they allow a large number of source/destination nodes to share network resources economically to take advantage:
  - of the fact that demands are not perfectly correlated in time and
  - of the scale economies that are common with network technologies
- Also, larger networks that provide more options for routing traffic between a larger number of source and destination nodes are typically more valuable to each subscriber, or in economic terms, exhibit positive network externalities

NETWORKS AND INTERCONNECTION: A GENERAL VIEW (3)

- The challenges for network sustainability:
  - *Technical issues*. The network needs sufficient capacity to meet the traffic demands of the source/destination nodes (otherwise the network can crash). These change over time and across nodes and applications, thereby complicating the network capacity provisioning challenge
  - *Economic issues*. The value of the network services has to exceed the costs of sustaining the network. Sufficient funds need to flow to network resource owners to allow them to recover the network investment and operating costs. Three problems:
    - The cost recovery challenge is further complicated because much of the investment is long-lived and must be put in place in advance of the demand that it is intended to serve
    - In some cases we are in front of big networks made of smaller networks. In this case the issue of interconnection emerges
    - The networks are often viewed as basic infrastructure by society. As a consequence there is a government interest in ensuring affordable and universal access through regulation

UNDERSTANDING INTERCONNECTION IN THE INTERNET (1)

- In telecommunications networks, there is a long history of interconnection regulations that established both the terms and prices for interconnection in ways that often embedded significant implicit and explicit subsidies
- In contrast to legacy telephone networks, interconnection in the Internet has been unregulated. This is largely a consequence of the Internet's evolution as an application that ran on top of the regulated public switched telephone network (PSTN). From an economic perspective, this means that the decision of whether to interconnect, and if so, how to interconnect, was left to the ISPs

UNDERSTANDING INTERCONNECTION IN THE INTERNET (2)

- From a technical perspective, the Internet is an end-to-end (e2e) packet delivery network that routes packets from source to destination Internet Protocol (IP) addresses.
- This often requires packets to traverse multiple networks. These networks, identified as autonomous systems (AS), control a range of IP addresses that they manage and that they are responsible for routing to. Most AS are associated with ISPs
- To route traffic between IP addresses on different AS or ISPs, the ISPs must be interconnected
- These interconnections define both physical/technical and business relationships

UNDERSTANDING INTERCONNECTION IN THE INTERNET (3)

- There are two basic ways in which ISPs may interconnect: at Internet exchanges where multiple ISPs interconnect, and directly via negotiated bilateral agreements
  - In the early days of the Internet, much of the traffic was exchanged at public interexchange points such as MAE-East and MAE-West where a large number of networks physically interconnected to exchange traffic without compensation
  - These early exchanges had relatively open interconnection policies and provided an inexpensive way for smaller ISPs to expand connectivity
  - However, they lacked mechanisms to provide incentives to expand capacity to deliver traffic from the public interexchange points. As a consequence, these public exchanges suffered from significant congestion
  - The bulk of Internet traffic subsequently shifted to bilateral interconnection agreements

HOW INTERCONNECTION IS REGULATED: TRAFFIC AND PEERING AGREEMENTS (1)

- In the early days of the commercialization of the Internet, two sorts of bilateral interconnection agreements among ISPs emerged: *transit* and *peering*
  - *Transit* is a traditional customer–supplier arrangement, in which one ISP purchased transit service from another, perhaps larger, ISP. An ISP that offers traditional transit service agrees to provide access to the entire Internet for its customers.
  - Peering is an arrangement in which two ISPs that each had traffic for the other agreed to interconnect to exchange that traffic directly. A peering arrangement does not give either ISP access to the entire Internet via the other. In other words, a peering agreement implies a routing restriction with respect to the traffic exchanged; the only traffic exchanged originates from the source ISP and its customers and terminates in the destination ISP and its customers
- *Transit* is an agreement between a *buyer* and a *seller*. Payments flow from the buyer to the seller as compensation for services rendered
- *Peering* shall, in principle, be an interconnection among approximate equals, with value to both parties and no a priori obvious direction for monetary payments to flow. In early negotiations among potential peering partners, it became clear that it would be very difficult to determine if the balance of values favored one or the other ISP, and the convention emerged that peering was ‘settlement free’ or ‘revenue neutral’

HOW INTERCONNECTION IS REGULATED: TRAFFIC AND PEERING AGREEMENTS (2)

- These two types of contractual arrangements should be based on a hierarchical identification of ISPs:
  - transit, a vertical contract in which smaller ISPs serving end users purchased services from larger backbone ISPs
  - peering, a horizontal contract between core or backbone ISPs to exchange traffic under a bill-and-keep or revenue-neutral arrangement
- For a number of years, these two options represented an informal norm or bargaining regime
- But now they are being replaced with more unconstrained bilateral negotiation among the parties. There are a number of reasons why this seems to be happening:
  - In the past, many ISPs were more or less similar. When similar ISPs established peering interconnections, the similarity made the assumption of balanced value plausible. Today, ISPs are more specialized, with some serving broadband residential customers (sometimes called access or 'eyeball' networks), some serving enterprise customers, some serving high-volume content providers and so on.
  - Networks that serve different sorts of customers may have very different internal cost structures
  - Furthermore, with the continued evolution of the Internet and the rise of multi-homing for both end users and ISPs which expands the range of routing options, it is increasingly feasible to implement more granular traffic treatments as part of interconnection agreements

HOW INTERCONNECTION IS REGULATED: PAID PEERING AND PARTIAL TRANSIT AGREEMENTS

- In these new circumstances, it is possible that one party or the other regards the legacy approach of revenue-neutral peering as unacceptable, and thus other approaches to interconnection emerge, including the option of *paid peering* or *partial transit*
- These latter models represent blends of the transitional pure transit or peering models that involve payments from one ISP to another and that restrict the range of addresses to which traffic may be delivered
  - In paid peering, one ISP pays the other ISP to accept the peering traffic
  - In partial transit, the ISP that is paying for transit is only purchasing delivery to a subset of the rest of the Internet
- These contracts are privately negotiated and are not standardized. The payment terms, the traffic (volume, address) restrictions, and other interconnection terms may be negotiated on a case-by-case basis

ACCESS NETWORKS AND CONTENT DELIVERY NETWORKS (1)

- Thinking at the shape of today's Internet, special attention should be paid to the sub-case of interconnection between residential broadband ISPs (*access networks*) and networks that deliver high-volume commercial content (*content delivery networks*)
- From a business point of view, these interconnection arrangements are of considerable interest because of the high volumes of data being exchanged, and the implications of this high volume for internal costs
- But problems arise also from the public regulation point of view, because high content-related value of this data can allow any actor with sufficient market power to extract rents from the value of the content, not its delivery
- In such case, to avoid distorsion, interconnection agreements should be based on *transport payments* (payments that cover the internal costs related to the delivery of flows of data), and should not include *content payments* (payments that relate to the value and costs of the commercial content itself, not its transport)
- But usually, access ISPs have market power and may seek to use it to earn monopoly profits from over-charging end users or content providers

A SCHEME OF INTERCONNECTION AND COST (1)

- Negotiation between a content delivery network and an access network about direct connection should normally be a peering negotiation, because the content delivery network is normally only trying to get access to the customers of the access network
- But because of the high volumes of traffic potentially involved and the fact that the traffic is highly asymmetric (from content to access), negotiation about payment may be commonplace.
- The carriage of this high-volume traffic will generate costs for access networks, which they will naturally attempt to recover; at the same time access to the content might create value for the access network's customers
- Based on anecdotal evidence, it appears that the most common outcome is for content delivery networks to make payments to access networks

A SCHEME OF INTERCONNECTION AND COST (2)

- Let us consider a single broadband access network A interconnecting with a content delivery network CD.
- Most of the traffic flows from network CD to network A, which adds usage-based costs to network A, designated as C_A
- Network CD also incurs costs to deliver this traffic, which include the costs of servers and communication links, and we denote these as C_{CD}
- There is a slight asymmetry to the situation of CD and A, which is worth keeping in mind
  - Network A has a substantial base cost which pre-exists to the growth of high-volume commercial content. For example, network A needs to have both the (last-mile) distribution network and backbone network to support basic communication and other services, regardless of whether this same infrastructure is shared to deliver video. By assumption, C_A comprises only the incremental, usage-based component of the total cost of A
  - CD, on the other hand, exists only to deliver this content, and will tend to view their total network-related costs as associated with the delivery of content.

A SCHEME OF INTERCONNECTION AND COST (3)

- The sources of dollar flows (payments) in this model include the payments that content owners/producers make to network CD for content delivery services, and the payments that end users pay to their broadband access provider, network A, for their access and usage
- Finally, there is the potential of a payment flow, P, between networks CD and A, as follows:
  - P = 0: Traditional revenue-neutral peering, where each network covers all its internal costs from its own customers.
  - P > 0: Payment from CD to A to help cover some of the internal costs of A. Such a payment would be part of C_{CD} and presumably these costs are then passed through by CD to the content producers, who pay more to CD and thus indirectly cover the cost of transporting their content across A. This outcome is fairly common in today's content delivery network market.
  - P < 0: Payment from A to CD. This payment arrangement seems uncommon, but makes sense in certain circumstances. Consider the case where A is a small, rural ISP. If there is no direct connection between CD and A, all of the content from the producers will come into A over a potentially very expensive transit link. Having CD make a direct connection to A may greatly reduce A's costs. However, if A is small, it may not be cost-effective for CD to connect to A; the connection might actually increase C_{CD}, not reduce it. In this case, A could pay CD.

A SCHEME OF INTERCONNECTION AND COST (4)

- The emergence of high-volume content (video) has generated substantial new costs C_A. There are only four ways that access network A can manage usage-related costs C_A:
  1. Lower the total costs: by careful design of their network, specifically with attention to where the content delivery networks interconnect with them, they may be able to reduce the actual CA incurred
  2. Payments from CD to A (P > 0): network A can negotiate to have the content network CD compensate them for some of these costs
  3. Raise retail prices for consumers:
     - ✓ allocating an equal share of CA to all customers;
     - ✓ creating mechanisms that will discriminate among users based on some proxy of usage, and increase the price of service for these users.
  4. Internally subsidize by accepting reduced margins: if the ISPs such as network A are competitive, however, then this is not a sustainable option

A SCHEME OF INTERCONNECTION AND COST: WHO PAYS WHOM?

- When CD networks connect to large access networks A, even if both CD an A have costs, payment seems to most commonly flow from CD to A. Two possible reasons:

1. network A has a better bargaining position, because it holds a terminating monopoly with respect to its customers
2. There is a common practice, older than the emergence of high-value commercial content, according to which money flows and packet flows go in the same direction (if X is delivering packets to Y, then X pays Y). How and why this practice originates?

    - When traffic in the two directions is balanced ISPs usually agree to revenue-neutral peering
    - If it becomes unbalanced, one or the other party may ask to renegotiate the agreement saying that the inbalance in traffic has caused an inbalance in costs.
    - Anecdotally, it is often the party receiving the excess traffic that complains. But increased traffic (say from X to Y) adds to the costs both for X and Y. So why would X pay Y?
    - There seems to be an unstated assumption that a transfer is of more benefit to the originator than the receiver, so the sender should be expected to cover more of the delivery costs. This assumption is not always true (es.: downloads of large open-source software package; download of open source videos on Youtube)

- Obviously, the parties will find it profitable to interconnect so long as the net benefit to each one (including the payments) is positive.

A SCHEME OF INTERCONNECTION AND COST: TRANSPORT PAYMENTS VS. CONTENT PAYMENTS?

- There are cases in which the negotiation between content delivery networks and access networks brings inefficient results
- If non-zero payments are possibile, one actor (e.g. A) might have enough market power (e.g., because it is a terminating monopoly with respect to its customers) to demand a payment from CD that exceeds its internal costs C_A
- In this case, the payment is not just a transport payment (Pt), but includes a content payment (Pc), as well.
- This would signal that the access networks have enough market power to impose unacceptable (inefficient) conditions to content delivery networks and, through them, to content producers, appropriating (part) of the, surplus of the upstream businesses.
- So the obvious question for regulators and industry observers is: if we allow non-zero values for P, how can we distinguish content payments from transport payments?

BOUNDING THE OUTCOME OF PEERING NEGOTIATIONS (1)

- One way to try to understand the context of negotiation between CD and A is to speculate about their relative market power
  - A has a terminating monopoly with respect to its customers, but
  - CD may be hosting valuable content that the customers of A demand
- Which network has the stronger bargaining position?
- Unfortunately, this question can be answered only on a case by case basis
- Moreover, ISPs and content delivery networks typically have multiple options for routing traffic from source to destination nodes. This limits the bargaining power of the two networks
- What can be done from a general perspective is, instead, to identify constraints or bounds on the outcome of negotiation between CD and A.
- If correctly identified, such constraints might allow policy-makers to infer whether the result of the interconnection negotiation was about:
  - a reasonable allocation of delivery-related costs
  - an unreasonable allocation of the surplus associated with end users' willingness to pay for content, above whatever it costs to efficiently deliver that content to the end users.

UPPER BOUNDS FOR PEERING NEGOTIATIONS: THE PRICE OF TRANSIT (1)

- Network CD may purchase transit from a third network T that subsequently either peers with network A directly or via its own transit agreements
  - Other peering agreements can affect the delivery of traffic from CD to the customers of network A. Network A has multiple peering partners, and may have its own transit provider to enable it to sustain its connectivity to the rest of the Internet and provide diverse routing options to enhance reliability and facilitate load balancing
  - In this case, if network A seeks to extract a large payment from network CD, then CD can choose to route the traffic to A indirectly via a transit connection to network T
- Because the market for transit services appears relatively competitive, with prices consistently declining over time, the availability of this alternate routing choice provides a loose upper bound on what CD might be induced to pay A to deliver content to A's customers

UPPER BOUNDS FOR PEERING NEGOTIATIONS: THE PRICE OF TRANSIT (2)

- The reason why transit prices represent a *loose upper bound* is that A might be able to offer valuable performance enhancements (e.g., caching services) or other quality assurances that CD desires, thereby making CD willing to pay a premium above the transit payment P_T
- Anyway, there is empirical and anecdotal evidence that this is not the case, because C_A seems to be (substantially) larger than current values of P_T
- For this reason, even if CD is persuaded by A to pay a premium over P_T as a part of a paid peering agreement, A will not recover from P_T all of its internal costs. Part of C_A must be recovered from A's customers
- It is therefore unlikely that there will be a content-related payment that is part of the negotiated payment from network CD to network A

UPPER BOUNDS FOR PEERING NEGOTIATIONS: THE SINGLE-HOP ACCESS

- 'Single-hop interconnection services' provides another way by which potential paid-peering payments might be moderated
- Under this model
  - another network K negotiates a peering agreement with network A, and then
  - K solicits network CD to interconnect with network K at the same physical location where K interconnects with A.
- This 'single-hop' interconnection arrangement imposes very few costs on K so it will still find it profitable to offer this interconnection option to CD at a very low mark-up over K's cost of handing off traffic to network A
- Those costs may be zero if K and A exchange traffic under a revenue-neutral peering arrangement

UPPER BOUNDS FOR PEERING NEGOTIATIONS: PARTIAL TRANSIT

- Single-hop interconnection cannot be organized by means of bilateral pure peering agreements
- Normally, a network would not agree to route traffic coming in from one peering partner out to another peering partner: it would be forwarding traffic without being paid by either partner
- We have seen that paid peering or transit agreements are possible options, but also other kinds of routing restrictions can be experimented
- For example, network CD might negotiate a partial transit agreement with network T which guarantees delivery only to a subset of addresses. That subset might include network A's end customers
  - Structurally, this would be like a paid-peering arrangement in that it includes a routing restriction
  - Network CD would expect to pay less than P_T for a guarantee of delivery to only a portion of Internet addresses

SOME GENERAL PRINCIPLES FOR NEGOTIATING INTERCONNECTION

- In summary, the previous discussion hints at two sorts of norms
- First are criteria by which one ISP would consider agreeing to revenue-neutral peering
  - One such criterion is balance of flows, in which the data rates between the two parties are roughly in balance (perhaps no more than 2 to 1 in the peak direction)
  - Balance of flows is a rather rough approximation for balance of value, but it can be used to impose limits on behavior such as single-hop access
  - If all parties understand up front the maximum amount of imbalance that A will tolerate, this can avoid the pain of after-the-fact attempts to renegotiate a peering agreement
- When revenue-neutral peering is not agreeable to both parties, we have speculated that a new norm might emerge to bound the price that might be charged for paid peering, which is that the rate for paid peering would be related to the price of transit
- A proposal for a paid peering fee that greatly exceeds the customary price of bulk transit would be seen as evidence that the network proposing that fee does indeed have market power that allows it to distort the market
- But a non-zero peering fee is not in itself a signal of such power

HOW THE CONSUMER IS CHARGED (1)

- There are three general approaches for structuring the subscriber usage fees
  1. *Flat-rate pricing.* The total portion of the C_A to be recovered is apportioned equally on a per capita basis. In flat rate pricing smaller users pay more than the (low) costs they impose to ISPs, and the larger users pay less than the (high) cost the impose on ISPs. Flat-rate pricing is discriminatory, since the smaller users subsidize the larger ones
  2. *Fees proportional to subscriber usage*. This latter option might be based on per GB pricing.
  3. *Usage tier pricing.* Users may choose among different usage tiers whose price depends on the traffic volume they allow. In this case, while the ex ante price is different according to the usage tier selected by the consumer, the ex post effective average price/GB depends on actual usage.

     Traffic that exceeds a tier's allotted volume may be priced at a pre-specified coverage rate, or the customer may be temporarily boosted into a higher-volume tier, or the traffic might be shaped (subject to reduced data rates) or even dropped.

HOW THE CONSUMER IS CHARGED: PROS AND CONS OF USAGE TIER PRICING

- Usage tier pricing seems to be a better and more efficient way to charge consumers because it allocate fees in (rough) proportion to costs, and this could reduces the distortions due to the discriminatory effects of flat-rate pricing , but:
  - it can help the ISPs to implement a second order price discrimination
  - it can significantly alter the balance of power in negotiation about interconnection fees. For example, in countries where residential broadband access is sold with rather low monthly usage caps, some ISPs are offering a ‘premium service’ to their content network partners. With this premium service (sometimes referred to as ‘zero rating’), the content delivery network CD pays a per GB fee to the access network A, in exchange for which the consumer can download the content without having it count against their monthly quota