7-28-2020

A Public Option for the Core

Yotam Harchol

Dirk Bergemann

Nick Feamster

Eric Friedman

Arvind Krishnamurthy

See next page for additional authors

Authors

Yotam Harchol, Dirk Bergemann, Nick Feamster, Eric Friedman, Arvind Krishnamurthy, Aurojit Panda, Sylvia Ratnasamy, Michael Schapira, and Scott Shenker

A PUBLIC OPTION FOR THE CORE

By

Yotam Harchol, Dirk Bergemann, Nick Feamster, Eric Friedman,
Arvind Krishnamurthy, Aurojit Panda, Sylvia Ratnasamy, and Michael Schapira

July 2020

COWLES FOUNDATION DISCUSSION PAPER NO. 2245

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

A Public Option for the Core

Yotam Harchol
EPFL
yotam@cs.berkeley.edu

Dirk Bergemann
Yale
dirk.bergemann@yale.edu

Nick Feamster
University of Chicago
feamster@uchicago.edu

Eric Friedman
ICSI and UC Berkeley
ejf@icsi.berkeley.edu

Arvind Krishnamurthy
University of Washington
arvind@cs.washington.edu

Aurojit Panda
New York University
apanda@cs.nyu.edu

Sylvia Ratnasamy
UC Berkeley
sylvia_ratnasamy@berkeley.edu

Michael Schapira
Hebrew University of Jerusalem
schapiram@cs.huji.ac.il

Scott Shenker
UC Berkeley
shenker@icsi.berkeley.edu

Abstract

This paper is focused not on the Internet architecture *– as defined by layering, the narrow waist of IP, and other core design principles – but on the Internet* infrastructure*, as embodied in the technologies and organizations that provide Internet service. In this paper we discuss both the challenges and the opportunities that make this an auspicious time to revisit how we might best structure the Internet's infrastructure. Currently, the tasks of transit-between-domains and last-mile-delivery are jointly handled by a set of ISPs who interconnect through BGP. In this paper we propose cleanly separating these two tasks. For transit, we propose the creation of a "public option" for the Internet's core backbone. This public option core, which complements rather than replaces the backbones used by large-scale ISPs, would (i) run an open market for backbone bandwidth so it could leverage links offered by third-parties, and (ii) structure its terms-of-service to enforce network neutrality so as to encourage competition and reduce the advantage of large incumbents.*

CCS Concepts

• **Networks** → **Public Internet**; **Network economics**;

Keywords

Internet transit, Network neutrality, Internet infrastructure

ACM Reference Format:
Yotam Harchol, Dirk Bergemann, Nick Feamster, Eric Friedman, Arvind Krishnamurthy, Aurojit Panda, Sylvia Ratnasamy, Michael Schapira, and Scott Shenker. 2020. A Public Option for the Core. In *Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication (SIGCOMM '20), August 10–14, 2020, Virtual Event, NY, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3387514.3405875

SIGCOMM '20, August 10–14, 2020, Virtual Event, NY, USA

ACM ISBN 978-1-4503-7955-7/20/08…$15.00
https://doi.org/10.1145/3387514.3405875

1 Introduction

1.1 Context

The structure of the current Internet infrastructure – as opposed to its architecture – is largely an accident of history, rather than a premeditated design. In fact, the Internet was originally a monolithic structure, with a uniform routing protocol operating across all nodes. As the Internet devolved into separate Autonomous Systems (each with many internal nodes) and became commercial in the 90s, it settled into a pattern of interconnections that, while complicated in its details, was conceptually simple. Each stub domain connected to the Internet through one or more Internet Service Providers (ISPs). Interconnections between ISPs were largely bilateral and typically fell into one of two categories – customer-provider or peering – and the interdomain routing protocol BGP was invented to give domains control over which paths to import and export. Because only the largest ISPs had substantial backbones, the notion of *transit* became necessary, where some ISPs carried packets that were neither from or to hosts in their domain. As a result, the path of a typical packet would start at the originating domain, continue through one or more transit domains, and then arrive at the destination domain.

In these early days of the Internet, wide-area bandwidth was extremely expensive, so the *backbones* or *cores* (we will use the terms interchangeably) – which allowed large ISPs to provide transit to other providers – were a sign of prestige and a source of dominance. This resulted in a small number of Tier 1 providers acting as the glue for the Internet as a whole. The primacy of these Tier 1 providers, and their key role in making wide-area transit possible, allowed the Internet to flourish.

However, over the past five years, long-haul bandwidth has become cheaper and easily leasable: median monthly lease prices across a selection of critical city-pairs declined an average of 27% and 24% (for 10Gbps and 100Gbps links, respectively), according to [53]. Of course, long-haul bandwidth is still far more expensive than bandwidth inside an enterprise campus or datacenter, but our point (which we elaborate on below) is that it no longer dominates the costs of ISPs [27].

The availability of leasable wide-area bandwidth has allowed companies such as Cato [7] and Aryaka [2] to create their own application-specific backbones. In fact, in the trans-Atlantic market, content providers accounted for 85% of the international demand in 2018 [53]. Several large cloud and application providers – such as

Google, Amazon, and Facebook – have gone further and built their own global high-bandwidth backbones to interconnect their datacenters and to reach various colocation facilities (such as Equinix). Since backbone bandwidth is now readily available for lease or purchase, and the expertise needed to run backbone networks has spread beyond the global ISPs, a large fraction of Internet traffic no longer relies on transit provided by the public Internet (see Section 2.4).

With core bandwidth becoming more plentiful, the attention of ISPs is turning to improving and expanding last-mile connectivity. The network edge now dominates the capital expenditures of ISPs (*e.g.,* see [27] for a discussion of the difference between core and access costs) and is also where innovation and expansion are most readily apparent.

After such sizable changes in the underlying costs and ISP priorities, now is a good time to rethink how we might want to re-architect the Internet's infrastructure. This first requires us to examine several problems with our current Internet infrastructure, which include outdated peering policies, threats to network neutrality, persistent lack of competition (in the US market), and increasing vertical integration. Then, motivated by these problems, we propose a restructuring of the Internet infrastructure that helps address them. Because the problems require a rather lengthy discussion, we first give a brief preview of our proposal.

1.2 Proposal

There are three key principles in our proposal. First, while transit is rapidly being privatized, it is essential that the public Internet continues to offer high-performance transit so that new content and service providers can emerge without having to lease or construct their own backbone. Second, the way to save public transit is to cleanly separate it from last-mile delivery. These two functions – transit and last-mile-delivery – are currently combined in global ISPs, but this paper describes the advantages of splitting them. Third, revenues should be more closely aligned with the value delivered and cost incurred; without such alignment, the Internet will continue to suffer from *tussles* [10] such as the network neutrality debate. What we are searching for is no less than the *economic architecture* of the Internet whose longevity can match that of its technical architecture.

For transit, we propose the creation of a global Public Option for the Core (POC).[1] The POC would be run by an international nonprofit organization that initially leases bandwidth from a set of Bandwidth Providers (BPs) and charges the users of its infrastructure to recoup these costs. That is, while the POC is a nonprofit, it is not a charity, so we expect it to break even financially. The nonprofit nature is necessary to ensure that it focuses on its mission of providing global transit, rather than moving into more lucrative markets (such as last-mile-delivery or content and services) or avoiding poorly served areas. The initial use of leased links allows the POC to start without massive capital expenditures, but the POC might eventually acquire some links of its own.

For servicing the last mile, a new generation of ISPs we call Last-Mile-Providers (LMPs) use the POC for their transit, so they need not build a core of their own nor use transit provided by a competing ISP. These LMPs could be existing access-oriented ISPs or newly created access providers; what is new is that they rely on the POC for transit. In addition, content and service providers (CSPs) could either directly attach to the POC (which would be the case for large CSPs), or make use of an LMP to reach the POC.

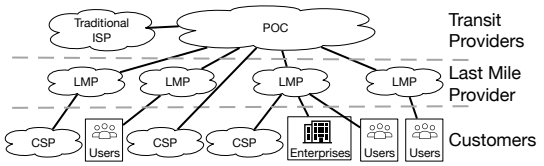

Figure 1: *Network connectivity in our proposal. The Public Option for the Core (POC) which we describe in this paper acts as the main transit provider, but it connects to one or more ISPs to provide connectivity to anything not accessible through the POC. Customers – which include content and service providers (CSPs), users, and enterprises – connect to last mile providers (LMPs), which in turn connect to the POC. Some large CSPs also connect to the POC directly.*

Within our proposed design, packets would originate in one LMP (or in a CSP), be carried by the POC directly to the destination LMP (see Figure 1). As such, the POC is functioning as a large IXP; the POC itself exercises no peering policies and merely acts as a transparent fabric, leaving it to the attached LMPs to decide whether and how to exchange traffic. This LMP-to-LMP peering decision raises the question of whether and how to enforce network neutrality, which is a topic we focus on later in the paper.

In terms of economics, in our proposal there is a clear delineation of who pays for what. As part of this delineation, we propose enforcing network neutrality (in a form we define later) through contractual obligations rather than regional legislation. This arrangement rewards those entities that bring value to consumers, while enabling new entities to enter the market without unfair competition from incumbents who, without the requirement of network neutrality, could easily extract revenue from the value being provided by others.

This POC and the LMPs attached to it are intended to operate in parallel with, and connected to, the current Internet infrastructure. Nothing in this proposal forces current ISPs to change their operations, and we expect the POC to interconnect with one or more traditional ISPs (and eventually become a Tier 1 provider itself). Our proposal is intended to open a parallel track through which transit can be provided (by the POC), and last-mile service established (through LMPs). However, the full benefits of this approach will come only when the POC becomes a large-scale provider.

We should note that there could be several coexisting (and interconnected) POCs, run by different entities but adopting the same basic principles (nonprofit, focusing on transit, enforcing network neutrality). This would allow innovation in network management practices while providing the same basic benefits. However, for ease of exposition, we will focus on the case where there is a single POC.

In the next section we delve into *why* we recommend this proposal, and then in Section 3 describe *how* this can be implemented. In Section 4 we analyze the economic implications of requiring network neutrality, and conclude in Section 5 with a frank discussion of the question most readers are asking themselves after reaching this point: *Are they crazy?*

[1] The term "Public Option" was previously used by Ma and Misra [39] to describe a quite different approach to improving network neutrality. We describe the technical differences between our proposal and theirs in Section 2.6.

1.3 Contributions

This is most definitely not a traditional SIGCOMM paper, which typically describes a concrete design, defines quantitative performance objectives, and then presents a thorough performance evaluation. Instead, our paper is far more general, considering a radical alternative for the Internet's infrastructure and an analysis of its impact on the overall ecosystem of ISPs and CSPs. Instead of detailed performance results, our contributions are more subtle and fall into three categories:

Structure: Motivated by the four troubling trends described in Section 2, we describe a proposal for (i) cleanly separating transit and last-mile-delivery in the Internet's infrastructure and (ii) structuring the flow of payments. This aligns the revenue flow with the value flow, which reduces economic distortions and allows growth and innovation to occur where needed. This overall structure, which we consider a new *economic architecture* for the Internet, is the most important aspect of our design. The centerpiece of this new structure is the POC, a global edge-to-edge transit network run by an international nonprofit.

Auction: To realize this POC, we propose an auction (in Section 3) that provides a strategy-proof mechanism for compensating BPs for their leased lines. The strategy-proof nature allows BPs to focus on evaluating their own costs, rather than analyzing the market to determine their bids.

Analysis of Network Neutrality: Finally, in Section 4 we discuss network neutrality, by which we mean the prohibition of termination fees, which we define in the next section. We present a novel economic model of CSPs and LMPs, and show how not requiring network neutrality would hurt social welfare (in the economic sense) and future innovation (by favoring incumbents). While our general conclusions are in line with some others in the literature [54], we are not aware of a similar model that captures the relevant aspects of network neutrality so cleanly.

1.4 Ethical Considerations

This work does not raise any ethical issues.

2 Why Change?

While the easy availability of wide-area bandwidth makes our proposal possible, it does not justify the need for major change. Here we first outline four troubling trends (similar to those cited in the 2005 FCC report [21]) that motivate our proposal, and then describe how the POC helps deal with them.

2.1 Peering Policies Are Outdated

The BGP policy mechanism is transitive, in the sense that a domain's policy choices (about which routes to import) are limited to the options exported by its neighbors, which are in turn limited to the choices presented to them by their neighbors. This works well because typical BGP policies are expressed in terms of customer/provider/peer relationships, which are themselves transitive (*i.e.,* the provider of my provider acts like my provider). Moreover, these customer/provider/peer relationships are often determined by the amount of traffic carried (*i.e.,* two ISPs exchanging roughly equal amounts of traffic can peer, while those sending more than they receive typically have to pay the other party).

While transit carries data across the Internet, the real value of the Internet is what happens at the end points where services and content are produced and consumed. As coined in [19], a seminal paper on this topic, we often refer to networks where services and content originate as *content* networks, and to networks where they are consumed as *eyeball* networks. Traffic flows primarily from content to eyeball networks, but the value flows both ways: eyeballs derive value from content and services, while content and service providers (CSPs) derive value (in terms of direct and/or advertising revenue) from eyeballs. This value chain does not fit BGP's transitive nature, nor is it based on the relative rates of traffic between domains (as some traffic has far more value-per-bit than others).

This is not merely an academic concern. For instance, the mismatch between value flow and current peering relationships has resulted in several disputes involving Netflix traffic. In one, Netflix contracted with Cogent for transit because of its low prices, but then Comcast complained when Cogent tried to transfer that data to Comcast's network. Comcast was seen as violating network neutrality, when the more relevant dynamic was a failure of modern peering policies and their transitive nature [18].

More generally, the problem is that ISPs often combine serving eyeballs (where value is consumed) with providing transit (where the costs are just per-bit) and offering their own content and services (where value is created). Even though traffic between ISPs can be binned into different categories, there is no way the simple transitive nature of Internet peering could possibly capture the resulting economic interactions (which are not, on the whole, transitive). Violating network neutrality is one way of getting around the transitive nature of interconnection; we feel that our proposal is a more constructive step forward.

In addition, there are two other trends that are further disrupting traditional interconnection arrangements: IXPs and CDNs. IXPs provide interconnection points where a wide variety of networks can directly peer with each other. As described in [5], they are gaining in popularity among smaller ISPs, while larger ones have mostly avoided them. CDNs have been around for decades, but with video dominating traffic CDNs have become more important than ever (we discuss CDNs briefly in Section 3.2).

2.2 Network Neutrality at Risk

The technical Internet community has long embraced the notion that the Internet should be application-neutral; that notion later became known as *network neutrality* (a term coined in [57]). In what follows, we use both of these terms – network neutrality and application-neutrality – with the former having more of a legal connotation than the latter. Wikipedia [56] defines network neutrality as follows: "Net neutrality is the principle that Internet service providers treat all data on the Internet equally, and not discriminate or charge differently by user, content, website, platform, application, type of attached equipment, or method of communication." A similar definition that explicitly mentions the topic of charging comes from [28]: "Net neutrality usually means that broadband service providers charge consumers only once for Internet access, do not favor one content provider over another, and do not charge content providers for sending information over broadband lines to end users."

Officials at several ISPs have made clear their opposition to this last point about not charging content providers for traffic reaching their customers. In 2005, Ed Whitacre (then CEO of SBC) said [43] "Now what they (content providers) would like to do is use my pipes [for] free, but I ain't going to let them do that because we have spent this capital and we have to have a return on it." Similar statements about charging content or application providers have been made more recently by the CEOs of Telefonica, Vodafone, and Deutsche Telekom [35, 45, 47].

The legal status of network neutrality has changed over time in the United States. In 2005, the US's FCC adopted network neutrality principles and in 2015 the FCC issued the Open Internet Order that gave them the right to enforce network neutrality. However, in 2018 these network neutrality regulations were repealed, although some states have since instituted their own network neutrality regulations. More globally, the European Union has strong network neutrality regulations [4], but elsewhere regulations vary from country to country, and there is no uniform global standard.

There is a large academic economic literature on network neutrality, which we briefly review in Section 4. The literature comes down on all sides of network neutrality: some explicitly in favor (*e.g.,* [23, 36]), some explicitly against (*e.g.,* [20, 58]), and some delivering a more mixed message of "it depends" (*e.g.,* [17]). The literature employs a set of simple models to make their case, and in Section 4 we introduce another model that we think better captures the relevant aspects of the situation. Our results indicate that without network neutrality, incumbent LMPs and CSPs have a significant competitive advantage, which would hinder innovation.

Given this result, we must answer the question of how such regulations could be enforced in an infrastructure like the Internet that transcends national boundaries. We return to this challenge later in this section.

2.3 Competition In The ISP Market

We restrict our comments on this topic to the US market, because the nature of ISP competition depends strongly on past and current regulatory frameworks. Most of the developed world has far more competition in the ISP market, largely due to loop unbundling, by which we mean regulations requiring telecommunication operators to allow other service providers to use (at a fair price) their last-mile lines into homes. This means that new service providers can enter the market without building their own last-mile infrastructure. However, in the US ISP competition is very thin, with many areas having only one or two viable high-bandwidth service providers. According to figures for December 2017 [32] (also see [48]), while over 95% of US census blocks have two or more providers supplying at least 25mbps downloads (and at least 3mbps uploads), only 26% of census blocks have one or more providers supplying at least 100mbps downloads (and at least 10mbps uploads), with only 5% of census blocks having three or more such providers.

The high capital and operational costs of reaching individual homes and businesses is one factor for why competition is so limited for high-bandwidth network service. However, a contributing factor is that such ISPs must either build their own core network (at significant cost and management complexity) or contract with an ISP to provide transit. In many cases (Cogent and Level3 are exceptions), these transit ISPs are competing for the same last-mile market, and can use their transit pricing to put new competitors at a disadvantage.

2.4 Vertical Integration

There are two forms of vertical integration, which we consider separately.

2.4.1 CSPs building their own network Some of the leading CSPs have built their own backbones (*e.g.,* Google, Facebook, Amazon, Akamai). Between CSPs that have their own backbone, and those that host their service on one of the CSPs that have their own backbone, it turns out that a very significant fraction of the Internet's traffic is immediately shunted into a private backbone after leaving their home domain. For instance, the results in [8] show that for traffic leaving GCE towards various BGP prefixes, weighted by the volume of requests from those prefixes in their CDN trace, roughly 66% of the requests went directly from one AS to another. Consistent with this, we were confidentially told by an operator that their estimate of the percentage of such direct-from-home-to-private-backbone was roughly 70% [44]. Regardless of the exact number, this trend is undeniable.

Geoff Huston comments on this trend in an article entitled "The Death of Transit?" [31] (see also [9]) where he notes that most traffic is first handled by CDN nodes at the edge, which then corral (according to [31]) "each client into a service 'cone' defined by a collection of local data centres." Thus, there is still tremendous bandwidth dedicated to moving bits on backbones, but much of the action has left the public Internet and is now carried on private networks. This runs the risk of creating several private Internet infrastructures that essentially cater to particular sets of services, leaving the public Internet to languish. This goes against the application-neutral spirit of the Internet, and might (in the long run) slow innovation.

2.4.2 ISPs also providing content and/or services Various large ISPs, such as Comcast and AT&T, are buying content providers [15]. This raises the risk of network neutrality violations, in that their network infrastructure can favor their own services over others. Prior work [37] has already shown that many cellular providers implement policies favoring some content providers to the detriment of others. If this trend is left unchecked, it could greatly impact the ability of CSPs to reach customers in LMPs who have competing offerings.

2.5 Why the POC?

As we have described, the current Internet infrastructure has an outdated mode of providing transit, based on transitive peering relationships that focus on the flow of packets but ignore the flow of value. This has caused many ISPs to contemplate charging so-called termination fees, where remote services (such as Netflix or Google) are charged for packets that flow into a last-mile's provider network. This would allow the ISPs to capture some of the value of these services, and they feel entitled because they are providing the access to customers.

While such termination fees are not yet implemented, there is no federal prohibition from doing so in the US (and in many other countries). Further, the growing trend of large ISPs entering the content and service market increases the temptation to violate network neutrality by favoring their own services over competing ones.

All of this is occurring in a landscape where there is little competition among ISPs (at least in the US), and traffic is increasingly funneled into private networks for transit. Thus, it seems a good time to step back and see if there is a way to reorganize the Internet's infrastructure to counter these trends.

Most fundamentally, our proposal is intended to provide an *economic architecture* that makes sense for the Internet. It scraps the traditional interconnection agreements and instead has all transit handled by the POC, with all LMPs (and CSPs) directly attached to the POC paying for transit based on traffic sent and received. CSPs and LMPs collect revenue directly from their customers. To enforce network neutrality, the POC's terms-of-service require that all attached LMPs peer freely with all others, with no termination fees and no differential service given to packets based on their source.

The POC would thus prohibit an LMP from giving their own content better service on their own network; LMPs would be free to acquire or develop such services as a business investment, but they would not be able to provide them with an unfair advantage. Similarly, the POC would not prevent large CSPs from building their own private backbones, but the POC would ensure that all CSPs, not just large ones, could have access to a high-performance transit service.

All of this helps improve competition in both the LMP and CSP markets: new entrants are not unfairly burdened by termination fees, and all LMPs and CSPs have access to high-performance transit that is not competing with them.

Note that loop unbundling, while definitely desirable, solves a different set of problems than the POC. Loop unbundling allows many LMPs to share the same last-mile infrastructure, making it easy for new entrants to arise, but unless they build their own network core these new entrants must contract with one or more transit carriers who might be competing with them. Moreover, as we observe in Section 4, the lack of network neutrality allows incumbent LMPs to charge higher termination fees than new entrants, thereby giving them an unfair advantage. Thus, the POC and loop unbundling are highly complementary solutions; one eases the construction of last-mile infrastructure, and the other ensures that new entrants need not build their own core or contract with potentially competing providers for transit and will not face unfair competition (via higher termination fees) from incumbent LMPs.

2.6 Is this really new?

There are already a variety of nonprofit and governmental networks in operation that provide transit. For instance, in Australia, NBN is a national governmental monopoly for wholesale Internet and telephony transit [38]. NBN owns its links and leases them using a uniform pricing mechanism regardless of where the service is delivered, so rural areas are cross-subsidized by metropolitan areas [52]. The budgeting and operation of NBN is regulated by law, and legislation also protects it from other ISPs undercutting its prices by cherry-picking low-cost markets [51]. In addition, many of the IXPs in Europe and elsewhere (but less so in the US) are nonprofits, and some require open peering between all connected networks. While these developments have some superficial similarity to our proposed POC, and are additional evidence that nonprofit and governmental networks can successfully carry commercial traffic, their limited geographic scope prevents them from having any significant impact on the broader structure of the Internet (and, to be clear, such impact was never their intent).

More academically, the notion of a "public option" has been previously raised in the research literature by Ma and Misra [39]. Despite the similarity of terminology, their paper addresses a very different problem than we do. First, they do not propose any changes to the core, only the last mile (what we call LMPs). Second, and far more fundamentally, they focus on service differentiation between content providers, not termination fees. They find that the presence of a nonprofit LMP that does not discriminate creates competitive pressure on commercial LMPs to not discriminate. However, note that the impact of service discrimination is visible to users, who then abandon LMPs that give worse service. In contrast, termination fees are invisible to users, and only reduce the profit margin of CSPs. Thus, their results would not apply to the problem we consider here.

3 Designing the Public Option for the Core

In this section we describe the overall design of the POC. We do not describe any of the low-level technical details of running a backbone network, as we assume that the POC uses industry best-practices for this. Instead, we address four basic questions on four different topics of concern:

Network Services: *What network services does the POC provide?*
Payment Structure: *Who pays who for what?*
Bandwidth Auction: *How are Bandwidth Providers compensated by the POC?*
Peering: *What are the peering arrangements between the LMPs that are attached to the POC?*

3.1 Network Services

At a minimum, the POC provides point-to-point connectivity between all connected LMPs and to the external ISPs to which the POC is connected. The POC can offer additional services; while these are not the main point of this paper, we do want to discuss some possible offerings. The first is offering different levels of quality-of-service (QoS). Some definitions of network neutrality disallow any QoS mechanisms on the basis that they could result in the Internet serving only those who could afford good service. We do not take a position on this debate in this paper (and therefore do not explicitly prohibit such measures), but there is nothing in our approach that would prevent the POC (or the attached LMPs) from offering different qualities of service to customers. What we *would* require is that these be openly offered, so that users could choose their desired level of service and pay the resulting price. We would not allow the POC or the attached LMPs to, on their own, decide to favor certain traffic over others. Thus, we make a distinction between service discrimination and QoS, and disallow the former while not prohibiting the latter.

In addition, the POC could support multicast and anycast delivery mechanisms, and any other standardized protocols that the IETF adopts. More generally, the POC could offer emerging services at the edge of its network, such as edge computing and network function virtualization (NFV). In fact, the presence of a neutral and nonprofit core might provide a place where such technologies – which are now struggling because end users need a uniform approach across ISPs, but there are no clear standards – could be tried out without worry

about proprietary advantages for one ISP over another (and LMPs could then offer these same technologies if they wanted).

3.2 Payment Structure

The entities we consider are the POC, a set of BPs, a set of LMPs, a set of CSPs, and customers (businesses and people). CSPs can either connect directly to the POC, or use an LMP to connect themselves to the Internet. The overall theme of the structure described below is that entities pay directly for what they receive.

- The POC pays the BPs to lease the links needed to create a backbone network. The POC also pays one or more ISPs for general access, so that the POC is connected to the rest of the Internet.
- Each LMP (and directly-attached CSP) pays the POC for network access.
- Each customer pays their LMP for network access. The customer also pays directly for any non-free CSP services they use.
- Each CSP using an LMP pays its LMP for network access.

This payment scheme is obvious, but by following the philosophy that entities pay directly for what they receive, we avoid the situation where ISPs turn to CSPs to pay for the bandwidth that the ISP's customers are consuming (which is the sentiment expressed by Ed Whitacre). Rather, in our scheme, if a user is ingesting too much traffic, it isn't the CSP who should pay, but rather the user herself.

If we allowed termination fees, then each CSP might have to pay each LMP for access to their customers. In the next section, we consider the implications of such charges, and reject them because they give an advantage to incumbents. We also disallow CSPs from paying remote LMPs to get priority service for their traffic when it arrives at the destination LMP, for the same reason; this would give an advantage to incumbent CSPs over emerging ones. One form of such priority service would be allowing some CSPs to pay for the right to install their own CDNs while disallowing others to do the same. LMPs (and the POC itself) are free to provide open CDN services (on a fee for service basis) or allow CSPs to install their own CDNs or similar network functions (for a set fee); what LMPs cannot do is only allow certain LMPs to use or deploy such services.

Our decisions about who pays whom for what does not dictate or restrict the nature of the pricing schemes (as long as they are not discriminatory) between any pair of entities. For instance, LMPs might charge home users a flat price, or a strictly usage-based charge, or some form of tiered service (a flat price up to a given level of usage). We understand that there is a tension between giving users some predictability in costs, while also charging based on usage so that LMPs (and the POC itself) can finance capacity expansion. Our proposal would allow the market to evolve over time to find practical solutions that meet both of these (and future) needs.

In addition, the question of what the POC pays BPs for their leased lines is conceptually separate from what the POC charges LMPs for their usage. As we describe below, the POC uses a strategyproof auction to pay BPs, but we leave open the question of how the POC charges LMPs. This could be, among many other options, a usage-based price just based on the sending/receiving rate, or it could be based on the costs incurred by where those packets flow. The only requirement is that the sum total of revenue from the LMPs is enough to cover the bandwidth (and other) costs of the POC.

3.3 Bandwidth Auction

Building a POC entirely out of links that it owned and operated would require a tremendous upfront expenditure of capital. We choose instead to initially create the POC's backbone network out of a set of leased lines, and use the interconnections to one or more ISPs as a fallback if the POC's backbone does not have sufficient connectivity. Of course, eventually the POC can transition to partially owning its infrastructure, but our concern here is to find a practical way to get started.

These leased lines can come from traditional sources, as there is already an active market for leased bandwidth. However, we also expect that others, such as the large CSPs that have already built their own backbone, would be eager to provide their excess bandwidth for lease for two reasons. First, the large CSPs have a problematic relationship with many large ISPs, in terms of who should pay for bandwidth. These large CSPs would like to displace the current large ISPs with a public backbone and a more innovative set of LMPs; the Google Fiber effort [24] is evidence of this. Offering up leased bandwidth to the POC would foster progress in this direction. Second, the large CSPs that are building their own backbone face hard provisioning choices, in terms of how much bandwidth to buy at any particular time: buy too much and they have wasted money; but buy too little and they run the risk of congestion on their backbone. The availability of the POC means that they can overbuy, and then lease out (on a temporary basis) their excess bandwidth but can quickly recall it from the POC when needed.

We think it is important that the reimbursement scheme for leased bandwidth be completely transparent, so that no one in the ecosystem feels that the POC is favoring certain BPs. In addition, we want to minimize the amount of effort devoted to strategic manipulation by the BPs (*i.e.,* rather than trying to figure out what the market will bear, they can just focus on what minimal price would cover their costs). To accomplish both of these goals, we use a strategy-proof auction mechanism for bandwidth that is a special case of the general class of VCG auctions [11, 26, 55].

VCG auctions are widely used in many market and allocation mechanisms. The use of VCG auction for electricity markets (in the US and elsewhere) is probably the closest to the bandwidth auction discussed here; see [13] for a recent survey. VCG auctions are also commonly used in financial markets such as in the US treasury auctions of bills and bonds, see [30].

For the POC auction, each BP (whose instances are denoted by α) offers a set of links for lease, with a specified minimal acceptable price for each subset of these links (assume this price is the monthly leasing charge). That is, each BP α provides a set of links L_α and a mapping C_α from the powerset 2^{L_α} to a minimal acceptable price for that subset of links (and if a subset is not offered, its price is set to be infinite). This allows the BP to offer discounts for multiple links, or other non-additive variations in pricing.

The external ISPs to which the POC is connected play two roles. First, they are needed to provide connectivity to the rest of the Internet, so packets whose destinations are not directly attached to the POC leave the POC through one of these ISPs. Second, we assume that these ISPs attach to the POC in multiple locations and thus they provide virtual links between these attachment points that go through the ISPs rather than the POC itself. We denote by VL the

set of these virtual links, and the cost of using them is dictated by the long-term contract between the external ISPs and the POC, not by the auction mechanism we describe below. We denote the cost of any subset of virtual links $L_v \subseteq VL$ to be $C_v(L_v)$. The presence of these virtual links VL provides the POC with a richer set of alternatives for conveying packets between LMPs. We define OL as the set of all offered links, both the virtual links provided by the external ISPs and the union of the links offered by BPs: $OL = VL \cup \{\cup_\alpha L_\alpha\}$.

We consider the set of attachment points to include all network locations where LMPs, directly-connected CSPs, and external ISPs are connected to the POC. We assume that the POC has some upper-bound estimate of its traffic matrix (how much traffic flows between each pair of attachment points). Given the set of offered links OL, the auction mechanism picks the lowest cost subset that (i) provides enough bandwidth to handle this traffic matrix, and (ii) obeys whatever other constraints the POC desires (such as requiring that it can still handle all the traffic even under some number of link failures). To make this more precise, given the various constraints (handling the traffic matrix, plus any additional constraints), we map the set OL to an *acceptable* subset of the powerset of OL that we will denote by $A(OL) \subseteq 2^{OL}$. Each element of $A(OL)$ represents a set of links that obeys the POC's various constraints. In what follows we assume that none of the external ISPs are also BPs, and the set of offered links is such that $A(OL - L_\alpha)$ is nonempty for all α; that is, these constraints can be met even if one of the BPs does not participate.

Define $C(L)$ for some acceptable subset $L \in A(OL)$ as the total cost of that subset of links:

$$C(L) = \sum_\alpha C_\alpha(L \cap L_\alpha) + C_v(L \cap VL)$$

The POC then picks the lowest cost member of the set $A(OL)$: that is, the set of selected links SL is given by

$$SL = \text{argmin}\, C(\tilde{L}) \text{ where } \tilde{L} \in A(OL)$$

This just says that the POC selects the lowest cost solution that obeys its constraints. The real question is how does it then pay the BPs for their links. If the POC just pays their costs as determined by C_α then each BP has an incentive to inflate their declared minimal price. So, instead, we use a strategy-proof auction whereby BPs are incentivized to reveal the minimal acceptable payments (via C_α), which set the lower bound on what they will receive.

To define this mechanism, we need additional notation. For a given BP α define $SL_{+\alpha} = SL \cap L_\alpha$ as the subset of SL that is in the set L_α; note that with this notation, $C(SL) = \sum_\alpha C_\alpha(SL_{+\alpha}) + C_v(SL \cap VL)$. We further define

$$SL_{-\alpha} = \text{argmin}\, C(\tilde{L}) \text{ where } \tilde{L} \in A(OL - L_\alpha)$$

as the set of links that would be selected if BP α did not offer any links.

With these definitions, we can define the payoff to each normal BP α as:

$$P_\alpha = C_\alpha(SL_{+\alpha}) + (C(SL_{-\alpha}) - C(SL))$$

This VCG mechanism (which is essentially the Clarke pivot rule [42]) selects the lowest cost options, and obeys the constraint that the payoffs are no less than the value $C_\alpha(SL_{+\alpha})$. This lower bound is obeyed because the cost cannot go down if you reduce the number of links available, so $C(SL_{-\alpha}) \geq C(SL)$. The strategyproofness follows from the fact that the values in this expression do not depend on any of the prices set by α (*i.e.,* the value of $C_\alpha(SL_{+\alpha})$ in the first and last terms cancel out).

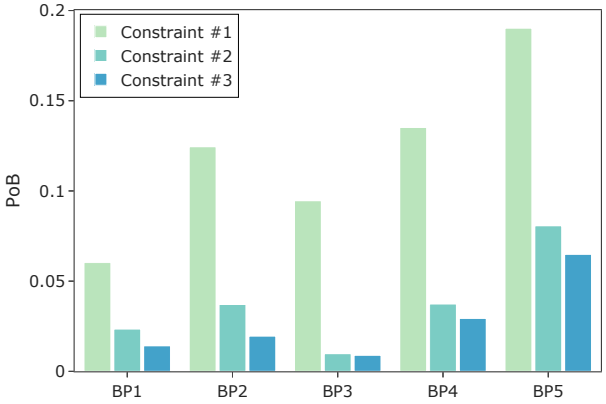

Figure 2: *We show the payment-over-bid margins (PoB) for the five largest BPs (ordered in decreasing size).*

While VCG mechanisms have been applied to routing before, our application here is more general. For instance, the approach in [22] applied VCG mechanisms to a simplified form of routing where all links were strategically independent (they were actually BGP domains) and only lowest cost routing is used without capacity constraints or resilience considerations. Here, we allow very general route computations that can take these factors into account.

Like all VCG mechanisms, this auction is vulnerable to collusion, but the presence of the connections to external ISPs sets an upper bound on the costs of alternate paths, so any of the manipulations mentioned below can only have limited impact. If the BPs can guess in advance what the set SL is, they can decide to not offer any links not in this set without changing their own payoff, but possibly changing that of others. For instance, removing links $L_\beta - SL$ from OL cannot make $C(SL_{-\alpha})$ smaller, and can make it substantially bigger, thereby increasing the payoff to BP α. At the same time, doing so does not change SL nor P_β, so BP β is unhurt by doing so. If all the BPs do this, they could potentially all gain (even without side payments). However, doing so requires them to know SL in advance (because if they withdraw links in SL they can hurt their own payoffs).

To illustrate how this auction might work in practice, we took network data from TopologyZoo [33], filtered out some of the small networks, combined some networks to form 20 BPs, and then placed POC routers at points where there were four or more BPs closely colocated. The resulting POC network has 4674 point-to-point connections between POC routers; we call these connections logical links because they may involve several physical links. The BPs vary in size, contributing from roughly 2% to roughly 12% of the logical links in the POC. We then generated a synthetic traffic matrix between all POC routers, and ran our auction with three different constraints (always looking for the cheapest solution that satisfies each constraint). Constraint #1 only requires that the set of links can handle the offered load. Constraint #2 requires that it could do so assuming that any single path between a pair of routers has failed. Constraint #3 requires that it do so assuming that a path between each pair of routers has failed. For each of these scenarios, we computed C_α and P_α, and in Figure 2 we show the resulting

payment-over-bid margins (PoB), $PoB = \frac{P_\alpha - C_\alpha}{C_\alpha}$, of the five largest BPs. One aspect to note is the high variation in the PoB, which is a good reason for the POC to use an open algorithm so that it cannot be accused of favoritism.

3.4 Peering Arrangements

We now focus on the POC and the attached LMPs, asking whether and how they exchange traffic, and ignore the rest of the Internet for the moment. The nature of peering relationships among the attached LMPs, and specifically whether network neutrality would dictate an answer, depends greatly on how the POC fits into the overall Internet ecosystem. In particular, is it considered an ISP, or an IXP?

If one thinks of the POC as a large ISP (with all those connecting to it being customers in the traditional provider-peer-customer categorization) then network neutrality requires that each LMP accepts whatever traffic arrives from the POC, not discriminating based on the original source of packets. Moreover, network neutrality requires that there be no termination fee for packets arriving from one of the other LMPs, because there is no direct peering relationship between the LMPs. In short, if the POC is seen as an ISP, and network neutrality continues to hold, then LMPs treat all arriving packets equally. The only payments are from the LMPs to the POC for providing core network service.

If one thinks of the POC as a global IXP – that is, essentially a policy-free interconnection infrastructure that allows all LMPs to peer (or not) with each other directly – then each LMP must decide with which other LMPs to peer (and exchange traffic), and whether any fees should be paid in doing so. One might think that enabling direct peering between all POC-connected LMPs would help foster network neutrality, because there are no intermediate transit ISPs to discriminate against traffic, but the opposite is true! Network neutrality has never constrained an individual domain's choice of peering: if peering is direct, discrimination based on the nature of the connecting LMP is easy to implement and does not violate network neutrality. For instance, if Netflix directly attaches to the POC (and acts as an LMP), and we consider the POC to be an IXP, then all other LMPs would be free (even under network neutrality regulations) to refuse to peer with Netflix without payment for doing so; in short, they could demand that Netflix pay them to be their "provider" in the traditional sense of customer/provider peering arrangements, and refuse to interconnect if Netflix refused to pay.

However, such a decision to not interconnect would be harmful, because it would disconnect pairs of LMPs, which would lead to fragmentation of the Internet. In today's Internet, each customer (whether enterprise, home, or CSP) is responsible for finding some ISP who will connect them to the rest of the Internet; there are many available ISPs that offer this service, so one relies on the market to provide acceptable solutions. However, for LMPs directly connected to the POC and only connected to the POC, the POC is the only way to interconnect with other LMPs, so there is no other choice.

Whether one considers this a network neutrality question or a peering question, the criteria for making this choice should be to choose the option that is best for the long-term health of the Internet. We consider this question in great detail in the next section by analyzing a formal model for these peering decisions, and come to the conclusion that it is best that LMPs be required to peer without termination fees with all other LMPs.

One might argue that this is unfair to LMPs that must carry traffic to their customers from bandwidth-heavy video services, without compensation from those video services. However, we are not proposing that LMPs cannot be compensated for the service they provide, we merely require that they obtain their compensation from their directly-paying customers rather from the LMPs they peer with. A key point in our design is that there is no need for compensation at a distance, except between CSPs and users (where the interaction occurs at the application-level). The LMP customers of the POC are paying for all traffic carried from and to them by the POC. Similarly, the customers of each LMP are paying for all traffic carried from and to them by that LMP.

If one of the endpoints on the POC is a CSP like Netflix, then the customers pay this service directly, which can cover Netflix's payments to its LMP (or to the POC if it is connected directly) for the bandwidth it uses. And the customers of Netflix are also paying their LMPs for the bandwidth they receive from Netflix. Thus, in our proposed arrangement, every LMP or CSP who is incurring large bandwidth costs recoups those costs directly from their customers (who are the ultimate cause of that traffic), rather than expecting non-customers to help defray those costs. By having the bandwidth costs collected by the parties causing them (the CSP and the end user), and payments for the service made directly to the CSP by the customers, this leads to a natural splitting of the revenues between LMPs and CSPs in a way that is driven by customer's willingness to pay and the presence of competitive alternatives, rather than through painful negotiations filled with brinkmanship as in the Netflix-Cogent-Comcast case.

However, this does require that users pay for their bandwidth usage. Historically, imposing bandwidth limits and usage-based charging has resulted in significantly bad press for the ISPs proposing such measures. However, we think the economics on this are clear; it is better to have costs borne by the entities that caused those costs. We should also note that whatever you think of our proposal, termination fees are not the right way of dealing with this problem. Allowing LMPs to impose termination fees is a mechanism that can improve LMP profits, but in no way guarantees that this additional revenue was needed to expand capacity or that the LMPs will indeed spend it on expanding capacity. As we will see in the next section, when termination fees are allowed, the LMP can extract revenue from CSPs independent of whether it is needed to cover bandwidth costs.

Before continuing, we make the peering conditions precise, with the caveat that exceptions should be made for security concerns (which may require blocking) or internal maintenance traffic (which may require priority or other special handling). The peering conditions we impose are that a POC-connected LMP must not: (i) differentially (in terms of priorities or blocking) treat incoming traffic based on the source or application, nor differentially treat outgoing traffic based on the destination or application; or (ii) differentially provide CDN or other application-enhancement services based on the source (for incoming packets) or destination (for outgoing packets); or (iii) differentially allow third-parties to provide CDN or other application-enhancement services that only target a subset of traffic (where this last condition prevents an LMP from allowing, say, Netflix to install services that enhance their traffic but disallowing others from installing such services). These conditions also apply to traffic

arising within the LMP providing content or services to customers. Also, note that the LMP can offer application-enhancement services or QoS for a given price, and only provide those services to those who pay; what they cannot do is arbitrarily discriminate between traffic.

These are included in the terms-of-service. If widespread cheating is anticipated, the POC could only forward encrypted traffic making in harder for LMPs to discriminate, but this would require a software change by users, which is unlikely.

4 The Case for Network Neutrality

4.1 Preliminaries

The peering policy presented in Section 3.4 is a crucial aspect of our proposal. To make this fundamental decision, we had to first consider three preliminary questions:

By what criteria do we make this fundamental decision? Here we take an economic perspective and focus on two goals: maximizing social welfare and fostering competition. Social welfare is the total utility (which for us will be the sum of utilities over all users); it ignores payments made by users because payments merely transfer some of that utility to others, and social welfare does not measure how utility is distributed within society, it merely measures the total. To be clear, we also care about the distribution but, as we will note later, if we succeed in fostering competition then the distributional issues will work themselves out.

As for fostering competition, we want an environment where the large incumbents do not have unfair advantages over new entrants in the market (where the operational definition of unfair will become clear later in this section). Fair competition is what allows new and innovative CSPs and LMPs to gain a foothold in the market, which in turn (because of their innovation which hopefully leads to better services that create more user utility) can lead to increases in future social welfare. So in some sense our goals are both about social welfare; one focuses on the current social welfare and the other on the future social welfare.

How do we model the economic interactions? To determine which peering policies would best achieve our goals, we must analyze a simple model that attempts to capture the relevant economic interactions. For issues as complicated as commercial transactions in the Internet ecosystem, *any* model we consider will be oversimplified. Choosing a good model is more art than science, in that the model must be both *tractable* (so we can derive results) and *representative* (yielding results that are suggestive of what would happen in the real world). The former is easy to characterize, but the latter is a matter of judgement.

What peering policies do we consider? We obviously cannot consider all possible policies that could guide how LMPs interact. For instance, one LMP might provide different levels of service (*i.e.,* QoS) to traffic from various other LMPs, or enter into joint marketing agreements, or collude to raise prices in a geographic area. Here, we restrict ourselves to two possible forms of peering: (1) freely peering, where every LMP accepts incoming traffic from all other LMPs and CSPs without any termination fees and (2) for-fee peering, where LMPs can charge termination fees to CSPs, or block their traffic if they refuse to pay. In both cases, we assume no traffic discrimination (by which we mean prioritization based on source, rather than different QoS levels which are charged differently independent of source). Thus, we want to consider two scenarios, one where network neutrality (NN) is contractually enforced by the POC, and an unregulated (UR) scenario where LMPs are free to impose termination fees and block traffic if they are not paid. While this may be limiting, it does yield conclusions that intuitively (but not quantitatively) apply to traffic discrimination in that imposing poor QoS on incoming traffic reduces the value of that traffic to users, so it can be seen as a form of termination fee.

4.2 Our Model

We consider a model that only contains the POC and the LMPs and CSPs directly connected to it. For ease of exposition we assume all LMPs are eyeball networks (this does not alter our analysis); the directly attached CSPs sell services to the users of the LMPs. We assume that the CSPs have no marginal costs for adding customers (as these are online services, and the costs are small enough to ignore in our simple model). In the NN scenario, the LMPs freely accept traffic from all CSPs. In the UR scenario, they can impose a termination fee t on a CSP or block its traffic.

There is a complicated competitive process between LMPs, who face high capital costs to enter the market, and lower marginal costs; these conditions are a recipe for natural monopolies, but even so some competition survives. Rather than model this directly, which would necessarily fall short of reality, we merely assume that the competition between LMPs results in one or more LMPs for each region, and that customers have chosen a single LMP for their connectivity. We assume this choice changes slowly, so that in the short term each LMP is the monopoly connectivity provider for its customers. Thus, while LMPs may coexist in a region, and the competition between them involves many different factors (cost structure, bandwidth, SLAs, etc.), we assume that this competition results (over the short term) in a static partitioning of customers among LMPs.

Similarly, there is a complicated market for content and services on the Internet, with all the products competing for user attention and with many products serving as partial substitutes for others. Again, instead of modeling this competition directly, which would be impossible to do with any accuracy, we merely consider a set of independent goods which are not substitutes for each other. This is unrealistic, but gives the best case for termination fees (*i.e.,* it minimizes their harm): if the CSP market is fully competitive, then LMPs have all the market power (and can set the termination fees to maximize their own revenue, which we model in Section 4.4), whereas if the goods are not competitive the market power of the LMPs is limited so, as we describe in Section 4.5, the setting of the termination fees can be modeled as a negotiation. In addition, we assume that each CSP charges a uniform price for their service independent of a user's LMP, and that the distribution of demand for a CSP is the same for customers of each LMP.

To express this model mathematically, we consider a mass of consumers, most conveniently described as a unit mass on the unit interval. Each consumer has a choice among a variety of CSPs, $s = 1, ..., S$. The value that a specific consumer attaches to service s is given by v_s, also called its "willingness to pay". The demand for each CSP is determined by the cumulative distribution of these values

in the population, denoted by $F_s(v_s)$, which quantifies the fraction of consumers with willingness to pay less than v_s. We assume that $F_s(v_s)$ also describes the distribution of demands of the customer populations in each LMP.

Any consumer who has a value v_s weakly larger than the posted price p_s will buy that CSP's service. The demand D_s at price p_s for the service s is therefore

$$D_s(p_s) = 1 - F_s(p_s),$$

which is monotone decreasing in the price p_s.

Similarly, there are a variety of LMPs, $l = 1, ..., L$. Each consumer only has a single LMP. Each LMP l charges a price c_l to their customers. In what follows, we ignore any welfare derived from merely having connectivity, and only focus on welfare arising from the CSPs.

4.3 Network Neutrality

We begin our analysis by assuming that we are in the network neutrality (NN) regime where no termination fees are allowed. Being connected to the network, each consumer can choose to purchase as many of the CSP services as she deems valuable. Since we have assumed the CSP products are not substitutes for each other, each CSP can set p_s to maximize its revenue:

$$p_s^* = \underset{p_s}{\operatorname{argmax}} \{p_s D_s(p_s)\}$$

In the NN regime, this is the end of the story: LMPs have their customers, CSPs set their prices to maximize revenue, and there are no complications. The resulting social welfare (the sum over user utilities) is merely:

$$\sum_s \int_{p_s^*}^{\infty} v_s dF(v_s)$$

Note that the social welfare is monotonically decreasing in the prices p_s^*; every increase in price p_s potentially causes some consumers to not purchase the service s.

4.4 LMPs unilaterally set fees

We now turn to the unregulated (UR) scenario where termination fees are allowed. We consider two possible ways these fees can be set: unilaterally (in this subsection) and through bargaining (in the next subsection).

We have assumed that, in the short term, each LMP is a monopoly provider for its users. One way of modeling these fees is to assume that each LMP can unilaterally set the fees for each CSP to reach its customers (since, in the short term, there is no other way of reaching them). This fee-setting behavior of the LMP results in the so-called "double marginalization" process [49]. Charged a fee t_s per customer, CSP s chooses a revenue-maximizing price $p_s^*(t_s)$. Given that the revenue per customer is now $p_s - t_s$, the revenue-maximizing price is given by:

$$p_s^*(t_s) = \underset{p_s}{\operatorname{argmax}} \{(p_s - t_s) D_s(p_s)\} \quad (1)$$

Note that with sufficient smoothness and convexity conditions, the maximizing price $p_s^*(t_s)$ can be shown to be strictly increasing in t_s.

LEMMA 1. *If $D_s(p_s)$ is strictly positive with continuous first and second derivatives, is strictly decreasing ($D_s'(p_s) < 0$), is strictly convex ($D_s''(p_s) > 0$), and asymptotically vanishes ($\lim_{p_s \to \infty} D_s(p_s) = 0$), then $p_s(t_s)$ is monotonically increasing in t_s: $p_s'(t_s) > 0$.*

Proof: Since $p_s(t_s)$ maximizes $(p_s - t_s)D_s(p_s)$ it must satisfy the equations (i) $p_s(t_s) > t_s$, (ii) $D_s(p_s(t_s)) + (p_s(t_s) - t_s)D'(p_s(t_s)) = 0$, and (iii) $2D_s'(p_s(t_s)) + (p_s(t_s) - t_s)D''(p_s(t_s)) \leq 0$. Taking the derivative of equation (ii) and rearranging yields

$$p_s'(t_s)[2D'(p_s(t_s)) + (p_s(t_s) - t_s)D_s''(p_s(t_s))] = D_s'(p_s(t_s))$$

Equation (iii) tells us that the left hand bracket is negative, and we know the right hand side is negative, so $p_s'(t_s)$ must be positive. ∎

Thus, as the termination fees t_s increase, the prices p_s increase, so the social welfare decreases. We can therefore conclude that termination fees strictly decrease social welfare.

Returning to the unilateral scenario, knowing how CSP s will set its price, the LMP chooses the fee t_s to maximize its revenue:

$$t_s^* = \underset{t_s}{\operatorname{argmax}} \left\{ t_s D_s\left(p_s^*(t_s)\right) \right\}.$$

Of course, each LMP independently chooses the fee that they charge, but they all do the same calculation (for each CSP), so the result is uniform termination fees t_s across all LMPs. This process is referred to as "double marginalization" because the CSP and then the LMP are maximizing revenue in sequence.

4.5 Bilateral Bargaining

Modeling the LMP as imposing fees unilaterally on every content provider neglects the fact that both the LMP and the CSP may have some degree of bargaining power. The LMP can ask for a termination fee t_s or else they will block the CSP. The CSP can threaten to walk away from the deal, leaving the LMP's customers without the services offered by that CSP. This situation is similar to many other bilateral monopolies where a bilateral bargaining approach has provided sharp insights; for example, cable providers bargain with content providers over how much they pay to show a particular channel (*e.g.,* Comcast and ESPN have to negotiate the fees that Comcast pays to ESPN, see [14]). Another situation is the negotiation between health insurers and local hospitals, where the insurance company acts as a gatekeeper between hospital and patient, see [29]. In such settings, two parties negotiate and eventually arrive at an agreement.

Rather than explicitly modelling the entire extensive form of the game that represents the complex strategic environment involved in such negotiations, the economic analysis has typically adopted a cooperative solution concept that satisfies a number of axiomatic requirements. We shall follow this approach here and adopt the Nash bargaining solution (NBS) [41]. This cooperative solution concept can be given non-cooperative and fully strategic foundations, as established by [46]. More relevant for our purposes here, the Nash bargaining solution can be fully extended to bargaining environments with many participants and externalities across environment and agreements, as established recently in [12]. For the present purpose, it will suffice to restrict attention to the bilateral bargaining solution.

The NBS can be defined as follows. There are two agents with utility functions u and v, and a feasible set of outcomes F and a disagreement point d that represents the outcome when the players do not come to agreement. Then the NBS outcome is the outcome that maximizes the product of the utility differences between agreement

and disagreement:

$$\underset{x \in F}{\operatorname{argmax}} \{(u(x) - u(d))(v(x) - v(d))\}.$$

We consider a series of three models using NBS, with increasing levels of complexity. First, we consider a single CSP s and a single LMP l who are bargaining over the termination fee. Since this is a bilateral negotiation, it does not affect the fees t_s being charged by other LMPs, so we assume that CSP s keeps its price p_s fixed regardless of the outcome of this negotiation (recall that we assume CSPs charge global prices that do not depend on the LMP of the user). In what follows, we focus only on the revenue per customer in LMP l. If they come to agreement on a fee t_s, then s obtains $D_s(p_s)(p_s - t_s)$ and l obtains $D_s(p_s)t_s$. If they disagree, then s gets no revenue from customers of l, and l loses some fraction r_l^s of customers who used to also be customers of s (those who were never customers of s are presumably unaffected by the fact that s is no longer offered on l's network), and recall that these customers were paying an access charge c_l to the LMP. The loss s suffers at the disagreement point is $D_s(p_s)(p_s - t_s))$ and the loss suffered by l is $D_s(p_s)(t_s + r_l^s c_l)$.

Thus, the quantity the NBS maximizes is:

$$[D_s(p_s)(p_s - t_s)][D_s(p_s)(t_s + r_l^s c_l)]$$

Taking the derivative with respect to t_s of this expression and setting the result to zero shows that the transfer payment that maximizes the product of the gains from agreement is:

$$t_s = \frac{p_s - r_l^s c_l}{2},$$

We therefore take this as the negotiated fee t_s. Note that the fee is decreasing in the rate r_l^s at which customers leave l when negotiations with service s break down. Moreover, the fee can be negative (l pays s) when the loss suffered at the disagreement point by l is greater than the loss suffered by s. However, in what follows we assume we are in the regime where the termination fees are positive.

The key parameter here is r_l^s, which is the rate at which the LMP l loses customers when s is no longer offered on its network. For a given s, r_l^s will presumably be smaller if l is a well-established incumbent than if it is a newly established LMP with a smaller market share. This means that well-established LMPs can extract more in termination fees than smaller ones, giving them a substantial competitive advantage.

Similarly, for a given l, r_l^s will presumably be larger if s is a well-established CSP than if it is a newly established one. This again gives a significant competitive advantage to CSPs with large market share, because they can pay less in termination fees.

While the exact values of r_l^s and the specific nature of the demand curves D_s are empirical matters that will determine the quantitative impact of allowing termination fees, it is clear that such fees will systematically favor established incumbents in both the LMP and CSP markets.

The above model focused only on one bilateral negotiation, but all the LMPs will want to extract what they can from each CSP. In our second model applying the bargaining approach, we account for the presence of these other fees. We find that the weighted average fee t_s^{ave} (normalized by number of customers n_l^s of s in each LMP l) charged to service s is given by:

$$t_s^{ave} = \frac{p_s - < rc >_s}{2}.$$

where $< rc >_s = \frac{\sum_l n_l^s r_l^s c_l}{\sum_l n_l^s}$ is the average of $r_l^s c_l$ over all l weighted by population. Thus, our previous result about bilateral negotiation applies even when all LMPs are charging fees.

In our third and final model based on bargaining, we note that when faced with a set of termination fees from all LMPs, each CSP s will modify its price p_s so as to maximize revenue given these fees. This revenue-maximizing price is given by $p_s^*(t_s^{ave})$ in Equation 1. After changing the price p_s the termination fees t_s will be renegotiated, and so on. Based on our earlier analysis of NBS, we eventually reach an equilibrium where the following equation holds:

$$t_s^{ave} = \frac{p_s^*(t_s^{ave}) - < rc >_s}{2}.$$

The core result here is that when we allow termination fees, the prices p_s increase with the imposed fees, which then decreases social welfare. While the price increase (due to termination fees) faced by the consumer will likely be less under bilateral bargaining than under unilateral fee setting, it will still result in a lower social welfare than the NN case. More importantly, incumbent CSPs and ISPs will have a significant competitive advantage over emerging ones: incumbent ISPs can negotiate higher termination fees than emerging ones, and incumbent CSPs can negotiate smaller termination fees than emerging ones. This is the reason we strongly favor the network neutrality regime over the unregulated scenario.

These conclusions are intuitive, but we are not aware of a result in the literature that derives them cleanly from such a simple model.

4.6 Related Literature on Network Neutrality

While the past two decades has seen an active debate about net neutrality, the economics literature on net neutrality is still in development, see [25, 34] for quick introductions and [3, 16, 54] for further reading. In particular, we lack a solid empirical understanding of the key economic trade-offs being made by LMPs and CSPs, so most of the theoretical investigations involve simple models similar to what we have just presented.

The network neutrality literature is too broad and varied to give a complete review of it here, so instead we review the aspects where our treatment differs from major portions of the general literature. We start by describing ways in which our general approach differs, and then describe some more technical differences.

Our first significant difference from a major portion of the network neutrality literature is that we focus on maximizing social welfare, whereas optimizing consumer welfare is sometimes used in arguments for network neutrality [34]. Social welfare is (in our setting) the total utility users derive from network services, ignoring the payments they have made for those services (since those payments just increase the utility of others), while consumer welfare takes those payments into account because it focuses on the net welfare (utility minus payments) of the users.

Thus, we are ignoring the distribution of that welfare (between users and CSPs and LMPs). We do this because we view innovation as the most important way to grow social welfare in the long-term, and enabling a more competitive market does that. As a byproduct, vigorous competition in the LMP and CSP market tends to drive

most of the value into consumer welfare (since payments decrease). Thus, while NN has both higher social welfare and fairer competition than UR, we view the latter as more important than the former.

Our second important difference is that many works on network neutrality focus on service differentiation, whereas we only consider termination fees. We view termination fees as the most fundamental violation of network neutrality, whereas (as mentioned earlier) offering different qualities of service, if done on an open basis with posted prices, is a very different matter which we do not address (and take no stand against).

In addition, there are concerns that ex ante network neutrality regulation (*i.e.,* regulations imposed before significant violations occur) might interfere with the Internet's evolution [20]. Because we do not take a hard stance against QoS, we are not limiting future design choices, only specific revenue mechanisms. Moreover, once termination fees start being imposed it will be hard to claw them back, so ex ante regulation seems to be exactly what is called for.

Also, it has sometimes been argued that the last mile provider needs to share (via termination fees) in the profits of the content provider to maintain adequate bandwidth. But this argument neglects two important points. First, such termination fees can be applied even in the absence of bandwidth needs, so the solution (termination fees) seems ill-suited to solve this particular problem (incentives for bandwidth expansion). Second, the last mile provider can, even in the presence of network neutrality, charge its consumer for the traffic volume. See [40] for recent work that establishes how better adapted pricing policies by the last mile providers can substantially improve the usage of broadband networks.

At a more detailed level, two significant ways we differ from the existing literature is that (i) we do not model CSP-CSP or ISP-ISP competition at all (as opposed to, for example, [6]); and (ii) we model the CSP-ISP interaction as a Nash bargaining problem rather than as a non-cooperative game with a Nash equilibrium (which is the dominant approach in the literature). The former decision is because Nash equilibria of toy models of the CSP and ISP markets often depend on small details in the modeling, and thus are not sufficiently representative; moreover, we don't see such competition as changing the essence of our results here (though it may change the values of the parameters r_l^s). For the latter decision about modeling CSP-ISP interactions via NBS, the closest paper to ours in this spirit is [1] but in that work the bargaining power was given exogenously, where here it was derived from the nature of the disagreement point.

5 *Are We Crazy?*

We end by considering two questions.

Is major change needed? The problems we are addressing all stem from a single cause: the Internet was not designed with a business model in mind. The peering practices that emerged from the early days of the commercial Internet were never able to produce revenue flows that were tightly tied to the way value is delivered and cost is incurred in the modern Internet. The resulting debate about network neutrality was doomed to intellectual failure, because it operated within a model where value and revenue would always be somewhat disconnected. We approached this problem more fundamentally by asking how could we redesign the infrastructure so that the value and revenue chains could be cleanly aligned.

The motivation for our approach is quite simple. Today's large ISPs combine the tasks of transit and last-mile-delivery, and often play the role of both CSP and ISP. There is no way the Internet's current transitive method of passing on traffic and recouping cost could possible encompass this kind of complexity. Our alternative is to cleanly separate transit and last-mile-delivery, and then have the LMPs and CSPs turn directly to their own customers to recoup costs and extract revenue from the value delivered.

We are proposing this as the fundamental *economic architecture* of the Internet, one that we think can continue to apply even under massive changes in the services provided, the content shared, and the resulting costs incurred.

Is such a change possible? We first note that while the POC is radically different from the status quo, it is incrementally deployable. We are not engaging in clean-slate fantasies here, but instead hoping that a radically different way of structuring the Internet could start off almost as a demonstration project, and then grow over time into a true alternative that competes with the current Internet infrastructure. If more and more LMPs find the POC attractive (and the success of IXPs in Europe suggest that this is not farfetched) then over time the POC can have a substantial impact on the Internet.

But does the POC have a chance of becoming real? From a straight business perspective, it is a dubious proposition, since it would take years to gain the confidence of LMPs, who would be risking their own financial future on the fate of the POC. However, there is another strong force at work here, one that is captured in the phrase "Commoditize Your Complement". This is a strategy first identified by Joel Spolsky [50] that we now briefly explain. In today's technology marketplace, delivering value to consumers typically requires several different products to work together: for instance, displaying a movie on a phone requires the mobile device itself, the cellular infrastructure, the Internet infrastructure, and the movie service. All these products are *complements* of each other. Spolsky's insight [50] was simple: "Demand for a product increases when the prices of its complements decrease." Simply put, a consumer is willing to pay a certain amount to watch movies on her phone; as the other components of this process become cheaper the remaining components can charge more.

Currently the ISPs and the CSPs are complements of each other in a hugely lucrative business, and the network neutrality debates are nothing more than an argument about how to share the proceeds. The POC would be a way for the CSPs to commoditize the ISP market, by explicitly turning transit into a nonprofit and creating a competitive market in LMPs. As such, the success of the POC would be extremely valuable to the large CSPs. They could easily provide the initial set of leased links, along with some guarantees for the POC's longevity, which would supply the much-needed credibility.

While this does not answer the question about our sanity, we hope it provides some hope for the Internet's future stability.

Acknowledgements

We would like to thank our shepherd David Wetherall, our anonymous referees, and Peter Cramton for helpful comments. We also thank Lloyd Brown and Ian Rodney, UC Berkeley students who helped bring this paper to fruition. We gratefully acknowledge financial support from Intel, VMware, and Ericsson, as well as from NSF grant 1817115.

References

[1] E. Altman, M. K. Hanawal, and R. Sundaresan. Regulation of off-network pricing in a nonneutral network. *ACM Trans. Internet Techn.*, 14(2-3):11:1–11:21, 2014.

[2] Aryaka. The Cloud-First WAN. https://www.aryaka.com/.

[3] G. S. Becker, D. W. Carlton, and H. S. Sider. Net neutrality and consumer welfare. *Journal of Competition Law and Economics*, 6(3):497–519, 2010.

[4] BEREC. All you need to know about Net Neutrality rules in the EU. https://berec.europa.eu/eng/netneutrality/, 2020.

[5] T. Böttger, G. Antichi, E. L. Fernandes, R. di Lallo, M. Bruyere, S. Uhlig, G. Tyson, and I. Castro. Shaping the internet: 10 years of ixp growth. *arXiv preprint arXiv:1810.10963*, 2018.

[6] M. Bourreau, F. Kourandi, and T. Valletti. Net Neutrality with Competing Internet Platforms. CEIS Research Paper 307, Tor Vergata University, CEIS, Feb. 2014.

[7] Cato Networks. Global Private Backbone. https://www.catonetworks.com/cato-cloud/global-private-backbone-3/.

[8] Y.-C. Chiu, B. Schlinker, A. B. Radhakrishnan, E. Katz-Bassett, and R. Govindan. Are we one hop away from a better internet? In *IMC*, 2015.

[9] Y.-C. Chiu, B. Schlinker, A. B. Radhakrishnan, E. Katz-Bassett, and R. Govindan. Are we one hop away from a better internet? In *Internet Measurement Conference*, 2015.

[10] D. D. Clark, J. Wroclawski, K. R. Sollins, and R. Braden. Tussle in cyberspace: defining tomorrow's internet. In *SIGCOMM 2002*, 2002.

[11] E. H. Clarke. Multipart pricing of public goods. *Public choice*, pages 17–33, 1971.

[12] A. Collard-Wexler, G. Gowrisankaran, and R. Lee. "nash-in-nash" bargaining: A microfoundation for applied work. *Journal of Political Economy*, 127:163–195, 2019.

[13] P. Cramton. Electricity market design. *Oxford Review of Economic Policy*, 33(4):589–612, Nov 2017.

[14] G. Crawford, R. Lee, M. Whinston, and A. Yurukoglu. The welfare effects of vertical integration in multichannel television markets. *Econometrica*, 86:891–954, 2018.

[15] J. Dunn. Trump's new FCC boss could make it easier for Internet providers to play favorites. *Business Insider*, Jan 2017.

[16] N. Economides. Economic Features of the Internet and Network Neutrality . In *The Oxford Handbook of the Economics of Networks*. Oxford University Press, 2016.

[17] N. Economides and J. Tåg. Network neutrality on the internet: A two-sided market analysis. *Information Economics and Policy*, 24:91–104, 2012.

[18] M. C. Erickson, E. Stallman, D. J. Kalt, A. W. Guhr, C. Libertelli, and C. Wright. Petition to deny of Netflix Inc. Web, Aug 2014. https://ecfsapi.fcc.gov/file/7521819696.pdf.

[19] P. Faratin, D. D. Clark, S. Bauer, W. Lehr, P. W. Gilmore, and A. Berger. The growing complexity of Internet interconnection. *Communications & strategies*, page 51, 2008.

[20] G. R. Faulhaber. Economics of net neutrality: A review. *Communications & Convergence Review*, 3(1):53–64, 2011.

[21] Federal Communication Commission. Internet Policy Statement. https://docs.fcc.gov/public/attachments/FCC-05-150A1.pdf, Sep 2005.

[22] J. Feigenbaum, C. H. Papadimitriou, R. Sami, and S. Shenker. A BGP-based mechanism for lowest-cost routing. *Distributed Computing*, 18(1):61–72, 2005.

[23] B. M. Frischmann and B. Van Schewick. Network neutrality and the economics of an information superhighway: A reply to professor yoo. *Jurimetrics*, pages 383–428, 2007.

[24] Google. Google Fiber. https://fiber.google.com/.

[25] S. Greenstein, M. Peitz, and T. Valletti. Net neutrality: A fast lane to understanding the trade-offs. *Journal of Economic Perspectives*, 30:127–150, 2016.

[26] T. Groves. Incentives in teams. *Econometrica: Journal of the Econometric Society*, pages 617–631, 1973.

[27] GSMA. Comparison of fixed and mobile cost structures. https://www.gsma.com/publicpolicy/wp-content/uploads/2012/09/Tax-Comparison-of-fixed-and-mobile-cost-structures.pdf, 2012.

[28] R. W. Hahn and S. J. Wallsten. The economics of net neutrality. *The Economist's Voice*, June 2006.

[29] K. Ho and R. Lee. Insurer competition in health care markets. *Econometrica*, 85:379–417, 2017.

[30] A. Hortaçsu, J. Kastl, and A. Zhang. Bid shading and bidder surplus in the us treasury auction system. *American Economic Review*, 108(1):147–169, Jan 2018.

[31] G. Huston. The death of transit? Web, Oct 2016. https://blog.apnic.net/2016/10/28/the-death-of-transit/.

[32] Industry Analysis Division, Federal Communications Comission. Internet Access Services: Status as of December 31, 2017. https://docs.fcc.gov/public/attachments/DOC-359342A1.pdf, Dec 2017.

[33] S. Knight, H. X. Nguyen, N. J. G. Falkner, R. A. Bowden, and M. Roughan. The internet topology zoo. *IEEE Journal on Selected Areas in Communications*, 29:1765–1775, 2011.

[34] J. Krämer, L. Wiewiorra, and C. Weinhardt. Net neutrality: A progress report. *Telecommunications Policy*, 37(9):794–813, 2013.

[35] P. Lambert. Vodafone and Telefonica are overplaying their hand with Google. https://telecoms.com/opinion/vodafone-and-telefonica-are-overplaying-their-hand-with-google/, Feb 2010.

[36] R. S. Lee and T. Wu. Subsidizing creativity through network design: Zero-pricing and net neutrality. *Journal of Economic Perspectives*, 23(3):61–76, 2009.

[37] F. Li, A. A. Niaki, D. R. Choffnes, P. Gill, and A. Mislove. A large-scale analysis of deployed traffic differentiation practices. In *SIGCOMM*, 2019.

[38] N. C. Ltd. About NBN Co. Web, accessed Jun 2020. https://www.nbnco.com.au/corporate-information/about-nbn-co.

[39] R. T. B. Ma and V. Misra. The public option: A nonregulatory alternative to network neutrality. *IEEE/ACM Trans. Netw.*, 21(6):1866–1879, Dec 2013.

[40] J. Malone, A. Nevo, and J. Williams. The tragedy of the last mile: Economic solutions to congestion in broadband networks. *NET Institute Working Paper*, 2017.

[41] J. Nash. The bargaining problem. *Econometrica*, 18:155–162, 1950.

[42] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*. Cambridge University Press, USA, 2007.

[43] O'Connell, P. Online Extra: At SBC, It's All About "Scale and Scope". *Bloomberg Businessweek*, 11 2005.

[44] Omitted for Double Blind. Personal communication, 2019.

[45] A. Parker and R. Waters. Google Accused of YouTube "Free Ride". *Financial Times*, April 2010.

[46] A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50:97–109, 1982.

[47] G. Schneibel and C. Farivar. Deutsche Telekom moves against Apple, Google and net neutrality. DW, https://www.dw.com/en/deutsche-telekom-moves-against-apple-google-and-net-neutrality/a-5439525, Apr 2010.

[48] S. Segan. Exclusive: Check Out the Terrible State of US ISP Competition. *PCMag UK*, Dec 2017.

[49] J. J. Spengler. Vertical integration and antitrust policy. *Journal of Political Economy*, 58(4):347–352, 1950.

[50] J. Spolsky. Strategy Letter V. Joel on Software, https://www.joelonsoftware.com/2002/06/12/strategy-letter-v/, June 2002.

[51] J. Taylor. NBN amendments clarify cherry-picking. Web, Mar 2011. https://www.zdnet.com/article/nbn-amendments-clarify-cherry-picking/.

[52] J. Taylor. Senate passes NBN bills with amendments. Web, Mar 2011. https://www.zdnet.com/article/senate-passes-nbn-bills-with-amendments/.

[53] TeleGeography. The state of the network 2020 edition. Web, 2020. https://www2.telegeography.com/hubfs/assets/Ebooks/state-of-the-network-2020.pdf.

[54] B. van Schewick. *Internet architecture and innovation*. MIT Press, 2012.

[55] W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance*, 16(1):8–37, 1961.

[56] Wikipedia contributors. Net neutrality. https://en.wikipedia.org/wiki/Net_neutrality, 2020.

[57] T. Wu. Network neutrality, broadband discrimination. *J. on Telecomm. & High Tech. L.*, 2:141, 2003.

[58] C. S. Yoo. Beyond network neutrality. *Harvard Journal of Law & Technology*, 19:1, 2005.