Combining multihoming with overlay routing
(or, how to be a better ISP without owning a network)

Yong Zhu, Constantine Dovrolis, and Mostafa Ammar
Networking and Telecommunications Group
Georgia Institute of Technology
Email: {yongzhu, dovrolis, ammar}@cc.gatech.edu

***Abstract*— Multihoming and overlay routing are used, mostly separately, to bypass Internet outages, congested links and long routes. In this paper, we examine a scenario in which multihoming and overlay routing are jointly used. Specifically, we assume that an Overlay Service Provider (OSP) aims to offer its customers the combined benefits of multihoming and overlay routing, in terms of improved performance, availability and reduced cost, through a network of multihomed overlay routers. We focus on the corresponding design problem, *i.e.,* where to place the overlay routers and how to select the upstream ISPs for each router, with the objective to maximize the profit of the OSP. We examine, with realistic network performance and pricing data, whether the OSP can provide a network service that is profitable, better (in terms of round-trip time), and less expensive than the competing native ISPs. Perhaps surprisingly, we find out that the OSP can meet all three objectives at the same time. We also show that the MON design process is crucial. For example, operating more than 10 overlay nodes or routing traffic through the minimum-delay overlay path, rarely leads to profitability in our simulations.**

I. Introduction

The most basic form of Internet access is *singlehoming*, where a stub network uses a single upstream ISP to reach all destinations. It has been shown that singlehoming can lead to poor availability and performance [1]. The single route from the source network to a destination network/prefix may not be always available, while routing policies and traffic engineering practices can (and often do) impose a heavy performance penalty on the resulting end-to-end performance [2].

To achieve improved reliability and performance, *multihoming* has become the mainstream service model for major content providers (see Figure 1(a)). In the more advanced form of this model, known as *intelligent route control*, the multihomed source network selects the upstream ISP for every significant destination prefix based on performance and cost considerations [3], [4].

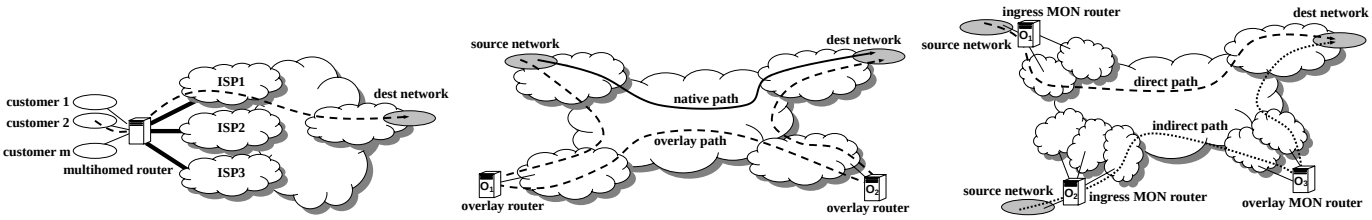
Another approach to improve end-to-end availability and performance is *overlay routing* (see Figure 1(b)). Here, the traffic between two networks is sent through one or more intermediate overlay nodes that are connected through IP tunnels [1], [5]. The advantage of overlay routing is that it typically provides a greater number of diverse paths to reach a destination network compared to the typical case of multihoming to 2-4 upstream ISPs [6]. On the other hand, overlay routing requires the deployment of a distributed routing/forwarding overlay infrastructure. A comparison of multihoming and overlay routing has been conducted in [7].

Given the previous two approaches, it is interesting to consider a scenario in which both models are used. Specifically, we envision a new type of Internet provider referred to as *Overlay Service Provider (OSP)* (to distinguish from an ISP) that attempts to offer its customers the combined benefits of multihoming and overlay routing in terms of improved performance, availability and reduced cost. The OSP operates a *Multihomed Overlay Network (MON)*, with each MON node being a multihomed router. MON nodes are placed at "key" Internet locations, mostly Internet Exchange Points (IXPs), and the OSP purchases Internet connectivity for each MON node from several locally present ISPs. An OSP customer can connect directly to a MON node if the former is collocated at the same IXP with that MON node. Major content providers are usually collocated at major IXPs to avoid the cost of leased lines. On the other hand, the OSP is responsible to route a customer's traffic with greater availability and higher performance than the customer's current *native* ISP. Note that this is similar to the InterNAP service and business model [3].

Furthermore, we envision that the OSP performs overlay routing, utilizing MON nodes as overlay routers. Based on the findings of [8], we limit the number of intermediate MON nodes in an overlay path to one. It is rarely the case that more intermediate nodes are needed to improve performance or availability significantly. Figure 1(c) shows that the MON network can utilize multihoming to form *direct paths* and overlay routing to form *indirect paths*. If a MON node is multihomed to K ISPs, there are K direct MON paths to choose from for each flow. With N MON nodes, the number of indirect MON paths for each flow increases to $K^2(N-1)$.

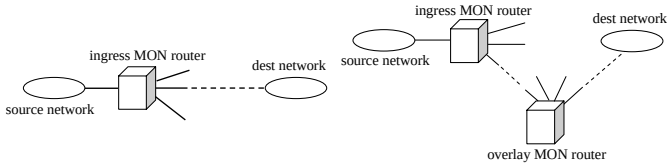
It is interesting that an OSP is "an Internet provider that does not own a network", in the sense that the OSP does not operate any long-distance links or a backbone. Its infrastructure is located at the network edges, and its long-distance communications are conducted from the underlying native-layer ISPs. In fact, early ISPs were often built in the same way, leasing long-distance trunks from telecommunication providers and placing IP routers at aggregation points at the network edge.

In this paper, we focus on the *MON design problem*, *i.e.,* where to place MON nodes and how to select the upstream ISPs for each node. We aim to examine, with realistic network performance and pricing data, whether an OSP can combine



(a) Multihoming (b) Overlay routing (c) MON architecture

Fig. 1. Multihoming, overlay routing, and the Multihomed Overlay Network (MON) architecture.



(a) Direct MON path (b) Indirect MON path

Fig. 2. Direct and indirect MON paths.

multihoming and overlay routing to provide a network service that is, first, profitable, second, better in terms of performance than the competing native ISPs, and third, less expensive than the competing native ISPs.[1] Perhaps surprisingly, we find out that the OSP can meet all three objectives. We also show, however, that the MON design process is crucial. For example, operating more than 10 MON nodes or routing traffic through the minimum-delay MON path, rarely leads to profitability in our simulations.

In more detail, we formulate an optimization problem where the OSP aims to maximize its profit by placing up to N MON nodes and connecting each node with up to K locally present ISPs. The OSP revenues come from subscribed customers while the costs are due to leased upstream capacity and node deployment. The optimization is constrained because a potential customer will only subscribe to the OSP if the latter can offer better performance than the competing native ISP, at least for a large fraction of the customer's traffic. As in any network design problem, we focus on large timescales, namely weeks or months. The reason is that both the deployment of MON nodes and the contractual agreements with native ISPs are hard to change in shorter timescales. Consequently, the performance metric we consider is the propagation delay between MON nodes. Other metrics, such as loss rate or available bandwidth, vary significantly in shorter timescales and so they would not be appropriate as inputs to a network design problem.

The rest of the paper is organized as follows. Section II presents our model and formalizes the MON design problem; we also prove that the problem is NP-hard. The MON design (and performance evaluation) requires a way to model native-layer propagation delays; we develop such a model in Section III. Section IV presents four MON design heuristics with different input requirements. Section V evaluates the previous heuristics under realistic network settings and pricing data, and examines the performance and profitability of the OSP. Section VI discusses related work and the paper concludes in Section VII.

[1]In terms of availability, we assume that the OSP can take advantage of its multihoming and overlay routing capabilities to provide higher availability than singlehoming or just multihoming.

II. Model and problem formulation

The MON design problem involves ISPs and the performance of the native network, the location and traffic matrix of potential customers, and the OSP routing strategy, pricing function and node deployment costs. In this section, we present a model for these components of the problem and formulate a MON design optimization framework. We also prove that the optimal MON design problem is NP-hard.

A. ISPs and the native network

Consider a geographical area that the OSP aims to cover. There are L possible locations where the OSP can place MON nodes (*e.g.,* IXP locations or network access points). The set of ISPs that are present at location $l \in L$ is denoted by I_l, while the union of all such sets is I. We use the term *POP* $p = (l, i)$ to identify the access point to ISP i at location l. $\mathcal{LOC}(p)$ and $\mathcal{ISP}(p)$ are the location and ISP that correspond to POP p, respectively. The set of all POPs is denoted by P.

We represent the native-layer performance with the matrix $T_{|P|\times|P|}$, where the entry $\tau_{p,q}$ represents the propagation Round-Trip Time (RTT) from POP p to q. We expect that most of the elements in this matrix remain practically constant for weeks, except during periods of interdomain routing instability. The matrix T can be measured directly, as long as the OSP can conduct simple measurements (*e.g.,* ping) between any pairs of POPs. If that is not possible, the matrix T can be estimated using the technique presented in Section III.

B. MON representation

MON consists of up to N nodes, with each node placed at a different location of L. Each MON node is multihomed to at most K locally present ISPs. We say that *a MON node is present at POP* p if the node is located at $\mathcal{LOC}(p)$ and connected to ISP $\mathcal{ISP}(p)$. The entire MON network can be represented with the *POP selection vector* $\mathcal{MON}$,

$$\mathcal{MON}(p) = \begin{cases} 1 & \text{if a MON node is present at POP } p \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Given a POP selection vector, we can identify the locations of all MON nodes with:

$$\mathcal{NODE}(l) = \begin{cases} 1 & \text{if } \sum_{\mathcal{LOC}(p)=l} \mathcal{MON}(p) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

C. Customers and OSP-preferred flows

We denote the set of potential OSP customers as U. A customer u is present at location $\mathcal{LOC}(u)$ and, before subscribing to the OSP, is connected to one or more locally present ISPs. To attract a customer, the OSP needs to provide better performance (lower RTT) to most of that customer's traffic. Specifically, the workload of customer u is a set of *flows* $F(u)$. A flow $f = (s_f, d_f, r_f, \tau_f)$ is defined as a large traffic aggregate from one of u's source POPs towards a destination POP, where s_f and d_f are the flow's source and destination POPs, respectively, r_f is the flow's average rate, and τ_f is the RTT in the flow's native path. The set of all flows is F. The flow f is *OSP-preferred* if the OSP can route f, through a direct or indirect path, with RTT $\hat{\tau}_f < \tau_f$. The corresponding MON path is referred to as *OSP-preferred* path for flow f. Customer u *subscribes* to the OSP if at least a fraction H (say 70-80%) of its traffic is in OSP-preferred flows.

D. OSP routing strategy

As mentioned earlier (and shown in Figure 2), an OSP can utilize either one of the direct paths from the ingress MON node to the destination, or one of the indirect paths through an intermediate MON node. The OSP uses a certain *routing strategy* to select the path to each destination. One routing strategy is that the OSP always selects the path, direct or indirect, with the minimum native RTT.

In most of this paper (except for Section V-D) we consider a more economical strategy, referred to as *Direct-Routing-First (DRF)*. With DRF, the OSP first attempts to route a flow f through the direct path with the minimum RTT. If that path does not result in lower RTT than τ_f, and there exists an OSP-preferred indirect path, the OSP selects the indirect path with the minimum RTT. The reasoning behind DRF is that indirect paths are more costly because the OSP has to pay for upstream capacity at two MON nodes (rather than at a single node for the case of direct paths). So, with DRF, the OSP assigns higher priority to direct paths than to indirect paths.

If $\mathcal{PATH}(f)$ is the sequence of POPs for an OSP-preferred flow f using a certain routing strategy, then the following function identifies whether flow f passes through the MON node at POP p,

$$\mathcal{RTE}(p, f) = \begin{cases} 1 & \text{if } p \in \mathcal{PATH}(f) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Given a POP selection vector and an OSP routing strategy, as well as the set of flows $F(u)$, it is easy to determine whether each flow of u can be OSP-preferred, and thus whether customer u would subscribe to the OSP or not,

$$\mathcal{SUB}(u) = \begin{cases} 1 & \text{if customer } u \text{ subscribes to OSP} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

E. OSP revenues and costs

The OSP generates revenue from subscribed customers. Let $\hat{P}(r)$ be the OSP pricing function, where r is the total traffic rate from a customer, say in the period of a month. Then the total OSP revenue in that period is:

$$R(\mathcal{MON}) = \sum_{u \in U} \mathcal{SUB}(u) \cdot \hat{P}\Big(\sum_{f \in F(u)} r_f\Big) \quad (5)$$

The OSP has two types of costs: first, a recurring cost for each deployed node (*i.e.,* monthly fee at IXPs) and second, the cost of upstream capacity from native ISPs. In general, different ISPs have different capacity pricing functions, and these functions may vary across different locations. Therefore, we calculate the OSP capacity cost on a per-POP basis. Specifically, the required upstream capacity at POP p is:

$$c(p) = \sum_{u \in U} \mathcal{SUB}(u) \cdot \sum_{f \in F(u)} \mathcal{RTE}(p, f) \cdot r_f \quad (6)$$

and so the total capacity cost is:

$$C_{CP}(\mathcal{MON}) = \sum_{p} \mathcal{MON}(p) \cdot P_p\left(c(p)\right) \quad (7)$$

where $P_p(\cdot)$ is the pricing function used by the ISP at POP p.

The total node deployment cost can be modeled as:

$$C_{ND}(\mathcal{MON}) = \sum_{l \in L} \mathcal{NODE}(l) \cdot d(l) \quad (8)$$

where $d(l)$ is the cost of deploying a MON node at location l.

The OSP pricing function $\hat{P}(r)$, as well as the ISP pricing functions $P_p(r)$, are assumed to be non-decreasing and concave, which is the typical case in practice [9]. The concavity is important because it implies *economies of scale*, *i.e.,* the price per Mbps decreases as the purchased capacity r increases. The OSP can exploit this property of the capacity market to offer less expensive services than the competing native ISPs by aggregating the traffic from many customers. Specifically, suppose that the pricing ratio R_p between the OSP and ISP pricing functions is constant,

$$R_p = \frac{\hat{P}(r)}{P_p(r)} \quad (9)$$

If $R_p < 1$, the OSP is less expensive than the ISP at POP p. The OSP can still be profitable, however, because the aggregation of traffic from several customers means that the OSP can purchase upstream capacity at a lower unit price than what it charges to its customers.

F. Problem statement

We can now state the MON design problem as the following constrained nonlinear optimization problem. Given the following inputs:

Native network information: Set of locations L and ISPs I; Delay matrix T; ISP pricing functions $P_p(\cdot)$ for all $p \in P$;

OSP information: OSP routing strategy; OSP pricing function $\hat{P}(\cdot)$; MON node deployment cost $d(l)$ for all $l \in L$;

Customer information: Set of customers U with their flow descriptors $F(u)$ for all $u \in U$; Subscription threshold H;

Determine the POP selection vector $\mathcal{MON}$ to maximize the profit:

$$\Pi(\mathcal{MON}) = R(\mathcal{MON}) - C_{CP}(\mathcal{MON}) - C_{ND}(\mathcal{MON}) \quad (10)$$

under the following constraints:
1) At most N MON nodes: $\sum_{l \in L} \mathcal{NODE}(l) \leq N$
2) Maximum multihoming degree K: For all $l \in L$, $\sum_{p \in P, \mathcal{LOC}(p)=l} \mathcal{MON}(p) \leq K$

G. NP-hardness

We next give a sketch of the NP-hardness proof based on a reduction from the set covering problem. Consider a set of sets S, where $\bigcup_{s_i \in S} s_i = X$. The set covering problem is to find a minimum-size set $C \subseteq S$ such that $\bigcup_{c_i \in C} c_i = X$. We construct an instance of the MON design problem in which there is a set of ISPs S available at a single location l and the MON design constraints are N=1 and K=$|S|$. Let X be the set of customers at location l, with one flow per customer, and suppose that all flows can be OSP-preferred, *i.e.,* there is at least one ISP in S that can make each flow OSP-preferred. Assume that the OSP has a constant pricing function $\hat{P}(r) = p_{osp}$, and all ISPs have the constant pricing functions $P_p(r) = p_{isp}$, with $p_{osp} \gg p_{isp}$ and $p_{osp} \gg d(l)$. Under these constraints, the OSP should deploy a single MON node at location l, and determine the minimum-size set of ISPs that maximizes its profit. We can determine (in polynomial time with $|S|$ and $|X|$) the set of flows s_i that become OSP-preferred when the MON node is connected to the i'th ISP. Because $p_{osp} \gg p_{isp}$ and $p_{osp} \gg d(l)$, the OSP can maximize its profit by connecting to just enough ISPs so that *all* flows are OSP-preferred. In other words, the OSP needs to find a minimum-size set of ISPs $C \subseteq S$ such that $\bigcup_{c_i \in C} c_i = X$. So, any instance of the set covering problem can be reduced in polynomial time to the previous instance of the MON design problem. Therefore, given that the set covering problem is NP-hard, the MON design problem is also NP-hard.

III. Estimating the native RTT matrix

In this section, we describe a methodology for estimating the native RTT matrix T. Recall that this matrix consists of the POP-to-POP pairwise RTTs in the native network, and it is a required input for one of the four heuristics of the next section. Also recall that the RTT we aim to estimate is the propagation delay RTT, which mostly depends on the physical distance between two POPs and the routing at the underlying native network; queuing delays, in particular, are not included in these RTTs.

We distinguish between intradomain RTTs, where both POPs are in the same Autonomous System (AS), and interdomain RTTs, otherwise. As shown next, an intradomain RTT can be modeled as proportional to the physical driving distance between the two POPs, at least for the networks we measured. An interdomain RTT, on the other hand, further increases with the number of AS's in the route between the two POPs. So, in the interdomain case, the ratio between RTT and driving distance depends on the number of AS hops.

A. Intradomain case

To develop an accurate model for intradomain RTTs, we examined the correlation among various distance metrics and the measured RTTs between POPs of various network providers in the US. The highest correlation resulted from the "highway driving distance", as reported from Google-map. This is probably because most backbone optical fiber is laid along highways or railroads. In the following, we analyze RTT pairwise measurements between 15 ping servers (located at POPs) in the US Level3 network. For each source/destination POP pair, we conducted 1,000 consecutive ping measurements, reporting the minimum RTT as the best estimate of the propagation delay RTT. Figure 3 shows that the minimum measured RTT varies almost linearly with the physical driving distance, with a correlation coefficient of about 95%. Therefore, the OSP could model the intradomain RTT between two POPs p and q of the Level3 network as:

$$\tau_{intra}(p,q) = 0.02349 \cdot L(p,q) \tag{11}$$

where $\tau_{intra}(p,q)$ and $L(p,q)$ are the intradomain RTT (in milliseconds) and the driving distance (in miles) from p to q, respectively. The same procedure should be followed for each ISP, because the proportionality constant between RTT and distance may be different across ISPs.

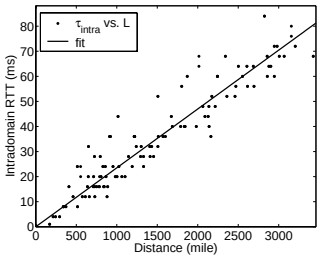
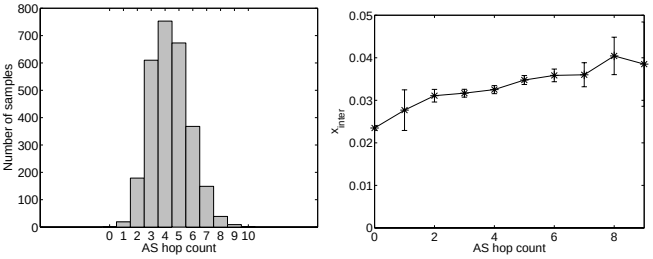


Fig. 3. Intradomain RTT versus driving distance in the US Level3 network



(a) AS hop-count histogram (b) Ratio between RTT and driving distance for various AS hop-counts

Fig. 4. Relation between interdomain RTT and AS hop-count

B. Interdomain case

We observed that in the interdomain case, the RTTs are not so highly correlated with the physical distance between POPs. The reason is that interdomain routes are commonly dictated by policy constraints, and so they often follow suboptimal paths. We observed, however, that there is a high correlation between RTT and driving distance when we consider routes with the same number of AS hops (*i.e.,* the same number of traversed networks). Consequently, we can model the interdomain RTT between two POPs p and q as:

$$\tau_{inter}(p,q) = x_{inter}(h) \cdot L(p,q) \tag{12}$$

where $x_{inter}(h)$ is a constant that depends on the number of AS hops h in the native route between POPs p and q.

To estimate $x_{inter}(h)$ for various values of h, we measured 3,136 interdomain paths from 32 PlanetLab US nodes to 98 web servers (of museums and newspapers) with two servers per continental state. For each path, we conducted 1,000 ping measurements (to estimate the propagation delay RTT) and a traceroute measurement. The latter was used to estimate the AS hop-count of the path, using the IP-to-AS mapping database of [10]. Figure 4(a) shows the histogram of AS hop-counts in the measured paths. Note that most paths have 2-7 AS hops, including the destination AS (but not the source AS).

Figure 4(b) shows the 95% confidence interval of $x_{inter}(h)$ as a function of h. Note that h=0 corresponds to the intradomain case. The results show clearly an increasing trend in $x_{inter}(h)$ as h increases. Furthermore, the confidence intervals are narrow, indicating that the RTT can be modeled as proportional to the driving distance for a given AS hop-count, except the single hop and 7-9 hop cases (for which we do not have enough measurements though). In summary, the OSP can estimate the RTT between two interdomain POPs as long as it can determine the number of AS's between the two POPs (*e.g.*, through BGP routing feeds) and also construct a measurement-based graph such as Figure 4(b).

IV. MON DESIGN HEURISTICS

TABLE I

INPUT REQUIREMENTS FOR EACH HEURISTIC

Heuristic	Customer info	Traffic info	Performance info
RAND			
CUST	√		
TRFC	√	√	
PERF	√	√	√

The MON design problem involves two major tasks: 1) select up to N locations for placing MON nodes; 2) select up to K upstream ISPs for each deployed MON node. In this section, we present four heuristics for the MON design problem. The heuristics differ in terms of their inputs, ranging from a simple random heuristic (RAND) that does not require any customer or performance data, to the most complex heuristic (PERF) that utilizes information for the location of customers, the traffic matrix of each customer, and the native delay matrix (see Table I). The heuristics assume that the node deployment cost is the same at all locations, and the ISPs at the same location have identical pricing functions. The latter is a reasonable assumption for a stable and competitive ISP market.

A. Random (RAND)

This heuristic represents a naive approach in which we select N random locations, and connecting node at location l to a random set of $\min(K, |I_l|)$ locally present ISPs.

B. Customer-driven (CUST)

In RAND, different locations and ISPs have equal probabilities of being selected. Obviously, this strategy will not perform well when customers are not uniformly distributed. CUST utilizes information about the location of each customer and places N MON nodes at the locations with the maximum number of customers. Placing MON nodes at those locations enables more customers to connect to the OSP and therefore it increases revenues. At each selected location l, CUST then selects the $\min(K, |I_l|)$ locally present ISPs with the maximum coverage, *i.e.*, the ISPs that are present at the largest number of locations. The intuition here is that larger-coverage ISPs can typically reach traffic destinations through fewer AS's, and so they are more likely to provide lower RTTs.

C. Traffic-driven (TRFC)

Although CUST utilizes the profile of customers locations, it does not consider the traffic that each customer generates, nor the distribution of traffic destinations. The traffic-driven heuristic uses the aggregated traffic rate that originates from all potential customers at a location. TRFC places N MON nodes at the locations with the largest volume of aggregated customer traffic. By placing MON nodes at locations where "traffic-heavy" customers are located, more traffic can subscribe to the OSP generating more revenue than CUST. At each selected location l, TRFC then selects the $\min(K, |I_l|)$ locally present ISPs that receive the maximum traffic rate from customers. The intuition here is that an ISP that can deliver traffic directly to its destination will probably result in lower delays than an ISP that delivers traffic through other AS's.

D. Performance-driven (PERF)

Note that the CUST and TRFC heuristics do not utilize any native delay information. If there are OSP-preferred direct paths then these two heuristics perform quite well in identifying a good set of locations, because the DRF routing strategy does not need to consider indirect paths in that case. If, however, many customer flows only have indirect OSP-preferred paths, then the previous two heuristics cannot choose good locations for placing intermediate MON nodes. This motivates the performance-driven heuristic (PERF). PERF requires an estimate of the native delay matrix T. The key idea in PERF is to select locations and upstream ISPs that will turn as many flows to OSP-preferred as possible.

The location selection process is performed iteratively. PERF keeps track of the set $\bar{L}$ of locations that are not yet selected and the set $\bar{F}$ of flows that are not yet OSP-preferred. Initially, $\bar{L} = L$ and $\bar{F} = F$. During each iteration, PERF associates a weight $W_L(l)$ with each location l to represent the amount of traffic that can become OSP-preferred if l is selected. At the beginning of an iteration, all weights are set to zero. PERF then processes the flows in $\bar{F}$ sequentially. For each flow $f \in \bar{F}$, PERF first finds all OSP-preferred MON paths for that flow based on the currently chosen locations, assuming that these locations are multihomed to all locally present ISPs (the assumption will be refined during the ISP selection phase of the algorithm). Then, PERF updates the weight $W_L(l)$ as follows: If l is the ingress location of a direct path for flow f, $W_L(l)$ is increased by the rate r_f. If l is either the ingress or the intermediate location of an indirect path for f, $W_L(l)$ is increased by $r_f/2$. After all flows have been

processed, the location with the highest weight is selected and removed from $\bar{L}$, and the flows that are now OSP-preferred are removed from $\bar{F}$. This process repeats until we either select N locations or all flows are OSP-preferred. After the MON node locations have been selected, the OSP-preferred flows are routed based on the given OSP routing strategy. The set of flows assigned to location l is denoted by $F_A(l)$.

Next, PERF enters the ISP selection phase for each selected location l. For each locally present ISP i, we calculate the weigh $W_I(i)$ as the traffic rate of all flows in $F_A(l)$ that use ISP i in their OSP-preferred path. The ISP with the highest weigh is selected and the process is repeated until either K ISPs are selected or all flows in $F_A(l)$ are OSP-preferred. The pseudo code for the PERF heuristic is shown in Algorithm 1.

Algorithm 1 Performance-driven heuristic

Location selection:
  Initialize $\bar{L} = L$, $\bar{F} = F$;
  Repeat until N locations are selected or $\bar{F} = \varnothing$;
  Begin
    Initialize weights $W_L(l) = 0$, $\forall l \in \bar{L}$;
    For each flow $f \in \bar{F}$, find all OSP-preferred paths, and update $W_L(l)$ as follows:
      If l is on a direct OSP-preferred path:
        $W_L(l) = W_L(l) + r_f$;
      If l is on an indirect OSP-preferred path:
        $W_L(l) = W_L(l) + r_f/2$;
    Select location $\hat{l} = \arg\max_l W_L(l)$;
    Update $\bar{L}$=$\bar{L}$-$\{\hat{l}\}$ and remove OSP-preferred flows from $\bar{F}$;
  End
  Assign OSP-preferred flows to selected location l, calculate $F_A(l)$;
ISP selection (for each selected location l):
  Initialize $\bar{I} = I_l$, $\bar{F} = F_A(l)$;
  Repeat until K ISPs are selected or $\bar{F} = \varnothing$;
  Begin
    Calculate the weight $W_I(i)$ as the total traffic in $F_A(l)$ that can be OSP-preferred if ISP i is selected;
    Select ISP $\hat{i} = \arg\max_i W_I(i)$;
    Update $\bar{I}$=$\bar{I}$-$\{\hat{i}\}$ and remove the OSP-preferred flows from $\bar{F}$;
  End

V. Performance evaluation

We conducted extensive simulations to compare the MON design heuristics, study the OSP profitability and performance depending on several key parameters (number of MON nodes, degree of multihoming, node deployment cost, OSP/ISP pricing ratio), and examine various OSP routing strategies.

A. Simulation setup

We randomly chose 51 cities in the continental US as the set of POP locations L. The population of these locations varies from 29,000 to 8,100,000.[2] These locations are served, overall, by 100 ISPs. The average number of ISPs per location is 10, and the number of ISPs per location is proportional to the logarithm of that locations's population.

[2]We have also experimented with the 50 largest US cities and the results are very similar.

Unless otherwise specified, we assume a *population-based customer distribution* (CUST-POPUL), *i.e.,* the number of customers at each location is proportional to that locations's population. Furthermore, 70% of the customers are multihomed to 2-4 ISPs (as long as there are enough ISPs at that location). For a multihomed customer, each flow originates from a single ISP at that location; the assignment of flows to ISPs is random. We model only the 10 largest flows of each customer; recall that a "flow" in this context is a traffic aggregate from a customer to a destination POP. The flow destinations are uniformly distributed across all POPs. The average flow rate follows the *gravity model* (RATE-GRVTY), *i.e.,* the rate between a pair of POPs is proportional to the product of the population at the two locations [11]. The flow rates are normalized by a constant factor so that the average flow rate is 1Mbps. We set the customer subscription threshold to H=70%. Unless otherwise specified, the maximum multihoming degree is K=2, and the total number of potential customers is 500.

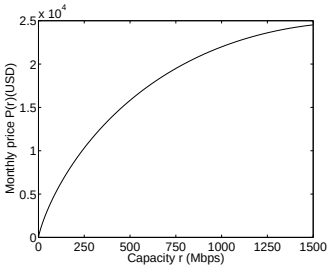


Fig. 5. Internet capacity pricing function

An important input to the MON design problem is the ISP pricing function $P_p(r)$ at each POP p and the OSP/ISP pricing ratio $R_p = \hat{P}(r)/P_p(r)$. It has been observed that, at least during the last 10 years or so, ISP capacity pricing shows economies of scale, *i.e.,* the price per Mbps drops almost logarithmically with the purchased capacity r [9]. Based on data from [12], and using the previous logarithmic relationship, we model the ISP pricing function as:

$$P_p(r) = [118 - 13.9 \cdot ln(r)] \cdot r \tag{13}$$

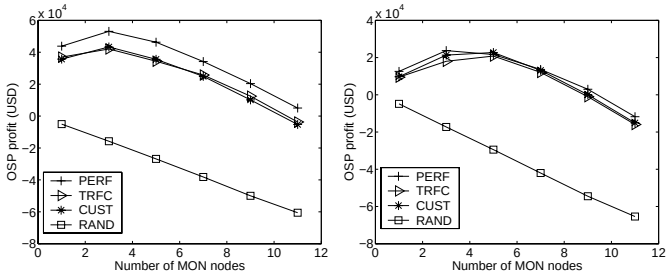
where r is measured in Mbps and the price is a monthly fee in USD (see Figure 5). Note that ISP pricing is sometimes done at discrete capacity values, and so the pricing function can be a discrete step-like function. Here, we assume a usage-based pricing model where both ISPs and the OSP charge based on the routed traffic volume. For simplicity, we also assume that the pricing function is the same at all POPs, and that the pricing ratio R_p is constant with r, the same at all POPs. If these assumptions do not hold in practice, the MON design process can still use the proposed optimization framework but the evaluation will be more cumbersome. Unless otherwise specified, the deployment cost per MON node at any location l is $d(l)$=\$5,000, based on [13], and the pricing ratio R_p=0.8.

B. Comparison of MON design heuristics

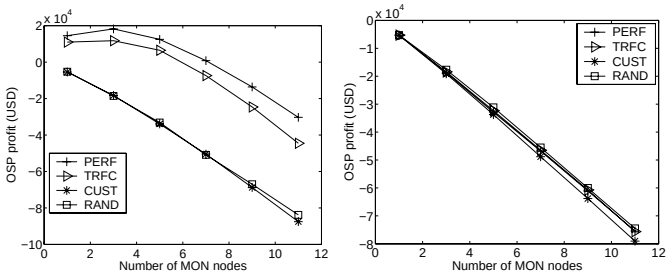
We first compare the four MON design heuristics in terms of OSP profitability as we increase the number of MON nodes. Together with the default models, CUST-POPUL and

RATE-GRVTY, we also examine models of uniform customer distribution across all locations (CUST-UNFRM), and uniform traffic rate distribution across all flows (RATE-UNFRM). The average flow rate is adjusted in each model to maintain similar maximum traffic load at the MON nodes.

Figure 6 shows that, as expected, heuristics that utilize more information about customers, traffic, or the native network perform better. Specifically, PERF outperforms all other heuristics, while RAND performs so poorly that it never leads to positive profit. In the more realistic combination of models, CUST-POPUL and RATE-GRVTY, PERF performs significantly better than other heuristics, providing a maximum monthly profit of about $50,000 instead of $40,000 with TRFC or CUST (Figure 6(a)). Clearly, the OSP would have a strong motivation to optimize the MON design process, even if that requires the collection of more input data. In the following sections we show results only for PERF.



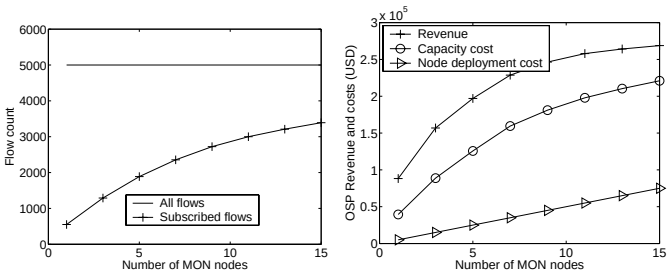
(a) CUST-POPUL, RATE-GRVTY (b) CUST-POPUL, RATE-UNFRM



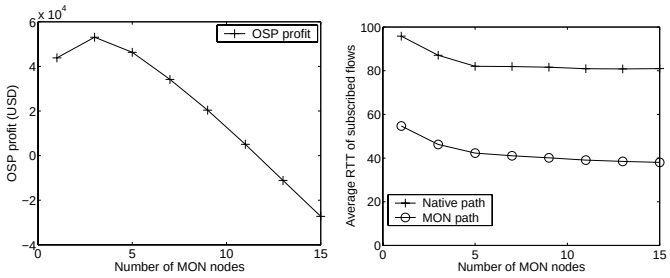
(c) CUST-UNFRM, RATE-GRVTY (d) CUST-UNFRM, RATE-UNFRM

Fig. 6. OSP profitability under four customer and traffic distribution models

With the CUST-POPUL and RATE-UNFRM models (Figure 6(b)), the OSP has lower profit, compared to the RATE-GRVTY model, because traffic tends to be more dispersed. So it becomes harder to aggregate large traffic volumes destined to the same POP and route them through direct paths. Furthermore, TRFC and PERF do not perform much better than CUST because the additional information they have about the traffic is not so useful with the RATE-UNFRM model. With the CUST-UNFRM and RATE-GRVTY models (Figure 6(c)), CUST performs equally bad with RAND, because in this case placing nodes where most customers are located is no different than placing nodes randomly. TRFC and PERF perform better, but the profits are still significantly lower than the CUST-POPUL model because the OSP cannot place nodes in just a small number of locations where most customers are. Finally, the CUST-UNFRM and RATE-UNFRM models (Figure 6(d)) do not lead to OSP profitability, with any of the four heuristics, under the default OSP/ISP pricing ratio $R_p = 0.8$. Achieving substantial traffic aggregation with just a few MON nodes is much harder in this case. At the same time, the OSP has to pay the node deployment costs and so it does not end up with profit.



(a) Number of subscribed flows (b) OSP revenues and costs



(c) OSP profit (d) Average RTT (subscribed flows)

Fig. 7. Effects of number of MON nodes

C. Optimal number of MON nodes

Here we investigate the effect of the number N of MON nodes on OSP performance and profitability. Figure 7(a) shows that the total number of flows from subscribed customers increases as more nodes are deployed. This is because more customers become collocated with MON nodes and the OSP can provide more paths towards each destination POP. As a result, the number of subscribed customers also increases with N.

Figure 7(b) shows that both the OSP revenues and costs increase with N. The reason that the revenue increase rate slows down as N increase is that, gradually, there are fewer new subscribed customers per node. The increase in capacity costs also follows a concave shape, because the OSP traffic volume follows a similar pattern. But as N increases, the amount of traffic per MON node grows more slowly, reducing the economic benefit of aggregation. On the other hand, the deployment costs increase linearly. The net result, at least in these simulations where K=2, is that the OSP achieves maximum profit when N is only 3-4 nodes (as shown in Figure 7(c)). Comparing Figures 7(a) and 7(c), we note that deploying more nodes attracts more customer flows, but that does not increase the OSP profit due to increased costs.

Figure 7(d) shows the average native RTT, as well as the average MON RTT, for all flows from subscribed customers. The results show that by subscribing to the OSP, customers reduce the average RTT of their flows significantly, by about 40msec in these simulations. This is despite the fact that only

H=70% of the customer traffic is guaranteed to have lower MON RTT than native RTT. The reason for this large decrease is because OSP customers are not a random sample of the customer population. Instead, most of their traffic goes through long interdomain routes, allowing the OSP to offer significant RTT improvements through multihoming and overlay routing.

The reason both native and MON RTTs decrease with N is as follows. As N increases, more flows become subscribed to the OSP, and thus included in the calculation of the average RTT. These flows, however, tend to have lower native RTTs because they become OSP-preferred only after the OSP has deployed more than a number of nodes. For the same reason, the average MON RTT also decreases with N.

D. Effect of OSP routing strategy

We now examine the effect of the OSP routing strategy, comparing Direct-Routing-First (DRF) with 1) Minimum-MON-Delay-Routing (MDR), *i.e.*, the OSP routes flows through the MON (direct or indirect) path with the minimum RTT, and 2) Direct-Routing-Only (DRO), *i.e.*, the OSP routes flows through direct MON paths only. With the DRO strategy, the OSP operates similarly to InterNAP, a commercial network provider that utilizes intelligent route control and multihoming, but not overlay routing.

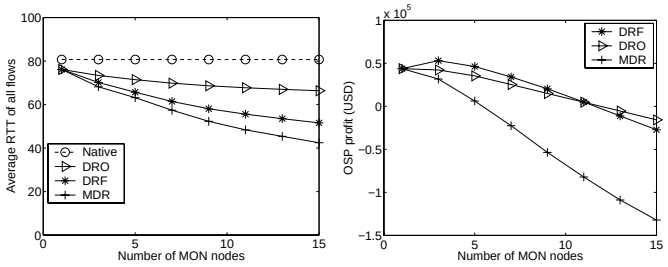
Figure 8(a) shows the average RTT of all flows using these three routing strategies; the average native RTT is also given for comparison. The results indicate that the OSP can reduce the RTT relative to native routing with any of the previous routing strategies. However, by combining multihoming with overlay routing, both DRF and MDR perform significantly better than DRO. Of course, by definition, MDR results in lower RTT than DRF, especially for larger N because the number of possible paths increases quickly with N.

On the other hand, Figure 8(b) shows that the DRF strategy leads to significantly higher profit than MDR. The reason is that MDR makes extensive use of overlay routing and indirect paths, and so the OSP often has to pay for upstream capacity at two locations instead of one to route a flow. The comparison between DRO and DRF is less clear and consistent, making the two strategies roughly equivalent in terms of profit. The previous results show that the DRF strategy achieves a good tradeoff between OSP performance and profitability: DRF is close to MDR in terms of average RTT, which is much better than DRO. At the same time DRF is much more profitable than MDR and it is as profitable as DRO.

E. Effect of pricing ratio

When the pricing ratio R_p is less than one, the OSP can offer a less expensive service than the native ISPs for the same traffic volume. If this is a profitable scenario, then the OSP can be both better and less expensive than native ISPs. Setting R_p too high can also hurt OSP's profitability because customers may not be willing to subscribe to the OSP despite the performance improvements.

In this simulation, we vary R_p from 0.4 to 2, monitoring the OSP profit as a function of the number of OSP customers. Figure 9 shows that the OSP can be profitable even when the OSP charges 40% of the ISP price, as long as there are enough customers. The reason is that, with enough customers, the OSP can achieve large traffic aggregation, and so it can purchase capacity from upstream ISPs at a lower price per Mbps than what it charges to its customers. An OSP with lower R_p needs more customers to be profitable. For example, the OSP only needs 50 customers to break even when R_p=2, but it needs more than 300 customers to make profit when $R_p = 0.8$.



(a) Average RTT (all flows) (b) OSP profit

Fig. 8. Effect of OSP routing strategy

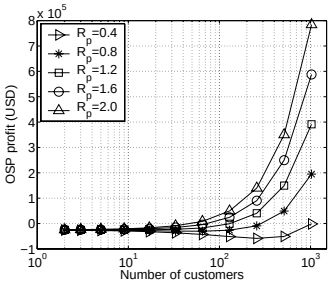
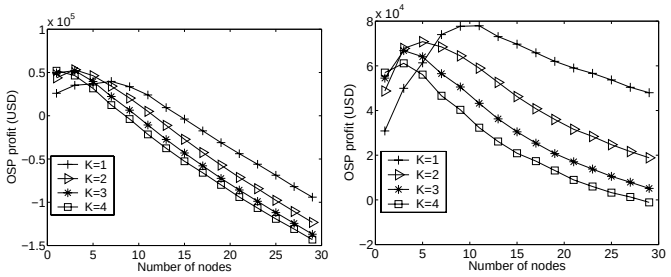


Fig. 9. Effect of OSP/ISP pricing ratio

F. Effect of maximum multihoming degree

The maximum multihoming degree K is another important parameter because it determines the local path diversity at each MON node. In the simulations so far, K was set to 2. Now we examine four values of K, from one (singlehoming) to four ISPs, as we increase the number of MON nodes N. Figure 10 shows the OSP profit under two node deployment costs: 1) $d(l)$=\$5,000, representing a new OSP that creates a MON from scratch, and 2) $d(l)$=\$100, representing an OSP that already has a distributed infrastructure in place (*e.g.*, Akamai).

Figure 10(a) shows that, with the higher deployment cost, the best strategy is to place a small number of MON nodes (3-5 nodes in our simulations) and connect each node to 2-3 ISPs. Singlehoming is clearly suboptimal, and deploying more than 5-6 nodes quickly makes the OSP unprofitable. In contrast, Figure 10(b) shows that, with the lower deployment cost, placing several (around 10) singlehomed nodes is a more profitable OSP configuration. More generally, there is certainly a trade-off between the number of MON nodes and their multihoming degree. The most profitable point in this trade-off depends on the node deployment cost relative to what the OSP has to pay for capacity at each POP.



(a) Deployment cost: $5,000 per node (b) Deployment cost: $100 per node

Fig. 10. Effect of maximum multihoming degree

VI. Related work

Most previous work on overlay network design focused on optimizing the performance or redundancy of overlay paths. Han *et al.*proposed a topology-aware overlay design to maximize path diversity without degrading latency-based performance [14]. Using a binning-based algorithm, Ratnasamy *et al.*addressed the overlay node placement problem based on network proximity information [15]. Lao *et al.*examined the problem of node and link selection for a multicast overlay network [16]. Andersen *et al.*proposed the MONET (Multihomed Overlay Network) architecture that combines multihoming with a cooperative network of proxies to obtain diverse paths between clients and Web servers [17]. Cha *et al.*studied the problem of intradomain relay node placement aiming to maximize path diversity [18].

The node selection problem has also been studied in the context of content delivery networks (CDNs), or web proxy/cache placement, where a set of nodes is selected to replicate popular content or to serve users. Qiu *et al.*explored the problem of web server replica placement to improve CDN performance [19]. Cahill and Sreenan proposed replica placement algorithms to minimize the CDN resource costs [20]. Li *et al.*investigated the optimal proxy placement policy to minimize access latency [21].

There is also some previous work in the multihoming design problem. Wang *et al.*proposed algorithms to select a set of upstream ISPs with a monetary cost-minimization objective [22]. Intelligent route control products use the idea of dynamic switching among upstream ISPs to select the best path for any prefix in real time [23]. Dhamdhere and Dovrolis proposed methodologies for the provisioning of egress routing at a multihomed content provider, taking into account monetary cost, interdomain level performance and path diversity [24].

VII. Conclusions

We examined the effectiveness of combing multihoming and overlay routing from the pragmatic perspective of a network provider (OSP) that attempts to be both profitable and also offer better and less expensive Internet access to its customers. Interestingly, we found out that it is possible to meet all previous objectives, as long as the OSP can follow the following basic guidelines: 1) use a performance-aware MON design heuristic (such as PERF) even if that requires additional inputs and measurements, 2) deploy nodes at few locations where significant traffic aggregation is possible, 3) connect each MON node to ISPs that can directly reach traffic-heavy destination POPs, 4) give direct paths higher priority than indirect paths, 5) charge less than the competing native ISPs for the same traffic rate to attract more customers, and 6) determine the best trade-off between the number of MON nodes and multihoming degree based on the node deployment cost.

References

[1] D. G. Andersen, H. Balakrishnan, M. F. Kaashoek, and R. Morris, "Resilient overlay networks," in *Proc. ACM SOSP*, 2001.

[2] A. Akella, B. Maggs, S. Seshan, A. Shaikh, and R. Sitaraman, "A measurement-based analysis of multihoming," in *Proc. ACM SIGCOMM*, 2003, pp. 353–364.

[3] InterNAP, http://www.internap.com.

[4] D. K. Goldenberg, L. Qiu, H. Xie, Y. R. Yang, and Y. Zhang, "Optimizing cost and performance for multihoming," in *SIGCOMM*, 2004.

[5] Z. Li and P. Mohapatra, "QRON: QoS-aware routing in overlay networks," *IEEE JSAC*, vol. 22, no. 1, pp. 29–40, 2004.

[6] H. Rahul, M. Kasbekar, R. Sitaraman, and A. Berger, "Towards realizing the performance and availability benefits of a global overlay network," in *Proc. Passive and Active Measurements (PAM) conference*, 2006.

[7] A. Akella, J. Pang, B. Maggs, S. Seshan, and A. Shaikh, "A comparison of overlay routing and multihoming route control," in *Proc. ACM SIGCOMM*, 2004.

[8] Y. Zhu, C. Dovrolis, and M. Ammar, "Dynamic overlay routing based on available bandwidth estimation: A simulation study," *Computer Networks Journal (Elsevier)*, vol. 50, pp. 739–876, 2006.

[9] P. M. Ferreira, "Implications of decreasing bandwidth price on allocating traffic between transit and peering agreements," Master's thesis, Massachusetts Institute of Technology, 2002.

[10] Z. M. Mao, J. Rexford, J. Wang, and R. H. Katz, "Towards an accurate as-level traceroute tool," in *SIGCOMM*, 2003.

[11] M. Roughan, M. Thorup, and Y. Zhang, "Performance of estimated traffic matrices in traffic engineering," in *Proc. ACM SIGMETRICS*, 2003.

[12] http://merit.edu/mail.archives/nanog/2004-08/msg00269.html.

[13] "Collocation: More than just a real estate play," *Equinix white paper*, 2001.

[14] J. Han, D. Watson, and F. Jahanian, "Topology aware overlay networks," in *Proc. IEEE INFOCOM*, 2005.

[15] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, "Topologically-aware overlay construction and server selection," in *Proc. IEEE INFOCOM*, 2002.

[16] L. Lao, J.-H. Cui, and M. Gerla, "Multicast service overlay design," in *Proc. International Symposium on Performance Evaluation of Computer and Telecommuni cation Systems (SPECTS)*, 2005.

[17] D. G. Andersen, H. Balakrishnan, M. F. Kaashoek, and R. Rao, "Improving web availability for clients with monet," in *Proc. Symposium on Networked Systems Design and Implementation (NSDI)*, 2005.

[18] M. Cha, S. Moon, C.-D. Park, and A. Shaikh, "Placing relay nodes for intra-domain path diversity," in *Proc. IEEE INFOCOM*, 2006.

[19] L. Qiu, V. N. Padmanabhan, and G. M. Voelker, "On the placement of web server replicas," in *Proc. IEEE INFOCOM*, 2001.

[20] A. J. Cahill and C. J. Sreenan, "An efficient CDN placement algorithm for the delivery of high-quality tv content," in *Proceedings of the ACM international conference on Multimedia*, 2004.

[21] B. Li, M. J. Golin, G. F. Italiano, and X. Deng, "On the optimal placement of web proxies in the Internet," in *Proc. IEEE INFOCOM*, 1999.

[22] H. Wang, H. Xie, L. Qiu, A. Silberschatz, and Y. R. Yang, "Optimal ISP subscription for Internet multihoming: Algorithm design and implication analysis," in *Proc. IEEE INFOCOM*, 2005.

[23] R. Gao, C. Dovrolis, and E. W. Zegura, "Avoiding oscillations due to intelligent route control systems," in *Proc. IEEE INFOCOM*, 2006.

[24] A. Dhamdhere and C. Dovrolis, "ISP and egress path selection for multihomed networks," in *Proc. IEEE INFOCOM*, 2006.