Sticky Consumers and Cloud Welfare *

Chuqing Jin† Sida Peng‡ Peichun Wang§

July 19, 2022

Abstract: We estimate welfare benefits of the public cloud and study the impact of customer inertia on welfare. We develop a novel demand model that allows for both multiple product choices and continuous usage, and estimate the model using proprietary customer-level data. We find the average consumer surplus from cloud usage to be 216% of its cost, and that smaller customers disproportionately benefit from public cloud. We also find significant inertia on the cloud, reducing welfare benefits by 62%. Finally, we show that cloud migration services and introductory discounts can improve both consumer surplus and provider revenue.

*We thank Juan Camilo Castillo, Ulrich Doraszelski, Marc Rysman, numerous colleagues at Microsoft, and seminar participants at Boston University, Nanyang Technological University and Zhejiang University for helpful comments. All remaining errors are our own.

†Department of Economics, Boston University, Email: cjin@bu.edu

‡Office of the Chief Economist, Microsoft, Email: sida.peng@microsoft.com

§Unity Technologies, Email: willw@unity3d.com

1 Introduction

The public cloud is now an important part of the economy, and becoming more so. The Infrastructure-as-a-Service (IaaS) market alone has exceeded $59 billion in revenue in 2020 and is forecast to grow to $82 billion in 2021 (Gartner 2022). Moreover, the rest of the economy has increasingly moved to the cloud, making it a critical infrastructure of the digital economy.[1] A key benefit of the cloud is that it allows firms to access computing resources without owning physical hardware, and thus firms can easily adjust their usage of computing products to suit their needs. However, in practice, cloud customers may fail to make optimal product adjustments, as they may develop inertia on existing products due to the lack of maintenance needs on the cloud, thus undermining cloud's benefit.[2] Despite its rapid growth, the cloud's overall welfare impact remains an open question.

In addition, the cloud may have a distributional effect across small and large firms. There is growing concern over decreasing dynamism in the global economy as well as rising industry concentration (Decker et al. 2016; De Loecker, Eeckhout and Unger 2020). Evidence has pointed to access to information technology (IT) as a potential cause (Bessen 2020; Tambe et al. 2020). Small firms, with budget constraints and demand uncertainty, are less likely to own computing hardware that are sufficient for peak demand and have the latest technology. By lowering the upfront setup costs of buying physical hardware, the cloud may disproportionately benefit small firms and alleviate concerns over dynamism and concentration. Finally, this distributional impact depends on whether small and large firms face different levels of inertia.

In this paper, we estimate both the overall and distributional welfare impact of cloud usage, taking into account the impact of inertia. To that end, we develop a structural model

[1] 92% of organization's IT environment is at least partially on the cloud. IDG. 2020. 2020 Cloud Computing Study. https://www.idg.com/tools-for-marketers/2020-cloud-computing-study/, accessed Dec. 12, 2021.

[2] Inertia is observed in many other low-maintenance service markets such as mobile, broadband, cash savings, insurance, and mortgages (see Citizens Advice (2018), a super-complaint filed to the UK Competition and Markets Authority).

of cloud demand, separately for small and large firms, and estimate the model using data from a major public cloud provider. We then conduct counterfactuals to compare the welfare benefit of cloud usage with and without inertia, as well as potential remedies for inertia.

The challenge in modeling cloud demand is that cloud customers use multiple products simultaneously and incur continuous usage on each product. To keep this demand pattern for welfare evaluation, we posit that in each period, each customer draws multiple computing tasks that are suited for different products. Then, the customer chooses the best product for each task. The econometrician observes the total usage of each customer on each product but not the number of tasks or the size of each task. Identification of the number of tasks comes from the sparsity of customers' product choices: if a customer has many tasks, then it is unlikely that we observe zero usage on any product for this customer. Task sizes are then estimated from the amount of usage.

Our data provides full usage history at the customer-product level for 32 months. We focus on virtual machine (VM) products —the main part of IaaS—which are bundles of different combinations of computing resources such as CPU, RAM, storage, and network. In the most popular VM product family, which we study in detail, we observe two product launches during our sample. These product launches, together with customer-level usage histories, identify customer inertia from unobserved heterogeneity in customer taste. Moreover, there are 24 geographic locations (regional data centers) customers can choose from to deploy their VMs. We define combinations of these geographic regions and two operating systems as markets. Variation in prices and product availability across markets identify other taste parameters.

Before introducing our structural estimates, we first document three patterns in the data that emphasize the importance of modeling inertia in the cloud market. First, new VMs are often launched with better performance and lower prices (Kilcioglu and Rao 2016), while the older and dominated VMs continue to be offered by providers.[3] However, there

[3]For example, AWS launched the M5 VMs in late 2017 at lower prices than the previous generation M4 VMs, which continued to be offered.

remains significant demand for the dominated VMs, suggesting the existence of customer inertia. Second, cloud providers sometimes launch promotional products of existing regular VMs with identical characteristics but discounted prices, while the regular VMs continue to be offered. One such promotional VM is launched in our sample. We find that 90% of existing customers from the regular VM did not move to the new promotional VM in the 12 months after its launch, even though these customers would have saved 22% of their total cloud spending. Third, using customers' usage histories, we can distinguish between existing customers who are affiliated with the regular VM, versus new customers who have to make an active choice between all VMs. We find that new customers are significantly more likely to adopt the new promotional VM than existing customers.

We model inertia with an adoption cost for new products that customers have not previously used and remain agnostic about its exact mechanism.[4] As Handel and Schwartzstein (2018) point out, identification of the specific mechanism of inertia is often difficult and may be irrelevant for the welfare evaluation of many policies.[5] We study the overall welfare impact of inertia and explore remedies that directly steer customers towards adopting new products.

Estimates from our demand model reveal several results about cloud usage. First, we find low price elasticities for most VMs from the cloud provider. Conditional on adopting this cloud provider, customers require large price changes to change their usage patterns. Second, cloud customers face significant costs in adopting new products. The average estimated cost of adopting a new VM is equivalent to increasing the price of an average VM by 12 times for small customers and 7 times for large customers for an average task they face. Third, small

[4]This specification of inertia ignores customers' switching costs to products they have previously used. We do not model switching costs between products for two reasons. First, switching costs are typically identified under the assumption that each customer chooses one product at a time so switching behaviors are clearly observed, which does not apply in our setting. Second, modeling continuous usage also adds to the difficulty: when a customer's usage on a product is decreased, we cannot distinguish whether her total demand has decreased or she has switched to another product.

[5]Handel and Schwartzstein (2018) argue that it is sufficient to identify the combined effect of inertia to evaluate policies that strongly steer customers to specific actions ("allocation policies"). Identification of the specific mechanism is needed only for policies that target specific mechanisms ("mechanism policies").

customers are estimated to have both lower number of tasks and smaller task sizes, but their number of tasks grows faster over time. While small customers have higher disutility from prices per unit of usage, they are less price elastic compared to large customers once usage is taken into account.

With our model and estimates, we first simulate customer product choice and usage histories from the beginning of the sample and show that the model and estimates capture the rich heterogeneity in usage and product choices, and fit the data well. We then compute consumer surplus and find large welfare gains from cloud usage. On average, we estimate cloud customers' return on investment (ROI), defined as the ratio between consumer surplus and their cloud spending (i.e., provider revenue), to be 216%, or roughly 2.2x. Average ROI for small customers is 2.7x and for large customers is 2x, consistent with the hypothesis that the public cloud disproportionately benefits small firms.

We then re-compute customers' cloud usage when there is no inertia. We find that customers lose 62% of consumer surplus due to inertia. The provider also loses 58% of revenue from customer inertia, due to the overall slower adoption of cloud products. To dissect whether the changes in consumer surplus are real welfare loss, we decompose the loss in consumer surplus into a direct adoption cost and an indirect cost from sub-optimal product choices. We find that sub-optimal product choices due to inertia account for 98% of the loss in consumer surplus, whereas the direct adoption cost only accounts for 2%. While the direct adoption cost is mechanism agnostic and may include real costs needed to adopt new VMs (thus not welfare relevant), the indirect cost from sub-optimal product choices provides a lower bound for the true welfare loss due to inertia.[6]

In our counterfactual analyses, we explore two remedies to customer inertia. First, we implement a full subsidy for customers to adopt new products: customers choose VMs as if there were no inertia, and their adoption costs are fully subsidized whenever they choose a

[6]This is conditional on customers being myopic in our model, which we assume given the complexity of the demand model. If customers were forward-looking, our adoption cost estimates would be higher and the direct adoption cost would likely account for a higher percentage of the loss in consumer surplus.

new VM they have not used before. Consumer surplus under this scenario is equivalent to the one without customer inertia. In practice, the subsidy can be implemented via "white-glove" services that help customers migrate to new cloud products. If such a subsidy is funded by the social planner, we estimate that each dollar spent would generate about 2 dollars in total welfare. If the provider were to fund the subsidy, we find it to be unprofitable for the provider in the short run, but profitable in the long run due to increased revenue from faster adoption of cloud products.

Another common practice to encourage customers to adopt new products is introductory discounting, e.g., offering discounts for new product launches ("new product preview") or for customers' first-month usage on any product ("personalized product trial"). We explore both forms of introductory pricing using the latest product launch in our sample. We further explore the effect of allowing the provider to target large and small customers with different discounts. We find that, in the long run, personalized product trial yields 12.8% higher revenue for the provider *and* 16.5% higher consumer surplus than the baseline with no discounts, compared to a 1.3% and 1.9% improvement in revenue and consumer surplus from new product preview, respectively. Moreover, the provider's optimal discount for personalized product trial is 225%, suggesting that more subsidy than free first-month usage (i.e., 100% discount) can improve provider revenue. Finally, when the provider can offer different discounts, we find that the optimal discount for small customers is higher than that of large customers, contrary to what managers typically do in practice and despite small customers' lower price elasticity in our setting. This is because smaller customers' usage grows faster than large customers and it is thus relatively cheaper for the provider to discount small customers upfront and increase their cloud usage in the long run.

This paper is related to a long literature on measuring the benefits of IT adoption. An extensive literature attempts to directly measure the effect of IT adoption on firm performance (see Brynjolfsson and Hitt (2000) for a review). Our paper instead follows an approach in the industrial organization literature to infer firms' gains from cloud usage by estimating their willingness to pay (i.e., demand curve) with transaction data of cloud products. Using the same approach, Bresnahan (1986) estimates a 3.3x-6.1x ROI from the adoption of mainframe computers in the financial services industry. Hendel (1999) estimates a 0.92x ROI from personal computer (PC) adoptions in the same industry. Our paper is the first to assess welfare benefits of cloud usage. Furthermore, our sample covers cloud customers from all sectors, and we study different gains from cloud by small and large firms separately.[7]

Our demand model fits in the literature on discrete choices. The extant literature mostly models each consumer purchasing one unit of a single product (Berry 1994; Berry, Levinsohn and Pakes 1995). However, in many settings, customers choose more than one products and use multiple units of each product, e.g., credit card payment, grocery shopping, and firm procurement. To estimate welfare and conduct counterfactual experiments, it is important to retain these demand patterns. There are a few exceptions in the literature that also model multiple discrete choice with continuous usage: Compared to Burda, Harding and Hausman (2012), we do not require the econometrician to observe either the number of tasks or the size of each task. An identification challenge is how to disentangle determinants of product choice from those of usage. Hendel (1999) and Koulayev et al. (2016) both rely on exclusion restrictions, assuming that some observable characteristics only affect product choice but not usage. In contrast, we observe that if a customer has zero usage on a product, it must be that the product is never chosen. This sparsity pattern in the choice data can be used to infer the number of discrete choices a customer has to make, hence identifying choice from usage without additional data.

Finally, our estimation of inertia is related to the literature on frictions in product choice. While most of this literature study consumers (Hortaçsu and Syverson 2004; Handel and Kolstad 2015; Erdem and Keane 1996; Bartoš et al. 2016; Bhargava, Loewenstein and Sydnor 2017; Hanna, Mullainathan and Schwartzstein 2014), we focus on firms as customers, who

[7] Other studies on the heterogeneous impact of IT adoption on different firms include Forman (2005), Tambe and Hitt (2012), Jin and McElheran (2017), and Peng et al. (2021). In particular, Jin and McElheran (2017), using expenditures on outsourced IT services from the US Census, argue that young plants benefit more from the cloud as it reduces their learning costs.

may have additional frictions: when changing between products, there may be engineering costs (Greenstein 1993) or other types of switching costs (Burnham, Frels and Mahajan 2003); there may also be organizational slacks (Cyert and March 1963) such as how budgets are planned and spent (Liebman and Mahoney 2017). Inertia in firms' IT procurement may also have implications for their downstream competition, and may become more prevalent as IT products are offered as services with low maintenance needs. On the supply side, the inertia we study is at the product level rather than firm level, creating new strategic incentives for multi-product providers. In terms of estimation, Farrell and Klemperer (2007) discuss the challenge of identifying inertia from unobserved consumer heterogeneity. With market-level data, the literature typically needs to assume that a consumer is affiliated with a single product at any time (MacKay and Remer 2022). In contrast, we allow customers to use multiple products simultaneously, and leverage customer-level usage history combined with product launches to identify inertia.

The rest of the paper is organized as follows. Section 2 gives an overview of the public cloud market and describes our data and descriptive evidence. Section 3 presents our demand model and discusses identification. Section 4 describes our estimation strategy and Section 5 presents the results. Section 6 discusses welfare and conducts counterfactual analyses. Section 7 concludes.

2 Data and Descriptive Evidence

In this section, we begin by providing an overview of the public cloud market and our data. We then turn to describing patterns in the data that motivate our model and identification.

2.1 The Public Cloud Market

The public cloud broadly includes three sets of markets: Infrastructure-as-a-Service (IaaS, e.g., VMs), Platform-as-a-Service (PaaS, e.g., App Services), and Software-as-a-Service (SaaS, e.g., Office 365). This paper focuses on the IaaS market, specifically VMs, which is the main infrastructure for the rest of cloud services and the digital economy. The IaaS market has four major providers globally, Amazon Web Services, Microsoft Azure, Google Cloud Platform, and Alibaba Cloud, accounting for almost 80% of market share.[8] These providers build large data centers globally with clusters of computing hardware purchased from upstream component makers such as Intel, AMD, and Nvidia. These hardware, with CPU, RAM, storage, and network bundled together, are then rented out via virtualization technology as VMs on an hourly basis. There are three purchase models of VMs: on-demand, by reservation, and preemptible. We focus on the most popular model: on-demand VMs. As the name suggests, customers can start or stop an on-demand VM at any time with a high level of performance guarantee and will be charged based on the amount of resources used and the duration of usage.[9]

Cloud customers, mostly firms, purchase and configure VMs to substitute or complement traditional on-premise computing resources managed by their IT departments. The public cloud allows these firms to minimize upfront setup costs and reduce maintenance costs, as they no longer need to purchase and manage physical machines. Auto-scaling technology then helps these firms efficiently scale the amount of computing resources they need based on their fluctuating demand. As a result, cloud customers can achieve higher utilization of the computing resources they purchase and afford scaling to higher demand with the latest technology.

2.2 Data

Our data comes from a major cloud provider and includes full histories of customer-SKU level usage for 32 months from November 2015 to June 2018. For confidentiality, we retain

[8]Gartner. 2020. Gartner Says Worldwide IaaS Public Cloud Services Market Grew 37.3% in 2019. https://www.gartner.com/en/newsroom/press-releases/2020-08-10-gartner-says-worldwide-iaas-public-cloud-services-market-grew-37-point-3-percent-in-2019, accessed on Dec. 24, 2021.

[9]In contrast, reservations typically require one year or three years of commitment in exchange for a per-hour discount. Preemptible VMs do not have any performance guarantees, e.g., customers' workloads can be preempted by the provider at any time.

a random sub-sample of customers. Furthermore, we drop customers who use the cloud for fewer than 6 months out of the total 32 months, leaving us a total of 3,233 customers who account for 90% of total usage in our sample.

In addition to usage data, we observe the price and characteristics of each SKU. A SKU is defined by the provider hierarchically as follows: (i) region: location of the data center where a VM is hosted; (ii) operating system (OS): Linux or Windows; (iii) VM series: product category defined by function, e.g., Azure's D series includes SKUs optimized for CPU-intensive applications; (iv) VM family: group of SKUs within a VM series defined by the underlying hardware, e.g., in Azure's D series, the 3rd generation Dv3 family is based on Intel's Broadwell chips and the 2nd generation Dv2 family is based on the older Haswell chips; (v) VM size: SKUs within a VM family with specified number of virtual CPU cores, gigabytes of RAM and storage, and other resources, e.g., the D2 v3 SKU comes with 2 virtual CPU cores, 8 gigabytes of RAM, and 50 gigabytes of temporary storage.

A market is defined as a region-OS combination, because most customers consistently choose VMs within same regions running the same OS.[10] There are 47 markets in our sample. We define a product at the VM family level, grouping different VM sizes within the same family as one product because they share the same technology and same price per unit of computing resource (e.g., D4 v3 has double the amount of CPU cores and gigabytes of RAM and storage than D2 v3, and is charged twice as much). We further restrict attention to the most popular VM series (henceforth X for anonymity), which accounts for 42% of total usage. There are four products, X1-X4, and one "other" product which we define as the group of VM series other than X. Finally, X1 and X2 exist throughout our sample, whereas X3 and X4 are launched during the sample, with timing variations across markets.

Prices mainly vary across markets and products. Over time, there is only one major price drop that lowered prices for X2 in about one third of the markets at the same time. Across regions, prices vary significantly. The most expensive region has a 56% average price uplift

[10] Reasons for little substitution for VMs across regions may include data sovereignty regulations, latency, or simply where the bulk of customers' data are stored.

compared to the cheapest region. There is also a price premium for Windows VMs compared to Linux VMs. Across products, the most expensive X-series product is priced 35% higher than the cheapest on average. We leverage the variation across products and regions to identify cloud customers' price elasticities. To that end, we supplement our data with three cost shifters obtained from the cloud provider: (i) electricity prices across regions; (ii) power consumption ratings across product-regions; and (iii) hardware costs across product-regions (more details in Appendix C).

Finally, we supplement data from the provider with industry reports on the total market size to calculate customers' computing needs outside of the provider. According to Cisco (2018) and Gartner (2017), public cloud usage accounts for 58% of total cloud usage, which accounts for 83% of total computing workloads worldwide. Multiplying these with the provider's share in the public cloud market gives us the proportion of computing done by the provider. For each customer in our data, we divide their observed usage by this proportion by year to calculate their total computing demand.[11]

2.3 Customer Usage Patterns

We now turn to patterns in the data that motivate our model and identification. Customer usage on the cloud is measured in core-hours. Core-hour measures the number of CPU cores times the number of hours used for a VM. To compare performance of different types of CPUs, cloud providers use compute units to normalize core-hours for different VMs.[12]

First, we compare trends of cloud usage between large and small customers. We define customer size based on whether a customer's first six-month's usage is above or below the median. On average, large customers' cloud usage is five times that of small customers. Both grow significantly over our sample period, with large customers growing by 237% and

[11] If a customer does not have any usage in a month with the provider, we attribute all her usage to her outside option (on premise or other cloud providers), where her total usage is interpolated from her usage before and after that month.

[12] For example, see documentation for Azure compute unit: https://docs.microsoft.com/en-us/azure/virtual-machines/acu, accessed Dec. 26, 2021.

small customers by 452%. Small customers' cloud usage grows faster, and by the end of our sample, the average small customer's cloud usage exceeds 24% of large customers.

Second, there is rich heterogeneity in customers' cloud product choice and usage. Table 1 shows the frequency of customers using different number of products and their total usage (anonymized) within the same market-month. There is significant heterogeneity in both the number of distinct products a customer uses as well as their total usage. In particular, 16.6% of customers use more than one products in a market-month and they account for 45.6% of total usage. Customers with more usage typically use more distinct products, possibly due to a larger number of diverse computing tasks. To estimate welfare from cloud usage and conduct counterfactuals, our model needs to retain these rich patterns of demand.

[Table 1 about here.]

2.4 Evidence of Inertia

In this section, we provide descriptive evidence of inertia on cloud. We proceed in three steps, looking first at the aggregate transition from X1 to X2, then X2 to X3, and finally the impact of usage history on product choices using a regression discontinuity design.

In the cloud market, new products are typically offered with higher performance and lower prices than the previous generation, while these old products continue to attract demand. For our provider, X2 is launched one year after X1 (both before our sample), with better CPU technology and at least 10% lower prices in the same markets in our sample. However, we find that, for customers who exist from the beginning of our sample, X1 usage accounts for 22% of total usage at the beginning of our sample, and 11% even at the end of the sample, while X2's usage share increases from 7% to 18%.

A more direct illustration lies in the transition from X2 to X3. X3 is launched two years after X2 as its promotional product. That is, X3 shares the same technical characteristics as X2, except for its lower prices. Although marketed as a promotional product, X3 is launched with an indefinite end date and ends up being available well beyond the launch of the next product X4. Thus, X3 should be a perfect substitute for X2 and dominates X2 due to its lower prices. However, in the 12 months after the launch of X3, 90% of X2 customers have not adopted X3, and 91% of total usage between X2 and X3 are from X2. Moreover, these choices are not financially trivial: these customers would have saved 22% of their total spending had they converted their usage from X2 to X3. These facts suggest that customers may face significant inertia in adopting new products.

One may argue that a promotional product can be perceived as lower quality by customers, and thus the low adoption rate can not be attributed to inertia. To alleviate this concern, we further compare adoption rates of X3 between existing X2 customers—those who have used X2 before the launch of X3—and new customers of the provider after X3's launch, using a regression discontinuity (RD) design.

Figure 1 presents the comparison in a standard RD plot. The running variable is each customer-cohort's month of first adopting the provider and centered around X3's launch month. Customer cohorts to the right of the vertical line are new customers after X3's launch, whereas customers to the left are existing X2 customers. We calculate the outcome variable as the probability of adopting X3, conditional on adopting X2 or X3, in the 12 months after X3's launch. We find that new customers, who are otherwise similar but do not have a natural affiliation with any existing products and thus face similar adoption costs for all products, have a significantly higher probability of adopting X3 compared to existing customers who have used X2.[13] Moreover, the overall adoption rate of X3 is low for all customers, suggesting that the perceived quality of the promotional product may indeed be low.

[Figure 1 about here.]

[13]We do not observe any significant change in the total number of new customers or those who use X2 or X3 after the launch of X3, so it is unlikely that the new customers are different from existing customers due to X3's launch.

3 Model and Identification

In this section, we first present a hierarchical demand model that allows for multiple product choices and continuous usage on each product. Then we discuss the identification of this model. For tractability, the model is static. We refer to market-months simply as markets and denote them by m. Customers are indexed with i, and model parameters with subscript i vary by customer size to capture their usage differences and welfare implications.

Each month, cloud customer i faces a random number of computing tasks n_{im}, drawn i.i.d. from a Poisson distribution, i.e., $n_{im} \sim Poisson(\lambda_{im})$. λ_{im} is the expected number of tasks in market m and has a time trend to capture growing computation needs, i.e., $\lambda_{im} = \lambda_{i0} + \lambda_{i1} t_m$. Each task is endowed with task size q_{im} drawn from an i.i.d. exponential distribution, i.e., $q_{im} \sim Exp(\gamma_i)$, where γ_i is the scale parameter and is equal to the expected task size. The task size represents the amount of computing resources needed to complete the task, which is in normalized compute units and thus the same regardless of the product it runs on. One difference between our model and Burda, Harding and Hausman (2012) is that we allow usage to be continuous, so our model applies to other settings with continuous usage such as credit card payment or most usage-based online services.

In the rest of this section, we first describe customer utility and product choice at the task level, and then discuss the aggregation to the product level in the next section.

For each task, the customer chooses between J available VMs and the outside option of either computing on premise or using other cloud providers. The customer's utility from using VM j in market m for task k is:

$$u_{ijmk} = \alpha_i p_{jm} q_{im} + X_j \beta_i + \delta_i New_{ijm} + \xi_{ijm} + \epsilon_{ijmk}, \tag{1}$$

where p_{jm} is the per compute unit price for VM j in market m and α_i is the price coefficient; $\alpha_i p_{jm} q_{im}$ measures the customer's dis-utility from the cost of the task. X_j is a vector of VM j's observable characteristics and $X_j \beta_i$ measures the customer's utility from these characteristics. In our setting, X_j is comprised of dummies for each product in the X series and the "other" product. Because X3 shares the same technical characteristics as X2, we specify the utility for X3 as the sum of the X2 dummy plus an indicator $Promo_j$, which equals 1 only for X3.

To model inertia, we define New_{ijm} as a new product indicator that equals 1 if customer i has never used product j in market m, i.e., the product is new to the customer. This specification agnostically captures inertia that prevents customers from adopting new products. Depending on the specific source of inertia captured, δ_i may be interpreted as information frictions, organizational slacks, or real engineering and operational costs associated with adopting a new IT product. In our model, δ_i captures the joint effect of these different types of inertia, and we call it the adoption cost for new products.

ξ_{ijm} is an unobservable demand shock for customer i using VM j in market m. While there is little unobservable characteristics at the VM j level, ξ_{ijm} captures any demand shocks that vary at the customer-VM-market level that are unobservable to the econometrician. The unobservable demand shocks may thus create endogeneity in prices, which we account for using the control function approach following Petrin and Train (2010) in the next section.

Finally, ϵ_{ijmk} is an idiosyncratic preference shock, distributed i.i.d. type-I extreme value, for all $j = 0, 1, ..., J$. If the outside option is chosen, the customer receives utility $u_{i0mk} = \epsilon_{i0mk}$. We normalize the mean utility of the outside option to be zero, so u_{ijmk} measures incremental utility from VM j compared to the outside option.

Putting it together, the task-level choice probability is given as follows. Let $d_{ijmk}(q_{im})$ be a dummy variable, which equals one if VM j is chosen for task k, i.e., $u_{ijmk} > u_{ilmk}$ for all $l \neq j$. Then, the probability of choosing VM j for task k of size q_{im} is given by

$$\mathbb{P}(d_{ijmk}(q_{im}) = 1) \equiv P_{ijm}(q_{im}) = \begin{cases} \frac{\exp(\alpha_i p_{jm} q_{im} + X_j \beta_i + \delta_i New_{ijm} + \xi_{ijm})}{1+\sum_{l=1}^{J} \exp(\alpha_i p_{lm} q_{im} + X_l \beta_i + \delta_i New_{ilm} + \xi_{ilm})} & \text{if } j \neq 0 \\ \frac{1}{1+\sum_{l=1}^{J} \exp(\alpha_i p_{lm} q_{im} + X_l \beta_i + \delta_i New_{ilm} + \xi_{ilm})} & \text{otherwise.} \end{cases} \tag{2}$$

3.1 Identification

To build intuition, we first discuss the identification of a simplified model without customer heterogeneity, inertia, or time trend in the number of tasks. We then conclude this section by adding them back. In the simplified model, there are at most $J+3$ parameters to identify: at most J parameters (β's) for product characteristics, one price coefficient (α), and two parameters from the task generation process (λ, γ).

Denote y_{ijm} as the total usage for customer i on product j in market m across all her tasks, i.e., $y_{ijm} = \sum_{k=1}^{n_{im}} q_{im} d_{ijmk}(q_{im})$. To identify the model, we require that for any customer i, her total usage on every product $(y_{i0m}, y_{i1m}, \cdots y_{iJm})$ is observed. This requirement is natural in most markets with continuous usage and more flexible than Burda, Harding and Hausman (2012), as we do not require the number of tasks to be observable. The likelihood function for our hierarchical model, however, unlike Burda, Harding and Hausman (2012), cannot be easily constructed. This is because our observable y_{ijm} is a continuous variable and, without observing task sizes, its density depends on an infinite sum of convoluted integration (see Appendix A for details). Therefore, we instead use generalized method of moments (GMM) and consider the following two sets of moments for identification.

Proposition 1 (Zero Usage Probability and Expected Usage Moments)**.** *The probability of customer i having zero usage on product j in market m is given by*

$$\mathbb{P}(y_{ijm}=0) = \sum_{n_{im}=0}^{\infty} \underbrace{\frac{\exp(-\lambda)\lambda^{n_{im}}}{n_{im}!}}_{Pr(Number\ of\ Tasks)} \cdot \int_{q_{im}} \underbrace{(1-P_{ijm}(q_{im})\ )^{n_{im}}}_{Pr(Product\ Choice)} \underbrace{\frac{1}{\gamma}\exp(-\frac{1}{\gamma}q_{im})}_{Pr(Task\ Size)} dq_{im}; \tag{3}$$

and the expected usage is given by

$$\mathbb{E}(y_{ijm}) = \underbrace{\lambda}_{\mathbb{E}(Number\ of\ Tasks)} \underbrace{\int_{q_{im}} q_{im} P_{ijm}(q_{im}) \frac{1}{\gamma}\exp(-\frac{1}{\gamma}q_{im})\, dq_{im}}_{\mathbb{E}(Task\ Size)}. \tag{4}$$

Proof. See Appendix B. □

In Proposition 1, the first moment $\mathbb{P}(y_{ijm} = 0)$, the probability of zero usage on each product, captures the sparsity of customers' product choices and helps identify the number of tasks. To first see how it is calculated, for any given number of tasks n_{im} and task size q_{im}, $\mathbb{P}(y_{ijm} = 0)$ has three components, as shown in equation (3). First, *Pr(Number of Tasks)* is the probability of receiving n_{im} tasks. Second, *Pr(Product Choice)* is the probability of not choosing product j in any of these n_{im} tasks of size q_{im}. Third, *Pr(Task Size)* is the probability of receiving a task of size q_{im}. Multiplying the three probabilities gives the probability of zero usage on a product for a given number of tasks and task size. We then integrate over all task numbers and sizes to obtain the zero usage moment.

To build intuition for why the sparsity of customers' product choices help identify the number of tasks, consider an extreme case of zero usage on *all* products. The probability that customer i in market m has zero usage on all products is given by

$$\mathbb{P}(\sum_{j=0}^{J} y_{ijm} = 0) \equiv \mathbb{P}(n_{im} = 0) = \exp(-\lambda).$$

If the sample analogue of this probability is non-zero, we can invert it to back out the average number of tasks λ directly. However, because we do not observe customers' total computation needs (including outside option) in our data, we do not use the probability of zero usage across all products $\mathbb{P}(\sum_{j=0}^{J} y_{ijm} = 0)$ as a moment. Instead, we use the probability of having usage on each product $\mathbb{P}(y_{ijm} = 0)$ as moments.

The second moment in Proposition 1 is the expected usage on each product $\mathbb{E}(y_{ijm})$ and helps identify task size. As shown in equation (4), product j's expected usage is calculated as the product of the expected number of tasks and the expected task size if product j is chosen. To see how each product's expected usage helps identify task size, we again take the sum of this moment across all products for customer i in market m:

$$\sum_{j=0}^{J} \mathbb{E}(y_{ijm}) \equiv \mathbb{E}(y_{im}) = \lambda\gamma,$$

which depends only on the task generation process. So if the average number of tasks λ is identified (by the first moment), this sum pins down the average task size γ.

Besides the task generation process, the zero usage probability and the expected usage on each product also help identify taste parameters α's and β's in the utility function via the market share function. Together, these two sets of moments jointly identify all parameters in the simplified model. To see that, in a market with J products and an outside option, Proposition 1 gives $J+1$ moments each for the zero usage probability and the expected usage moments. These $2(J+1)$ moment conditions are sufficient to parametrically identify the at most $J+3$ parameters of the simplified model for any $J \geq 1$.

Finally, going back to our full model, in order to identify the adoption cost for new products, customer heterogeneity, and the time trend in the number of tasks, we construct the zero usage probability and expected usage moments conditional on customer and market characteristics. Conditioning the moments on whether a product is new to a customer identifies the adoption cost parameter. Conditioning on customer size and each month in the sample identifies customer heterogeneity and the time trend in the number of tasks. We discuss these conditional moments in detail in the next section.

4 Estimation

Following the identification argument, we estimate the model with GMM. We begin by describing all the moments used in estimation. We then discuss how we account for potential price endogeneity and finally our estimation procedure.

4.1 Moments

We start with the two moments, zero usage probability $\mathbb{P}(y_{ijm} = 0)$ and expected usage $\mathbb{E}(y_{ijm})$, from Proposition 1. We then condition both moments on customer and market characteristics to further identify heterogeneity in the model. Table 2 shows all the moment conditions we use in estimation.

[Table 2 about here.]

We first condition both moments on each combination of region, customer size, and choice set. Conditioning on regions helps identify the price coefficient, and conditioning on customer size identifies heterogeneity between small and large customers. Choice sets reflect variation in the products available to customers across markets. For example, the same market's choice set is different after a product launch, so we compute moments before and after separately. Given different product availability across markets, there are eight unique choice sets in total. We then drop the zero usage probability moment for the outside option given the limitation of our data.[14] We also drop moments with few observations or outliers.[15] Second, we condition both moments on whether a cloud product (excluding outside option) is new to a customer in a market. These moments reflect differences in choice probabilities and usage with and without inertia, thus identifying the adoption cost parameters. Finally, we condition the sum of the expected usage moment across all products on each month in our sample and customer size, identifying the time trend in the number of tasks for small and large customers separately.

4.2 Price Endogeneity

A common identification challenge in demand estimation is that prices may be set based on unobserved demand shocks, i.e., in our setting, $p_{jm} \not\perp \xi_{ijm}$. To account for potential price endogeneity, we use the control function approach with cost shifters as instruments.[16]

Let z_{jm} denote cost shifters at the product-market level, which affect prices of the corresponding product-markets and are independent of demand shocks ϵ_{ijmk} and ξ_{ijm}. Following

[14] We only observe aggregate shares of outside option usage in the industry reports and apply them to each customer in each market. As a result, the outside option always has positive usage in our data.

[15] We drop moments with fewer than 200 observations. We also drop one region-product with the smallest average usage.

[16] The characteristics being conditioned on in Table 2 are sometimes called GMM "instruments". To avoid confusion, we only refer to the control function instruments as instruments.

Petrin and Train (2010), we assume price is additive in its observed and unobserved covariates, i.e.,

$$p_{jm} = W(z_{jm}, \zeta) + \theta_{jm}, \tag{5}$$

where θ_{jm} is the unobserved covariate that is independent of z_{jm} but correlated with ξ_{ijm}. We write $\xi_{ijm} = \kappa_i \theta_{jm} + \tilde{\xi}_{ijm}$ and substitute ξ_{ijm} in equation (1). Then, rewriting θ_{jm} as $p_{jm} - W(z_{jm}, \zeta)$ following equation (5), we can derive the utility function with control function as

$$u_{ijmk} = \alpha_i p_{jm} q_{im} + \beta_i X_j + \delta_i New_{ijm} + \kappa_i \left(p_{jm} - W(z_{jm}, \zeta)\right) + \tilde{\xi}_{ijm} + \epsilon_{ijmk}, \tag{6}$$

where $p_{jm} - W(z_{jm}, \zeta)$ is the control function, and $\tilde{\xi}_{ijm}$ is an independent demand shock.

4.3 Estimation Procedure

As $W(z_{jm}, \zeta)$ is unknown to the econometrician, we estimate the model in two stages. In the first stage, we regress prices p_{jm} on the instruments z_{jm}, and take the residuals, $\hat{\theta}_{jm} = p_{jm} - W(z_{jm}, \hat{\zeta})$, as estimates for the unobserved covariate. Our first two instruments include electricity prices varying across regions and power consumption ratings varying across products and regions given how VMs are supported on different hardware in different regions. Besides electricity costs, hardware procurement costs, which the provider calculates as part of variable costs, are included as the third instrument and also vary across products and regions. Detailed construction of the instruments and results of the first stage are presented in Appendix C.

We assume that the part of the unobserved demand shock that is independent of prices, $\tilde{\xi}_{ijm}$, is normally distributed with mean zero and standard deviation σ_j. The standard deviation is allowed to vary across products to capture the different degrees of demand heterogeneity. This is similar to adding a random coefficient to each product dummy in X_j.

In the second stage, we treat $\hat{\theta}_{jm}$ as a product characteristic according to equation (6). To initialize New_{ijm} for existing customers at the beginning of our sample, we hold out the first six months of data in our sample from estimation, leaving us with 26 months in our estimation sample. Furthermore, we hold out every customer's first-month data from estimation given our focus on cloud customers' product choices and usage rather than their cloud adoption decisions.

The rest of the second stage follows the standard GMM procedure using moment conditions listed in Table 2. We first compute these moments in the data. Then, for any set of parameter values, we use numerical integration to compute the model moments. Finally, we run a quasi-Newton algorithm to minimize the weighted sum of the squared differences between data moments and model moments. Details of the numerical integration and the optimization algorithm are presented in Appendix D. The weighting matrix is a diagonal matrix containing the inverse of the sample variance of the moments. We compute standard errors using the GMM asymptotic variance and taking the square root of its diagonal. In Appendix E, we also show convergence of this estimator in finite samples with Monte Carlo simulations.

5 Results

Table 3 presents our estimated demand parameters. Parameters governing product choice (both mean and standard deviation of the random coefficients) are presented separately from parameters governing task generation. The two columns show estimates separately for small and large customers.

[Table 3 about here.]

The estimates of the product dummies are in line with the evolution of the cloud products' technical specifications. Quality increases from X1 to X2 and X3 significantly due to significant improvements of the underlying CPU technology. On the other hand, the CPU improvement from X2 and X3 to X4 is on a minor Intel release, resulting in similar performance between these products.[17] Moreover, X4 is launched with hyperthreading, which reduces its cost but slightly hurts its performance. Consistent with the earlier descriptive evidence, the perceived quality of a promotional product is significantly lower. Note that our estimated product dummies are all negative, because during our sample period, cloud usage still only accounts for a small fraction of total computing needs. Finally, the estimated standard deviations of the random coefficients are statistically significant and sizeable for X1-X4, suggesting significant heterogeneity in preferences for the X-series products across customers and markets.

The price coefficients are estimated to be negative, i.e., cloud customers dislike higher prices. The significant control function parameter estimates suggest that prices are indeed endogenous and positively correlated with unobserved demand shocks in a market (first stage results reported in Appendix Table C1).

Conditional on the same amount of computing needs, we find small customers to be more price sensitive than large customers. Taking into account differences in computing needs, however, yields higher price elasticities for large customers (Table 4). The own-price elasticity for VM j is computed as the change in expected usage on VM j in response to a unit change in its price, i.e., $\frac{\partial \mathbb{E}(y_{ijm})}{\partial p_{jm}} \frac{p_{jm}}{\mathbb{E}(y_{ijm})}$ (detailed formula in Appendix F). The estimated own-price elasticities are generally inelastic, which are lower than most consumer products, though slightly higher than estimates of short-run elasticity for electricity (Deryugina, MacKay and Reif 2020). Our estimates of low price elasticities are consistent with other studies in emerging digital markets, where providers are often concerned with expanding the overall market size and long-run growth objectives and do not increase prices despite inelastic demand (Castillo 2020).

[Table 4 about here.]

[17] Intel has a tick-tock model of CPU innovation, alternating between major CPU improvements ("tock") and minor cost-reducing die shrinks ("tick") in its product launches. See https://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html, accessed on Feb. 28, 2021.

We find that customers face significant inertia when adopting a new product. The estimated adoption cost is equivalent to increasing the average product price for an average task size by 12 and 7 times for small and large customers, respectively, or 13 and 21 times the significant quality improvement from X1 to X2 and X3.[18] These large adoption costs prevent cloud customers from adopting new products, which are typically more powerful and cheaper, thus limiting the potential welfare benefits that could be realized from cloud usage. We illustrate more on this point in the counterfactual section. Finally, small customers face lower adoption costs than large customers when adopting a new product.

Turning to the task generation process, we find that small customers have lower number of tasks as well as smaller task sizes than large customers. On average, at the beginning of our estimation sample in May 2016, we estimate small customers have 43 tasks per month while large customers have 81 tasks. These tasks represent the number of distinct computing projects for a customer in a region-OS per month. The number of tasks for small customers, however, grows about 4 times as fast as that of the large customers over time, capturing faster growth rates of small cloud customers. By June 2018, at the end of our sample, the average small customer has 89 tasks per month while the average large customer has 92 tasks, suggesting that size differences are shrinking on the cloud. Small customers also have smaller task sizes than large customers, on average 71% smaller in compute units for each task.

5.1 Model Fit

To show how our model fits the data, we initialize customers' adoption history using each customer's first month usage in our data, and then simulate the model forward for each customer with the rest of the data. Importantly, whether a product is new to any customer New_{ijt} after the first month is simulated within the model. We compare the simulated usage

[18] The equivalent price increases are calculated as $\frac{\delta_i}{0.56 \cdot \alpha_i \gamma_i}$, where 0.56 is the average price across product-markets in the data and i varies between small and large. The equivalent quality increases are calculated as $\frac{\delta_i}{\beta_i^2 - \beta_i^1}$.

with the usage in the data.

The simulation proceeds as follows: (i) take each customer's first month's data as given and initialize which products the customer has used; (ii) starting in the second month, for every customer in every market, draw their number of tasks and task sizes from the estimated distributions; (iii) for each task, draw i.i.d. type-I extreme value shocks and random coefficients for each product, and then make product choice based on estimated utilities; (iv) aggregate usage for each customer-product, and every time a customer uses a new product, update the new product indicator in her utility function; (v) iterate until the end of the sample.

Figure 2 shows that our estimated model fits the data well. For each simulation, we calculate the distribution of average usage for each product-market, and then average across 200 simulations. The simulated usage captures well the variation in usage across products and markets. For example, the simulated usage tracks the low usage of X4 in the data despite its high estimated utility, suggesting that our model captures the effect of customer inertia on newer products. The model also predicts well zero usage for each product in the sample. Moreover, Appendix Figure H3 shows that we fit the aggregate time series of total computing demand well, and Appendix Figure H4 shows the fit of cloud usage over time by product and customer size.

[Figure 2 about here.]

6 Counterfactuals

With our model and estimates, we conduct two sets of counterfactual analyses. First, we compute the welfare benefits of cloud usage under current market conditions and compare it to a model where customer inertia is eliminated. Second, we explore two potential remedies that incentivize customers to overcome inertia and adopt new products.

6.1 Welfare

Using our model and estimates, we first compute the welfare benefits of cloud usage for customers during our sample with our estimated inertia. To do that, we follow the same procedure described in Section 5.1 when evaluating model fit by simulating each customer's product choices and usage forward over time. We convert the utility received for each task into dollars by dividing the price coefficient, and then aggregate across all tasks to obtain consumer surplus. Provider revenue is calculated as the total usage on each product multiplied by the corresponding prices. We repeat these simulations 200 times for each customer and compute the averages.

The first row in Table 5 shows our estimates of welfare benefits from cloud usage. The dollar amounts are on a normalized scale due to data anonymization, so we focus on interpreting cloud customers' return on investment (i.e., ROI ≡ consumer surplus / provider revenue). We find large welfare gains from cloud usage. The average ROI for all cloud customers is 216% (= 51.6/23.9) or roughly 2.2x. Small customers have higher ROIs (2.7x) than large customers (2.0x), even though their total consumer surplus is lower given their lower total usage. This is consistent with the hypothesis that cloud usage disproportionately benefits small firms by allowing them to easily scale their computing needs and access the latest technology.

[Table 5 about here.]

Our estimated ROI from cloud usage is comparable to ROI estimates of other computing technologies: it is higher than that of PCs (0.92x, Hendel 1999) and lower than that of mainframes (3.3-6.1x, Bresnahan 1986). Like previous studies, our estimate should be taken as an upper bound because it only reflects the welfare of using cloud for existing cloud customers — there may be other costs (e.g., labor training) associated with transitioning to cloud that are not accounted for in our model. Moreover, cloud providers, similar to providers in other digital markets, often optimize for consumer surplus for long-run growth rather than short-run profits (Castillo 2020), and thus early cloud adopters may extract more surplus.

We then re-do the simulations with our estimated model but with the adoption cost eliminated to evaluate the welfare benefits that could be realized from cloud usage in a world without inertia. The second row in Table 5 shows the results. The third row then shows the difference between the first and the second row, illustrating the welfare cost of inertia. Consumer surplus is 62% (= 85.3/136.9) lower than what it could be if there were no inertia. Correspondingly, if there were no inertia, cloud customers' ROI would be 2.4x compared to 2.2x with inertia. Interestingly, provider revenue is also 58% (= 33.0/56.9) lower in the presence of inertia, despite customers using older and more expensive products.[19] This is mainly due to differences in overall cloud usage: without inertia, customers would adopt new products from the cloud provider faster and move more usage from on-premise and other cloud providers. As the last column shows, total cloud usage would more than double if inertia is fully eliminated. Appendix Figure H5 further decomposes time series of cloud usage by product and customer size with and without inertia: while all cloud products' usage increase without inertia, new product launches—X3 and X4—gain popularity much more quickly.

How much of the 62% consumer surplus lost due to inertia are true welfare loss? Because we agnostically estimate inertia as adoption costs for new products, we cannot speak to specific mechanisms driving customer inertia. It is possible that some of the adoption costs we estimate are real costs firms need to incur when adopting a new VM (e.g., engineering cost), and thus they do not represent true welfare loss. However, we can bound the true welfare loss from inertia by decomposing the consumer surplus difference with and without inertia into a direct adoption cost and an indirect cost from sub-optimal product choices. While the former is mechanism agnostic, the latter provides a lower bound for the true welfare loss due to inertia.

[19] The average price per unit of cloud usage is $23.9/35.8 = 0.67$ with inertia and $56.9/99.7 = 0.57$ without inertia.

To first calculate the direct adoption cost, in the simulations without inertia, every time a new product is chosen, we sum up the adoption cost it would have incurred, and average across simulations. We then attribute the rest of the consumer surplus lost from inertia to sub-optimal product choices. Table 6 presents the results separately for small and large customers, which are similar. On average, only 2% of the loss in consumer surplus come from direct adoption costs, and the rest 98% are lost due to customers not choosing their most preferred products had there been no inertia. These results show that the vast majority of consumer surplus lost from inertia are true welfare loss, and thus we may be able to improve total welfare by incentivizing customers to adopt new products.

[Table 6 about here.]

Finally, one may argue that the result of this decomposition depends on the length of the sample and results from Table 6 only apply to our particular sample. As a robustness check, in Appendix Table H3, we re-do the decomposition exercise for different lengths of samples by simulating our sample forward with the same set of customers, choice sets, and prices. We find that the direct adoption cost consistently accounts for a similar percentage of total consumer surplus lost from inertia.

6.2 Remedies

Given the large welfare cost of inertia, in this section, we explore two remedies that incentivize customers to adopt new products.

6.2.1 Subsidy for Cloud Migration

Cloud providers and third-party service providers often provide cloud migration services to help customers move their computing workloads. These services range from basic information provision to automated tools, or in some cases white-glove migration services for typically the largest customers. We first explore a counterfactual scenario where such services fully subsidize and eliminate customer inertia during our sample period. Welfare under this scenario is thus the same as if there were no inertia as shown in the second row of Table 5.

One way to implement this subsidy of adoption cost is via public financing. Table 7 shows how much it would cost (in anonymized dollars) to fully subsidize adoption costs for each product and customer segment. The subsidy cost is calculated in the same way as the direct adoption cost by adding up the adoption cost every time a customer chooses a new product in the simulations without inertia. This assumes that all of the adoption costs are real costs, and thus the subsidy costs in Table 7 should be viewed as upper bounds. So the total cost to fully subsidize adoption costs for all products and customers amounts to at most \$59.2k. The total welfare gain (consumer surplus plus provider revenue) from eliminating adoption costs is \$118.3k (=85.3k + 33.0k, from Table 5). In other words, the social planner would gain at least 2 dollars in welfare for each dollar spent on the subsidy. Besides welfare, this subsidy would also more than double the total amount of cloud usage (Table 5).

[Table 7 about here.]

Another way to implement this subsidy is via cloud providers themselves since revenue also goes up in the absence of inertia. However, at least during our sample, the cost of this subsidy (\$59.2k) exceeds the revenue benefits (\$33.0k) cloud providers would gain from eliminating inertia. This is also true for any specific product and customer segment. This suggests that, left on their own, cloud providers should not subsidize migration services. However, it is possible that cloud providers stand to gain more revenue benefits from eliminating inertia in the long run because they pay for customers' adoption costs early on and receive revenue benefits from more cloud usage over time. In the next section, we compare long-run benefits of the full subsidy together with other remedies.

6.2.2 Introductory Discount

One reason subsidizing adoption costs may not be profitable for providers is that an adoption cost is paid for regardless of the size of the benefits to providers (e.g., usage or revenue).

Another common approach online service providers use to help customers overcome inertia and adopt new products is in the form of introductory discounts for new products or new users of a product. Introductory discounts also reduce providers' short-run revenue and potentially increase their long-run revenue by incentivizing customers to adopt new products, but do so by only subsidizing part of adoption costs proportional to revenue.

We first study a product-level introductory discount, often referred to as "public new product preview" or "open beta", which discounts a new product after its initial launch. A second form of introductory discount we explore is a "personalized product trial" that discounts a product for any customers who have not used it before. For both types of discounts, we assume they only last for one month. Finally, we explore how much finer level of targeting by customer segments may improve the outcome of these introductory discounts.

Our goal is to first assess whether different forms of introductory discounts benefit the provider in the long run, and if so, what the optimal discounts are for the provider. Taking those optimal discounts, we then evaluate their impact on consumer surplus and total welfare. To do so, we take the launches of X4 across markets as an example. We consider the provider's problem of finding optimal discounts for long-run revenue.[20] To solve the provider's problem, we conduct a grid search over a wide range of discounts for X4 for the first month after its launch, holding everything else constant. For each discount, we fix characteristics of each market (customers, choice sets, and prices) and simulate demand forward, allowing the number of tasks to grow following estimated time trends. Finally, we average across 1000 simulations to find the provider's optimal discounts and evaluate their impact on welfare and cloud usage. Because the optimal discounts depend on how far out we simulate as the provider trades off short-run costs vs. long-run benefits, in Appendix G, we show that the optimal discounts increase as the provider's optimization horizon increases, and stabilize once we simulate 96 months after the initial X4 launch, which we use as our main results in this section.

[20] We assume a monthly discount factor of 0.99.

Table 8 presents the results, relative to a baseline when there are no discounts at all. In the baseline, we simply simulate our model with inertia for 96 months, holding market characteristics fixed from the launch of X4, and calculate consumer surplus, provider revenue, and total cloud usage.

In the first row of Table 8, we find the optimal discount for a uniform product-level introductory discount ("new product preview") at the launch of X4 to be 250%, i.e., the provider pays back to customers (possibly in store credits) 1.5 times their cloud spending on X4 in the first month after its launch.[21] The optimal discount is high because of the low price elasticity and high adoption cost we estimate in the model. With this product-level introductory discount, we find that the provider can improve its revenue by 1.3%, and at the same time, increase consumer surplus by 1.9%, improving total welfare by 1.7%. So the introductory discount results in a Pareto improvement of total welfare. Total cloud usage also goes up by 1.9%.

[Table 8 about here.]

In the second row, we allow the provider to give targeted differential discounts for different customer segments. We find that the optimal discount for small customers is 325% vs. 250% for large customers. There are two reasons for why the provider should give a higher discount to smaller customers in our setting, which is contrary to what managers typically do in practice. First, small customers are less price elastic after taking into account their usage and thus need a higher discount to overcome similar adoption costs. Second, because small customers have lower usage in the beginning but grow faster, it is cheaper for the provider to discount small customers upfront to encourage new product adoptions and reap the benefits of their higher usage and revenue later. Finally, the targeted discounts only yield a small increase in provider revenue, consumer surplus, total welfare, and cloud usage compared to the uniform discount.

[21] Another interpretation of the magnitude of the discount is that it represents the combined discount in one month from a multi-month campaign.

We now turn to introductory discounts at the customer-product level ("personalized product trials"). The product-level introductory discounts only yield small improvements in revenue for the provider due to its restrictive timing (i.e., only during the month of initial product launch). Personalized product trials relax this restriction by giving discounts to customers whenever they try a new product for the first time, i.e., the timing of discount eligibility is personalized.

The third row in Table 8 shows the results for a uniform customer-product level introductory discount. We find that the optimal discount for this more flexible introductory discount is 225%, slightly lower than the optimal product-level introductory discount. In other words, under this optimal discount, the provider would pay its customers 1.25 times their first month's cloud spending on X4 if they have not previously used it. While the optimal discount is lower, the flexibility of the customer-product level discount leads to a significant 12.8% revenue increase compared to the baseline. Consumer surplus also increases by 16.5%, and total welfare by 15.4%. Thus, the uniform personalized product trial dominates new product previews in terms of both consumer surplus and provider revenue. Total cloud usage also increases by 16.7%.

In the next and fourth row, we again show the results for a more targeted and differential personalized product trial discounting for different customer segments. The optimal discounts are, again, higher for small customers than large customers. While the provider revenue (necessarily) increases under the targeted discounting compared to the uniform personalized product trial, consumer surplus decreases, and so do total welfare and cloud usage. This is because with differential discounts the provider optimally chooses a smaller discount for large customers than under uniform discounting, which slows down cloud adoption and decreases consumer surplus and total welfare.

Finally, to facilitate comparison, we repeat the exercise from the previous section and simulate demand and welfare under a full subsidy of adoption costs for X4 for 96 months after its initial launch (rather than during our sample period in the previous section). The last row in Table 8 shows the results. Consumer surplus is the highest compared to other remedies, representing a 24.5% increase compared to the baseline. Provider revenue net of the costs of subsidizing X4's adoption costs is lower than those under the personalized product trials but higher than new product previews, representing a 11.8% increase compared to the baseline. So while it is not profitable for the provider to fully subsidize adoption costs during our sample period, it is profitable in the long run as the provider continues to receive revenue from customers who overcome their inertia to adopt new cloud products. This may explain the various cloud migration and management services offered in the market. Total cloud usage, similar to consumer surplus, is also the highest under the full subsidy. Taken together, we find that fully subsidizing adoption costs yields the highest total welfare, and although also profitable for the provider, it generates lower revenue compared to personalized product trials.

7 Conclusion

The public cloud is one of the most important technological innovations in the 21st century. It has grown to be a significant industry and enabled much of the rest of the economy to digitize. Both digitization and the cloud itself are still fast growing and becoming increasingly important. Cloud alleviates customers' needs to buy and maintain physical computing hardware and thus lowers the upfront cost of owning computing resources and affords customers of all sizes access to the latest technology. At the same time, lack of maintenance needs and continuous feature updates from cloud providers may induce customers to develop inertia that prevents them from adopting new and the best products.

In this paper, we first estimate significant (2.2x) returns to customers using cloud, comparable to estimates for the significant inventions of mainframes and PCs in the 20th century. Different from those technologies, we find that cloud disproportionately benefits small customers due to its nature of democratizing computing resources from ownership to access.

Despite the significant welfare benefits, we estimate that more than half of the potential welfare benefits from cloud are lost because customers face inertia in adopting new products, which lowers both consumer surplus and provider revenue, as customers are slower to adopt cloud products. To help customers overcome inertia and adopt new products, we find that both subsidizing adoption costs (e.g. offering cloud migration services) and introductory discounts (e.g., new product preview or personalized product trials) are effective. The subsidy, if financed publicly, can generate substantial returns, and, while not profitable for the provider during our sample if financed privately, is profitable in the long run, highlighting the short vs. long-run tradeoff for any policy. Personalized product trials offering first-month discounts for customers to adopt a new product yields the highest revenue gain for the provider, while increasing consumer surplus at the same time, resulting in a Pareto improvement compared to the baseline.

One major contribution of the paper is the multiple-choice continuous-usage demand model itself. The model only requires the researcher to observe customer-product level usage, and thus can be easily applied to other markets with multiple product choices and continuous usage on each product. One limitation of the model in our setting is that demand is static, whereas customers facing inertia may exhibit forward-looking behavior, depending on their level of sophistication. We do not capture such dynamics given the complexity of the static model, in line with the literature on consumer inertia.

Another limitation of our paper is that we only focus on existing customers' demand for products from one cloud provider. We thus abstract away from customers' choices over cloud providers, due to our data limitation. While this abstraction has limited impact on our welfare estimates, for future research, it would be important to understand migration frictions between cloud providers and how cloud providers' pricing and product entry strategies shape the industry dynamics.

Finally, an alternative way to measure welfare benefits of cloud is to analyze the impact of cloud adoption and usage on direct measures of firm performance, which is out of the scope of this paper due to data limitations but would be complementary to our paper and an interesting avenue for future research.

References

Bartoš, Vojtěch, Michal Bauer, Julie Chytilová and Filip Matějka. 2016. "Attention discrimination: Theory and field experiments with monitoring information acquisition." *American Economic Review* 106(6):1437–75.

Berry, Steven, James Levinsohn and Ariel Pakes. 1995. "Automobile prices in market equilibrium." *Econometrica* 63(4):841–890.

Berry, Steven T. 1994. "Estimating discrete-choice models of product differentiation." *The RAND Journal of Economics* 25(2):242–262.

Bessen, James. 2020. "Industry concentration and information technology." *The Journal of Law and Economics* 63(3):531–555.

Bhargava, Saurabh, George Loewenstein and Justin Sydnor. 2017. "Choose to lose: Health plan choices from a menu with dominated option." *The Quarterly Journal of Economics* 132(3):1319–1372.

Bresnahan, Timothy F. 1986. "Measuring the spillovers from technical advance: mainframe computers in financial services." *American Economic Review* 76(4):742–755.

Brynjolfsson, Erik and Lorin M Hitt. 2000. "Beyond computation: Information technology, organizational transformation and business performance." *Journal of Economic Perspectives* 14(4):23–48.

Burda, Martin, Matthew Harding and Jerry Hausman. 2012. "A Poisson mixture model of discrete choice." *Journal of Econometrics* 166(2):184–203.

Burnham, Thomas A, Judy K Frels and Vijay Mahajan. 2003. "Consumer switching costs: a typology, antecedents, and consequences." *Journal of the Academy of Marketing Science* 31(2):109–126.

Castillo, Juan Camilo. 2020. "Who benefits from surge pricing?" *Available at SSRN 3245533.*

Cisco. 2018. Cisco Global Cloud Index: Forecast and Methodology, 2016-2021. White paper Cisco.

Citizens Advice. 2018. Excessive prices for disengaged consumers. Technical report.
URL: *https://www.citizensadvice.org.uk/Global/CitizensAdvice/Consumer publications/Super-complaint - Excessive prices for disengaged consumers (1).pdf*

Cyert, Richard M and James G March. 1963. *A behavioral theory of the firm.* Englewood Cliffs, N.J.: Prentice-Hall.

De Loecker, Jan, Jan Eeckhout and Gabriel Unger. 2020. "The rise of market power and the macroeconomic implications." *The Quarterly Journal of Economics* 135(2):561–644.

Decker, Ryan A, John Haltiwanger, Ron S Jarmin and Javier Miranda. 2016. "Declining business dynamism: What we know and the way forward." *American Economic Review* 106(5):203–07.

Deryugina, Tatyana, Alexander MacKay and Julian Reif. 2020. "The long-run dynamics of electricity demand: Evidence from municipal aggregation." *American Economic Journal: Applied Economics* 12(1):86–114.

Erdem, Tülin and Michael P Keane. 1996. "Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets." *Marketing Science* 15(1):1–20.

Farrell, Joseph and Paul Klemperer. 2007. Coordination and lock-in: Competition with switching costs and network effects. Vol. 3 of *Handbook of Industrial Organization* Elsevier.

Forman, Chris. 2005. "The corporate digital divide: Determinants of Internet adoption." *Management Science* 51(4):641–654.

Gartner. 2017. "Gartner says worldwide IaaS public cloud services market grew 31 percent in 2016.".

URL: *https://www.gartner.com/en/newsroom/press-releases/2017-09-27-gartner-says-worldwide-iaas-public-cloud-services-market-grew-31-percent-in-2016*

Gartner. 2022. "Gartner Says Worldwide IaaS Public Cloud Services Market Grew 41.4% in 2021.".

URL: *https://www.gartner.com/en/newsroom/press-releases/2022-06-02-gartner-says-worldwide-iaas-public-cloud-services-market-grew-41-percent-in-2021*

Greenstein, Shane M. 1993. "Did installed base give an incumbent any (measureable) advantages in federal computer procurement?" *The RAND Journal of Economics* 24(1):19–39.

Handel, Benjamin and Joshua Schwartzstein. 2018. "Frictions or mental gaps: what's behind the information we (don't) use and when do we care?" *Journal of Economic Perspectives* 32(1):155–78.

Handel, Benjamin R and Jonathan T Kolstad. 2015. "Health insurance for" humans": Information frictions, plan choice, and consumer welfare." *American Economic Review* 105(8):2449–2500.

Hanna, Rema, Sendhil Mullainathan and Joshua Schwartzstein. 2014. "Learning through noticing: Theory and evidence from a field experiment." *The Quarterly Journal of Economics* 129(3):1311–1353.

Hendel, Igal. 1999. "Estimating multiple-discrete choice models: An application to computerization returns." *The Review of Economic Studies* 66(2):423–446.

Hortaçsu, Ali and Chad Syverson. 2004. "Product differentiation, search costs, and competition in the mutual fund industry: A case study of S&P 500 index funds." *The Quarterly Journal of Economics* 119(2):403–456.

Jin, Wang and Kristina McElheran. 2017. "Economies before scale: survival and performance of young plants in the age of cloud computing." *Rotman School of Management working paper* (3112901).

Kilcioglu, Cinar and Justin M Rao. 2016. Competition on price and quality in cloud computing. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee pp. 1123–1132.

Koulayev, Sergei, Marc Rysman, Scott Schuh and Joanna Stavins. 2016. "Explaining adoption and use of payment instruments by US consumers." *The RAND Journal of Economics* 47(2):293–325.

Liebman, Jeffrey B and Neale Mahoney. 2017. "Do expiring budgets lead to wasteful year-end spending? Evidence from federal procurement." *American Economic Review* 107(11):3510–49.

MacKay, Alexander and Marc Remer. 2022. "Consumer inertia and market power." *Available at SSRN 3380390*.

Peng, Sida, Peichun Wang, Eric Auerbach, Hongwei Liang and Andy Wu. 2021. "Digitization and employment in the pandemic: Evidence from seventy billion emails." Working Paper.

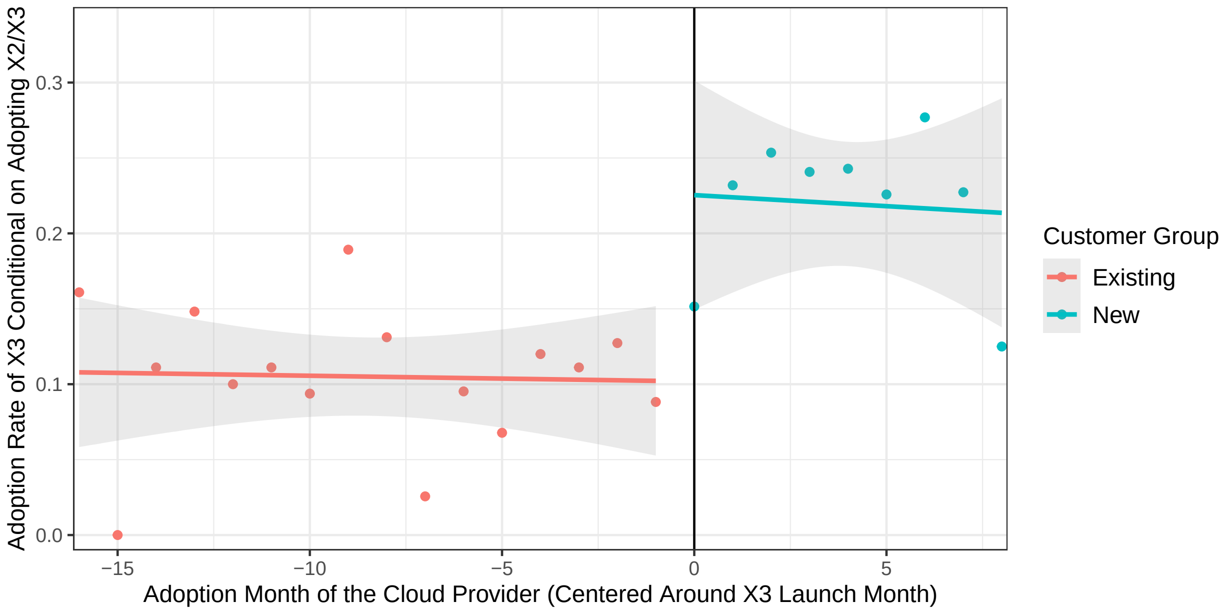
Petrin, Amil and Kenneth Train. 2010. "A control function approach to endogeneity in consumer choice models." *Journal of Marketing Research* 47(1):3–13.

Tambe, Prasanna, Lorin Hitt, Daniel Rock and Erik Brynjolfsson. 2020. Digital capital and superstar firms. Working Paper 28285 National Bureau of Economic Research.

Tambe, Prasanna and Lorin M Hitt. 2012. "The productivity of information technology investments: New evidence from IT labor data." *Information Systems Research* 23(3-part-1):599–617.

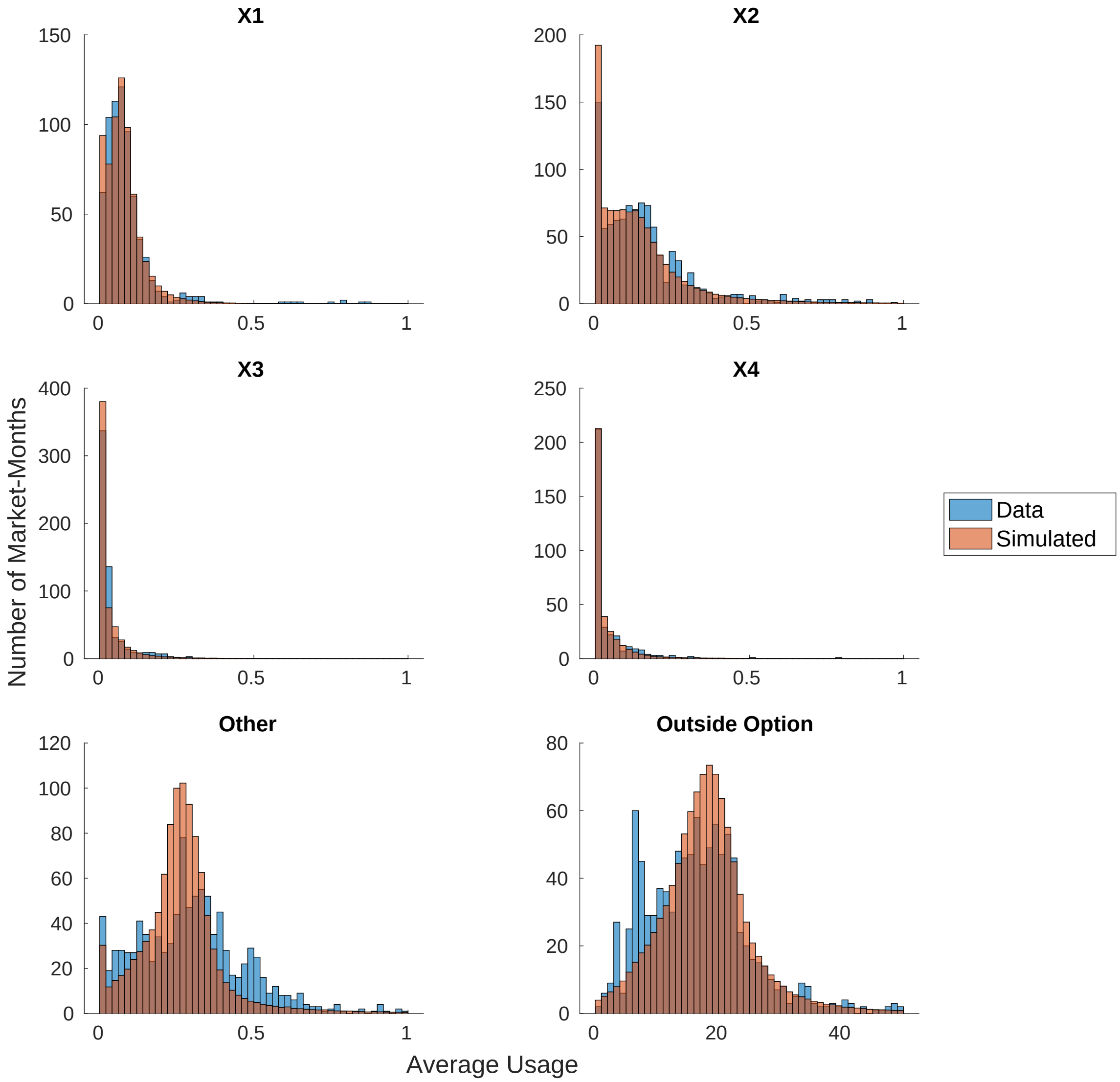
Figures

Figure 1: X3 adoption rates: New vs. existing customers



Notes: Regression discontinuity plot comparing adoption rates of X3 between existing X2 customers and new customers of the provider after X3's launch. The running variable is each customer-cohort's month of first adopting the provider and centered around X3's launch month. The outcome variable is each customer-cohort's probability of adopting X3, conditional on adopting X2 or X3, in the 12 months after X3's launch. Lines are linear regressions of the adoption rates on the running variable. Gray shaded area represents 95% confidence interval.

Figure 2: Model fit: Distribution of average usage for each product-market-month



Notes: Figure shows histograms of average usage in a market-month for each product based on data vs. simulations. The blue histogram are based on the data, and the orange histograms are based on simulations. Each bar represents the number of market-months that fall within the corresponding average usage for that product. Results are averaged across 200 simulations.

Tables

Table 1: Number of customers by product choices and total usage

Usage / Products	1	2	3	4	5	Total
0-0.1	10316	479	73	3	1	10872
0.1-1	38469	4469	523	17	1	43479
1-10	5002	3613	1130	159	18	9922
10-100	63	97	93	17	3	273
Total	53850	8658	1819	196	23	64546

Notes: Table shows the number of customer-market-months tabulated by the number of distinct products chosen and total usage. Usage is measured in compute units and re-scaled for confidentiality. Products include X1-X4, and the "other" product.

Table 2: Moment conditions

Moments	Characteristics conditioned on	# Moment conditions
Zero usage probability	- region, customer size, choice set	178
	- New_{ijm} (for $j \neq 0$)	20
Expected usage	- region, customer size, choice set	222
	- New_{ijm} (for $j \neq 0$)	20
	- month (t_m), customer size	52

Table 3: Demand estimates

	Variables	Coefficient estimates	
		Small	Large
Product choice	X1 (β_i^1)	-4.324	-3.460
		(0.063)	(0.041)
	X2/X3 (β_i^2)	-3.884	-3.174
		(0.059)	(0.040)
	X4 (β_i^4)	-4.020	-3.134
		(0.089)	(0.059)
	Other (β_i^O)	-2.609	-2.343
		(1.472)	(1.466)
	Promo (β_i^P)	-0.850	-0.303
		(0.072)	(0.039)
	Price $\times$ task size (α_i)	-7.857	-4.226
		(0.947)	(0.264)
	Adoption cost (δ_i)	-5.535	-6.053
		(0.077)	(0.118)
	Control function (κ_i)	3.359	0.911
		(0.346)	(0.162)
Random coefficient	X1 (σ_1)	1.149	
		(0.061)	
	X2 (σ_2)	1.018	
		(0.034)	
	X3 (σ_3)	1.311	
		(0.095)	
	X4 (σ_4)	1.381	
		(0.075)	
	Other (σ_O)	0.002	
		(8.620)	
Number of tasks	Intercept (λ_{i0})	43.135	80.848
		(1.388)	(2.906)
	Time trend (λ_{i1})	1.846	0.460
		(0.141)	(0.127)
Task size	Exponential scale (γ_i)	0.106	0.371
		(0.003)	(0.008)
Observations		172,223	167,244

Notes: Table presents estimates of demand parameters. The standard errors are computed using estimated GMM asymptotic variance.

Table 4: Own-price elasticities

Product	Avg.	Small customers	Large customers
X1	-0.824	-0.671	-0.988
X2	-0.732	-0.585	-0.887
X3	-0.629	-0.487	-0.766
X4	-0.744	-0.588	-0.887
Other	-0.941	-0.788	-1.102

Notes: Table presents estimated own-price elasticities for different products and customer sizes.

Table 5: Welfare benefits of cloud usage and cost of inertia

		Consumer Surplus (10^3)		**Provider Revenue (10^3)**		**Customer ROI (CS/Revenue)**		**Cloud Usage (10^3)**	
		(Small)	(Large)	(Small)	(Large)	(Small)	(Large)	(Small)	(Large)
Welfare: With inertia	By size	13.2	38.4	5.0	18.9	2.7	2.0	6.7	29.2
		(0.1)	(0.3)	(0.04)	(0.1)	(0.02)	(0.01)	(0.1)	(0.2)
	Total	51.6		23.9		2.2		35.8	
		(0.3)		(0.1)		(0.01)		(0.2)	
Counterfactual: Without inertia	By size	33.4	103.5	11.0	45.9	3.0	2.3	17.5	82.2
		(0.2)	(0.5)	(0.1)	(0.2)	(0.02)	(0.01)	(0.1)	(0.4)
	Total	136.9		56.9		2.4		99.7	
		(0.5)		(0.2)		(0.01)		(0.4)	
Δ Welfare (without - with)	By size	20.2	65.1	6.0	27.0	0.3	0.3	10.8	53.1
		(0.2)	(0.6)	(0.1)	(0.2)	(0.03)	(0.02)	(0.1)	(0.5)
	Total	85.3		33.0		0.2		63.9	
		(0.6)		(0.3)		(0.02)		(0.5)	

Notes: Table shows consumer surplus, provider revenue, and total cloud usage in the estimated model and in a model where inertia is eliminated. Consumer surplus and provider revenue are shown in anonymized dollars. Cloud usage is shown in anonymized compute units. Results are based on 200 simulations, with standard deviations in parentheses.

Table 6: Decomposition of consumer surplus lost due to inertia

	Small	Large
Δ Consumer surplus (10^3)	20.2	65.1
- Direct adoption cost	2.4%	1.6%
- Indirect cost from sub-optimal product choices	97.6%	98.4%

Notes: Table shows the difference between consumer surplus without inertia and with inertia, and decomposes this difference into direct adoption costs and indirect costs from sub-optimal product choices.

Table 7: Cost of subsidizing adoption costs (10^3)

Product	All Customers	Small	Large
X1	9.9	2.5	7.4
	(0.2)	(0.1)	(0.2)
X2	10.8	3.1	7.7
	(0.3)	(0.1)	(0.2)
X3	15.2	3.1	12.1
	(0.4)	(0.1)	(0.4)
X4	18.7	4.5	14.2
	(0.5)	(0.1)	(0.5)
Other	4.5	0.8	3.7
	(0.1)	(0.03)	(0.1)
Total	59.2	14.0	45.1
	(0.7)	(0.2)	(0.7)

Notes: Table presents costs (in anonymized dollars) of fully subsidizing adoption costs during our sample, broken down by products and customer sizes. Results are based on 200 simulations, with standard deviations in parentheses.

Table 8: Effect of different remedies for X4 launch in comparison to the baseline

		% Difference from Baseline			
	Optimal Discount	Provider Revenue	Consumer Surplus	Total Welfare	Cloud Usage
1. New product preview	250%	1.3%	1.9%	1.7%	1.9%
2. New product preview (by customer size)	Small: 325% Large: 250%	1.3%	2.1%	1.9%	2.0%
3. Personalized product trial	225%	12.8%	16.5%	15.4%	16.7%
4. Personalized product trial (by customer size)	Small: 275% Large: 200%	13.2%	16.0%	15.2%	15.8%
5. Subsidizing X4 adoption costs		11.8%	24.5%	20.8%	21.4%

Notes: Table compares welfare and cloud usage in the baseline (i.e., no discount) vs. product-level introductory discounts (i.e., new product preview) vs. customer-product level introductory discounts (i.e., personalized product trial) vs. fully subsidizing adoption costs whenever a customer chooses X4 for the first time. Optimal discounts are calculated based on grid searches over a wide range of discounts and 96 months of simulation after X4's initial launch. Provider revenue, consumer surplus, total welfare, and cloud usage are shown in percentage improvements relative to the baseline. All results are averages across 1000 simulations.

Appendices

A Maximum Likelihood Estimation

In this section, we discuss challenges of maximum likelihood estimation for our model. Suppose for now that task sizes are known and equal to one, i.e., $q_{im} = 1$. Then our model collapses to that of Burda, Harding and Hausman (2012), and the likelihood of $(y_{i0m}, y_{i1m}, ..., y_{iJm})$ can be written as

$$\begin{aligned} & L_i(\lambda_{im}, \alpha_i, \beta_i, \delta_i | y_{i0m}, y_{i1m}, \cdots y_{iJm}) \\ =& \frac{\exp(-\lambda_{im}) \lambda_{im}^{\sum_{j=0}^{J} y_{ijm}}}{(\sum_{j=0}^{J} y_{ijm})!} \cdot \prod_{j=0}^{J} P_{ijm}(1)^{y_{ijm}}, \end{aligned}$$

which is a product of two components: 1) the probability of receiving $\sum_{j=0}^{J} y_{ijm}$ tasks; and 2) the probability that VM j is chosen for exactly y_{ijm} tasks, for all j. This likelihood is easy to compute because the task size is known. However, if the task size is unknown, the likelihood becomes the sum of the probabilities of all scenarios that generate usages $(y_{i0m}, y_{i1m}, ..., y_{iJm})$. For example, if we observe usage of $y_{ijm} = 10$, it could be generated by one task of size 10, two tasks of sizes $(1, 9)$, $(2, 8)$, and so on, amounting to infinite possibilities. In other words, the likelihood becomes an infinite sum of convoluted integration, which is computationally infeasible. Therefore, we take the GMM approach instead.

B Proof of Proposition 1

Proof. For the simplified model without customer heterogeneity, inertia, or time trend in the number of tasks, let the probability of customer i choosing product j in market m for a task of size q_{im} be denoted by $P_{ijm}(q_{im})$, i.e.,

$$P_{ijm}(q_{im}) = \begin{cases} \frac{\exp(\alpha p_{jm} q_{im} + X_j \beta)}{1+\sum_{l=1}^{J} \exp(\alpha p_{lm} q_{im} + X_l \beta)} & \text{if } j \neq 0, \\ \frac{1}{1+\sum_{l=1}^{J} \exp(\alpha p_{lm} q_{im} + X_l \beta)} & \text{otherwise.} \end{cases}$$

Zero Usage Probability Moment For a given number of tasks n_{im} and a task size q_{im}, the probability of customer i not using product j in market m is equal to the probability of customer i not choosing product j in market m for any of the n_{im} tasks. Because tasks are independent, this probability is equal to $(1 - P_{ijm}(q_{im}))^{n_{im}}$, which we call Pr(Product Choice). Now, to obtain the probability of zero usage at the customer-product level, we first integrate the probability of product choice over the exponential distribution of task size, and then the Poisson distribution of the number of tasks:

$$\mathbb{P}(y_{ijm} = 0) = \sum_{n_{im}=0}^{\infty} \underbrace{\frac{\exp(-\lambda)\lambda^{n_{im}}}{n_{im}!}}_{Pr(\text{Number of Tasks})} \cdot \int_{q_{im}} \underbrace{(1 - P_{ijm}(q_{im})\ )^{n_{im}}}_{Pr(\text{Product Choice})} \underbrace{\frac{1}{\gamma}\exp(-\frac{1}{\gamma} q_{im})}_{Pr(\text{Task Size})} \, dq_{im}.$$

Expected Usage Moment Again, first fix the number of tasks to be n_{im}. By integrating over the exponential distribution of task size and the probability of product j being chosen, we obtain the average task size on product j:

$$\int_{q_{im}} q_{im} P_{ijm}(q_{im}) \frac{1}{\gamma} \exp(-\frac{1}{\gamma} q_{im}) dq_{im}.$$

Then, we multiply it by n_{im} to get the expected usage for each product given n_{im} tasks, because task sizes are the same across tasks for the same customer in the same market. Finally, integrating over the Poisson distribution of the number of tasks, we obtain the expected usage at the customer-product level:

$$\begin{aligned}
\mathbb{E}(y_{ijm}) &= \sum_{n_{im}=0}^{\infty} \frac{\exp(-\lambda)\lambda^{n_{im}}}{n_{im}!} \cdot \left(n_{im} \int_{q_{im}} q_{im} P_{ijm}(q_{im}) \frac{1}{\gamma} \exp(-\frac{1}{\gamma} q_{im}) dq_{im} \right) \\
&= \left(\sum_{n_{im}=0}^{\infty} \frac{\exp(-\lambda)\lambda^{n_{im}}}{n_{im}!} \cdot n_{im} \right) \int_{q_{im}} q_{im} P_{ijm}(q_{im}) \frac{1}{\gamma} \exp(-\frac{1}{\gamma} q_{im}) dq_{im} \\
&= \underbrace{\lambda}_{\mathbb{E}(\text{Number of Tasks})} \underbrace{\int_{q_{im}} q_{im} P_{ijm}(q_{im}) \frac{1}{\gamma} \exp(-\frac{1}{\gamma} q_{im})\, dq_{im}}_{\mathbb{E}(\text{Task Size})},
\end{aligned}$$

where λ is exactly the expected number of tasks of the Poisson distribution.

□

C Prices, Instruments, and First Stage Results

The price for the "other" VM, which includes all of the provider's VM products other than the X series, is constructed as a demand-weighted average price. Specifically, in each market, we calculate the price index as the average price weighted by the demand of each product from 2017, so that the variation of the price index over time is driven by changes in product availability and price changes, rather than demand.

We use three cost shifters as instruments for prices. Cloud providers' variable costs vary across regions and products due to regional differences in electricity costs and how VMs are supported on different hardware across regions. Our first instrument uses electricity prices in each region in 2019 to capture the provider's regional variation in costs.

The second instrument uses the power ratings of different types of hardware and the third instrument uses costs of procuring an additional cluster of different types of hardware. Because different VMs can run on different hardware and the allocation of VMs onto different hardware clusters differ across regions, we transform the hardware costs and power ratings at the hardware type level to be at the VM level, weighting them by the demand from 2017. As a result, both the power rating and the hardware cost instruments vary at the region-product level.

The first stage results of the control function, introduced in Section 4, are shown in Table C1. We regress prices of all cloud products (X1-X4 and Other) at the product-market level on the three cost shifters, as well as the product characteristics X_j. The estimated coefficients of the cost shifters are statistically significant and have the expected signs: VM prices increase with electricity prices and VM hardware costs. VM prices also increase with VM hardware power rating, i.e., VMs with higher energy consumption (higher rating) are more expensive.

Table C1: Control function first stage results

	VM price
Electricity price	0.092^{***}
	(0.015)
VM (hardware) power rating	0.766^{***}
	(0.078)
VM (hardware) cost	1.151^{***}
	(0.065)
X1	0.195^{***}
	(0.025)
X2/X3	0.136^{***}
	(0.027)
X4	-0.050^{***}
	(0.018)
Promo	-0.112^{***}
	(0.007)
Constant	-0.933^{***}
	(0.106)
Observations	3,790
R^2	0.600

Notes: Table presents results from the first stage in the control function approach. We regress VM prices at the product-market level on the three cost shifters as well as product characteristics. *p<0.1; **p<0.05; ***p<0.01

D Computation Details

To calculate the GMM objective function, we numerically integrate over the distributions of task size, the random coefficients, as well as the number of tasks. For task size, we use the Gauss-Laguerre quadrature with 10 nodes. For the random coefficients, we use the Halton sequence with 20 nodes. For the number of tasks, we compute moments for any k number of tasks, for $k = 0, 1, 2, ..., 120$. For large customers at the end of our sample (i.e., largest number of tasks), 120 tasks cover 99.8% of the distribution of the number of tasks.

To speed up the computations, we parallelize $k = 0, 1, 2, ..., 120$ across 28 cores. In total, this integration procedure yields $10 \times 20 \times 121/28 = 864$ points to compute per core. On a 2.6 GHz Intel Gold 6132 CPU with 28 cores, one evaluation of the objective function with gradient takes about 162 seconds, using 508G of memory. We are thus limited by the curse of dimensionality to further increase the number of nodes for the numerical integrations. Finally, we minimize the GMM objective function using the BFGS quasi-Newton algorithm supplied with the analytical gradient.

E Monte Carlo Simulations

To illustrate identification of our model and our estimation procedure in the finite sample, we conduct Monte Carlo simulations for the simplified model with product characteristics and the mean number of tasks and task size.

We now specify the data generating process. To generate tasks, we let both the Poisson distribution for the number of tasks and the exponential distribution for the task size to be parameterized by one parameter each, i.e., their means λ and γ. For product choices, we consider markets each with four products and an outside option, where the utility of customer i choosing product j in market m for task k is given by

$$u_{ijmk} = \alpha p_{jm} q_{im} + X_j \beta + \epsilon_{ijmk},$$

where the product characteristics X_j include product dummies, and the idiosyncratic shocks ϵ_{ijmk}'s are distributed type-I extreme value. Mean utility for the outside option is normalized to zero.

To generate estimating samples, we first fix parameters at some true values. We generate two samples, both with 100,000 customers per market, but one with 100 markets and the other with 500 markets. We then draw prices independently from a normal distribution at the product-market level. With prices (and the price coefficient) and product dummies, we first draw each customer's number of tasks and the size of each task, and then simulate their product choices and usage.

Finally, we take the resulting estimating samples through our GMM estimation procedure and present the results in Table E2. We accurately recover true values of the parameters when initial values are different from the truth. Moreover, the accuracy slightly improves with more markets (and thus price variation).

Table E2: Monte Carlo simulation results

		True Value	**Initial Value**	**Spec1**	**Spec2**
	No. Customers			100k	100k
	No. Markets			100	500
Product Choice	Product1FE	1.000	1.000	1.002	1.001
	Product2FE	1.500	1.000	1.503	1.500
	Product3FE	2.000	1.000	2.001	2.000
	Product4FE	2.500	1.000	2.501	2.501
	Price	-10.000	-15.000	-9.995	-9.998
Usage	Mean	0.300	0.100	0.300	0.300
Number of Tasks	Mean	2.000	1.000	1.999	2.000
Runtime (s)				3285	18567

Notes: Table presents results from Monte Carlo simulations. We generate two samples, one with 100k customers spread across 100 markets and the other with 100k customers spread across 500 markets. Both samples are generated based on true values of the parameters. Prices are drawn from an i.i.d. normal distribution at the product-market level. Estimation is based on our GMM procedure and started at the initial values of the parameters.

F Own-Price Elasticity

In this section, we derive the formula for the own-price elasticity, $\frac{d\mathbb{E}(y_{ijm})}{dp_{jm}}\frac{p_{jm}}{\mathbb{E}(y_{ijm})}$. Following Proposition 1, the expected usage of customer i on VM j in market m is

$$\mathbb{E}(y_{ijm}) = \lambda_{im}\int_{q_{im}} q_{im}P_{ijm}(q_{im})\cdot\frac{1}{\gamma_i}\exp(-\frac{1}{\gamma_i}q_{im})dq_{im}.$$

Differentiating the expected usage with respect to its price:

$$\frac{d\mathbb{E}(y_{ijm})}{dp_{jm}} = \lambda_{im}\int_{q_{im}} \alpha_i q_{im}^2\cdot P_{ijm}(q_{im})(1-P_{ijm}(q_{im}))\frac{1}{\gamma_i}\exp(-\frac{1}{\gamma_i}q_{im})dq_{im}.$$

Finally, the own-price elasticity is given by

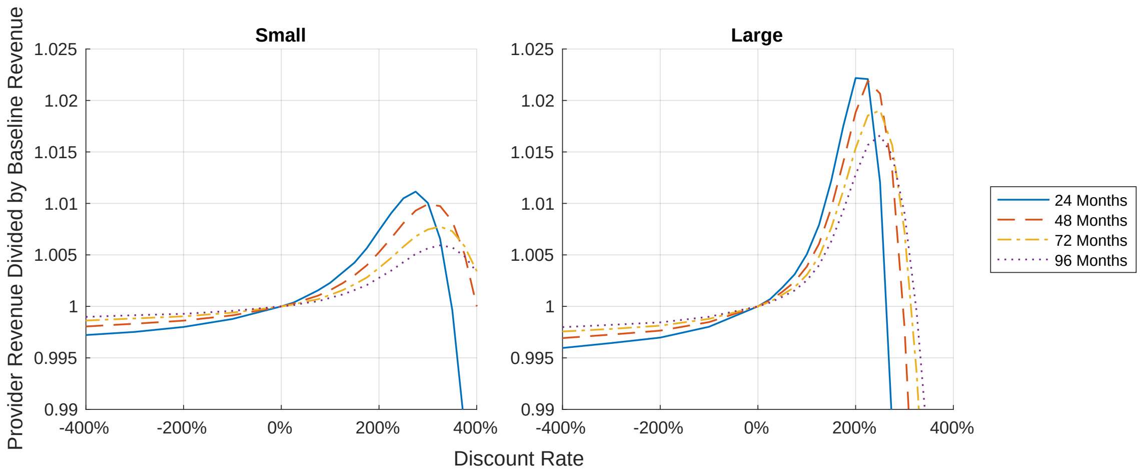
$$\frac{d\mathbb{E}(y_{ijm})}{dp_{jm}}\frac{p_{jm}}{\mathbb{E}(y_{ijm})} = \frac{\alpha_i p_{jm}\int_{q_{im}} q_{im}^2\cdot P_{ijm}(q_{im})(1-P_{ijm}(q_{im}))\exp(-\frac{1}{\gamma_i}q_{im})dq_{im}}{\int_{q_{im}} q_{im}P_{ijm}(q_{im})\exp(-\frac{1}{\gamma_i}q_{im})dq_{im}}.$$

G Optimal Introductory Discounts

In this section, we introduce how we find the provider's optimal introductory discounts and compare the short-run vs. long-run optimal discount rates. For each type of discount (new product preview or personalized product trial), we conduct a grid search over discount rates from -1000% to 1000%, where a discount rate of x on price p_{jm} implies a discounted price of $(1-x)p_{jm}$. We first conduct a coarse search over this range with 100% intervals to narrow down to a 200% range for the optimal discount. Then, within this 200% range, we refine the grid search to 25% intervals. For each discount, we fix characteristics of each market (customers, choice sets, and prices) and simulate demand forward following the same procedure as in Section 5.1 to compute consumer surplus, provider revenue, and cloud usage. We average across 1000 simulations to find the provider's optimal discount rates.

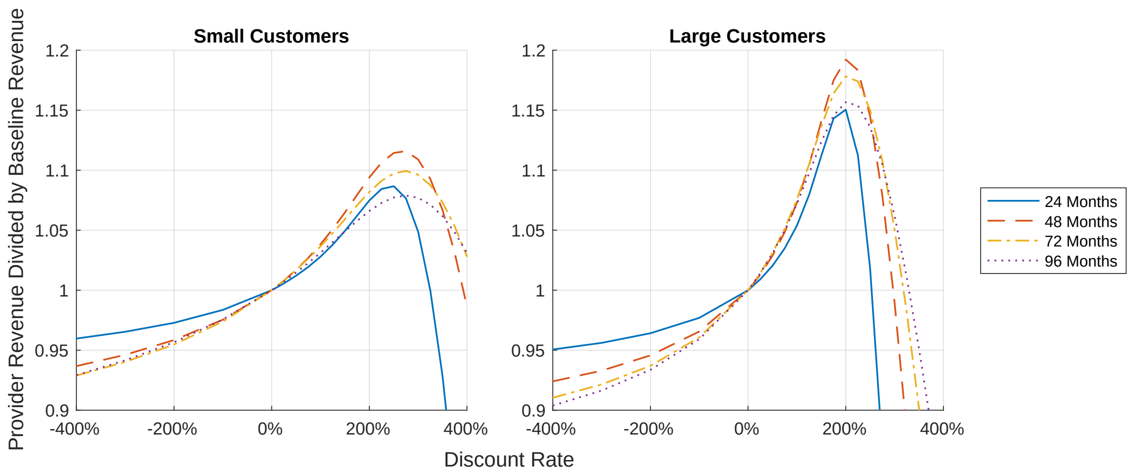
To highlight the difference between the short-run vs. long-run trade-off for the provider, in Figures G1 and G2, we present one line each representing the relationship between the introductory discount rates and the provider's revenue, for when the provider's total revenue objective is defined for 24, 48, 72, and 96 months. Peak of each curve represents the corresponding optimal discount. As the provider's horizon becomes longer, the optimal discount becomes larger as the benefit of overcoming customer inertia increases. The optimal discount stabilizes between 72 and 96 months as later months do not contribute as much to the provider's objective.

Figure G1: New product preview: Different horizons



Notes: Figure shows the ratio of provider revenue with different introductory discounts relative to the baseline, separately for small and large customers. Different lines represent different horizons in the provider revenue calculations. Results are averages across 1000 simulations. Discount rates, shown on the x-axis, are in absolute levels: A discount rate of x means that, in the first month after X4's launch, customers would pay (1-x) times their original spending on X4.

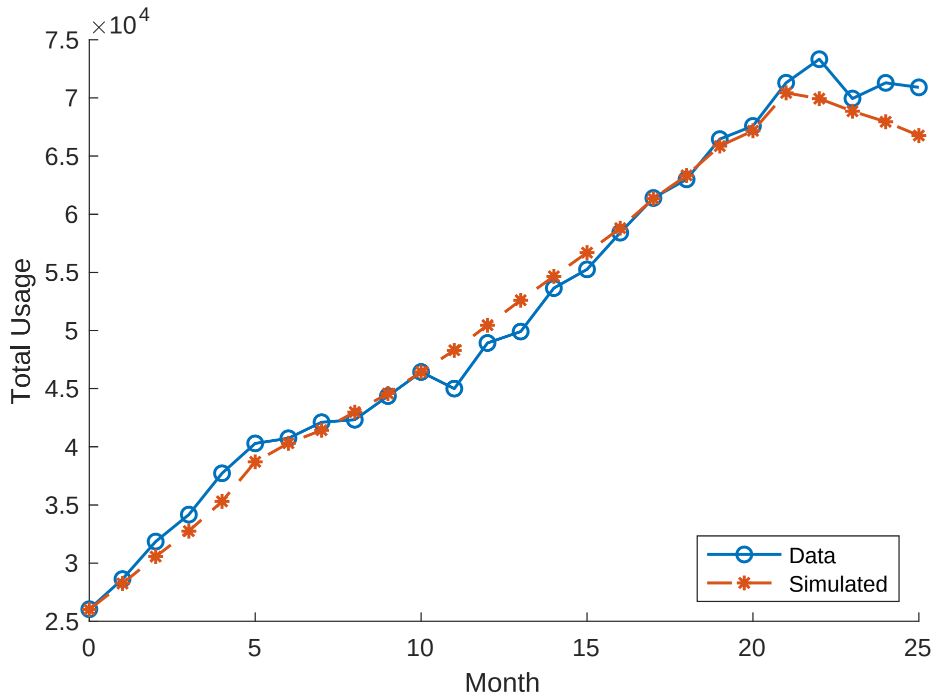
Figure G2: Personalized product trial: Different horizons



Notes: Figure shows the ratio of provider revenue with different introductory discounts relative to the baseline, separately for small and large customers. Different lines represent different horizons in the provider revenue calculations. Results are averages across 1000 simulations. Discount rates, shown on the x-axis, are in absolute levels: A discount rate of x means that, in the first month that a customer uses X4, the customer would pay (1-x) times her original spending on X4.

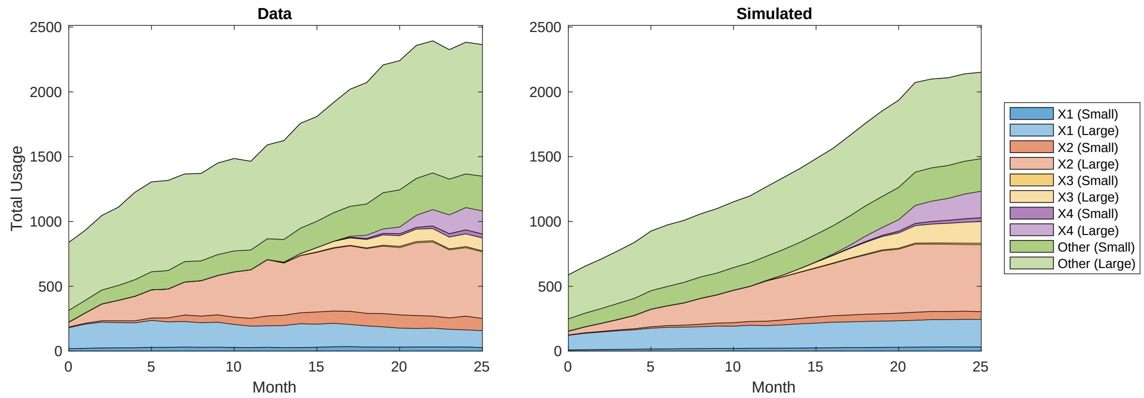
H Additional Figures and Tables

Figure H3: Time series fit of total computing demand



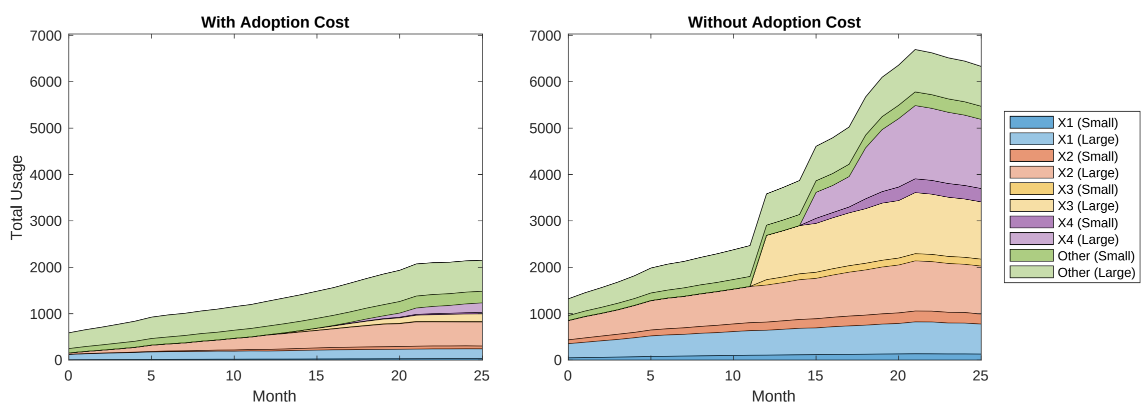
Notes: Figure shows total computing demand (including outside option) over time (months from beginning of sample) from the data (blue solid line with circles) and from simulations at our estimates (red dashed line with stars).

Figure H4: Time series fit of cloud demand by product and consumer size



Notes: Figure shows cloud usage over time (months since beginning of sample) from the data (left) and from simulations at our estimates (right), by product and customer size.

Figure H5: Impact of inertia on usage by product and customer size



Notes: Figure shows cloud usage over time (months since beginning of sample) simulated with adoption cost (left) and without adoption cost (right), by product and customer size.

Table H3: Share of direct adoption cost by different sample lengths

# of months after X4 launch	Small	Large
12	2.40%	1.65%
24	2.40%	1.65%
36	2.40%	1.65%
48	2.39%	1.64%
60	2.38%	1.64%
72	2.38%	1.64%
84	2.37%	1.63%
96	2.37%	1.63%

Notes: Table shows the share of direct adoption cost as a percentage of total consumer surplus lost from inertia when the sample is simulated forward for different lengths, separately for small and large customers. The sample is simulated with the same market characteristics (customers, choice sets, and prices) from the time of the X4 launch. Results are averages across 200 simulations.