A Taxonomy of Communications Demand

Steven G. Lanning, Shawn R. O'Donnell, W. Russell Neuman [1]

ABSTRACT

Demand forecasts are an essential tool for planning capacity and formulating policy. Traffic estimates are becoming increasingly unreliable, however, as accelerating rates of use and new communications applications invalidate conventional forecasting assumptions.

This paper presents an alternative approach to the study of telecommunications demand: build aggregate estimates for demand based on the elasticity of demand for bandwidth.

We argue that price elasticity models are necessary to grasp the interaction between Moore-type technological progress and non-linear demand functions.

Traditional marketing models are premised on existing or, at best, foreseeable services. But in a period of sustained price declines, applications-based forecasts will be unreliable. Dramatically lower prices can cause fundamental changes in the mix of applications and, hence, the nature of demand.

We consider the option of posing demand theoretically in terms of service attributes. Our conclusion is that the positive feedback loop of technology-driven price decreases and high-elasticity demand will quickly make it possible to base forecasts on bandwidth elasticity alone.

Industry analysts and policymakers need models of consumer demand applicable under dynamic conditions. We conclude by drawing implications of our demand model for network planning, universal service policies, and the commoditization of communications carriage.

---

[1] Steve Lanning: (slanning@lucent.com) Lucent Technologies, Bell Labs, 600 Mountain Ave 2D461, Murray Hill, NJ 07974.
Shawn O'Donnell: (sro@rpcp.mit.edu) Research Program on Communications Policy, E40-218, MIT, Cambridge MA 02139
W. Russell Neuman (rneuman@asc.upenn.edu) Annenberg School for Communications, University of Pennsylvania, 3620 Walnut Street, Philadelphia PA 19104-6220.

1. Demand forecasting under conditions of exponential growth

This paper presents a somewhat counter-intuitive approach to forecasting demand for telecommunications capacity: rather than attempting to make the right guesses about the types of applications that users will want in the future, we abandon our crystal balls and look to aggregate demand elasticity as a guide for forecasting.

This paper represents one component of work in progress by the authors on network business planning, characterizing demand for telecommunications services, and new media economics and policy. We are at an early stage of this work, just formalizing the models and testing against market data.

The paper begins with a discussion of the bandwidth forecasting problem. After presenting the elasticity-based modeling concept itself, we suggest a number of implications for network and policy planning.

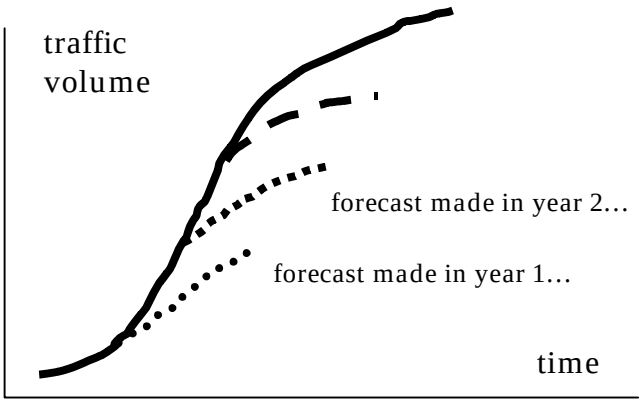
1.1. Problems with marketing science forecasts

Conventional forecasting methods take two forms: statistical and structural. In marketing science-style structural modeling, we estimate the demand response that will result from price decreases. This method works so long as new services are not introduced. When new services are introduced, projections that are based on the demand response of existing service to price changes will underestimate total usage. Forecasting voice traffic was once a very accurate activity because the application space was so stable. The demand response as measured by elasticity was also very close to one.

Forecasting for newer technologies involves more uncertainty. A typical history of forecasts looks like the graph in Figure 1. The branch curves in Figure 1 represent forecasts for total capacity in successive years. The solid curve represents the actual history of traffic growth. The problem with the marketing

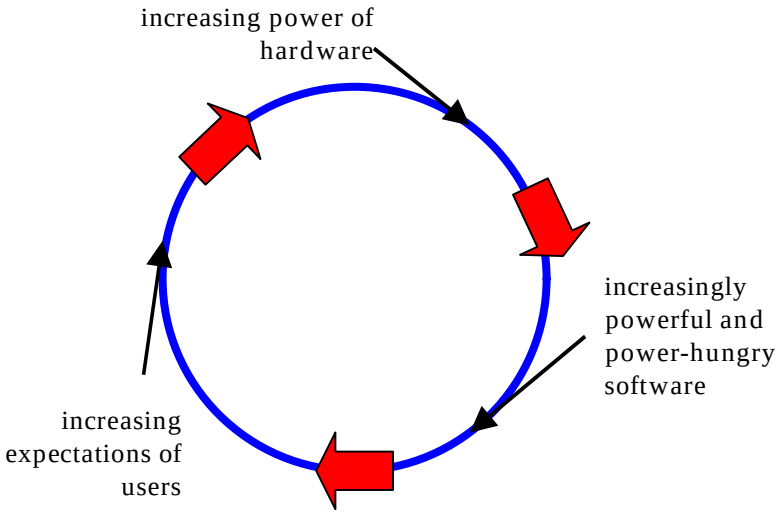
science approach is that it attacks the problem at the wrong level. A better approach is to estimate total capacity usage as a response to price changes. Although we cannot describe in detail the characteristics of the applications space, we can estimate total capacity far more accurately.

Figure 1. Chronic underestimation



In the 1940s, IBM's Thomas Watson, Sr. saw governments and research centers using computers. He predicted "a world market for maybe five computers." DEC's Ken Olson thought there was "no reason anyone would want a computer in their home." And Bill Gates thought that 640K of RAM ought to be enough for anybody. Estimates based on current applications can be very wrong.

Figure 2. Pumping the technology/market feedback loop



Hardware manufacturers, software publishers and users are driving a feedback loop that helps explain the consistent growth in the power of computing technologies. End users demand smarter, faster applications; software publishers write new applications with the expectation that hardware manufacturers will provide the additional processing power necessary to run them, and hardware manufacturers offer systems with greater and greater power, increasing the expectations of users and completing the loop. (Figure 2.)

Estimation of aggregate capacity begs the question of quality of service (QoS). When we speak of voice, we usually mean a circuit switched service with call setup delays measured in fractions of a second, for which blocking rates are quite low and for which reliability is very high. When we speak of data services, we tend to think in terms of best effort service, TCP slow-start algorithms that slow our transmission rates, DiffServ access routers and other protocols that allow bandwidth hungry applications to peacefully co-exist.[2]

---

[2] The TCP slow-start algorithm imposes a limit on the speed TCP connections can attain immediately after being established. The slow-start flow control algorithm guarantees that impatient hosts won't flood underpowered or overloaded hosts with too much traffic. In the absence of information about the status of the network, slow start is the accepted method of

We believe we can boil total demand down to aggregate bandwidth with a blocking quality of service constraint. At the applications layer, other characteristics such as security, interoperability, systems management / ease of use latency and delay continue to exist. However, these can be provisioned with additional bandwidth. Protocols that are meant to gracefully throttle down their transmission rates in the face of congestion allow peak to be spread, but they do not solve blocking problems, or else they tend to assume the network has not been adequately provisioned.[3]

In the end, a service provider selects bandwidth capacity and pricing packages (including protocols) which generate revenues. These pricing packages are called service level agreements. A service provider will lower prices just to the point of filling its network if demand is elastic. Of course, the nature of demand varies from sub-market to sub-market. There is some evidence, for example, that demand for toll services is inelastic, but strong evidence that demand for data services is very elastic.

We believe we can boil total demand down to aggregate bandwidth with a blocking quality of service constraint

There is a widely publicized technology explosion in optics. The rate of innovation is even faster in optics than it is in microprocessors. Microprocessors, according to Moore's Law, double computing capacity every 18 months. In optics, bandwidth capacity doubles every 12 months. In communications, as in microprocessors, the minimum efficient scale of operations increases as well. For microprocessors, demand has kept up with this expansion of minimum efficient scale and so the trend continues. Our examination of component equipment

---

insuring that a host's connection to the network will be set at a speed appropriate to conditions. A DiffServ access router is a differentiated services server, or a network element designed to offer varying QoS to different classes of users or services.

[3] For example, see discussions in Cahn[1998] and Verma[1999].

markets for core networks suggest the same is true for optics and the services they support.In communications, as with microprocessors, it is not possible to estimate demand by projection of existing applications. Instead, we do better to estimate the aggregate demand response under the assumption that applications will be written that will create the demand for cheaper computation

This computation/communication analogy is actually very closely linked. Amdahl observed that inside a computer we require about 1 megabit per second for input/output for each MIPS of processing power. As computing becomes increasingly distributed, more I/O per MIPS will be required in transport networks for input and output. The readily apparent increase in the number of desktop computers pales in comparison to the number of MIPS those computers represent. When measured as total MIPS, it is clear that billions of installed MIPS are, by design, underserved by communications.[4]

Future generations of these devices may be served by communications channels as the costs continue to fall. The introduction of Electro-Absorption Optical Modulator allows for the connection of fiber directly to a DECT (1.52Mb/sec) RF antenna. Electro-Absorption Optical Modulators require no bias voltage, meaning that they are passive optical devices. They are also very cheap. These devices can be used for home networks, micro-cameras, micro-microphones, wearable devices, palm-like devices and high-resolution displays, to name but a few possibilities. The beauty of such devices is that they eliminate the problem solved by fiber to the curb, sharing expensive electronics to convert optical signals to electronic signals.

---

[4] Amdahl's law is an interesting starting point for modeling in era of distributed computing. What is the appropriate amount of I/O for a computer, given the relative share of distributed vs. local computing?

2. Forecasting aggregate demand response for bandwidth

The demand for bandwidth capacity does not grow at a constant rate. We attribute variation to supply factors such as the price and performance of new generations of optical equipment. We take a mixed approach to estimation of bandwidth capacity. We advocate estimating aggregate price elasticity and incorporate this into a network planning solution. The primary input to this techno-economic bandwidth forecast is demand response and not a statistical demand forecast.

For the purposes of this paper, establishing the elasticity of demand for capacity is sufficient. Our conclusions are drawn from demand elasticity estimates and not from particular demand forecasts[5]. We want to get away from forecasting for individual, and in particular, existing services. We propose a method based on analogous markets, using indirect measures with partial corroboration from direct measures reported by other carriers.

Many estimates for broadband access indicate inelastic demand response. These estimates are often flawed by not taking into account the effect of two-part tariffs[6]. It is not sufficient to fit a curve for monthly access fees against the demand response for bandwidth or the number of subscribers to a high bandwidth system without including terms for transport cost. Flat rate pricing often makes such a regression impossible. However, flat rate pricing really does not exist in the market. Almost all service providers offering a flat rate service insert a high usage clause in their service agreements. Whenever users exceed a threshold, they pay more. This 'abuse' penalty must be included in the regression along with actual usage to obtain meaningful estimates of elasticity.

---

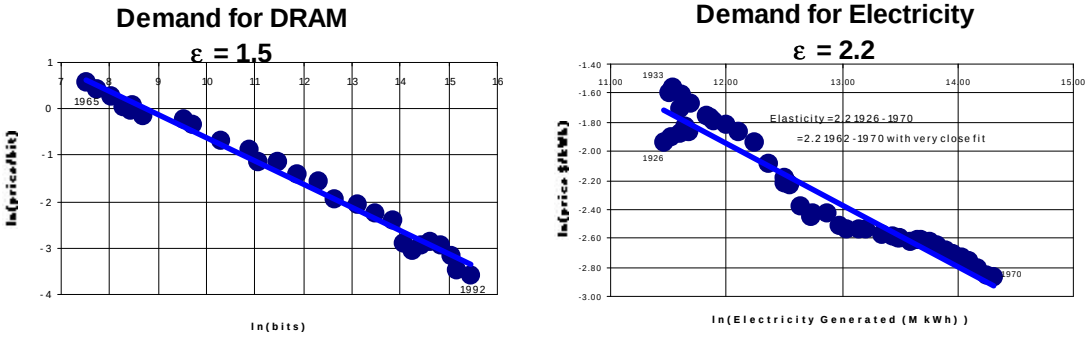
[5] cf. traditional telecommunications demand analysis (de Fontenay, et al. [1992], Taylor [1994].)

[6] Two-part tariffs: separate rates for local access and transport.

Without the transport cost term, the regression is based only on partial information. Since we do not think the transport costs in the core of the network are changing at the same rate as transport costs at the edge of the network (sometimes called access costs), we find it difficult to accept the findings from some carriers that demand response is inelastic. It is particularly difficult when we see observe a very elastic demand response for the capacity new equipment carries.

2.1. Demand Elasticities: Examples

The demand for capacity in industries related to telecommunications can be very elastic. In computing, the demand for capacity is quite high. In the DRAM market, we find that elasticity is 1.5. In electricity production, we find an elasticity of 2.2 when we fit a constant elasticity demand curve against price alone.



Charts 1 & 2. Demand Elasticity for DRAM & Electricity

The fit for DRAM is very good. The fit for electricity matches the fit for DRAM in the years that the time series overlap. In the case of electricity, the fit is good despite a depression and a world war. It is interesting that the long run elasticity over such a turbulent period is the same as the relatively stable 60s and 70s. This

suggests that short term inflation in demand or conservation in usage does not have a long term effect on demand growth.[7]

Both examples demonstrate that new applications arise to use new capacity. We once turned lights off before leaving a room, we now leave them on for security. We used to turn the TV off before leaving a room, now there is a TV on in almost every room. We used to use fans and swamp coolers, now we use air conditioners. We would not do this if the price of electricity had not fallen by an order of magnitude. We once shared our computers, we now have one in the office and another at home. The penetration of desktop and laptop computers continues to increase as does the percentage of homes with online connections. This would not have happened if prices for computer components such as DRAM, microprocessors and disk drives did not fall by orders of magnitude. Rather than devalue the industry, the innovations that allowed the market to drop prices precipitously increased the value of the market because the demand response is elastic.

New applications arise to fill new capacity. We once turned lights off before leaving a room, we now leave them on for security.

The incentive to innovate at rates as high as DRAM must be linked to its elastic demand response. It seems hard to believe that Intel would want to build a new fabrication plant every other year at prices that are growing exponentially (a fabrication plant cost $1B in 1998 and about half that in 1996) unless the demand was elastic. These new fabs (or chip foundries) represent substantial innovation in production processes to allow for continued miniaturization of the transistors in these devices.

---

[7] The careful reader will note that our electricity data ends just before the energy shortages of the 1970s. Alas, the data series we were able to find changed formats at that point. The post-embargo data plots a relatively straight line, too, but on a different line based on different parameters.

The fact that there is a similar rate of innovation in optics is evidence of a similarly elastic demand response for bandwidth in the core of networks. Such an elastic response could not exist without end-to-end expansion in communications. Although there may be bottlenecks for some classes of users, they cannot be widespread or else there could be no elastic demand response for bandwidth in the core of the network.

Table 1. Elasticities for various network elements[8]

	Original Source	Bandwidth Elasticity					
		Equipment Estimates		Calibrated Service Estimates			
Circuit	NBI	Digital Circuit Switch	-1.28	-1.05	-1.10	-1.20	-1.34
ATM	In-Stat, IDC	WAN ATM Core Switch	-2.84	-1.33	-1.66	-2.31	-3.26
	Dell'Oro	LAN ATM Backbone Switch	-2.76	-1.31	-1.63	-2.25	-3.16
	In-Stat, IDC	WAN ATM Edge Switch	-2.11	-1.20	-1.40	-1.80	-2.37
Router	Dell'Oro, IDC, In-Stat	High End Router	-1.18	-1.03	-1.06	-1.13	-1.22
Switch Route	Dell'Oro	LAN L2 Fast Ethernet Switch	-3.02	-1.36	-1.72	-2.44	-3.49

Data Source: Rich Janow
Traditional estimates of long distance service vary, but tend to center around -1.1
Service estimates are calibrated to relation between digital circuit switch elasticity estimates and estimates for circuit switched long distance services as deviations from -1.0
Best estimate of transport bandwidth elasticity is in range 1.3 - 1.7 because FCC estimates are in range of 1.05 - 1.1

In the digital circuit switch, the equipment elasticity is 1.28. The service elasticity for capacity served by digital circuit switches, the elasticity is in the range of 1.05 to 1.1. If we assume the same cost ratio between digital circuit switches and the total cost of circuit service and WAN ATM core switches and the total cost of data services, then we may infer the elasticity of data services. We use an elasticity of 1 as the baseline because a profit maximizing service provider would not willingly operate in an inelastic portion of the demand curve. So we take the difference of 1.28 from 1 and the difference of 1.05 from one and apply this ratio to the equipment elasticity of 2.84 to estimate a data service elasticity of 1.33. If we think the benchmark voice service is more elastic, 1.1, then this method

[8] Source: FCC and Rich Janow of Lucent Technologies, Bell Laboratories.

would produce an estimate of 1.66 for data service. The relation between equipment costs and total costs for both circuit service and ATM/IP service would have to be studied more closely to derive an improved calibration to the circuit benchmark.

It is interesting that the elasticity of demand for the equipment is greater than the elasticity for the service. The reason for the disparity is that there are many fixed costs associated with land and labor operating network facilities. As these vary slowly relative to equipment, most of the price changes we see come from equipment and the capacity they allow a system to carry. If half the cost is in the equipment and equipment capacity is doubled for the same price, then service prices fall by a quarter. If the demand for the equipment is derived from the demand for the service, then the response to a cost change in equipment would seem to be muffled by the fixed costs. This is not the case when demand is sufficiently elastic. Instead, the investment in new equipment represents an opportunity to spread the fixed costs for right of way, some cable costs, land costs and labor costs over greater bandwidth capacity. The result is that the demand for equipment is more elastic than the demand for the service. The inference that the demand for data services is elastic remains, it is just less elastic than the demand for the underlying equipment.

2.2. Aggregating demand across applications and services

Is there a problem with aggregating demand and ignoring service features that differentiate one bit from another? There are sizeable differences in the price and cost of delivering bits through various channels today, but we argue that convergence of the telecommunications industry will allow consumers to see disparities in costs, forcing providers to eliminate or justify price differentials.

Currently, carriers differentiate their services from competitors' in a multi-dimensional attribute space. But technological and market logics will force a consolidation of service attribute dimensions. One dimension—bandwidth—will grow in significance at the expense of the others. While a larger number of service attributes will remain of interest to end users (for example, blocking probability, security, interoperability, systems management,) service providers will convert requirements for these attributes into additional allocations of bandwidth. So while a handful of quality of service issues might find their way into end users' SLAs with service providers, the SLAs that service providers arrange with bandwidth providers will concentrate on bandwidth alone.

In rough outline, the consolidation of service attributes might proceed as follows. (1) First will be an initial period in which telecommunications markets will remain separated along traditional boundaries, each with its own elasticity: voice, public data, private data, etc., (2) Next there will be a transitional period of advancing convergence and competition on the basis of service attributes. (3) Finally, a disjunctive transformation in the telecommunications markets caused by the dominance of optical technologies will greatly reduce the marginal cost of bandwidth in local access as well as the backbone. After this transformation, other service attribute dimensions will be projected onto bandwidth as it becomes easier to recast requirements in those areas into bandwidth. Firms will then compete on the basis of their ability to deploy and efficiently provision bandwidth.

The driving force behind the "generification" of bandwidth demand is optical communications technology. Again, it is the large elasticity of demand for bandwidth that makes possible continuous, profitable investment in capacity. Since optical technologies promise several more generations of continued growth in capability, the feedback loop driving down prices for bandwidth is the appropriate focus for forecasting telecommunications demand. Since such

attributes as blocking, security, interoperability, and so forth can be provided through the provisioning of modest increments in bandwidth, a market strategy of differentiating on those attributes is a recipe for modest growth in an exponentially expanding market. Service providers who choose to focus on other attributes rather than chasing the accelerating growth in bandwidth are likely to be left in the rear-view mirrors of carriers who aggressively build capacity.

Among service attributes, latency and blocking are special cases. Latency cannot be created out of additional bandwidth per se. That is, throwing bandwidth at latency will reduce some congestion-related delays, but not basic switching, routing or propagation delays. Nevertheless, latency will become less problematic as a side effect of the technologies that expand capacity. Only blocking—the probability that a customer cannot be provided with the desired level of service—will remain an independent concern to carriers, though it, too, will be of interest to the extent that blocking probability indicates whether sufficient bandwidth has been allocated to a service.

The driving force behind the generification of bandwidth demand is optical communications technology.

The most important service attribute that cannot easily be exchanged with bandwidth—latency—will likely disappear as a differentiable service attribute because of the integration of optical components into network elements and the unrelenting speed of optical networks.

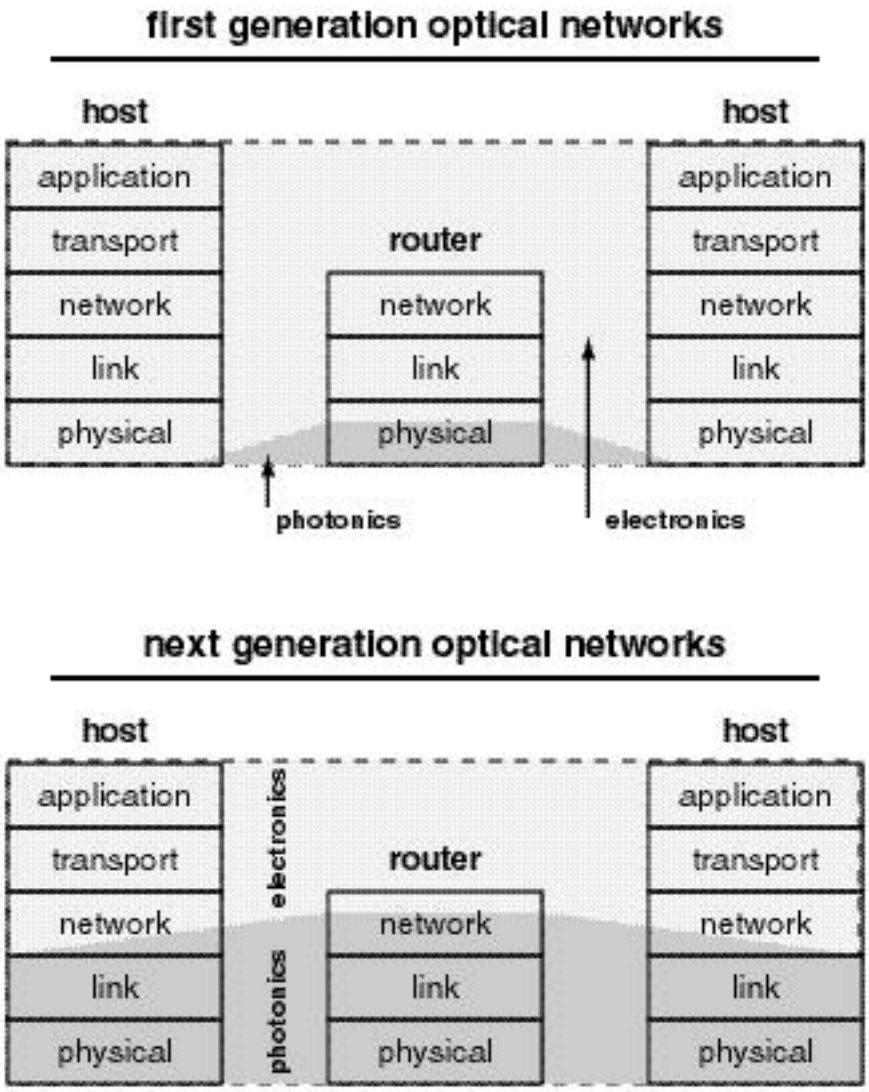
Photons don't wait

Contrast the role played by optical communications components in first- and second-generation optical networks.[9] In first generation networks, optics were

[9]See Ramaswami & Sivarajan [1998] for a characterization of first & second-generation optical networks.

used solely as a substitute for copper as a transmission media. As suggested by the upper set of network stacks in Figure 3, optical components were limited to the physical layer of networks. The real hard work of networks was performed in electronics, at the link and networking layers. Optical communications offered no speed advantage in early applications of fiber technology, since there is no speed advantage of fiber over copper in transmission.

Figure 3. Optical components in networking stacks



As cheaper photonics creep up the networking stacks of hosts and network elements, though, the cost for providing broadband services will drop as capacity explodes. The larger number of functions allocated to optical components within the network will substantially reduce bandwidth as a factor in the pricing of telecommunications services.

Note that the adoption of a technology into switches after its adoption in transport has precedent: in the 1970s, the development of the 4 ESS digital switch followed the widespread deployment of digital transmission systems. Just as optical links were first introduced on point-to-point links, earlier digital links substituted for analog on selected routes where efficiency was crucial. Only later was the ability (or necessity) to switch in the digital realm added to the network.[10]

The advance of photonics up networking stacks has implications not only for bandwidth, but also for latency. Electronic bits dart from router to router, pausing at each node along its path. Routers parse packet headers, buffer the packets, and queue them for transmission on the appropriate interface. In contrast, optically switched networks offer the equivalent of non-stop bit routes. So long as bits are cruising around the network in photonic form, the latency of communications approaches the propagation delay. Wavelength routing, wavewrapper technology[11], and optical cross-connect switches all replace electronic switching in the network and its resulting latencies.

If the ability of carriers to differentiate service on the basis of latency will decline, what about other service attributes? Until optical routing and switching

---

[10] Stern & Bala [1999, 665]

[11] WaveWrapper is Lucent's network management tool for optical networks. Wavewrapper will provide such functions as optical-layer performance monitoring, error correction and ring protection on a per-wavelength basis.

technologies penetrate the network completely, though, there will be period in which carriers will charge premiums for bandwidth and other bitway characteristics. As communications industry stakeholders struggle through the transition from regulated monopoly provision to true facilities-based competition there are numerous incentives to resist convergence, commodification and competition. We watch the marketing wars with interest as friends-and-family discounts compete with 5-cent Sundays and 2-cent Tuesdays. Will pin-drop sound quality ultimately compete with stereo telephony? We believe the economic incentives to try to convince the communications consumer of technical advantages may delay commoditization and the collapse of service qualities into bandwidth, but not prevent it. Bandwidth is fungible with other service attributes, and since the price of bandwidth will drop precipitously, the ability of carriers to charge for other service attributes will decline.

2.3. Where this is all leading…

We foresee a transformation in telecommunications services markets: starting from today's legacy system of inscrutable price differentials and non-market impediments, markets will move rapidly to services differentiated by service attributes. As price differentials based on distinct levels of those attributes become untenable, markets will move to distinctions based principally on bandwidth and price-per-bit. Price differentials will be based on fungible bandwidth for security, blocking, latency, delay and interoperability. Reliability may remain a system difference. For example, a SONET ring restoration takes approximately 50 milliseconds. Restoration numbers for mesh networks run 250-500 milliseconds.[12] As mesh networks are cheaper than (SONET) rings in many

[12] Restoration refers to the ability to respond to equipment failure or line cuts. The time to restore service on SONET rings is targeted at 66 ms.

cases, it follows that there may be a need for overlay networks and different prices based upon reliability. The implication is that some physical channelization of traffic qualities is likely to emerge around latency and reliability.

This historical premiums paid for voice service can be explained by the relatively low demand elasticity for voice which gives service providers an incentive to operate with considerably higher margins, and the fact that low demand elasticity does little to reward investment in more efficient high capacity equipment. This lack of interest in efficient equipment leads to a lack of interest by equipment providers to push hard for new innovation in the circuit service space. By contrast, high elasticity data service does reward investment in new equipment. High elasticity means thinner margins and lower prices. In addition, interest in more efficient high capacity equipment leads to a higher rate of innovation by equipment suppliers. It follows that costs fall more quickly in the data space as well. For a time, data was more expensive than voice service. There are many observations that this has changed with the higher rate of innovation in data.

Legacy Services

A profit maximizing firm or reasonably well-coordinated industry will charge a markup over unit costs proportional to demand elasticity. In the case of voice, this markup would be 10 – 20 times unit costs. In the case of data services, demand is far more elastic. If we take an elasticity of 1.5 for data, then the markup would be 3 times unit cost. If we accept that some call coordination is still required within an IP network, assume that half the cost is in call coordination and device service, and assume comparable unit costs for transport, then we could expect voice costs to fall by 40-60% in a converged market. If we

allow for lower transport costs in a data network, then the potential savings are even greater.

Transitional period: multiple-attribute bits

In the medium term, the cost per bit can be expected to reflect the type of service offered. Relevant characteristics of the service include bandwidth, latency, blocking, security, interoperability, systems management / ease of use, and availability. Most of these service attributes are mentioned, if not specified, in Service Level Agreements made today by large buyers of telecommunications services.

Incumbents are well aware of the coming transition. The jury is still out on whether or not it is cheaper for an incumbent to begin the transition now or in the future. The problem for incumbents is that they must operate and support two systems during the transition period, thus increasing their operating costs. Increased operating expenses may swamp any efficiency afforded by new equipment. Greenfield entrants do not face this obstacle. They are entering at an ambitious rate with plans to install tens of thousands of miles of fiber in the US alone, running very high capacity networks. This period of seemingly rational delay by incumbents gives entrants a chance to establish themselves before incumbents start to make their move into the high capacity data space.

As the killer voice application migrates to IP / data networks, incumbents will be forced to install new capacity or to lease capacity from newer network providers. It will not be possible to install capacity and milk it for 10 – 20 years. Network service providers will have to install new equipment on a regular basis to be competitive and hold their customer base.

Longer term: bandwidth, maybe latency and reliability

The innovation race in optics is leading very naturally to passive optics and optical switching. When it becomes available at sufficiently low price points, it will be adopted very quickly. The result will be new engineering solutions that no longer need to be sharply constrained by hop counts to reduce transmission latency. It remains to be seen whether or not the market settles on a thinner ring architecture for all service or converges on overlay ring and mesh networks. For sufficiently elastic demand, overlay networks can be easily supported by the market. The result would be higher cost for ring network service and its improved reliability though faster service restoration times.

If optical cross connects and optical switching remain the stuff of science fiction, then latency will remain a quality of service consideration at the level of network service. This quality of service is intimately connected to the provisioning of network with sufficiently few counts between each origin and destination pair that significant latency is not introduced.

If a single thin ring architecture becomes the industry standard, then there will be no network level difference in reliability. There may be service level agreement induced reasons for differences in reliability, but these will be the product of market conditions.

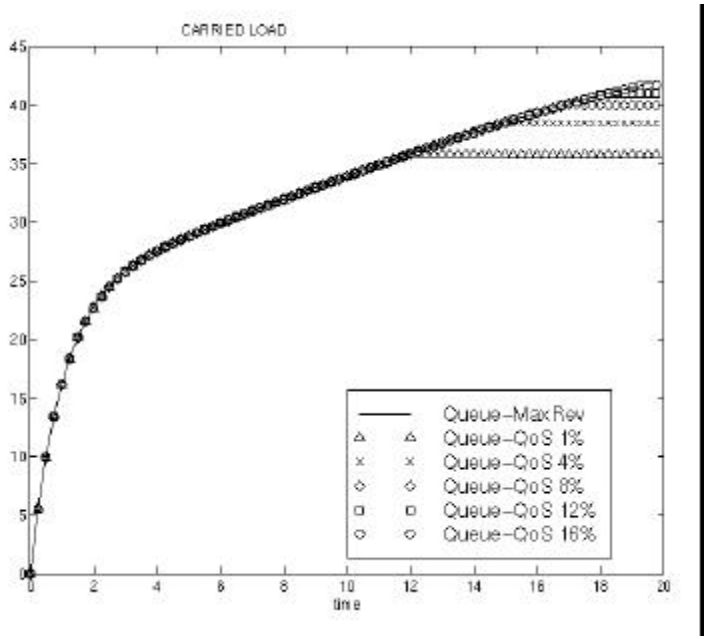
As for other QoS dimensions: Security is a bandwidth user roughly in proportion to the level of security desired. Interoperability is a software problem with some relation to the cost of MIPS. In the end there will be sufficient willingness to pay for strange protocols or these protocols will cease to be used. At the signaling device server levels, very elegant solutions are being developed within the Softswitch consortium. One may hope that elegant solutions for carriage may not be far behind. Operations are ripe for re-engineering and we can expect

these will also occur which will help further reduce the cost of all communication attributes.

Figure 4 illustrates an example of a scheme for offering graduated QoS over bandwidth-constrained networks in the face of heavy tailed traffic.[13] If we assume we have squeezed all the traffic possible into available channels, there may still be conditions of network congestion. These represent an opportunity to service providers to increase their revenue while rationalizing traffic to the benefit of all users. In this scheme, there is one price depending on the mean arrival rate up until traffic approaches a threshold defined by the blocking rate allowed for a given quality of service. To maximize revenue, the service provider must anticipate the likelihood of blocking prior to complete congestion. The lower the blocking rate (higher quality of service), the lower is the traffic load under which the service provider will increase the usage rate. This scheme plays off the proof by Eick, Massey and Whitt[1993] that peak arrivals come before peak traffic in the presence of heavy tailed holding times. Thus, it makes sense to increase rates prior to peak traffic because price operates on arrivals. This result contrasts with pricing in a voice network. It is well known that peak arrivals correspond to peak load in voice networks and so pricing by congestion to control arrivals makes sense.

[13] Lanning et al. [1999]

Figure 4. Pricing for QoS



3. Implications

If our elasticity-based model of communications demand reflects the dynamics of the market, how should network planners, regulators, and investors react?

3.1. Network planning

The usual practice of telecom network planners is to take traffic requirements as inputs and to produce a cost minimizing network. In the case of low demand elasticity, such an approach will reasonably approximate profit maximizing solutions. When technology innovation is fast, as it is in optics, and demand is elastic, as it is in data, then the practice needs to be modified. Demand response (elasticity) is the input and planned traffic and pricing is added to the usual solution of when, what and where to install network elements under

consideration. Solving this problem is far more difficult than solving the already difficult conventional network planning problem. In the case of a simple constant elasticity demand curve, this problem becomes a nonlinear mixed integer programming problem. Analytic results are not available to such a problem, though there have been significant improvements in computational methods for nonlinear optimization which make such problems tractable.

When demand elasticity is low, price reductions do not increase revenues. Under such conditions, the main objective of network planners is cost reduction. Cost-consciousness leads to using legacy equipment for a sufficiently long time to allow the service provider to spread costs over many years, lowering unit costs. Without a revenue incentive, the time to consider lower cost equipment is at the end of the useful life of legacy equipment. This explains the traditional 10-25 year lives in telecom equipment.

By contrast, when elasticity is high (1.3 or greater), there exists a sufficiently large revenue reward to price reductions that investment in equipment to justify lower prices makes financial sense almost every year. The notion of a shorter economic life becomes more salient than useful equipment life. The cost of maintaining equipment that carries a fraction of the traffic of newer equipment leads to a justification to retire equipment in 3-5 years given the current rate innovation in optics and reasonable assumptions regarding operations costs.

3.2. Universal service

The American universal service tradition took shape in a world in which voice was king. Voice service, with its low elasticity (roughly -1.05 – 1.1) and high markups over cost, left a comfortable margin for supporting universal service obligations. So long as everyone used the same type of telephony service, and so long as all carriers contributed to universal service funds, the model was feasible.

But consider what happens when data becomes king. Digital networks can carry all types of traffic, including voice. Once data networks have been engineered to provide the same low latency of switched voice networks (through the incorporation of optical switching technologies,) what incentive will consumers have to direct their traffic to higher cost, higher markup voice networks? The attraction of data is not only in price: voice-over-data consumers will enjoy a broader range of service options, greater innovation, and greater options for integrating services.

Consumers—at least those able to make the jump—will move their business to the digital infrastructure. But those voice network customers able to make the transition to general purpose data networks are likely to be those in areas that are net contributors to universal service plans. As the number of contributing customers declines, either the universal service fees must be increased or the fund will go bankrupt.[14]

This much of our argument is familiar to anyone following the IP telephony regulatory debate. As we see it, however, the problem is not simply a matter of chasing down old POTS customers at their new IP addresses to collect universal service fees. The hazards facing the universal service model are more fundamental. Casting the tax net more widely will fix only the budgetary problem of the universal service model, but not the deeper, more vexing problem of investment strategy.

At the core of the issue is impact of elasticity on infrastructure investment. The greater the price elasticity, and the greater the decline in unit cost per unit of investment, the greater the incentives for bandwidth providers to invest in infrastructure. Additional investments will lower unit costs—and prices—but

---

[14] On universal service, see Weinhaus, et al. [1994], Noam [1995], Mueller [1993, 1997].

because of the magnitude of elasticity for data services (roughly 1.5,) revenues will increase despite lower costs. In relatively inelastic markets, however, investments that lower costs and prices have virtually no impact on revenues. There is considerably less incentive to invest in such markets.

A perverse side effect of the exodus to data amplifies the effect of elasticity: as the more price-elastic market segments flee to data, the remaining customers will be more inelastic than average. The segregation of users into enhanced, broadband data vs. POTS-only would become entrenched, as it would become increasingly difficult to encourage operators of legacy POTS-only networks to abandon their old investments.

As an alternative to the existing universal service model, we envision a policy to encourage small, rural telcos to rebuild on a new technological foundation. We would impose a fixed horizon on current universal service subsidies—say 2-5 years. Within that time, rural telcos would be encouraged to shift their operations to Moore-type technologies that benefit from rapid cost declines and high demand elasticities. Interim subsidies could help during the transition to services based on new technologies. Telcos choosing not to make the move could choose to cash out of the business or allow competitors to apply for investment subsidies in their areas.

The universal service concept was originally conjured up by Theodore Vail in 1907 in his effort to protect AT&T from increasingly debilitating competition after the Bell patents ran out. His ploy worked. And to his successors credit, AT&T did live up to its side of the bargain in the Kingsbury Commitment of 1913 to build out the network even to the less profitable neighborhoods. Most business and political leaders now accept the legacy of the Universal service concept as an ideological force to be reckoned with despite its self-interested origins and its awkward applicability in the age of digital convergence. That is understandable political realism. But if our calculations about declining costs,

declining cost differentials between large and small bandwidth customers, and explosive demand are even partially correct, this legacy of redistributive taxation and service mandates is likely to lead to poor public policy. We need to develop a new model of universal access based on a technically and economically realistic assessment of the evolving network architecture.

3.3. Is carriage a commodity? / The first mover advantage?

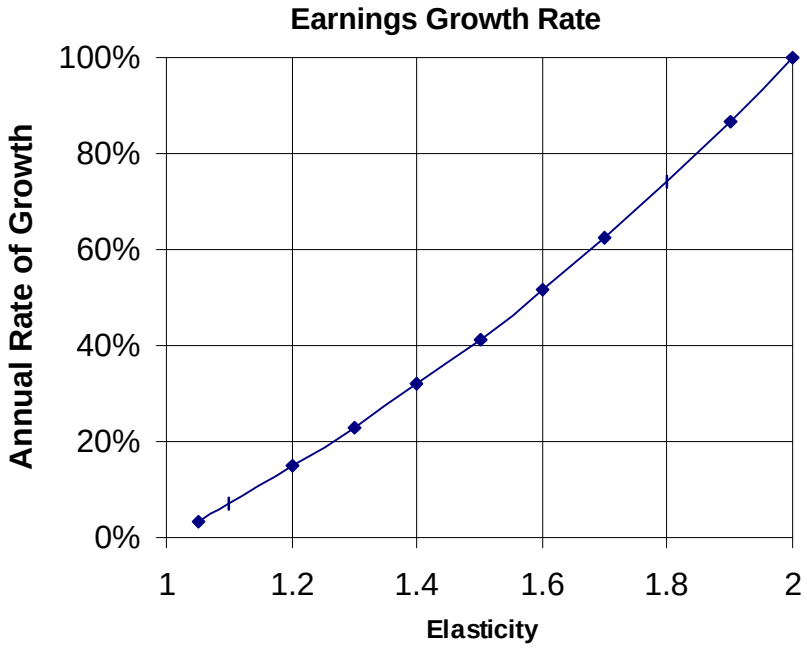
A driving motivation for telecom carriage in the 1980s and 90s has been to avoid being a commodity. Product differentiation to support high margins has been the objective. This objective is not necessarily warranted in an environment of high elasticity and cost reductions. Operating in a region of high margins is equivalent to operating in a region of low elasticity for a profit maximizing firm. In the case of a constant elasticity demand curve and declining costs, it is easy to prove that low margins and high elasticity are good.

The relation is easy to derive in the presence of a constant elasticity demand curve, $y = Ap^e$. Let the cost of producing capacity y be given by $C(y) = cy$, then profit is given by $\pi(p) = Ap^{e+1} - cAp^e$. Choose price, p, to maximize profit yields the first order condition, $p = ce/(e+1)$. The markup over the unit cost, c, is given by $e/(e+1)$. In economics 101 we teach that as the demand curve confronting a firm becomes increasingly elastic, price is driven to cost. We do not refute this story, but observe that it misses the relation of elasticity to profit growth potential through cost reduction.

We generalize this pure static analysis in a spreadsheet which accumulates productive capacity over a 3 year economic life and treats investment as an expense. A plot of the relation between earnings shows that earnings growth is higher, the higher is demand elasticity. (Chart 3.) Given a choice between low elasticity, high current earnings but low earnings growth, and high elasticity, low

current earnings but high earnings growth it can easily be the case that the net present value is higher for high elasticity.

Chart 3. Price elasticity & earning growth



rates

In the long run, it may be preferable to engage in a commodity business under conditions of high elasticity and a high rates of innovation, rather than a product-differentiated high margin business with low rates of innovation.

In an environment of high demand elasticity and investment in every year there is a weak technological first mover advantage. The incumbent has an incentive to purchase the same equipment as a new entrant. However, the incumbent has the advantage of past investments which are not yet retired. The total capacity in a frictionless game will be as high or higher than a new entrant would choose. Since lower prices generate greater revenues in what is necessarily a repeated capacity game, the incumbent will want to charge lower prices than the entrant. This technological effect coupled with the usual sluggishness in customer transitions may confer a significant advantage to the first firm to adopt this

strategy. Further analysis is required to investigate the tradeoff between the rate at which a firm can win customers to fill a large network and lock in its advantage. In the case of the US, it is greenfield entrants such as Qwest, Enron, Williams and Level3 that have adopted this strategy first. If the strategy had been adopted by an incumbent, first mover advantages would be far easier to prove. Under normal circumstances, the Cournot oligopoly model is relatively uninteresting, but in experience curve technology markets, firms wouldn't back-off production to lower prices, they would likely seek higher outputs instead. So there could be an advantage to being a capacity / price leader. Whether any such advantage would be sustainable is another question.

4. Conclusion

This paper presents a telecommunications demand analysis technique that abstracts away specific applications and services in favor of an aggregate demand model based solely on the elasticity of demand for telecommunications services.

Additional work on the relationship between equipment price elasticities and service price elasticities will improve our inferences about the price elasticity for bandwidth based on equipment sales. As mentioned above, it is difficult to obtain information that would help us sort out the real costs of providing service. Such information would simplify the task inferring service elasticities from equipment elasticities.

The demand model raises important questions about the competitive strategy for carriers. How aggressively should carriers invest in fiber? When should firms invest, given that next year's investment will yield more bandwidth per dollar than this year's? How should firms react to competitors' investments? Models of multiple player games under rapid, competition-fueled, technology-driven price declines will be required to determine what strategies make most sense,

and whether there is any sustainable competitive advantage to early movers or incumbents. The question of early mover/incumbent advantage hinges on whether the bandwidth market is a natural monopoly. It is theoretically possible, of course, that it is. But additional work is required to confirm our intuition that bandwidth markets are not a natural monopoly.

Finally, if experience suggests that the elasticity dynamic really is driving growth in demand, then policy makers will have to consider whether their interventions in markets expedite or impede the availability of low-cost services to all consumers. If investment in infrastructure is a better engine for lowering prices than subsidization, then universal service policies will have to adapt to more dynamic, higher elasticity markets.

Bibliography

Aldebert, Marc, Marc Ivaldi, and Chantal Roucolle. 1999. Telecommunication Demand and Pricing Structure : An Economic Analysis. In *Proceedings of 7th International Conference on Telecommunications Systems: Modeling and Analysis*:254-267. Nashville, TN.

Cahn, Robert S. 1998. *Wide Area Networks: Concepts and Tools for Optimization*. San Francisco: Morgan Kaufmann.

Eick, S., W. A. Massey, and W. Whitt. 1993. The Physics of the M(t)/G/infinity Queue. *Operations Research* 41: 400-408.

de Fontenay, Alain, Shugard, and Sibley, ed. 1992. *Telecommunications Demand Modeling: An Integrated View*. New York: Elsevier Science Pub.

Lanning, S. , W. A. Massey, B. Rider, and Q. Wang. 1999. Optimal Pricing in Queuing Systems with Quality of Service Constraints. In 16th International Teletraffic Congress.

Mueller, Milton. 1993. "Universal Service in Telephone History: A Reconstruction." *Telecommunications Policy* 17(5):352-370.

Mueller, Milton. 1997. *Universal Service: Competition, Interconnection, and Monopoly in the Making of the American Telephone System*. Washington: AEI Press.

Mukherjee, Biswanath. 1997. *Optical Communication Networks.* New York: McGraw-Hill.

Noam, Eli. 1995. "Economic Ramifications of the Need for Universal Telecommunications Service," in National Research Council, ed., *The Changing Nature of Telecommunications/Information Infrastructure*, pp. 161-164. Washington, DC: National Academy Press.

Ramaswami, Rajiv and Kumar N. Sivarajan. 1998. *Optical Networks: A Practical Perspective*. San Francisco: Morgan Kaufmann.

Stern, Thomas E. and Krishna Bala. 1999. *Multiwavelength Optical Networks: A Layered Approach*. Reading: Addison Wesley.

Taylor, Lester D. 1994. *Telecommunications Demand*. Amsterdam: Kluwer Academic Publishers.

Verma, Dinesh. 1999. *Supporting Service Level Agreements on IP Networks*. Indianapolis: Macmillan Technical Publishing.

Weinhaus, Carol, Sandra Makeeff, and Peter Copeland. 1994. "Redefining Universal Service: The Cost of Mandating the Deployment of New Technology in Rural Areas." Telecommunications Industries Analysis Project.