Collaborative Research: ITR/ANIR: 100 Mb/sec For 100 Million Households

Editors: Hui Zhang, David A. Maltz, Edward W. Knightly
Contributors: John C-I. Chuang, Alexander Fraser, Nick McKeown, Michael K. Reiter, Raj Reddy

March,2003

Contents

ITR Collaborative Research: 100 Mb/sec For 100 Million Household

PROJECT SUMMARY

This project is driven by the vision of providing *100 Mb/sec to 100 Million households and small businesses*. By providing a greater-than-two-order-of-magnitude increase in end-to-end interconnection speed, such an infrastructure will provide a platform that enables radically new content, applications, and services to emerge over the next decades. This 100×100 network has unprecedented speed, scale, and services representing a fundamental departure from today's Internet. Consider a U.S. city with 1 million households and small businesses, each with 100 Mb/sec access to switching centers where they are multiplexed with universities and businesses connected at 1 to 10 Gb/sec speeds. The bandwidth demands are astounding. Up to 100 Terabits per second of residential bandwidth combined with 100s of Terabits per second for enterprise and educational traffic, and a single city will approach the 1 Petabit per second threshold. Despite the scale, the network must offer dependable, secure, and guaranteed services to its users.

Thirty years ago, the Internet pioneers had the foresight to see the importance of data communication and its fundamentally different requirements from telephony. Rather than trying to enhance the successful telephony network, they started from a clean state and invented a technology that has since changed the world. We are at a similar historical crossroads. The design assumptions and requirements that underlie a nation-wide 100×100 network are vastly different than those considered by the designers of the original Internet 30 years ago. Consequently, a revolutionary approach, as opposed to an evolutionary one, is required to realize the 100×100 vision.

Achieving the 100×100 goal requires a three pronged approach: (1) a clean-slate architecture design that overcomes fundamental limits of today's Internet, (2) fundamental research that addresses the design of an economical, robust, secure and scalable 100×100 network, and (3) proof-of-concept network implementations to demonstrate how the network of the future can be built.

With this approach, the outcome of this project will result in formulation of the cross-cutting design *principles* for 100×100 including development of simple, structured, and resilient logical overlays and a technology-trend driven design methodology. It will result in a new architecture and protocols that jointly provide (1) cost-effective and resilient *access networks* that utilize both fiber-to-the-home and wireless last-hop access, (2) a scalable, fault-tolerant *backbone network* having simple logical structure and predictable performance, (3) *economic efficiency* that ensures sustained competition, and (4) *security* that strikes a balance between accountability vs. anonymity, as well as connectivity vs. isolation.

The 100×100 researchers form an interdisciplinary team with expertise in networking architecture, protocols, switch/router design, network management, traffic analysis, network operation, security, and economics, and are uniquely positioned to undertake the definition and accomplishment of the 100×100 vision.

The **intellectual merit** of this research includes the development and evaluation of the 100×100 architecture and protocols through the fundamental research identified above. The **broader impact** of this research includes new applications enabled by 100×100, societal impact, and education and outreach impact, ensured by engaging a large number of undergraduate and graduate students in the project, development of classroom software tools, and alliance with a women's college.

1 Introduction

The 100×100 network will have three breakthrough properties: (1) true broadband to the home at 100 Mb/sec (2) large-scale and economically viable deployability to 100 million residences and small businesses, and (3) unprecedented reliability, performance predictability, manageability, and security. With this infrastructure, radically new content, applications, and services will emerge and lead us into an era in which secure, reliable, and high-speed communication is an integral part of our social and business fabric.

Unfortunately, there are three key reasons why 100×100 cannot be achieved by following today's evolutionary path. **(1) Access.** Just as the once seemingly fast-paced rise in modem speeds came to a screeching halt at 56 kb/sec, today's copper-based "broadband" access technologies (cable modems and DSL) are rapidly approaching their transmission limits of several Mb/sec. **(2) Backbone.** Even with today's low-speed access links, Internet traffic is doubling every 12 months [24, 84], and routers, following technology trends enabled by Moore's law, are struggling to keep up. Indeed, projecting forward even with current access networks, in 10 years there will be a factor of 5 disparity between offered traffic and router capacity. With the access link speeds and wide-scale deployment of 100×100, traffic will increase by orders of magnitude, and this disparity will increase dramatically so that the basic foundations of backbone networking must be revisited. **(3) Architecture and Protocols.** Today's denial-of-service attacks [41], long-time-scale outages due to faults and misconfigurations [45, 66], periods of poor and unpredictable performance [1], and devastating economic failures [12], are inseparable from architectural and protocol design decisions. To achieve the above properties of 100×100 requires not only capacity scaling, but also rethinking the network's services and basic architectural and protocol building blocks.

Thus, we advocate a *clean-slate* design and a *revolutionary* approach based on four premises. First, access networks will be fiber-based, necessitating a re-build of the access infrastructure and having a deployment cost that will dwarf the cost of the backbone network. Second, core networks must be rebuilt to scale to future capacity demands. Third, a clean state design has intrinsic value by establishing a solution that is unblemished by constraints due to legacy design decisions and by providing a compass to guide the path of future development. Finally, current technologies and architectures are reaching fundamental limits as described above that will, indeed, *preclude* success of continuing with an incremental evolutionary approach.

With an inter-disciplinary research team consisting of economists, security experts, and networking researchers with expertise ranging from wireless to backbone networking, utilizing methodologies ranging from theoretical modeling to VLSI protocol implementation, our goal is to design, implement, and study the theoretical foundations of the protocols and architecture for the 100×100 network. Addressing access and backbone networks, we will design protocols and architectures that:

- ensure high resilience, bandwidth-scalability, performance predictability, and manageability via construction of simple, yet strategically chosen rapidly reconfigurable logical overlays built upon the naturally occurring "organic" physical topology;
- explicitly incorporate the advantages and constraints of technology trends such as Moore's law and optical switching advances;
- provide a powerful service interface for the network that enables new functionality and services to emerge;
- ensure a competitive, efficient, and economically sustainable 100×100 network by utilizing economic modeling of the Internet's history combined with an economically inspired design of its future; and
- enable society to rely on our communications infrastructure by preventing denial of service attacks and assuring the non-repudiatable identification of attackers.

With these advances, we decompose the end-to-end path into two subcomponents and propose the following technical contributions.

- **High-Performance, Resilient, Economically Viable Access Networks.** We will exploit fiber-to-the-home combined with fiber-to-the-neighborhood and beam-formed wireless-to-the-home as critical building blocks for a high-performance "last hop." First, we will leverage our recent multiplexer-on-a-chip design to develop a first-of-its-kind low-cost pole-mounted access switch that exploits statistical multiplexing deep inside the access network, thereby achieving efficiency and economic viability. Second, we will design pole-mounted "metropolitan access points" that exploit multiple unlicensed frequency bands to provide high-quality beam-formed wireless links to residences that cannot be cost-effectively reached by fiber. Finally, we will design resilient access networks that utilize structured logical overlays to rapidly and transparently recover from faults.
- **A Simple, Scalable, Robust Backbone.** The 100×100 goals demand a resilient, scalable, ultra-high capacity backbone network operating at several Petabits per second. Careful analysis of technology trends shows that it will be increasingly difficult to scale packet-switching routers to achieve this speed. Circuit switches[1] (either electronic or optical) can achieve much higher switching capacity at a lower cost than packet switches, at the expense of losing the benefit of statistical multiplexing. Both packet-switching and circuit-switching will need to cooperate to achieve the demand of 100×100 network. We will revisit the very foundations of circuit vs. packet switching and develop intelligent *dynamic* circuit switching protocols and hybrid packet/circuit solutions designed to exploit the vast capacity scaling, economic viability, and simplicity of circuits. Next, with our architectural design based on simple logical overlay topologies, we will jointly develop overlay strategies and protocols that are scalable, have predictable performance, and are resilient to node and link failures. Here, our key objective is to exploit the simplicity and structure of the logical overlay such that the data and control paths of routing are simplified, and protocols are optimized for logical topologies such as a fully-connected optical mesh, versus today's protocols which are technology and topology agnostic.

To demonstrate the capabilities of our solution, we will build a first-of-its-kind prototype implementation and testbed including (1) a wireless access testbed in Houston with prototype pole-mounted beam-forming access points, (2) fiber access testbeds in Princeton and Pittsburgh with combinations of prototype pole-mounted access-multiplexers-on-a-chip and off-the-shelf multiplexers, and (3) a prototype backbone node and a nationwide testbed to demonstrate the feasibility of 100 Mb/sec end-to-end. These implementations and testbeds will not only provide a proof-of-concept of our vision, but will also provide data and insights that will be critical to a nation-wide deployment of 100×100.

Thus, our goal is revolutionary: to create an economically sound 100×100 network with unprecedented scale, security, ubiquity, and performance.

2 Design Principles for the 100×100 Network

The success of the Internet was predicated on fundamental technical breakthroughs such as the understanding of packet-switching by Baran and Kleinrock [6, 61], the design of the internetworking protocol by Cerf and Kahn [18], and the framing of the Internet design principles and architecture guidelines by Clark et al [85]. However, the revolutionary goals of the 100×100 network demand a fundamental new look at the basic principles of network architecture and protocol design.

2.1 Principles

Simple and Structured Logical Overlay Topologies Enable Resilient, High-Performance Protocols. Network nodes *physically* connect to each other in a seemingly "organic" way. For example, there is compelling evidence that networks naturally become connected according to power laws, in which a small

[1] By circuits, we refer to true circuits in the style of TDM circuits or WDM circuits — not virtual circuits in the style of ATM. The key difference is that a circuit switch does not have queueing buffers.

number of nodes are densely connected while most nodes are sparsely connected [36]. Indeed, there are valid economic and physical reasons for such topologies to arise [101]. However, as a consequence of this inherent complexity and lack of structure, the failure modes of such topologies are innumerable and the paths through such networks are complex and of unpredictable performance, resulting in difficult-to-tune and non-robust protocols (see [45, 66] for example).

In contrast, networks with simple and structured logical overlays can employ highly simplified fault recovery, forwarding, and control protocols, thereby enabling simplicity and scalability of the network nodes themselves, as well as performance predictability and network analyzability [34, 80, 100]. Examples of *structured* topologies are trees, rings, hypercubes, and fully connected meshes. In densely interconnected access networks, such logical topologies are constructed by forwarding traffic only over a structured subset of the physical links; in more sparsely interconnected backbone networks, a structured topology such as a fully connected logical mesh is achieved by establishing multihop physical circuits (e.g., optical paths) among nodes.

Consequently, logical overlays will enable radically new protocols to be designed that exploit the topology's structure and achieve performance not possible if protocols must account for all possible performance and failure scenarios of an organic mesh. For example, the topological simplicity of a structured logical overlay enables design of protocols that can identify and provision redundant paths, rapidly detect the location of failures, and reroute traffic by activating a new physical topology to implement the same logical topology. Thus, nodes can have *fast preplanned and coordinated recovery* to faults. In Section 2.2, we present an example of a structured logical overlay.

Re-Evaluate Networking Design Tenets in the Context of Technology Trends. The ARPANET architects were technology agnostic and ensured that IP could be layered over a diverse set of link technologies ranging from Ethernet to dialup modems. Moreover, statistical multiplexing via packet switching was essential when costs were dominated by expensive shared T1 links between cities. While a wise decision at the time, technology trends in high-speed optics and integrated circuit design provide clear guidelines for protocol and architectural design moving forward. We illustrate the principle with two examples.

Fundamentals of Packet and Circuit Switching. Achieving bandwidth scalability is needed to build an ultra-high capacity core that supports pervasive, high-speed user connectivity. However, this scalability requires a new look at the fundamentals of packet and circuit switching. At the Pb/sec scale required for the core, the classical arguments about circuit and packet switching break down, since the very flexibility of packet switching necessitates electronic components that are not feasible in the Pb/sec range, even when accounting for Moore's law. Similarly, a brute-force static circuit mesh among core nodes would be highly inefficient with the dynamic and unbalanced traffic demands of the Internet.

Availability of Multiplexers-on-a-Chip. Enabled by Moore's law and "system-on-a-chip" designs, low-cost, low-power multiplexers on a single chip are now feasible. Consequently, active and intelligent components can be pushed deeper into the neighborhoods to create an access structure that is far more observable, scalable and manageable and which will incur much lower operating costs due to reduced human intervention. Moreover, packet-based statistical multiplexing deep in the *access* network, e.g., in pole-mounted access switches, can provide a foundation for extracting a significant statistical multiplexing gain while maintaining predictable performance at this critical juncture of the end-to-end path.

Design a Powerful Client-Network Service Abstraction. The definition of the service interface that the network offers to its users is one of the most important network architecture decisions. On one side of the interface are hundreds of millions customer devices; on the other side is the network, in a sense the world's largest distributed computer system. A well-designed interface should allow the two sides to evolve independently, even in the presence of unforeseeable and continuously changing technologies and requirements.

The current Internet offers a single service interface: a datagram abstraction for hierarchically addressed packets. This simple service interface has been effective for deploying a wide range of applications, but also

has limitations. For example, the Internet service model uses a single entity, the IP address, to both locate and identify a device. With these two concepts overloaded, the identity of a device must change whenever its location changes – problematic for mobile and portable devices, and cumbersome for security and economic mechanisms. In contrast, an abstraction and mechanisms that enable the network to separate these two concepts will enable powerful new services such as protection against DoS attacks via non-repudiatable identification of the source of messages and attacks. As a second example, a basic service abstraction of a quality-of-service aware *flow*, enables the network to move beyond a best-effort datagram service and provide communication with predictable performance.

Integrate Economics into Architecture and Protocol Design. A firm foundation of the economic principles of network *access* and *interconnection* is essential for the design of a competitive, efficient, and economically sustainable 100×100 network. A sound understanding of both the economic forces at work, and the technical requirements and constraints, are necessary for the design of the interconnection architecture.

For example, *access* networks are currently characterized by *lock-in* because access providers have little incentive to facilitate subscriber changeover or provide number/address portability. The design of the 100×100 access architecture must ensure sustained market-driven investment in the infrastructure, competitive access to "unbundled network elements," and minimal switching costs for users. Moreover, network *interconnection* at "peering points" provides universal host reachability across network domains, and has a significant impact on network performance and reliability. Empirical studies have revealed an underprovisioning of interconnection in today's Internet, resulting in significant routing inefficiencies and major bottlenecks [95]. One way to facilitate interconnection is to build lightweight accounting and measurement tools into the infrastructure, allowing fair and efficient division of the economic gains from interconnection. Another is to provide scalable negotiation and settlement mechanisms to facilitate flexible, on-demand interconnection.

Predicate Architectural and Protocol Design on Security Considerations. The Internet protocols were originally designed with little concern for security [30]. Consequently, a hodge-podge of network and end-user protocols and devices (firewalls, secure sockets, etc.) have had only limited success in thwarting even simple-minded attacks. At this critical design juncture, we have a unique opportunity to integrate security principles into protocol design.

For example, security is fundamentally a trade-off between restrictiveness and risk. Most revolutionary, at every layer of the system interconnection can be disallowed by default, earned only through negotiation, and enabled only by explicit network support to allow it. At the other end of the spectrum, the network can allow ubiquitous connectivity by default, but add capabilities such as automatic screening methods to deflect unwanted traffic; dynamic network reconfiguration to quarantine attacks; and traffic tracing and isolation. With security-driven protocol design, we can harden the security of the network infrastructure while simultaneously addressing challenges of stable routing, verification of code and routing table updates, and screening methods to effectively protect routers and end sites against Denial of Service (DoS) attacks.

2.2 An Example 100×100 Overlay

We now describe a candidate logical topology of the 100×100 network to illustrate the principles and highlight the challenges. Of course, the logical topology that we design will be one *outcome* of the research.

We first separate the network into *access* and *backbone* motivated by their vastly different characteristics and to enable an optimized design of each. The access network is characterized by a large-scale user population which dictates that it be low-cost and easily extended. Moreover, traffic patterns in the access network are unpredictable, wildly fluctuating and hard to characterize. The access network is naturally arranged as a hierarchy, with residences at the bottom and backbone nodes at the top. The backbone network is characterized by its need for enormous aggregate capacity and extremely high availability. Traffic patterns

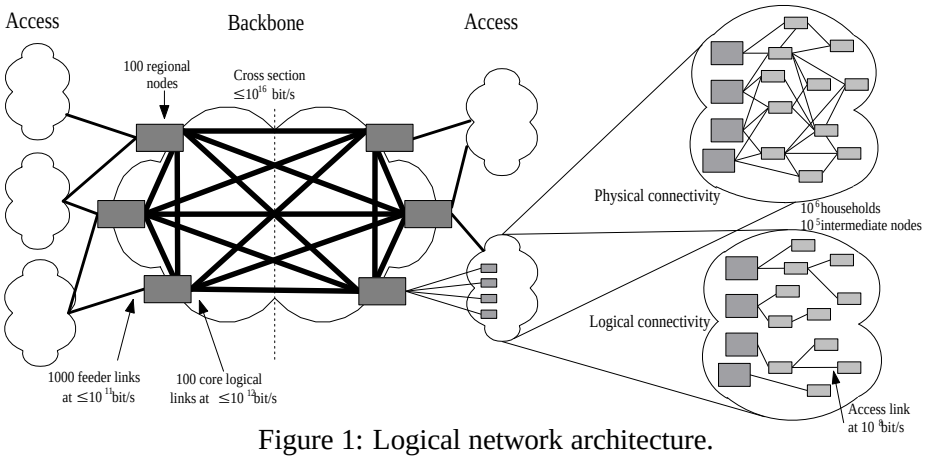

Figure 1: Logical network architecture.

in the backbone network are smoother and more predictable due to aggregation. The backbone network is naturally arranged as a rich interconnection of paths between access networks.

The access and backbone networks meet at a *regional nodes* as illustrated in Figure 1. Regional nodes are the roots of the access network and are interconnected by the backbone network. The backbone network considered here consists of approximately 100 regional nodes with each connecting to approximately one million users via multiple access networks. We believe that the number of regional nodes will be on the order of 100 (vs. 10 or 1000) as driven by population trends and in keeping with large backbone networks today comprised of 40 to 100 regional switching offices richly interconnected by high capacity circuits. The exact number and location of nodes will be determined by physical, political, administrative, and historic constraints.

Each access network is a highly reliable, low cost, low power, simple-to-manage logical overlay topology (a tree is illustrated) that connects 1 million households, and small businesses into a *regional node*. In contrast to the current Internet, the access network performs no local routing, but rather multiplexes traffic from the leaves toward the regional node, and demultiplexes traffic from the regional node back down to the leaves. Regardless of their destination, all packets from the access network are multiplexed up the tree to the regional node where they are processed and routed. Local traffic is sent back down to another leaf of the same tree, whereas long-haul traffic is routed over the backbone to a different access network. This optimizes for the most common case in which the two communication end points are in different access networks, and relieves the access network of significant complexity.

The logical backbone is built from a fully interconnected “mesh” of links created by an underlying network of physical circuits and circuit switches, in much the same way that SONET and DWDM networks are used to interconnect core routers today. The logical topology of the backbone consists of approximately 100 regional nodes, with each node logically connected to every other. A *fully-interconnected* logical backbone network has three key characteristics. First, all routing is 1-hop, which dramatically simplifies routing protocols. This should be contrasted with the current Internet in which packets are routed many times at successive routers and routing protocols must contend with a complex, dynamic and organically growing, multihop topology. The second way in which the 1-hop network is more reliable is that redundant paths are plentiful, enabling rapid routing around failure. If a link fails between two regional nodes, there are still many alternate paths through other regional nodes. Finally, as we discuss in Section 4, a fully-connected backbone can spread traffic uniformly over all links, eliminating hot-spots and giving a network-wide bandwidth guarantee: something not possible in today’s multihop network.

3 Fiber and Wireless Access with Resilient Logical Overlays

In the 1970s, the Internet pioneers could not have foreseen the revolution of the personal computer. Because ownership of computers was relegated to universities and government labs, the concept of "access networks" for residences and small business was then nonexistent. Residential access has since evolved from modem pools into the DSL lines and cable modems that characterize today's "broadband" access, and these copper-line-based solutions have limited prospects for growth beyond several Mb/sec.

The access network for the future (the next 50 to 100 years) will be fiber-based. Compared with copper, fiber offers virtually unlimited bandwidth potential and lower maintenance cost. Moreover, for a new network build-out, the cost of fiber is no longer an obstacle as the cost of new installation of fiber is comparable to legacy alternatives. Namely, advances in optical transceiver technologies such as VCSEL (850 nm and 1310 nm vertical cavity surface emitting lasers) enable 100 Mb/sec over sufficiently long distances (5-10 km) at extremely low cost. Likewise, innovative wireless access technologies can be combined with fiber access to achieve higher deployment flexibility. In particular, devices with beam forming antenna arrays and multiple air interfaces can achieve high performance with low cost by leveraging the attractive volume/cost curve of WiFi components and by using unlicensed spectrum.

While there have been a number of fiber-to-the-home activities including technology development (such as PON) and service trials/deployments [57, 76–78], their focus is on "last-mile" high speed access to the Internet. In contrast, we take a holistic view of 100×100 and develop an access network scalable to 1 million households, and providing secure, reliable, high-performance communication *end-to-end*. Thus, achieving the 100×100 goal requires a radically new look at access technologies, architectures, and protocols.

3.1 Access Architecture

Our study of access architecture is driven by the following core design principles. (1) *Exploit statistical multiplexing at the edges.* While traffic in the core is "smoothed" via high aggregation, traffic at the edges exhibits a high degree of burstiness allowing significant economies of scale via multiplexing. Thus, we will utilize pole-top access multiplexers to extract these efficiencies at the earliest possible point: inside the neighborhood. (2) *Ensure rapid failure recovery transparent to applications.* Because access networks naturally interconnect in complex meshes that reflect the complexity of physical geography and population distributions, we will utilize resilient structured overlays that rapidly recompute a new overlay whenever a node or link failure occurs. (3) *Jointly optimize for performance and economics.* This principle implies that the access architecture must be designed to minimize per subscriber cost [58]. Moreover, the architecture must be highly "future proof" to ensure that the capital expenditures of fiber and multiplexer deployment have a long-duration return on investment. (4) *Support network management and provisioning.* A large part of the networks operational expense is maintenance of failed lines, provisioning of new lines, and customer care. Automation via remotely manageable multiplexers in the neighborhoods can cut these expenses dramatically.

Access network design encounters numerous technology considerations, e.g., passive (shared medium) vs. active (switched) remote nodes, wireless vs. fiber to the home, deterministic multiplexing (TDM, WDM) vs. statistical multiplexing, statistical multiplexing via dynamic circuits vs. via packet switching. By addressing each issue with the above principles, we will develop models that incorporate performance and economics and lead to holistically designed architectures. For example, we will develop planning tools to map a physical metropolitan topology onto an optimal combined wired and wireless topology. Using state-of-the-art wireless propagation models, engineering cost models, and the performance constraints of individual and aggregate traffic, we will characterize the impact of the density of access points and fiber vs. wireless last-hops on the access network's economic and performance profile.

Likewise, to meet the scaling and cost requirements of the 100×100 network, the access network needs to be designed in a hierarchical fashion with combinations of passive and active aggregation/switching

devices. At the lower end of the hierarchy (toward the subscribers), we will investigate the design trade-offs (in terms of cost-effectiveness, manageability, security, etc.) between two competing architectures: (a) point-to-point dedicated fiber connection with active switches closer to homes; and (b) the shared medium architecture enabled by technologies such as PON (passive optical network). At the higher end of the hierarchy, we will investigate the trade-offs of using different multiplexing/switching technologies (packet, TDM, optical) at each stage of the multiplexing hierarchy.

3.2 Device and Protocol Design

In addition to architectural design, the 100×100 network introduces new challenges in the devices and protocols for both the fiber-based and wireless part of the network.

Fiber Access Network. To introduce access multiplexers into neighborhood poletops requires a low cost, low power, physically small, high performance access multiplexer. We have recently designed a micro-electronic access multiplexer in which the main data path of a packet switch is integrated into a DRAM memory chip. The key observation is that during the past 20 years the size of a DRAM memory chip has increased at an average rate approaching 60% per year whereas memory speed has only grown at about 11% per year. Consequently, to design a queue memory that can track the 60% annual performance increase for fiber-optic transmission requires transferring data to and from memory in words that match memory width. Because, word size is the square root of memory size, this technique allows us to obtain memory bandwidth which grows at about 40% per year. Utilizing these techniques, we will design and prototype both the neighborhood access multiplexer and the next-aggregation-level multiplexers to demonstrate capacity scaling in a low cost, low power, small-form-factor device.

There are two key protocol design challenges in the fiber access network. First, to achieve robustness, a structured topology must be overlaid over the naturally occurring random interconnection mesh. Today, tree structures are predominant in access networks and resilient rings in metropolitan backbones. We will take a broad new look at logical overlay structures such as trees of rings and structured grids and study fundamental tradeoffs between efficiency (the ability to highly utilize physical capacity) and resilience (the ability to quickly recover from failures), within the context of the required scale and device simplicity of the 100×100 access network. For each overlay structure, we will design fault recovery protocols that construct overlays such that the access network can rapidly and efficiently recover from node or link failures. Second, while congestion control has been widely studied (by us and others) in the context of TCP/IP [55, 72, 79], ATM [62, 93], and metropolitan networks [40], the unique structure of the 100×100 access network requires a new design. In particular, we will exploit the homogeneity of link capacities, the highly structured overlay, and the small round-trip-times within access networks, to design congestion control protocols that are sufficiently computationally simple to be integrated onto our single-chip access multiplexer devices, yet sufficiently responsive to provide higher throughput and lower latency than that achievable with a purely end-point (TCP-based) approach.

Wireless. There is a large gap between available wireless solutions and the 100×100 goals. For example, off-the-shelf WiFi access points placed in a dense urban mesh would result in nothing short of a throughput collapse due to excessive interference and contention [60]. Likewise, while 3G, LMDS, IEEE 802.16, etc. all have their niche applications, none can achieve high the 100×100 network's required per-user throughput, scalability, nor economically viability.

Node Architecture. The core design theme is an economically viable and performance scalable physical layer with *opportunistic* protocols that exploit any and all available resources and high-quality channels. To achieve this, we will design and implement a 100×100 Metro Access Point (MAP) and Residential Access Point (RAP) that uses antenna arrays to focus transmission energy into a directed beam aimed at the receiver. Together with multiple air interfaces, the 100×100 APs will simultaneously communicate with multiple

receivers without contention. Second, each air interface will be equipped with the ability to transmit within multiple frequency bands, including multiple channels within a sub-band (e.g., WiFi's 11 overlapping 22 MHz "channels" in the 83 MHz band at 2.5 GHz) as well as multiple unlicensed bands (e.g., 900 MHz, 2.5 GHz, and 5.7 GHz). By exposing diverse resources to higher layers, the architecture will provide every available opportunity to achieve the desired performance scaling.

Opportunistic Protocols. The key issue with today's medium access protocols is contention resolution: how to resolve collisions and multiple users contending for the same physical resource. In contrast, our key methodology is to design media access protocols that opportunistically seek out high quality channels with low contention and high potential throughput, and simultaneously communicating on *multiple* orthogonal channels. In this way, we will overcome the inherent scalability limitations of today's wireless protocols. For example, with multiple air interfaces, the 100×100 MAP can simultaneously beam-form to multiple RAPs without collision. Moreover, if two residences are located too close for separate beams, orthogonal frequencies can separate them, and multiple frequencies can be simultaneously used to ensure a sufficiently high data rate to the residence. The key challenge is therefore to design a media access protocol that selects from among all possible destinations (residences) and resources (e.g., channels) the set that will maximize throughput while simultaneously ensuring fair throughputs to each destination. Up-link traffic from RAPs (at residences) to MAPs introduces a fundamental *distributed* scheduling and media access problem. In particular, while MAPs act as a centralization point such that the above optimization problem can be solved with "global" information (albeit imperfect due to estimation error and resource variability), RAPs do not have knowledge of the state of other RAPs. Consequently, we will design distributed protocols that approximate the centralized solution with bounded deviation by sharing state among RAPs and coordinating scheduling decisions.

4 A Scalable Fault-Tolerant Network Backbone

Today's technology is inadequate to support the 100×100 backbone network for two reasons. First, today's backbone routers are based on packet-switching technologies which are increasingly difficult to scale. Second, today's backbone *topologies* are so complex and ad hoc that no network operator can predict or guarantee the performance of their network, nor rapidly recover from the wide variety of network failures. The complexity of the network has lead to complex routing protocols, which have hindered robustness, reliability, and manageability.

We will design the 100×100 network based on the following two observations. First, there is a growing disparity between packet and circuit switches in terms of capability and cost. We will design a backbone with a hybrid packet/circuit switched architecture that achieves high scalability and flexibility. Second, by deploying a rich interconnection of optical links and switches, the whole backbone can emulate a large, distributed router with predictable performance. Our backbone design will create a simple, structured overlay that will drastically simplify routing and recovery, enable bandwidth scaling, and sharply increase reliability and robustness of the network.

4.1 Fundamentals of Circuit Switching versus Packet Switching

With traffic growth exceeding Moore's Law, a packet-switched routers' ability to process packets diminishes over time. Our work therefore requires us to take a new look at the basic tradeoffs between packet switching (traditionally viewed as essential for the efficiency of statistical multiplexing) and circuit switching (traditionally viewed as inefficient for bursty data traffic).

First, we observe that for each class of backbone packet switched routers today, there is an equivalent circuit switch that costs about 75% less per Gb/s, consumes about 1/4 of the power, and is 1/4 the size. This is because a circuit switch does not process or buffer each packet; arriving data is simply mapped from an incoming to an outgoing circuit. The smaller size and reduced power of a circuit switch should be no

surprise. A typical 10 Gb/s router linecard today consists of 30 million gates, 300 Mbytes of buffers and consumes 200 Watts of power. A typical 10 Gb/s TDM linecard has fewer than 7 million gates, no packet buffers and consumes a fraction of the power.

Our second observation is that new optical switching technology lends itself to circuit switching, whereas all-optical *packet* switches are not feasible because we can't cost-effectively process or buffer photons. Optical switching technology (such as MEMS switches, tunable lasers and receivers, and DWDM) offer enormous switching capacity, well beyond the capabilities of electronic switches. Our thesis is that optical switching will allow the capacity of circuit switches to continually outstrip packet switches; and the difference will grow over time. Similar observations have led other efforts such as GMPLS [9, 68] to consider how IP traffic can control and be carried over an optical network. Yet, key issues regarding path selection and restoration remain open [11, 33].

Thus, our research task is to quantitatively and qualitatively revisit the basic arguments of packet and circuit switching from the perspectives of reliability, performance, cost, and power consumption.

4.2 Structured Logical Overlays

As illustrated in Figure 1, the backbone network can consist of approximately 100 regional nodes connected to each other by a *uniform mesh* of circuits. In other words, if there are N regional nodes, and each node is logically connected to the backbone with total rate R, each regional node is connected to every other node at rate R/N.[2]

Reliability and predictable throughput. The uniform fully-connected mesh backbone is more reliable because (1) Packets can be routed from ingress to egress node with simple and robust 1-hop routing, and (2) Upon failure of paths between nodes, it is easy to identify and use alternate redundant paths. While there is just one 1-hop path between every pair of nodes, there are $(N-1)$ two-hop paths. Furthermore, with such a simple topology, it is possible to make guarantees of robustness and availability; something that is not possible with - in fact, plagues - the current complex Internet topology. Moreover, the uniform mesh backbone has predictable throughput because of its non-blocking structure; a uniform mesh network has readily identifiable capacity.

2-Pass networks and unbalanced traffic. It is not immediately clear how a uniform mesh made from links running at R/N can support arbitrary traffic matrices; after all, the traffic matrix might contain elements larger than R/N. Yet, this can be overcome if all data traverses the backbone twice: on the first pass, traffic is spread from its ingress node across all the other nodes in the network. On its second pass, it is delivered to its correct egress node. This "2-pass network" has the critical property that it has guaranteed throughput for *all* traffic matrices; a mesh implemented today can support all future traffic matrices without incurring hot-spots or localized congestion. Moreover, as described below, this design has the flexibility to utilize packet- or circuit-switching among its nodes, with the ultimate choice made on merits of cost, reliability, ease of maintenance, and ease of implementation.

A circuit switched 2-pass mesh network. The logical network creates a large, distributed 3-stage Clos network as shown in Figure 2(a). Recall that in a Clos network with N nodes at each stage, the network is non-blocking if the nodes at each stage are connected to every node in the next stage by a link of rate R/N. We can make our N regional nodes function as a Clos network if each regional node acts as a first, second and third stage simultaneously, so that traffic traverses the backbone twice between ingress and egress. This 2-pass network is non-blocking, and can support any traffic matrix, so long as the nodes are connected at rate $2R/N$, where R is the total switching capacity of the regional node. An illustrative four node network is shown in Figure 2(b).

[2]Recall that circuits such as optical light-paths enable creation of the fully connected logical mesh from a non-fully-connected physical mesh as discussed in Section 2.1.

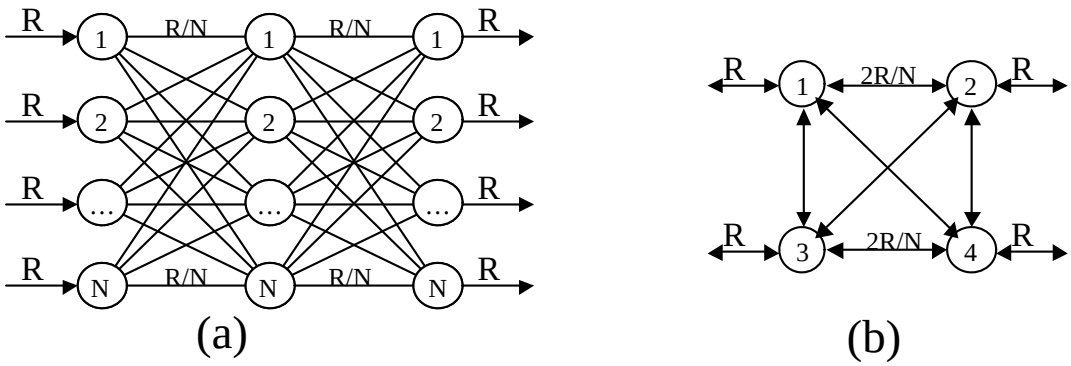

Figure 2: Conceptual organization of a Clos network (a) and its implementation using regional nodes (b).

With a *circuit-switched* 2-pass network, the key research challenge is the design of algorithms to decide when to dynamically create new circuits over the underlying R/N mesh, and to choose their rates. One possibility is for a regional node to monitor the average data rate or average queue occupancy of traffic destined to every other regional node. When rates increase, or queues build up, the regional node can allocate new circuits. The granularity of the circuit can range from individual application flows [73] to coarse DWDM channels based on a slowly changing traffic matrix [74].

A packet switched 2-pass mesh network. Alternatively, each regional node can spread traffic packet-by-packet across its outgoing links. When a packet arrives from the access network, it is first transmitted to another regional node, and then across the network again to the correct egress regional node. Packets are not buffered by the first or third nodes, but just once in the intermediate node. In this regime, each regional node operates as a packet switch, and buffers packets between the first and second time they cross the backbone.

Recent results have shown, somewhat surprisingly, that this approach has provably 100% throughput (the packet switching equivalent of non-blocking), with no hot-spots, for a broad class of traffic [19], provided that the data rates are the same as for the Clos network above. The technique is already being used *inside* commercial and research routers [4, 59]. In principle, the first stage of switching makes the non-uniform arrival process sufficiently uniform for the second stage to provide 100% throughput.

Design of the regional node. Regional nodes are the gateways between the packet switched access network and the circuit switched backbone network, and require enormous aggregate processing capability. Considering the basic service rate and size of the population to be served, it appears that the traffic carrying capacity of a fully populated regional node will lie in the range of 20 Tb/s to 100 Tb/s. Leveraging our work on design of a 100 Tb/s switch by introducing a low power optical switch fabric [4], we believe that it will be possible to build a regional node with this capacity in less than five years time.

The regional node in this network is simpler than a packet switched router with the same capacity. Although a regional node processes every packet in the access network (we envisage all packets being sent to the regional node, with no local switching between end users in the access hierarchy), and must map packets to and from the long-haul circuits, it does not need to process a single concatenated 100 Tb/s stream of packets. Thus, regional nodes do not need to be integrated into a single system, and their functionality can be achieved by building a distributed system consisting of several smaller nodes interconnected by a small, high-capacity, optical network.

5 Economics of 100×100

Architectural decisions made for communications networks have long-lasting and far-ranging impact on business models, industry structure, economic efficiency, and social welfare. Our multidisciplinary research

team will develop a comprehensive economics framework to learn from the Internet experience, and apply the findings to the design of the next generation network.

5.1 Retrospective Analysis

While the Internet has been an enormous technological success, the business record of the Internet industry has been mixed. This should come as no surprise, since interconnection and survivability were explicitly valued over accounting in the design of the ARPANET architecture thirty years ago [22]. On the one hand, the backbone providers have collapsed under the weight of intensive competition due to infrastructure over-investment. On the other hand, access networks have suffered from chronic under-investment under the incumbents. Despite legislative and regulatory efforts to bring about local competition, competitive local exchange carriers have largely failed. Meanwhile, ISPs' attempts at vertical expansion have met limited success, and many networking hardware and software firms are struggling to stay afloat. The fact that "broadband" penetration remains under 10 percent [25], despite the prospect of $500 billion annual benefit from universal broadband to the U.S. economy [29], suggests that there has been some market failure in the Internet industry.

To provide a foundation for an economic framework of the 100×100 network, a critical first step is to investigate the causes of this mixed business record. We will develop models of the industry:

- to clarify the role of federal (FCC) and state regulators (PUCs), the legislative intent of the 1996 Telecommunications Act, and actual market impact of "open access".
- to compare the evolution of the U.S. industry with that of other countries. How do differences in market structure, degree of regulation and competition, and pricing structures contribute to differences in outcome, in terms of infrastructure investment, service deployment, and market penetration?
- to assess the social and private benefits of networking, examining the extent to which "network effects" lead to a situation that is inimical to private appropriation, suggesting a major role for the public sector.

Armed with these fresh insights, we can turn to the challenge of designing the next generation network architecture based on sound economic principles and clear policy objectives.

5.2 Access Networks

Building access networks to 100 million households is a capital intensive undertaking, and dominates the cost of the 100×100 network. Once built, the access network infrastructure will serve the end users for many years to come. Therefore, it is critical that the access network architecture be designed to provide the economic incentives for continued investment and innovation by the network providers.

Access networks exhibit strong supply-side economies of scale, which implies a tendency towards market concentration and entrenched monopolies, and corresponding welfare loss. Furthermore, subscriber lock-in means incumbents have little incentive to invest and upgrade, and this in turn leads on infrastructure lock-in. We will explore the costs and benefits of different strategies for reducing user switching costs, such as protocols that enable an end user with a single network interface to seamlessly switch between access providers in real-time.

A different calculus must be performed for the case of municipality and community-owned access networks. While these networks are also subject to scale economies and infrastructure lock-in, they face vastly different investment and upgrade paths. Many of these networks serve rural communities, and therefore have very different cost models from urban networks. Building upon public finance economics, industrial organization theory, public and club goods theory [26], and telecommunications economics and policy [17, 70], we will propose an access network architecture that encourages sustained investment across ownership models, geography, and is compatible with universal access policy objectives.

5.3 Backbone Networks

Should the 100×100 backbone network be organized as a single national infrastructure run like a utility, or a market where multiple firms each deploy their own networks and compete for customers? The prevailing economic view favors a multi-provider market, since natural monopoly arguments do not generally apply to backbones, and competition provides good market discipline for backbone providers. On the other hand, having a single one-hop national backbone clearly offers significant availability and predictability gains at lower cost. There are historical precedents for both models in communications, utility, and transportation networks. We will identify and evaluate the various market organization alternatives for the 100×100 backbone.

Taking the shared infrastructure approach, is it possible to devise a backbone network architecture that supports multiple competing providers over a single infrastructure? For example, multiple competing regional nodes can interconnect over the single national network of circuits, and differentiate their services to the various access networks, data centers and other service providers that connect to them. Alternatively, the competition could take place inside the regional node among competing access network operators.

Taking the multi-backbone market approach, how can we design an interconnection architecture that supports seamless, flexible and on-demand interconnection and peering agreements, such that the network topology can dynamically adapt to traffic conditions and produce consistently high quality routes? How can we minimize strategic interactions between competing backbones that lead to suboptimal routing and performance bottlenecks? We believe that the key to optimal interconnection is to reduce transaction costs through the use of contracts. Building upon a small, but growing literature on automated negotiation and smart contracts [37, 86, 89, 99, 103], which are in turn grounded in mechanism design and contract theory, we will design mechanisms and protocols to support the specification, negotiation, execution and verification of contracts that are scalable to 100×100.

6 Network Security

6.1 Connectivity Versus Isolation

Today's networks were designed around the premise that arbitrary, efficient connectivity is an unquestioned goal. The growth of Internet-based attacks, however, gave rise to security "add-ons" such as firewalls and related technologies that protect one subnetwork from another by denying connectivity in an ad-hoc fashion. Unfortunately, by the time the attack reaches any defense the victim is in a position to deploy, it has already reached the victim's network and may already have done its damage. For example, as a central point where a network's access policy is applied, a firewall is itself vulnerable to denial-of-service via overloading, and may be filtering traffic too late to be of use.

The major thrust of recent research has been to deploy security defenses more pervasively in a network in hopes of placing defenses closer to the attackers. For example, *pushback* is a proposal by which filtering policies are pushed into the network so as to limit traffic flow to a victim under load [46, 65]. Another effort explores throttling denial-of-service traffic at the egress router of the network originating the traffic [71]. All of these approaches spring from the mindset that connectivity is the default posture that must be selectively "plugged" as unwelcomed traffic is discovered.

As we contemplate a clean slate redesign of the network, however, we question the fundamental premise that connectivity should be permitted by default. Specifically, we envision a network that can isolate parties from one another as easily as it can connect them. A default posture in this network might be one in which nearly all forms of communication are *disabled* by default. A new network, once physically connected to others, would remain unreachable until it grants permission for selected parties to send it certain types of traffic, or after it advertises its willingness to generally receive certain traffic. At the same time, any communication it initiates without first negotiating permission would either be dropped or delivered with

severely degraded service. This default network posture is the opposite end of the "connectivity versus isolation" spectrum from where we are today, and though extreme, could significantly impair attacks that exploit the permissive connectivity posture (including most worms and distributed denial-of-service attacks).

There are many research challenges that such a network vision suggests. First is to determine the types of policies that can be efficiently enforced in the network, particularly given the enormous bandwidth demands of 100×100. Second, although applying a network's protections at its edge is not sufficient as described above, applying every networks' protections pervasively at every node is not possible, either. Finding the right points in the network to enforce policies, which may differ per target network, requires balancing the granularity of the policies with the throughput of the node and carefully managing distributed state. We believe that these challenges can be overcome. Early work on distributed firewalls [47] and the automatic generation and deployment of filtering rules [7] has already started appearing, although not to the scale or with the constraints that the 100×100 network requires. These early results leave many important research questions unanswered, but provide hope that certain components of this vision are within reach.

6.2 Accountability Versus Anonymity

In hindsight, both accountability and privacy of users were inadequately considered in the design of today's networks. On the one hand, today's Internet is routinely abused with little recourse to hold the perpetrators accountable. This is at least partially due to the ease with which hostile users can cause computers they control to emit network packets with forged source addresses (and to have those packets delivered by the network). This is a frequent ingredient in Internet denial-of-service attacks of many varieties, so much so that in recent years, the research community has spent significant effort on the *traceback* problem, i.e., locating the true source of attack packets [2, 8, 16, 31, 67, 92]. Virtually without exception, this work focuses on approaches for tracing sources of attack packets with minimal (but nevertheless often significant) imposition on existing infrastructure.

On the other hand, the Internet infrastructure is frequently maligned for disclosing *too much* information about legitimate users, primarily by providing the source address of every packet it delivers. This is frequently touted as a violation of privacy — particularly since web sites can cross-correlate their web logs by client IP address. As a result, much research in *anonymous communication* is devoted to the goal of hiding the true source of requests from their destinations [39, 63, 82, 83].

Today's networks exhibit a particular choice on the axis between accountability and anonymity, in particular, a packet source address that is a *hint*. Unfortunately, this choice has proven to be inadequate for either purpose in practice — the source address is readily forgeable by attackers, but it violates the privacy of any honest user who forgoes the complicated measures required to hide it.

We envision a network that provides services by which communicants can negotiate a balance between accountability and anonymity, potentially on a per-flow basis. For example, a server may demand that the network strongly ensure the server's ability to hold the client accountable for misbehavior before it permits packets to reach the server. This may include the network making it possible for the server to request the identity of the client, isolate the client, or to otherwise sanction the client in the event of misbehavior. Going further, the network could hold payment from the client in escrow, as a disincentive for misbehavior by the client. As another example, a client could demand control over the release of identifying information to a server.

There are numerous benefits to integrating these mechanisms into the network. Today, for example, anonymous communication technologies are typically built as expensive application-level overlay networks that do not take the network topology into account. There is much performance to be gained by providing this service within the network infrastructure. At the same time, building the means to enforce strong accountability into the network would dramatically improve the manageability and resilience of the network and the community is supports.

7 Related Work

The above sections presented related work in specific research areas (access, core, economics, security, etc.) and contrasted 100×100 with today's Internet. In this section, we discuss other related efforts.

The NewArch project [13, 75] addresses network architectures and design principles focusing on issues such as trust and "tussles" [14, 23] that naturally emerge among competing players. In contrast, we pursue a holistic clean-slate design of the network, including access and backbone, the associated technologies, and the protocols and abstractions required to implement 100×100. While sharing many goals (e.g., security, economic incentives, etc.), our approach uniquely considers technology trends and network structure and targets a new nationwide infrastructure guided by our research and testbed deployments.

Active Networks [96, 97, 104] proposed to dramatically increase the flexibility and extensibility of the network layer via programmable elements. In contrast, our analysis of technology trends leads us instead to evaluate the power of simplifying and structuring the network layer in order to gain manageability and scalability.

Efforts in ATM represented an industry-lead consortium to utilize fixed-sized small-cell (53-byte) virtual-circuit switching as a basic construct for an integrated services network that some hoped would combine data, voice, and video traffic into a single network with a single end-to-end protocol. While ATM technology fostered many advances in areas such as switch design [98], congestion control [56, 62, 93], and traffic control [20, 27, 35, 42], ATM's failure to achieve its broad goals emphasizes the need for 100×100's tenets of highly simplified and scalable networks and protocols with ensured cost effectiveness.[3]

The IRIS project [5, 32, 48] aims to use "distributed hash tables" to develop a robust common framework and infrastructure for distributed overlay and peer-to-peer applications. 100×100 and IRIS are quite complementary, as we focus on access and backbone networks with unprecedented speed and scale, and IRIS focuses on overlays running on top of the network infrastructure. The 100×100 network provides the means to utilize a large number of high-bandwidth end systems in the circle of DHT nodes. In addition, DHT may prove to be an effective solution to naming and addressing that can be leveraged in 100×100 [28].

Finally, the Internet has been put to an uncountable number of uses and applications, ranging from the World Wide Web [10] to digital libraries [21, 81]. All of these will benefit from the increased bandwidth, reliability and stable performance that the 100×100 network will bring. More critically, 100×100's ubiquitous high-speed access will enable us to harness the vast untapped power of home computer resources to solve large-scale distributed computing problems as envisioned by eScience, grid computing [38, 43], the Smallpox Research Grid [44], and untold new applications.

References

[1] K. Adams and K. Bhalla. Quality of service over IP networks, October 2002. Gartner Report.

[2] M. Adler. Tradeoffs in probabilistic packet marking for IP traceback. In *Proceedings of* 34^{th} *ACM Symposium on Theory of Computing*, 2002.

[3] V. Anantharam, N. McKeown, A. Mekkittikul, and J. Walrand. Achieving 100% throughput in an input-queued switch. *IEEE Communications*, 47(8):1260–67, August 1999.

[4] Electrical Engineering Department at Stanford University. The Optical Router (OR) project. http://klamath.stanford.edu/or/.

[5] H. Balakrishnan, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica. Looking up data in P2P systems. *Communications of the ACM*, 46(2), February 2003.

[3] Note that virtual circuit switching is a form of packet switching quite different than both the datagram packet switching and circuit switching discussed in the context of 100×100.

[6] P. Baran. Introduction to distributed communications networks. Rand Corporation, Memorandum RM-3420-PR, August 1964.

[7] Y. Bartal, A. Mayer, K. Nissim, and A. Wool. Firmato: a novel firewall management toolkit. In *Proceedings of the 1999 IEEE Symposium on Security and Privacy*, pages 17–31, 1999.

[8] S. M. Bellovin. ICMP traceback messages. Internet Draft, March 2001.

[9] Editor L. Berger. Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description. Internet Request for Comments (RFC) 3471, January 2003.

[10] T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret. The World Wide Web. *Communications of ACM*, 37(8):76–82, 1994.

[11] G. Bernstein, J. Yates, and D. Saha. IP-Centric control and management of optical transport networks. *IEEE Communications Magazine, October 2000*, October 2000.

[12] P. Bernstein. What's wrong with telecom. *IEEE Spectrum*, 40(1):26–29, January 2003.

[13] R. Braden, D. Clark, S. Shenker, and J. Wroclawski. Developing a next-generation Internet architecture. Technical report, MIT/ISI, July 2000. Available at URL http://www.isi.edu/newarch/DOCUMENTS/WhitePaper.pdf.

[14] R. Braden, T. Faber, and M. Handley. From protocol stack to protocol heap – role-based architecture. In *Proceedings of HotNets-I*, Princeton, NJ, October 2002. Available at URL http://www.isi.edu/newarch/DOCUMENTS/hotrba.paper.pdf.

[15] L. Breslau, E. W. Knightly, S. Shenker, I. Stoica, and H. Zhang. Endpoint admission control: Architectural issues and performance. In *Proceedings of ACM SIGCOMM'00*, pages 57–69, Stockholm, Sweden, September 2000.

[16] H. Burch and B. Cheswick. Tracing anonymous packets to their approximate source. In *Proceedings of the* 14^{th} *USENIX Systems Administration Conference*, 2000.

[17] M. Cave, S. Majumdar, and I. Vogelsang, editors. *Handbook of Telecommunications Economics: Structure, Regulation and Competition*. ElSevier North-Holland, 2003.

[18] V. Cerf and R. Kahn. A protocol for packet network interconnection. *IEEE Transactions on Communications*, 22(5):637–648, May 1974.

[19] C.S. Chang, D.S. Lee, and Y.S. Jou. Load balanced Birkhoff-von Neumann switches, part I: one-stage buffering. In *IEEE HPSR Conference*, Dallas, TX, May 2001.

[20] G. Choudhury, D. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44(2):203–217, February 1996.

[21] M. Christel, T. Kanade, M. Mauldin, R. Reddy, M. Sirbu, S. Stevens, and H. Wactlar. Informedia digital video library. *Communications of the ACM*, 38(4):57 – 58, November 1995.

[22] D. Clark. The design philosophy of the DARPA internet protocols. In *SIGCOMM*, pages 106–114, Stanford, CA, August 1988. ACM.

[23] D. Clark, J. Wroclawski, K. Sollins, and R. Braden. Tussle in cyberspace: Defining tomorrow's internet. In *Proceedings of ACM SIGCOMM*, August 2002.

[24] K. Coffman and A. Odlyzko. Internet growth: Is there a 'Moore's Law' for data traffic? In J. Abello, P. Pardalos, and M. Resende, editors, *Handbook of Massive Data Sets*. Kluwer, 2001.

[25] Federal Communications Commission. FCC releases report on the availability of high-speed and advanced telecommunications capability, February 2002. Available at URL http://www.fcc.gov/Bureaus/Common_Carrier/News_Releases/2002/nrcc0201.html.

[26] R. Cornes and T. Sandler. *The Theory of Externalities, Public Goods and Club Goods*. Cambridge University Press, 1996.

[27] C. Courcoubetis, G. Kesidis, A. Ridder, and J. Walrand. Admission control and routing in ATM networks using inferences from measured buffer occupancy. *IEEE Transactions on Communications*, 43(2):1778–1784, 1995.

[28] R. Cox, A. Muthitacharoen, and R. Morris. Serving DNS using a peer-to-peer lookup service. In *Proceedings of IPTPS*, March 2002.

[29] R. Crandall and C. Jackson. The $500 Billion opportunity: The potential economic benefit of widespread diffusion of broadband Internet access, July 2001. Brookings Institute.

[30] National Research Council CSTB and CPSMA. *The Internet's Coming of Age*. National Academies Press, 2001.

[31] M. Franklin D. Dean and A. Stubblefield. An algebraic approach to IP traceback. *ACM Transactions on Information and System Security*, 5(2):119–137, 2002.

[32] F. Dabek, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica. Wide-area cooperative storage with CFS. In *Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP '01)*, Chateau Lake Louise, Banff, Canada, October 2001.

[33] R. Doverspike and J. Yates. Challenges for MPLS in optical network restoration. *IEEE Communications Magazine*, 39(2):89–96, 2001.

[34] R. D. Doverspike, G. Sahin, J. L. Strand, and R. W. Tkach. Fast restoration in a mesh network of optical cross-connect. In *Proceedings of OFC-99*, San Diego, CA, February 1999.

[35] A. Elwalid, D. Heyman, T. Lakshman, D. Mitra, and A. Weiss. Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing. *IEEE Journal on Selected Areas in Communications*, 13(6):1004–1016, August 1995.

[36] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of Internet topology. In *Proceedings of ACM SIGCOMM '99*, Cambridge, MA, aug 1999.

[37] J. Feigenbaum and S. Shenker. Distributed algorithmic mechanism design: Recent results and future directions. In *Proceedings of 6th International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, New York NY, 2002.

[38] I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of Supercomputer Applications*, 15(3), 2001. Available at URL http://www.globus.org/research/papers.html.

[39] M. J. Freedman and R. Morris. Tarzan: A peer-to-peer anonymizing network layer. In *Proceedings of the* 9^{th} *ACM Conference on Computer and Communications Security*, November 2002.

[40] V. Gambiroza, P. Yuan, L. Balzano, Y. Liu, S. Sheafor, and E. Knightly. Design, analysis, and implementation of DVSR: A fair, high performance protocol for packet rings," to appear in *IEEE/ACM Transactions on Networking*, December 2003.

[41] L. Garber. Denial-of-service attacks rip the Internet. *IEEE Computer*, 33(4):12–17, April 2000.

[42] R. Gibbens, F. Kelly, and P. Key. A decision-theoretic approach to call admission control in ATM networks. *IEEE Journal on Selected Areas in Communications*, 13(6):1101–1114, August 1995.

[43] Grid.org - grid computing. http://www.grid.org/home.htm.

[44] The smallpox research grid. http://www.grid.org/projects/smallpox/.

[45] T. Griffin and G. Wilfong. On the correctness of IBGP configuration. In *Proceedings of ACM SIGCOMM 2002*, Pittsburgh, PA, October 2002.

[46] J. Ioannidis and S. M. Bellovin. Implementing pushback: Router-based defense against DDoS attacks. In *Proceedings of the 2002 ISOC Network and Distributed System Security Symposium*, February 2002.

[47] S. Ioannidis, A. D. Keromytis, S. M. Bellovin, and J. M. Smith. Implementing a distributed firewall. In *Proceedings of the* 7^{th} *ACM conference on Computer and Communications Security*, November 2000.

[48] IRIS: Infrastructure for Resilient Internet Systems. http://project-iris.net/. (last visited 2003).

[49] S. Iyer, A. A. Awadallah, and N. McKeown. Analysis of a packet switch with memories running slower than the line rate. In *Proceedings of IEEE INFOCOM*, pages 529–538, Tel-Aviv, Israel, March 2000.

[50] S. Iyer, R. R. Kompella, and N. McKeown. Analysis of a memory architecture for fast packet buffers. In *IEEE Conference on High Performance Switching and Routing*, pages 368–373, Dallas, Texas, May 2001.

[51] S. Iyer and N. McKeown. Making parallel packet switches practical. In *Proceedings of IEEE INFOCOM*, volume 3, pages 1680–87, Alaska, USA, March 2001.

[52] S. Iyer and N. McKeown. On the speedup required for a multicast parallel packet switch. *IEEE Communication Letters*, 5(6):269–271, June 2001.

[53] S. Iyer and N. McKeown. Analysis of the parallel packet switch architecture. *IEEE/ACM Transactions on Networking*, April 2003. To appear.

[54] S. Iyer, R. Zhang, and N. McKeown. Routers with a single stage of buffering. In *Proceedings of ACM SIGCOMM*, Pittsburgh, USA, August 2002.

[55] V. Jacobson. Congestion avoidance and control. In *Proceedings of ACM SIGCOMM'88*, pages 314–329, Stanford, CA, August 1988.

[56] R. Jain. Congestion control and trac management in ATM networks: Recent advances and a survey. *Computer Networks and ISDN Systems*, 28(17):1723–1738, 1996.

[57] J. A. Jay. An overview of international fiber to the home deployment. In *Proceedings of the 2002 FTTH Conference*, October 2002. Read as www.ftthconference.com/pdf/101602-C1011-1.PDF.

[58] D. Johnson, M. Minkoff, and S. Phillips. The prize collecting steiner tree problem: Theory and practice. In *Proceedings of the* 11^{th} *ACM-SIAM Symp. on Discrete Algorithms*, pages 760–769, 2000.

[59] Juniper Networks. *IP infrastructure - M-series routers: Datasheet*. Available at URL http://www.juniper.net/products/ip_infrastructure/m_series/100042.html.

[60] V. Kanodia, A. Sabharwal, M. Xiao, and E. Knightly. Architectures and performance analysis of WiFi hot-spot scalability. Rice University Technical Report, March 2003.

[61] L. Kleinrock. *Queueing Systems*. New York, Wiley, 1976.

[62] H. T. Kung and R. Morris. Credit based flow control for ATM networks. *IEEE Network*, 9(2):40–48, March 1995.

[63] B. N. Levine and C. Shields. Hordes: A protocol for anonymous communication over the Internet. *ACM Transactions on Information and System Security*, 10(3):213–240, 2002.

[64] S. Machiraju, M. Seshadr, and I. Stoica. A scalable and robust solution for bandwidth allocation. In *Proceedings of IWQoS'02*, Miami Beach, FL, May 2002.

[65] R. Mahajan, S. M. Bellovin, S. Floyd, J. Ioannidis, V. Paxson, and S. Shenker. Controlling high bandwidth aggregates in the network. *Computer Communication Review*, 32(3), July 2002.

[66] R. Mahajan, D. Wetherall, and T. Anderson. Understanding BGP misconfigurations. In *Proceedings of ACM SIGCOMM 2002*, Pittsburgh, PA, October 2002.

[67] A. Mankin, D. Massey, C.-L. Wu, S. F. Wu, and L. Zhang. On design and evaluation of "intention-driven" ICMP traceback. In *Proceedings of the* 10^{th} *IEEE International Conference on Computer Communications and Networks*, October 2001.

[68] Editor E. Mannie. Generalized Multi-Protocol Label Switching (GMPLS) Architecture, February 2003. This is a working document and subject to change.

[69] M. Mathis, J. Heffner, R. Reddy, R. Raghunarayan, and J. Saperia. TCP extended statistics MIB. Internet Draft: draft-ietf-tsvwg-tcp-mib-extension-03bis.txt, March 2003. This is a working document and subject to change.

[70] L. McKnight and J. Bailey. *Internet Economics*. MIT Press, Cambridge MA, 1997.

[71] J. Mirkovic, G. Prier, and P. Reiher. Attacking DDoS at the source. In *Proceedings of the* 10^{th} *IEEE International Conference on Network Protocols*, pages 312–321, November 2002.

[72] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, October 2000.

[73] P. Molinero-Fernández and N. McKeown. TCP Switching: Exposing circuits to IP. *IEEE Micro Magazine*, 22(1):82–89, Jan/Feb 2002.

[74] P. Molinero-Fernández and N. McKeown. The performance of circuit switching in the Internet. *OSA Journal of Optical Networking*, 2(4):83–96, March 2003. Available at URL http://www.osa-jon.org/abstract.cfm?URI=JON-2-4-83.

[75] The NewArch Project. http://www.isi.edu/newarch.

[76] NTT. 100 Mbps FTTH Service Now Available. Press Release: available as http://www.ntt-east.co.jp/release_e/0204/020411c.html, 2003. Citation taken from from Dave Farber's IPers Newsletter 2/23/03, "A Personal Account: FTTH in Japan" by Naoki Yamamoto.

[77] City of Chicago. Chicago CivicNet. http://www.cityofchicago.org/CivicNet/Description.html. (last visited 2003).

[78] City of Palo Alto Utilities. CPAU Fiber-to-the-Home Project. http://www.cpau.com/fiberservices. (last visited 2003).

[79] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose. Modeling TCP throughput: A simple model and its empirical validation. In *Proceedings of ACM SIGCOMM 1998*, Vancouver, British Columbia, September 1998.

[80] S. Phillips, N. Reingold, and R. Doverspike. Network studies in IP/Optical layer restoration. In *Proceedings of OFC-2002*, Anaheim, CA, March 2002.

[81] Director Raj Reddy. The Universal Library. http://delta.ulib.org/html/vision.html.

[82] M. Reed, P. Syverson, and D. Goldschlag. Anonymous connections and onion routing. *IEEE Journal on Selected Areas in Communication*, 1998.

[83] M. K. Reiter and A. D. Rubin. Crowds: Anonymity for web transactions. *ACM Transactions on Information and System Security*, 1(1):66–92, November 1998.

[84] RHK. United States Internet traffic experiences annual growth of 100%, but just 17% revenue growth. Press release #157, RHK, Telecommunication Industry Analysis, May 2002.

[85] J. Saltzer, D. Reed, and D. Clark. End-to-end arguments in system design. *ACM Transactions on Computer Systems*, 2(4):277–288, November 1984.

[86] T. Sandholm and V. Lesser. Issues of automated negotiation and electronic commerce: Extending the contract net framework. In *Proceedings of ICMAS*, San Francisco CA, 1995.

[87] D. Shah, S. Iyer, B. Prabhakar, and N. McKeown. Maintaining statistics counters in router line cards. *IEEE Micro*, pages 76–81, Jan/Feb 2002.

[88] Matthew Siler and Jean Walrand. On-line measurement of QoS for call admission control. In *Proceedings of Internet Workshop on Quality of Service*, Napa, CA, May 1998.

[89] R. G. Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*, 29(12):1104–1113, 1981.

[90] I. Stoica, T.S.E. Ng, and H. Zhang. REUNITE: A recursive unicast approach to multicast. In *Proceedings of INFOCOM'00*, Tel-Aviv, Israel, March 2000.

[91] I. Stoica, H. Zhang, and S. Shenker. Self-verifying CSFQ. In *Proceedings of IEEE INFOCOM 2002*, New York, NY, June 2002.

[92] R. Stone. Centertrack: An IP overlay network for tracking DoS floods. In *Proceedings of the* 9^{th} *USENIX Security Symposium*, August 2000.

[93] C. Su, G. de Veciana, and J. Walrand. Explicit rate flow control for ABR services in ATM networks. *IEEE/ACM Transactions on Networking*, 8(3):350–361, June 2000.

[94] L. Subramanian, I. Stoica, H. Balakrishnan, and R. Katz. OverQoS: Offering QoS using overlays. In *Proceedings of HotNets 2002*, Princeton, NJ, October 2002.

[95] H. Tangmunarunkit, R. Govindan, D. Estrin, and S. Shenker. The impact of routing policy on internet paths. In *Proceedings of IEEE INFOCOM 2001*, Alaska, USA, April 2001.

[96] D. L. Tennenhouse, J. M. Smith, W. D. Sincoskie, D. J. Wetherall, and G. J. Minden. A survey of active network research. *IEEE Communications Magazine*, 35(1):80–86, 1997.

[97] D. L. Tennenhouse and D. J. Wetherall. Towards an active network architecture. *Computer Communication Review*, 26(2), 1996.

[98] J. Turner. Maintaining high throughput during overload in ATM switches. In *Proceedings of IEEE INFOCOM 1996*, San Francisco, CA, March 1996.

[99] H. Varian. Economic mechanism design for computerized agents. In *Proceedings of First Usenix Conference on Electronic Commerce*, New York NY, July 1995.

[100] Task Leader E. Varma. Task 3B - Mesh Based restoration for WDM Networks 1997 Q3 Report. Technical report, MONET Consortium, September 1997.

[101] D. Watts. *Small Worlds*. Princeton Press, 1999.

[102] Web100 project. http://www.web100.org. (last visited 2003).

[103] M. Wellman. A market-oriented programming environment and its application to distributed multicommodity flow problems. *Journal of Artificial Intelligence Research*, 1:1–23, 1993.

[104] D. Wetherall. Active network vision and reality: lessons from a capsule-based system. In *Symposium on Operating Systems Principles*, pages 64–79, 1999.