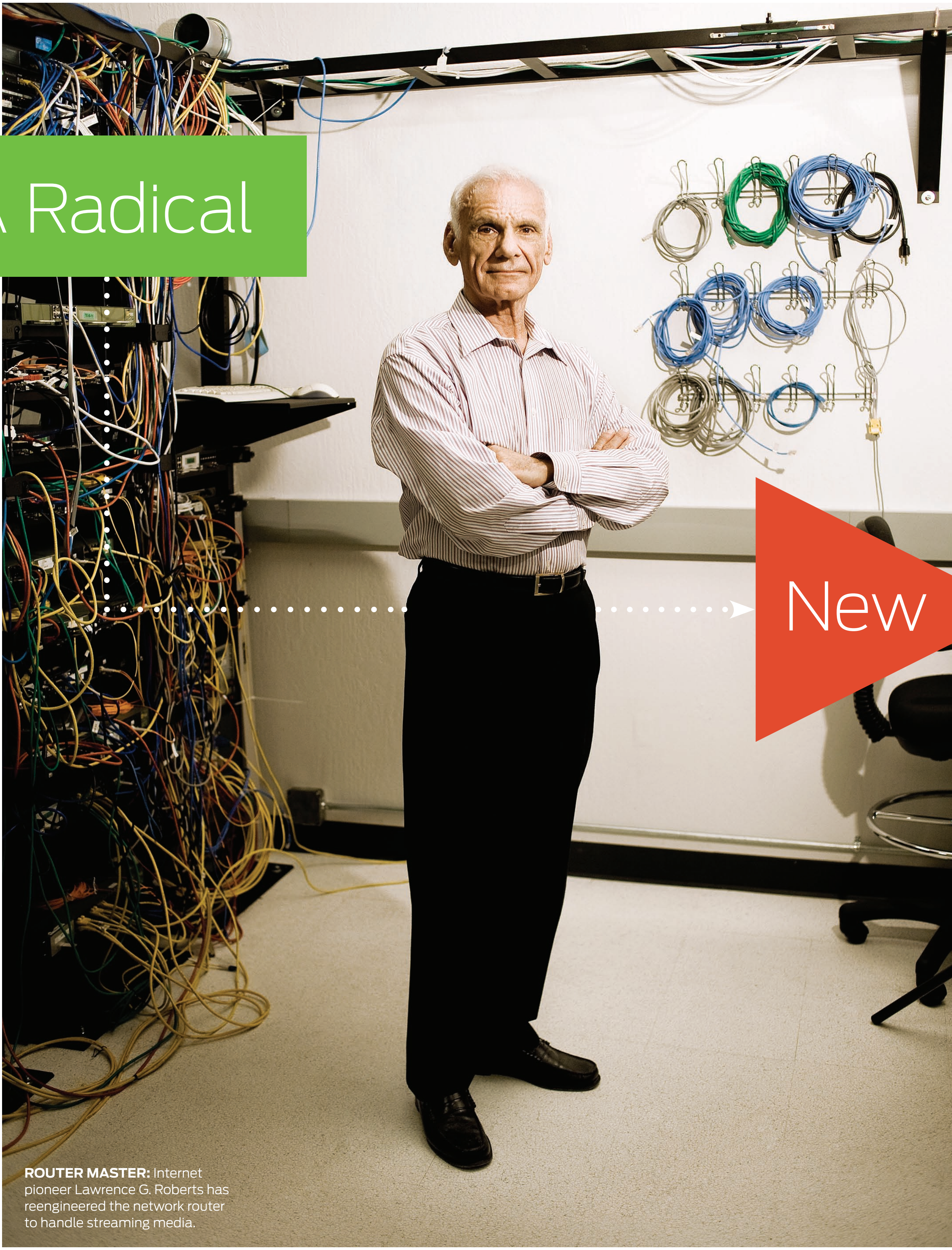
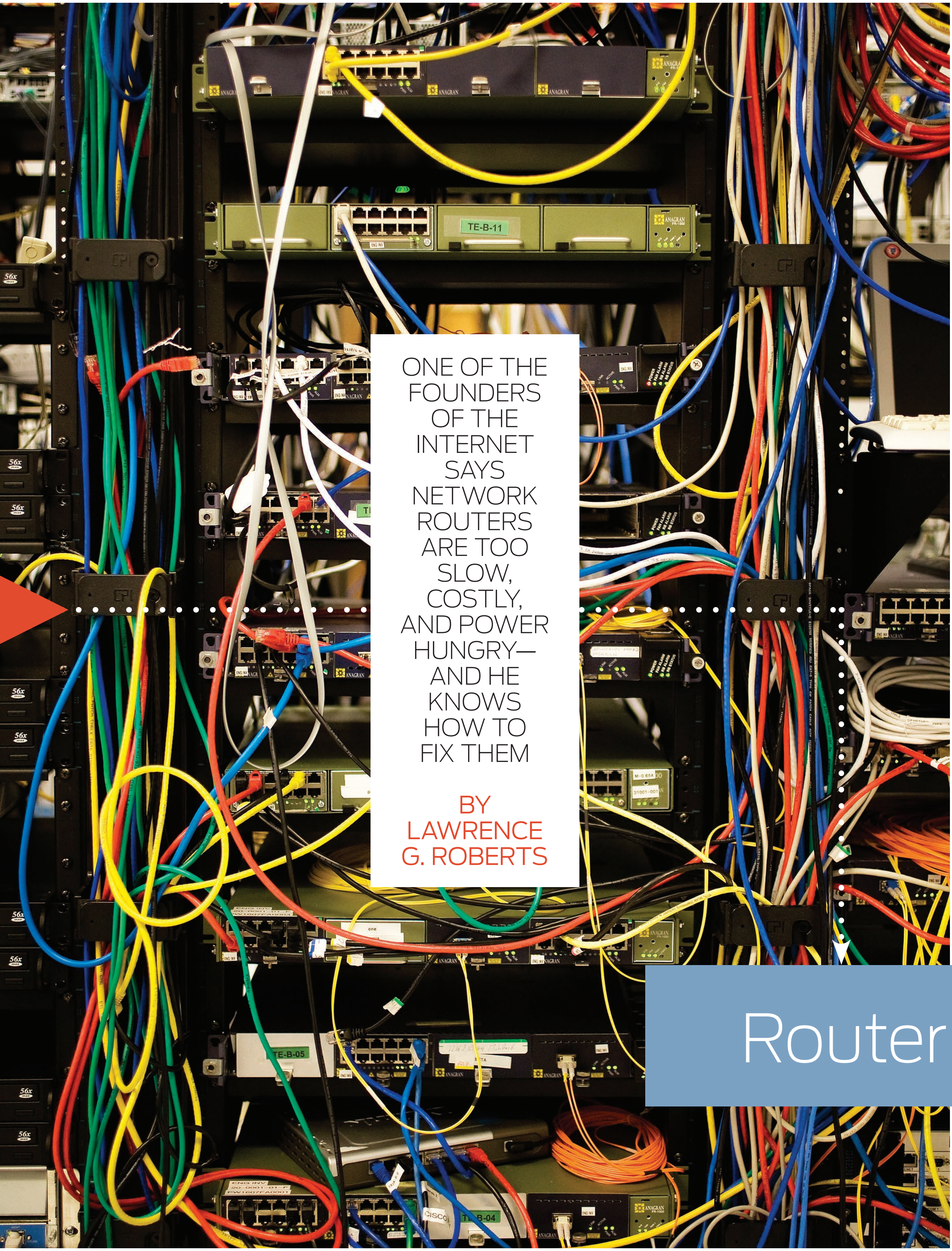
A Radical New



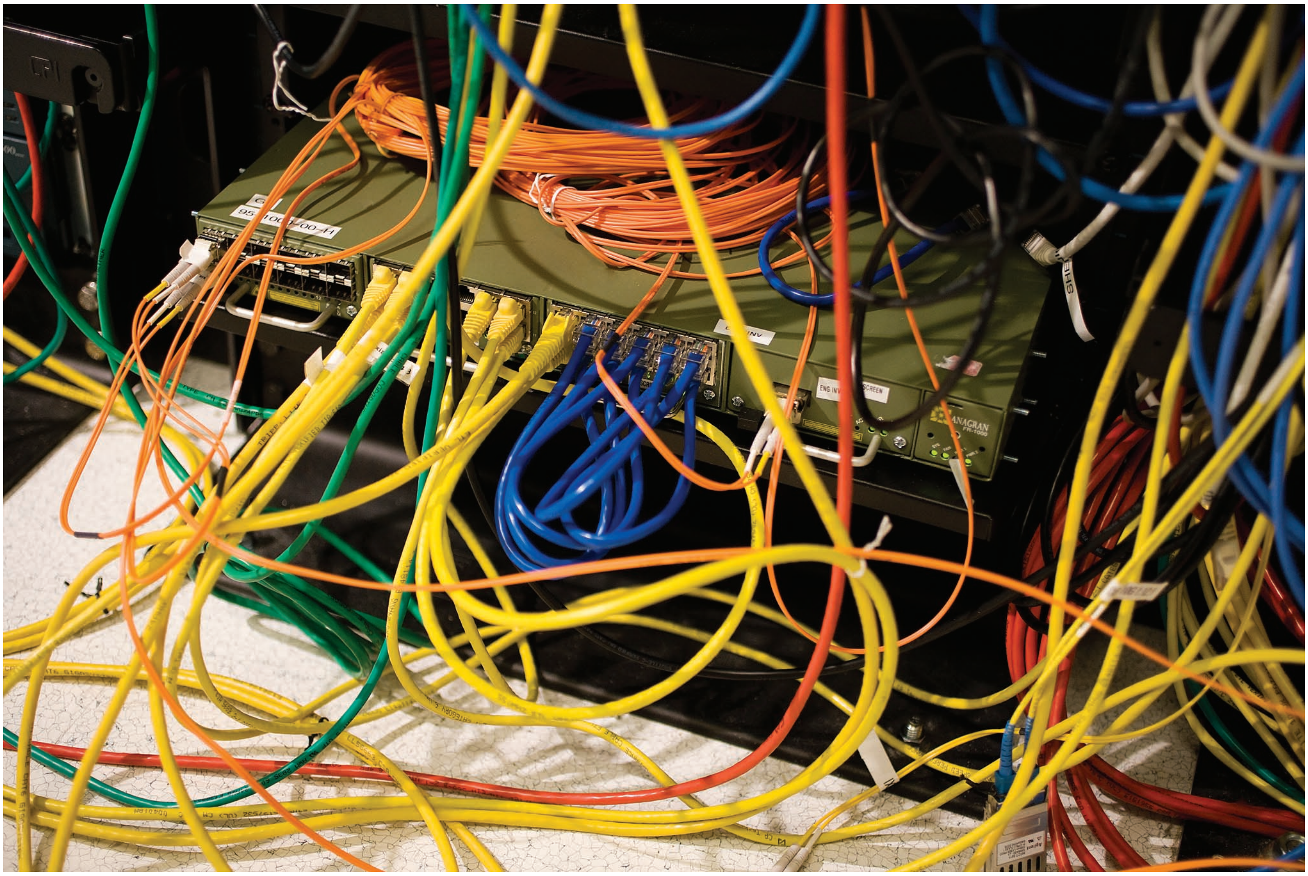
ROUTER MASTER: Internet pioneer Lawrence G. Roberts has reengineered the network router to handle streaming media.



Router

ONE OF THE FOUNDERS OF THE INTERNET SAYS NETWORK ROUTERS ARE TOO SLOW, COSTLY, AND POWER HUNGRY—AND HE KNOWS HOW TO FIX THEM

BY LAWRENCE G. ROBERTS



The Internet is broken.

I should know: I designed it. In 1967, I wrote the first plan for the ancestor of today's Internet, the Advanced Research Projects Agency Network, or ARPANET, and then led the team that designed and built it. The main idea was to share the available network infrastructure by sending data as small, independent packets, which, though they might arrive at different times, would still generally make it to their destinations. The small computers that directed the data traffic—I called them Interface Message Processors, or IMPs—evolved into today's routers, and for a long time they've kept up with the Net's phenomenal growth. Until now.

Today Internet traffic is rapidly expanding and also becoming more varied and complex. In particular, we're seeing an explosion in voice and video applications. Millions regularly use Skype to place calls and go to YouTube to share videos. Services like Hulu and Netflix, which let users watch TV shows and movies on their computers, are growing ever more popular. Corporations are embracing videoconferencing and telephony systems based on the Internet Protocol, or IP. What's more, people are now streaming content not only to their PCs but also to iPhones and BlackBerrys, media receivers like the Apple TV, and gaming consoles like Microsoft's Xbox and Sony's PlayStation 3. Communication and entertainment are shifting to the Net.

But this shift is not without its problems. Unlike e-mail and static Web pages, which can handle network hiccups, voice and video deteriorate under transmission delays as short as a few milliseconds. And therein lies the problem with traditional IP packet routers: They can't *guarantee* that a YouTube

ALL PHOTOS: JONATHAN SPRAGUE/REDUX

clip will stream smoothly to a user's computer. They treat the video packets as loose data entities when they ought to treat them as *flows*.

Consider a conventional router receiving two packets that are part of the same video. The router looks at the first packet's destination address and consults a routing table. It then holds the packet in a queue until it can be dispatched. When the router receives the second packet, it repeats those same steps, not "remembering" that it has just processed an earlier piece of the same video. The addition of these small tasks may not look like much, but they can quickly add up, making networks more costly and less flexible.

At this point you might be asking yourself, "But what's the problem, really, if I use things like Skype and YouTube without a hitch?" In fact, you enjoy those services only because the Internet has been grossly overprovisioned. Network operators have deployed mountains of optical communication systems that can handle traffic spikes, but on average these run much below their full capacity. Worse, peer-to-peer (P2P) services, used to download movies and other large files, are eating more and more bandwidth. P2P participants may constitute only 5 percent of the users in some networks, while consuming 75 percent of the bandwidth.

So although users may not perceive the extent of the problem, things are already dire for many Internet service providers and network operators. Keeping up with bandwidth demand has required huge outlays of cash to build an infrastructure that remains underutilized. To put it another way, we've thrown bandwidth at a problem that really requires a computing solution.

With these issues in mind, my colleagues and I at Anagran, a start-up I founded in Sunnyvale, Calif., set out to reinvent the router. We focused on a simple yet powerful idea: If a router can identify the first packet in a flow, it can just prescreen the remaining packets and bypass the routing and queuing stages. This approach would boost throughput, reduce packet loss and delays, allow new capabilities like fairness controls—and while we're at it, save power, size, and cost. We call our approach flow management.

TO UNDERSTAND HOW flow management works, it helps to describe the limitations of current packet routers. In these systems, incoming packets go first to a collection of custom microchips responsible for the routing work. The chips read each packet's destination address and query a routing table. This table determines the packet's next hop as it travels through the network. Then another collection of chips puts the packets into output queues where they await transmission. These two groups of chips—they include application-specific integrated circuits, or ASICs, as well as expensive high-speed memory such as ternary content-addressable memory (TCAM) and static random access memory (SRAM)—consume 80 percent of the power and space in a router.

During periods of peak traffic, a router may be swamped with more packets than it can handle. The router will then pile up more packets in its queue, establishing a buffer that it can discharge when traffic slows down. If the buffer fills up, though, the router will have to discard some packets. The lost packets trigger a control mechanism that tells the originator to slow down its transmission. This self-controlling behavior is a critical feature of the Transmission Control Protocol, or TCP, the primary protocol we rely on with the Internet. It's kept the network stable over decades.

Indeed, during most of my career as a network engineer, I never guessed that the queuing and discarding of packets in routers would create serious problems. More recently, though, as my Anagran colleagues and I scrutinized routers during peak workloads, we spotted two serious problems. First, routers discard packets somewhat randomly, causing some transmissions to stall. Second, the packets that are queued because of momentary overloads experience substantial and nonuniform delays, significantly reducing throughput (TCP throughput is inversely proportional to delay). These two effects hinder traffic for all applications, and some transmissions can take 10 times as long as others to complete.

As I talk to network operators all over the world, I hear one story after another about how the problem is only getting worse. Data traffic has been doubling virtually every year since 1970. Thanks to the development of high-capacity optical systems like dense wave division multiplexing (DWDM), bandwidth cost has been halved every year, so operators don't have to spend more than they did the year before to keep up with the doubling in traffic. On the other hand, routers, as pieces of computing equipment, have followed Moore's Law, and the cost of routing 1 megabit per second has decreased at a slower pace, halving every 1.5 years. Without a major change in router design, this cost discrepancy means that every three years a network operator will have to double its spending on infrastructure expansion.

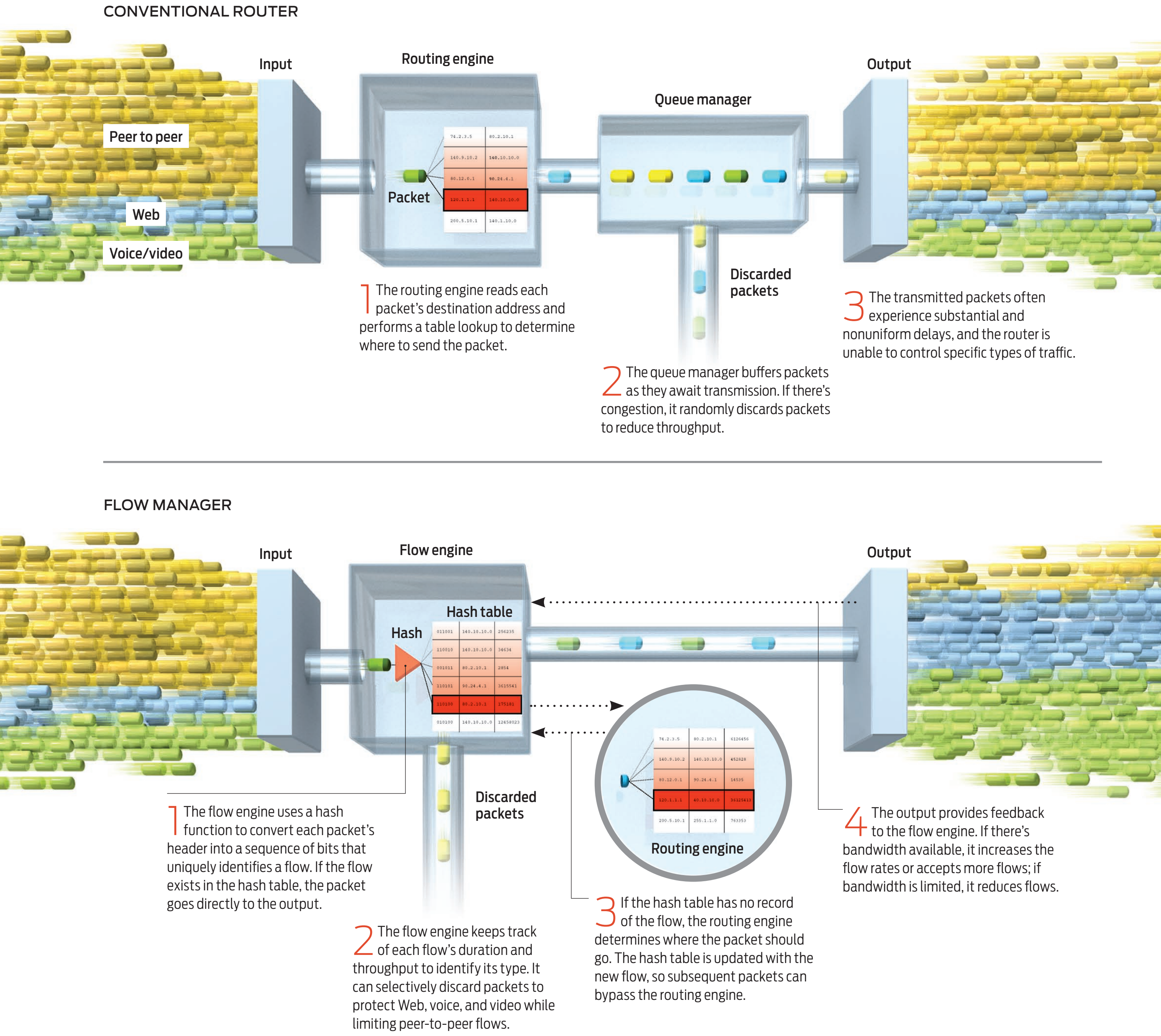
FLOW MANAGEMENT can solve this capacity crunch. The concept of data flow might be more easily understood in the case of a voice or video stream, but it applies to all traffic over the Internet. Key to our approach is the fact that each packet contains a full identification of the flow it belongs to. This identification, encapsulated by the packet's header according to the Internet Protocol version 4, or IPv4, consists of five values: source address, source port, destination address, destination port, and protocol.

All packets that are part of the same flow carry the same five-value identification. So in flow management, you have to effectively process—or route—only the first packet. You'd then take the routing parameters that apply to that first packet and store them in a hash table, a data structure that allows for fast lookup. When a new packet comes in, you'd check if its iden-

FLOW CONTROL: The Anagran FR-1000 can be plugged into existing networks and can manage up to 4 million simultaneous flows.

HOW FLOW ROUTING WORKS

Flow managers keep track of streams of packets and can protect voice and video transmissions while reducing peer-to-peer traffic.



1 The routing engine reads each packet's destination address and performs a table lookup to determine where to send the packet.

2 The queue manager buffers packets as they await transmission. If there's congestion, it randomly discards packets to reduce throughput.

3 The transmitted packets often experience substantial and nonuniform delays, and the router is unable to control specific types of traffic.

1 The flow engine uses a hash function to convert each packet's header into a sequence of bits that uniquely identifies a flow. If the flow exists in the hash table, the packet goes directly to the output.

2 The flow engine keeps track of each flow's duration and throughput to identify its type. It can selectively discard packets to protect Web, voice, and video while limiting peer-to-peer flows.

3 If the hash table has no record of the flow, the routing engine determines where the packet should go. The hash table is updated with the new flow, so subsequent packets can bypass the routing engine.

4 The output provides feedback to the flow engine. If there's bandwidth available, it increases the flow rates or accepts more flows; if bandwidth is limited, it reduces flows.

EMILY COOPER

tification is in the hash, and if it is, that means the new packet is part of a flow you've already routed. You'd then quickly dispatch—the more accurate term is "switch"—the packet straight to an output port, thus saving time and power.

If traffic gets too heavy, you'll still have to discard packets. The big advantage is that now you can do it intelligently. By monitoring the packets as they're coming in, you can track in real time the duration, throughput, bytes transferred, average packet size, and other metrics of every flow. For example, if a flow has a steady throughput, which is the case with voice and video, you can avoid discarding such packets, protecting these stream-based transmissions. For other types of traffic, such as Web browsing, you can selectively discard just enough packets to achieve specific rates without stalling those transmissions.

This capability is especially convenient for managing network overload due to P2P traffic. Conventionally, P2P is filtered out using a technique called deep packet inspection, or DPI, which looks at the data portion of all packets. With flow management, you can detect P2P because it relies on many long-duration flows per user. Then, without peeking into the packets' data, you can limit their transmission to rates you deem fair.

Since the early days of the ARPANET, I've always thought that routers should manage flows rather than individual packets. Why hasn't it been done before? The reason is that memory chips were too expensive until not long ago. You need lots of memory to store the hash table with routing parameters of each flow. (A 1 gigabit-per-second data trunk often carries about 100 000 flows.) If you were to keep a flow table on one IMP of 40 years ago, you'd spend US $1 million in memory. But about a decade ago, as memory cost kept falling, it started to make sense economically to design flow-management equipment.

In 1999, I founded Caspian Networks to develop large terabit flow routers, which I planned to sell to the carriers that maintain the Internet's core infrastructure. That market, however, proved hard to crack—the carriers seem satisfied with overprovisioning, as well as techniques like traffic caching and compression, which ameliorate congestion without addressing the roots of the problem. In early 2004, I decided to leave Caspian and start Anagran, focusing on smaller flow-management equipment to solve the overload and fairness problems. We designed the equipment to operate at the edge of networks, the point where an Internet service provider aggregates traffic from its broadband subscribers or where a corporate network connects to the outside world. Virtually all network overload occurs at the edge.

ANAGRAN'S FLOW MANAGER, the FR-1000, can replace routers and DPI systems or may simply be added to existing networks. It supports up to 4 million simultaneous flows—a combined 80 Gb/s in throughput. Its hardware consists of inexpensive, off-the-shelf components as opposed to ASICs, which increase development costs. We implemented our flow-routing algorithms in a field-programmable gate array, or FPGA, and the router's memory consists of standard high-speed DRAM. The FR-1000 sells in different models, starting at less than $30 000.

Like a regular router, the FR-1000 has input and output ports. But the similarities end there. Recall that in a traditional router the routing and queuing chips consume 80 percent of the power and space. By routing only the first packet of a flow, the FR-1000's chips do much less work, consuming about 1 percent of the power that a conventional router requires.

Even more significant, the FR-1000 does away entirely with the queuing chips. During congestion, it adjusts each flow rate at its input instead. If an incoming flow has a rate deemed too high, the equipment discards a single packet to signal the transmission to slow down. And rather than just delaying or dropping packets as in regular routers, in the FR-1000 the output provides feedback to the input. If there's bandwidth available, the equipment increases the flow rates or accepts more flows at the input; if bandwidth is scarce, the router reduces flow rates or discards packets.

By eliminating power-hungry circuitry, the FR-1000 consumes about 300 watts, or one-fifth the total power of a comparable router, and occupies one unit in a standard rack, a tenth of the space that other routers fill. We estimate that the equipment allows network operators to reduce their operating costs per gigabit per second by a factor of 10.

Measurements of the FR-1000 in our laboratories and by customers showed that networks equipped with the flow manager were able to carry many more streams of voice and video without quality degradation.

Another important capability we tested was whether the equipment could maintain quality of transmissions during congestion. The test involved a 100-Mb/s data trunk using a conventional router and another that included the Anagran flow manager. We progressively added TCP flows and measured the time required to load a specific Web page. The conventional router began to discard packets once traffic filled the trunk's capacity, and the time to load the Web page increased exponentially as we kept adding flows. The Anagran flow manager was able to control the rate of the flows, slowing them down to accommodate new ones, and the load time increased only linearly. The result: At 1000 flows, the flow manager delivered the page in about 15 seconds, whereas the conventional router required nearly 65 seconds.

Another capability we tested was fairness controls. Currently, P2P applications consume an excessive amount of bandwidth, because they use multiple flows per user—from 10 to even 1000. But services like cloud computing, which rely on Web applications constantly accessing servers that store and process data, are likely to expand the problem. We conducted measurements at a U.S. university whose wireless network was overwhelmed by P2P traffic, with a small fraction of users consuming up to 70 percent of the bandwidth. Early attempts to solve the problem using DPI systems didn't work, because P2P applications often encrypt packets, making them hard to recognize. The Anagran equipment was able to detect P2P by watching the number and duration of flows per user. And instead of simply shutting down the P2P connections, the flow manager adjusted their throughputs to a desired level. Once the fairness controls were active, P2P traffic shrank to less than 2 percent of the capacity.

The upshot is that directing traffic in terms of flows rather than individual packets improves the utilization of networks. By eliminating the excessive delays and random packet losses typical of traditional routers, flow management fills communication links with more data and protects voice and video streams. And it does all that without requiring changes to the time-tested TCP/IP protocol.

So is the Internet really broken? Okay, maybe that was an exaggeration. But the 40-year-old router sure needs an overhaul. I should know. ❏

TO PROBE FURTHER For more technical details on flow management and the Anagran FR-1000, visit http://www.anagran.com. See also the white papers "Flow Rate Management" and "TCP Rate Control With IFD [Intelligent Flow Delivery]," available at Lawrence G. Roberts's Web site: http://www.packet.cc.

For more on Internet routing, visit the Web sites of the Internet Research Task Force's Routing Research Group at http://www.irtf.org/charter?gtype=rg&group=rrg and the Internet Engineering Task Force's Routing Area at http://www.ietf.org/html.charters/wg-dir.html.