大学计算机教育丛书（影印版）网络互连技

IPv6
The New Internet Protocol
Second Edition

新因特网协议
IPv6
（第2版）

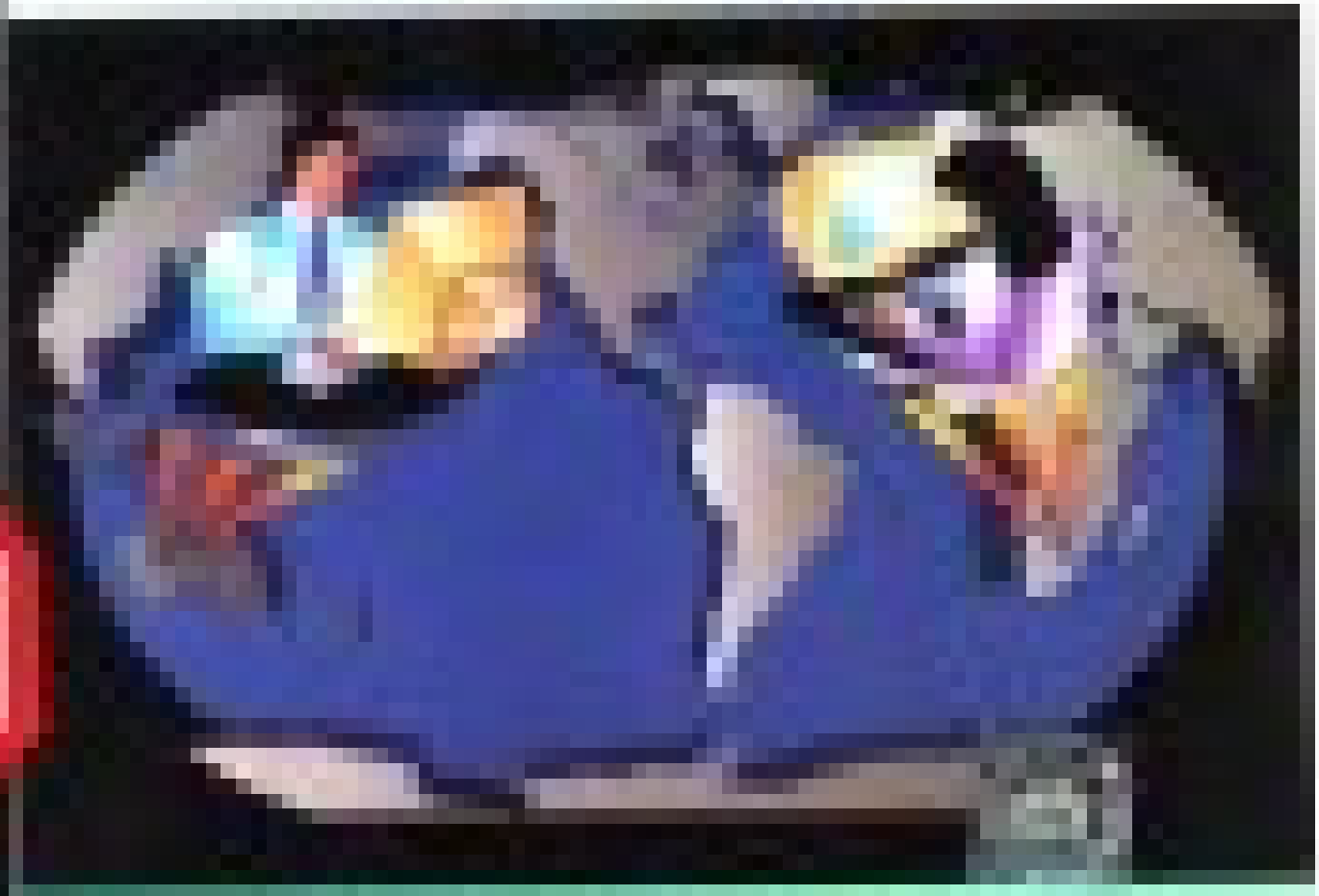
Christian Huitema

清华大学出版社 · PRENTICE HALL
http://www.tup.tsinghua.edu.cn

IPv6
The New Internet Protocol

Second Edition

新型网络协议
IPv6



IPv6

The New Internet Protocol

Second Edition

新因特网协议 IPv6

(第二版)

Christian Huitema

清华大学出版社

Prentice-Hall International, Inc.

(京)新登字 158 号

IPv6 —— the new Internet protocol second edition/Christian Huitema

北京市版权局著作权合同登记号:01-99-1741

图书在版编目(CIP)数据

新因特网协议 IPv6:第 2 版/(　)惠特马(Huitema,C.)著. —影印版. —北京:清华大学出版社,1999
(大学计算机教育丛书·网络互连技术系列)
ISBN 7-302-03547-4

Ⅰ. 新…　Ⅱ. 惠…　Ⅲ. 因特网-传输控制协议,IPv6　Ⅳ. TP393.4

中国版本图书馆 CIP 数据核字(1999)第 16044 号

出版者:清华大学出版社(北京清华大学校内,邮编 100084)
http://www.tup.tsinghua.edu.cn
印刷者:清华大学印刷厂
开　本:850×1168 1/32　印张:8.125
版　次:1999 年 5 月第 1 版　1999 年 11 月第 2 次印刷
书　号:ISBN 7-302-03547-4/TP·1948
印　数:5001～8000
定　价:14.00 元

出 版 前 言

清华大学出版社与 Prentice Hall 出版公司合作推出的“大学计算机教育丛书(影印版)”和“ATM 与 B-ISDN 技术丛书(影印版)”受到了广大读者的欢迎。很多读者通过电话、信函、电子函件给我们的工作以积极的评价,并提出了不少中肯的建议。其中,很多读者希望我们能够出版一些网络方面较深层次的书籍,这也就成为我们出版这套“网络互连技术系列”的最初动机。

众所周知,网络协议是网络与通信技术的关键组成部分。而今,因特网技术、移动通信技术的飞速发展,为网络协议注入了新内容。本套丛书以 Duglas Comer 教授的网络协议的经典名著 TCP/IP 网络互连技术系列为主干,并补充以论述新协议如 IPv6 和移动 IP 等国外最新专著,力求为从事网络互连技术研究与开发的人员以及大专院校师生提供充分的技术支持。

衷心希望所有阅读这套丛书的读者能从中受益。

清华大学出版社

Prentice Hall 公司

1998.9

Introduction

On a Saturday in June 1992, I took a plane from Osaka airport. I was leaving Kobe, where I had taken part in the first congress of the Internet Society. The Internet Activities Board (IAB) met in parallel with that congress. Shortly after the plane took off, I opened my portable computer and started to write the draft of the recommendation that we had just adopted. The choice of 32-bit addresses may have been a good decision in 1978, but the address size was proving too short. The Internet was in great danger of running out of network numbers, routing tables were getting too large, and there was even a risk of running out of addresses altogether. We had to work out a solution, we needed a new version of the Internet protocol, and we needed it quite urgently. During the meeting, we had managed to convince ourselves that this new version could be built out of CLNP, the Connection-Less Network Protocol defined by the ISO as part of the Open System Interconnection architecture. The draft that I was writing was supposed to explain all this: that we wanted to retain the key elements of the Internet architecture, that we would only use CLNP as a strawman, that we would indeed upgrade it to fit our needs, and that we hoped to unite the community behind a single objective—to focus the effort and guarantee the continued growth of the Internet.

1.1 Preparing for a Decision

I wrote the first draft on the plane and posted it to our internal distribution list the next Monday. The IAB discussed it extensively. In less than two weeks, it went through eight successive revisions. We thought that our wording was very careful, and we were prepared to discuss it and try to convince the Internet community. Then, everything accelerated. Some journalists got the news, an announcement was hastily written, and many members of the community felt betrayed. They perceived that we were selling the Internet to the ISO and that headquarters was simply giving the field to an enemy that they had fought for many years and eventually vanquished. The IAB had no right to make such a decision alone. Besides, CLNP was a pale imitation of IP. It had been designed 10 years before, and the market had failed to pick it up for all those years. Why should we try to resurrect it?

The IAB announcement was followed by a tremendous hubbub in the Internet's electronic lists. The IAB draft was formally withdrawn a few weeks later, during the July 1992 meeting of the Internet Engineering Task Force (IETF). The incident triggered a serious reorganization of the whole IETF decision process, revising the role of managing bodies such as the Internet Engineering Steering Group (IESG) or the Internet Architecture Board, the new appellation of the IAB. The cancellation of the IAB decision also opened a period of competition. Several teams tried to develop their own solutions to the Internet's crisis and proposed their own version of the new Internet Protocol (IP). The IESG organized these groups into a specific area, managed by two co-directors, Scott Bradner and Alison Mankin. In addition to the competing design groups, the area included specific working groups trying to produce an explicit requirement document or to assess the risk by getting a better understanding of the Internet's growth. A directorate was named. Its members were various experts from different sectors of the Internet community, including large users as well as vendors and scientists. The directorate was formed to serve as a jury for the evaluation of the different proposals.

The most visible part of the decision process was an estimation of the future size of the Internet. That effort started in fact in 1991, at the initiative of the IAB. We all agreed, as a basic hypothesis, that the Inter-

net should connect all the computers in the world. There are about 200 million of them today, but the number is growing rapidly. Vast portions of the planet are getting richer and more industrialized. There are reasons to believe that at some point in the near future, all Indian schoolboys and all Chinese schoolgirls will use their own laptop computers at school. In fact, when we plan the new Internet, it would be immoral not to consider that all humans will eventually be connected. According to population growth estimates available in 1992, it would mean about 10 billion people by the year 2020. By then, each human is very likely to be served by more than one computer. We already find computers in cars, and we will soon find them in domestic equipment such as refrigerators and washing machines. All these computers could be connected to the Internet. A computer in your car could send messages to the service station, warning that the brakes should be repaired. Your pacemaker could send an alarm message to your cardiologist when some bizarre spikes are noticed. We could even find microscopic computers in every light bulb so that we could switch off the light by sending a message over the Internet. A figure of a hundred computers per human is not entirely unrealistic, leading to a thousand billion computers in the Internet in 2020. But, some have observed that such a target was a bit narrow, that we wanted safety margins. Eventually, the official objectives for IPng (Internet Protocol, new generation) were set to one quadrillion computers (10 to the power 15) connected through one trillion networks (10 to the power 12).

A precise survey of the Internet growth quickly taught us that there was no real risk of running out of addresses in the next few years, even if 32-bit addresses only allow us to number four billion computers. We get estimates of the number of allocated addresses every month. If we plot them on a log scale and try to prolongate the curve, we see that it crosses the theoretical maximum of four billion somewhere between 2005 and 2015. This should give us ample time to develop the new protocol that we were at the time calling IPng (Internet Protocol, new generation). But we should take into account the limited efficiency of address allocation procedures. I proposed to estimate this efficiency through the *H* ratio:

$$H = \frac{\log(\text{number of addresses})}{\text{number of bits}}$$

The *H* ratio is defined as the division of the base 10 logarithm of the number of addressed points in the network by the size of the address, expressed in bits. If allocation were perfect, one bit would number two hosts, 10 bits would number 1024 hosts, and so on. The ratio would be equal to the logarithm of 2 in base 10, which is about 0.30103. In practice, the allocation is never perfect. Each layer of hierarchy contributes to some degree to the inefficiency. The logarithmic nature of the ratio tries to capture this multiplicative effect. Practical observation shows that *H* varies between 0.22 and 0.26 in large networks, reflecting the degree of efficiency that can be achieved in practice today.

If the *H* ratio may vary between 0.22 and 0.26, 32-bit addresses can number between 11 and 200 million hosts. We should keep this in mind. The current Internet protocol is adequate for connecting all the computers of the world today, but it will have almost no margin left at that future stage. Predicting a date of 2005 or 2015 simply means that we do not expect a rush into the Internet in the next few years. We may well be wrong. In fact, I hope that we are wrong—that there will indeed be a rush to connect to the Internet.

The other lesson that we can draw from the *H* ratio is that if we want to connect one quadrillion computers to the new Internet, addresses should be at least 68 bits wide for a ratio of 0.22 and only 57 bits wide for a ratio of 0.26. We used these figures when we made our final selection.

1.2 Two Years of Competition

When the IAB met in Kobe, there were only three candidate proposals for the new IP. The proposal to use CLNP was known as TUBA (TCP and UDP over Bigger Addresses). The main difference between IP and CLNP was CLNP's 20-octet Network Service Access Point addresses (NSAP). This would certainly suffice for numbering one trillion networks. The main argument for this proposal was its installed base. CLNP and its companion protocols, such as IS-IS for routing, were already specified and deployed. A side effect was convergency between the OSI and Internet suites. TCP, UDP, and the ISO transport would all run over CLNP; the protocol wars would be over. The main counterar-

guments were that this deployment was very limited and that CLNP is a very old and inefficient protocol. It is in fact, a copy of IP, the result of an early attempt to get IP standardized within the ISO. During this standardization process, many IP features were corrected, or rather changed, in a way that did not please the Internet community. A slower but more robust checksum algorithm was selected. The alignment of protocol fields on a 32-bit word boundary was lost, as well as some of the key services provided by ICMP. In the end, this proposal failed because its proponents tried to remain rigidly compatible with the original CLNP specification. They did not change CLNP to incorporate any of the recent improvements to IP, such as multicast, mobility, or resource reservation. They did not want to lose the "installed base" argument, even if that base was in fact quite slim.

In June 1992, Robert Ullman's proposal, called IP version 7, was already available. This proposal evolved between 1992 and 1994. The name was changed to TP/IX in 1993. The new name reflected the desire to change the Transport Control Protocol, TCP, at the same time as the Internet Protocol. It included hooks for speeding up the processing of packets, as well as a new routing protocol called RAP. The proposal failed however to gain momentum and remained quite marginal in the IETF. It evolved in 1994 into a new proposal called CATNIP, which attempted to define a common packet format that would be compatible with IP and CLNP, as well as with Novell's IPX. The proposal had some interesting aspects, but the IPng directorate felt that it was not sufficiently complete at the time of its decision, July 1994.

The third alternative available in June 1992 was called IP in IP. It proposed to run two layers of the Internet protocol, one for a worldwide backbone and another in limited areas. By January 1993, this proposal had evolved into a new proposal called IP Address Encapsulation, IPAE, that was then adopted as the transition strategy for Simple IP, or SIP, which Steve Deering had proposed in November 1992. SIP was essentially a proposal to increase the IP address size to 64 bits and to clean up several of the details of IP that appeared obsolete. It used encapsulations rather than options and made packet fragmentation optional. SIP immediately gathered the adherence of several vendors and experimenters. In September 1993, it merged with another proposal called Pip. With Pip, Paul Francis proposed a very innovative routing strategy based

on lists of routing directives. This allowed a very efficient implementation of policy routing and also eased the implementation of mobility. The result of the merging of SIP and Pip was called Simple IP Plus, SIPP. It tried to retain the coding efficiency of SIP and the routing flexibility of Pip.

The IPng directorate reviewed all these proposals in June 1994 and published its recommendation in July 1994. It suggested using SIPP as the basis for the new IP, but changed some key features of its design. In particular, they were unhappy with the lists of 64-bit addresses used by SIPP. The new IP would have 128-bit addresses. It will be version 6 of the Internet protocol, following version 4 that is currently in use. The number 5 could not be used because it had been allocated to ST, an experimental "stream" protocol designed to carry real-time services in parallel with IP. The new protocol will be called IPv6.

1.3 The New Specifications

A first version of this book was written in the fall of 1995 and published in December of that year. At that time, the working groups had worked for more than a year to finalize the specifications of IPv6, and I thought that the available drafts were almost definitive. It turns out that I was not entirely right. Some key elements changed between the writing of the first edition and the publication of the final specifications, notably the format of the source routing header. The specifications of the basic protocol were published in January 1996, the transition strategy in April, and the neighbor discovery and address configuration procedures in August. This second edition is based on this first set of publications, that have now reached the "proposed standard" stage in the IETF standardization process, with two exceptions: the routing protocols and the key negotiation procedures for authentication and encryption that are still being worked on by the IETF. The book is organized into eight chapters, including this introduction and a provisional conclusion.

Chapter 2 will present the protocol itself, as well as the new version of the ICMP, Internet Control Message Protocol. It will explain how Steve Deering and the members of the working group exploited the opportunity to design a new protocol. We avoided most of the second design syndrome effect, kept the proliferation of options and niceties to

a minimum, and in fact produced a new Internet Protocol that should be simpler to program and more efficient than the previous version.

In Chapter 3, we will analyze the evolution of addressing and routing, presenting the various address formats and the supports for multicast, and provider addressing.

The three following chapters will be devoted to the new capabilities of IPv6: autoconfiguration, security, and the support of real-time communication. All these functionalities could only be partially integrated in IPv4. They will be mandatory in all implementations of IPv6. Chapter 7 will describe the deployment strategy, explaining the transition of the Internet from IPv4 to IPv6.

1.4 Points of Controversy

In theory, the adoption of IPv6 was a miracle of consensus building. The debates were fair and everybody was supposed to smile after the decision. The members of the SIPP working group tried to play by the rules. They held a party shortly after the decision, but there was no mention of a victory. Officially, it was the "we can't call that winning" party.

In fact, the consensus was quite large. Many members of the TUBA working group joined the IPv6 effort and took part in the final discussions of the specifications. Ross Callon, the very person who forged the TUBA acronym, co-chaired the IPv6 working group with Steve Deering. But, a large consensus is not equivalent to unanimity. Many IETF members still believe that their pet ideas have not been taken into account. Many decisions were only adopted after long discussion, and some points are still being debated. I have tried to present these at the end of each chapter in a separate section, "Points of Controversy."

1.5 Further Reading

Each chapter ends with a list of references for further reading. Many of these references are Requests for Comments (RFCs). The RFC series is the electronic publication of reference of the IETF. RFCs are freely available from a number of repositories around the Internet. Some of the references have yet to be published. Provisional versions can be found in the Internet Draft repositories of the IETF.

The IETF decision itself is documented in RFC 1719. Scott Bradner and Alison Mankin, the chairs of the IPng area of the IETF, have been careful to also publish most of the discussion papers as RFCs. The TUBA proposal is documented in RFC 1347, 1526, and 1561; Pip in RFCs 1621 and 1622; TP/IX in RFC 147;, CATNIP in RFC 1707; and SIPP in RFC 1710. Contributions to the debate may be found in RFCs 1667 to 1683, 1686 to 1688, 1705, and 1715. Scott and Allison edited a book, *IPng Internet Protocol Next Generation*, published by Addison-Wesley, that provides an easy-to-read summary of these discussions.

Readers are expected to be familiar with TCP-IP. Many books have been written to present this technology, notably *Internetworking with TCP-IP* by Douglas E. Comer, published by Prentice Hall.

The Design of IPv6

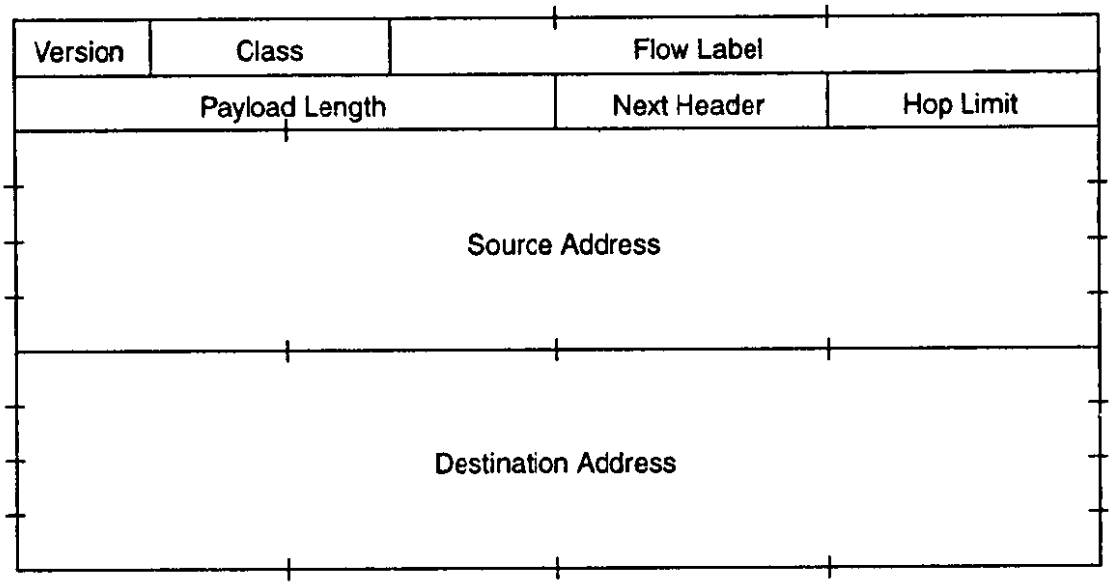
The new IP is based on a very simple philosophy: The Internet could not have been so successful in the past years if IPv4 had contained any major flaw. IPv4 was a very good design, and IPv6 should indeed keep most of its characteristics. In fact, it could have been sufficient to simply increase the size of addresses and to keep everything else unchanged. However, 10 years of experience brought lessons. IPv6 is built on this additional knowledge. It is not a simple derivative of IPv4, but a definitive improvement.

2.1 The IPv6 Header Format

Any presentation of the new IP has to start with a presentation of the IPv6 header format. It is composed of a 64-bit header, followed by two 128-bit IPv6 addresses for source and destination, for a total length of 40 bytes.

The initial 64 bits are composed of the following:

- Version field (4 bits)
- Class (8 bits)
- Flow label (20 bits)
- Length of the "payload" (16 bits)
- Type of the next header (8 bits)
- Hop limit (8 bits)

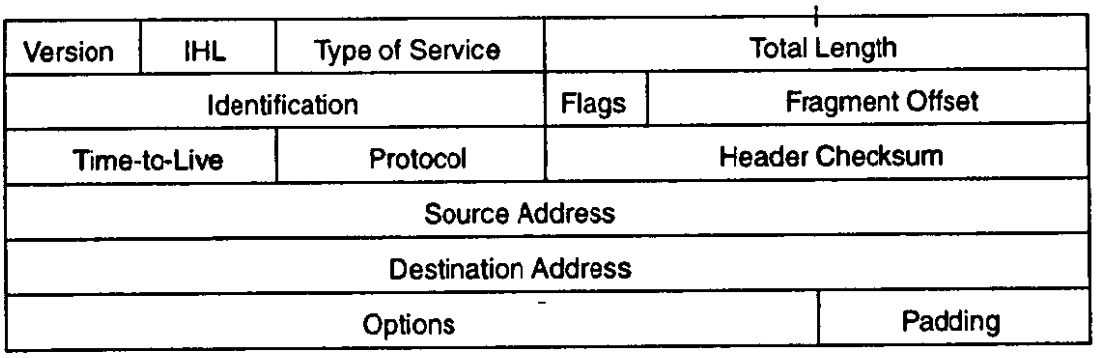


— The IPv6 Header —

Assuming that the reader is already somewhat familiar with "classic IP," we will start the analysis of the new IP, IPv6, with a comparison to the previous version.

2.1.1 A Comparison of Two Headers

The new header is in fact much simpler than that of classic IP. The new version counts only six fields and two addresses, while the old version had 10 fixed header fields, two addresses, and some options.



— The IPv4 Header —

The only field that kept the same meaning and the same position is the version number, which in both cases is encoded in the very first four bits. The original idea was to run IPv4 and IPv6 simultaneously on the same wires, on the same local networks, using the same encapsulations

and the same link drivers. The network program would use the initial version field to determine the packet's processing. If the version code is 4 (0100 in binary), it recognizes an IPv4 packet, while if the code is 6 (0110 in binary), it recognizes an IPv6 packet. This idea was in fact abandoned, or at least scaled down. Whenever possible, IPv4 and IPv6 will be demultiplexed at the media layer. For example, IPv6 packets will be carried over Ethernet with the content type 86DD (hexadecimal), instead of IPv4's 8000.

Six fields were suppressed: the header length, the type of service, the identification, the flags, the fragment offset, and the header checksum. Three fields were renamed, and in some cases, slightly redefined: the length, the protocol type, and the time-to-live. The option mechanism was entirely revised, and two new fields were added: class and flow label.

2.1.2 Simplifications

The IPv4 header was based on the state-of-the-art in 1975. We should not be surprised to learn that, about 20 years later, we know better. We could thus proceed with three major simplifications:

- Assign a fixed format to all headers
- Remove the header checksum
- Remove the hop-by-hop segmentation procedure

IPv6 headers do not contain any optional element. This does not mean that we cannot express options for special-case packets. But, we will see in the next section that this is not achieved with a variable-length *option field* as in IPv4. Instead, *extension headers* are appended after the main header. An obvious consequence is that there is no need in IPv6 for a header length field (IHL).

Removing the header checksum may seem a rather bold move. The main advantage is to diminish the cost of header processing, because there is no need to check and update the checksum at each relay. The obvious risk is that undetected errors may result in misrouted packets. This risk is, however, minimal since most encapsulation procedures include a packet checksum. One finds checksums in the media access control procedures of IEEE-802 networks, in the adaptation layers for

ATM circuits, and in the framing procedures of the Point-to-Point Protocol (PPP) for serial links.

IPv4 included a fragmentation procedure so that senders could send large packets without worrying about the capacities of relays. These large packets could be chopped into adequately sized fragments if needed. The recipients would wait for the arrival of all these segments and reconstitute the packet. But, we learned an important lesson from the experience with transport control protocols: The unit of transmission should also be the unit of control. Suppose that we try to transmit large packets over a network that can carry only small segments. The successful transmission of a packet depends on the successful transmission of each segment. If only one is missing, the whole packet must be transmitted again, resulting in a very inefficient usage of the network.

The rule with IPv6 is that hosts should learn the maximum acceptable segment size through a procedure called *path MTU discovery*. If they try to send larger packets, these packets will simply be rejected by the network. As a consequence, there is no need in IPv6 for the segmentation control fields of IPv4, that is, the packet identification, segmentation control flags, and the fragment offset. IPv6 includes, however, an end-to-end segmentation procedure, which will be described in the next section. Also, all IPv6 networks are supposed to be able to carry a payload of 536 octets according to the 1996 specification. Steve Deering would like to raise this size to 1500 octets in the 1997 version of IPv6. Hosts that do not want to discover or remember the path MTU can simply send small packets.

The last simplification of IPv6 is the removal of the Type Of Service (TOS) field. In IPv4, hosts would set the TOS to indicate preferences for the widest, shortest, cheapest, or safest paths. However, this field was not frequently set by applications. We will see in Chapter 6 how IPv6 provides mechanisms for handling these preferences.

2.1.3 Classic Parameters, Revised

Just like IPv4, the IPv6 header includes indications of the packet length, the time-to-live, and the protocol type. However, the definitions of these fields were revisited in the light of experience.

The total length of IPv4 is replaced by the *payload length* of IPv6. There is a subtle difference because the payload length, by definition, is