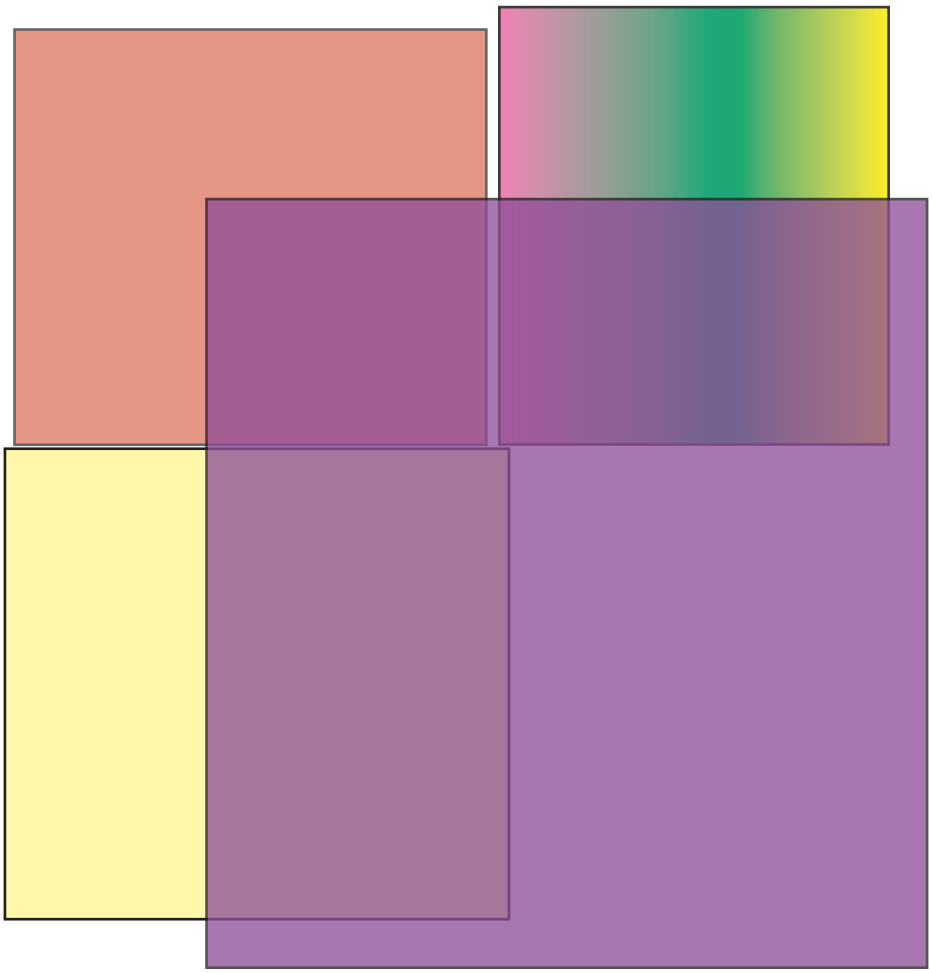
THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG. the quick brown fox jumped over the lazy dog

THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG. the quick brown fox jumped over the lazy dog

THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG. the quick brown fox jumped over the lazy dog

THE QUICK BROWN FOX JUMPED OVER THE LAZY DOG. the quick brown fox jumped over the lazy dog

これは、pdfのサンプルを印刷す



APPENDIX F

Linearized PDF

A Linearized PDF file is a file that has been organized in a special way to enable efficient incremental access in a network environment. The file is valid PDF in all respects, and is compatible with all existing viewers and other PDF applications. Enhanced viewer applications can recognize that a PDF file has been linearized and can take advantage of that organization (as well as added hint information) to enhance viewing performance.

The Linearized PDF file organization is an optional feature available beginning in PDF 1.2. Its primary goal is to achieve the following behavior:

- When a document is opened, display the first page as quickly as possible. The first page to be viewed can be an arbitrary page of the document, not necessarily page 0 (though opening at page 0 is most common).
- When the user requests another page of an open document (for example, by going to the next page or by following a link to an arbitrary page), display that page as quickly as possible.
- When data for a page is delivered over a slow channel, display the page incrementally as it arrives. To the extent possible, display the most useful data first.
- Permit user interaction, such as following a link, to be performed even before the entire page has been received and displayed.

This behavior should be achieved for documents of arbitrary size. The total number of pages in the document should have little or no effect on the user-perceived performance of viewing any particular page.

The primary focus of Linearized PDF is optimized viewing of read-only PDF documents. It is intended that the Linearized PDF be generated once and read many times. Incremental update is still permitted, but the resulting PDF is no

longer linearized and subsequently is treated as ordinary PDF. Linearizing it again may require reprocessing the entire file; see Section F.4.6, "Accessing an Updated File," for details.

Linearized PDF requires two additions to the PDF specification:

- Rules for the ordering of objects in the PDF file
- Additional data structures, called *hint tables*, that enable efficient navigation within the document

Both of these additions are relatively simple to describe; however, using them effectively requires a deeper understanding of their purpose. Consequently, this appendix goes considerably beyond a simple specification of these PDF extensions to include background, motivation, and strategies.

- Section F.1, "Background and Assumptions," provides background information about the properties of the Web that are relevant to the design of Linearized PDF.
- Section F.2, "Linearized PDF Document Structure," specifies the file format and object-ordering requirements of Linearized PDF.
- Section F.3, "Hint Tables," specifies the detailed representation of the hint tables.
- Section F.4, "Access Strategies," outlines strategies for accessing Linearized PDF over a network, which in turn determine the optimal way to organize the PDF file.

The reader is assumed to be familiar with the basic architecture of the Web, including terms such as URL, HTTP, and MIME.

F.1 Background and Assumptions

The principal problem addressed by the Linearized PDF design is the access of PDF documents through the Web. This environment has the following important properties:

- The access protocol (HTTP) is a transaction consisting of a request and a response. The client presents a request in the form of a URL, and the server sends a response consisting of one or more MIME-tagged data blocks.

- After a transaction has completed, obtaining more data requires a new request-response transaction. The connection between client and server does not ordinarily persist beyond the end of a transaction, although some implementations may attempt to cache the open connection to expedite subsequent transactions with the same server.
- Round-trip delay can be significant. A request-response transaction can take up to several seconds, independent of the amount of data requested.
- The data rate may be limited. A typical bottleneck is a slow modem link between the client and the Internet service provider.

These properties are generally shared by other wide-area network architectures besides the Web. Also, CD-ROMs share some of these properties, since they have relatively slow seek times and limited data rates compared to magnetic media. The remainder of this appendix focuses on the Web.

Some additional properties of the HTTP protocol are relevant to the problem of accessing PDF files efficiently. These properties may not all be shared by other protocols or network environments.

- When a PDF file is initially accessed (such as by following a URL hyperlink from some other document), the file type is not known to the client. Therefore, the client initiates a transaction to retrieve the entire document and then inspects the MIME tag of the response as it arrives. Only at that point is the document known to be PDF. Additionally, with a properly configured server environment, the length of the document becomes known at that time.
- The client can abort a response while the transaction is still in progress if it decides that the remainder of the data is not of immediate interest. In HTTP, aborting the transaction requires closing the connection, which interferes with the strategy of caching the open connection between transactions.
- The client can request retrieval of portions of a document by specifying one or more byte ranges (by offset and count) in the HTTP request headers. Each range can be relative to either the beginning or the end of the file. The client can specify as many ranges as it wants in the request, and the response consists of multiple blocks, each properly tagged.
- The client can initiate multiple concurrent transactions in an attempt to obtain multiple responses in parallel. This is commonly done, for instance, to retrieve inline images referenced from an HTML document. This strategy is not

always reliable and may backfire if the transactions interfere with each other by competing for scarce resources in the server or the communication channel.

Note: *Extensive experimentation has determined that having multiple concurrent transactions does not work very well for PDF in some important environments. Therefore, Linearized PDF is designed to enable good performance to be achieved using only one transaction at a time. In particular, this means that the client must have sufficient information to determine the byte ranges for all the objects required to display a given page of the PDF file so that it can specify all those byte ranges in a single request.*

The following additional assumptions are made about the PDF viewer application and its local environment:

- The viewer application has plenty of local temporary storage available. It should rarely need to retrieve a given portion of a PDF document more than once from the server.
- The viewer application is able to display PDF data quickly once it has been received. The performance bottleneck is assumed to be in the transport system (throughput or round-trip delay), not in the processing of data after it arrives.

The consequence of these assumptions is that it may be advantageous for the client to do considerable extra work to minimize delays due to communications. Such work includes maintaining local caches and reordering actions according to when the needed data becomes available.

F.2 Linearized PDF Document Structure

Except as noted below, all elements of a Linearized PDF file are as specified in Section 3.4, "File Structure," and all indirect objects in the file are numbered sequentially in two groups, based on their order of appearance in the file.

- The first group consists of the document catalog, certain other document-level objects, and all objects belonging to the first page of the document. These objects are numbered sequentially, starting at the first object number after the last number of the second group. (The stream containing the hint tables, called a