# Exploring Reconfigurable Technologies for the U.S. National Archives and Records Administration

Craig P. Steffen
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
csteffen@ncsa.uiuc.edu

June 28, 2007

## 1 Introduction

The National Archives and Records Administration (NARA) is tasked with the preservation of the records of the United States federal government. NARA's facility in College Park, Maryland, contains a large portion of all of the records generated by the federal government that must be stored indefinitely. With digital documents becoming ubiquitous, and the volume of the documents growing, the problem of storage, not to mention indexing, is growing at a geometric rate.

NARA has contracted with Lockheed Martin to build the Electronic Records Archive, which will index and store all documents coming into NARA when those documents become all digital starting in the year 2012. NARA is also looking at computing technologies that will improve the efficiency of digital document storage in the Electronic Records archive. The Innovative Systems Laboratory in the National Center for Supercomputing Applications at the University of Illinois has expertise in programming Field Programmable Gate Arrays. The ISL is investigating the possibility of using FPGA technology to assist in the ingest processing of data as it comes into the Electronic Records Archive.

The ISL is experimenting with difference file analysis techniques to enhance data ingest by the electronic records archive. The general thrust of this investigation is to maximize the value of a single pass of analysis as the data is brought into the archive. The plan is to use the parallel processing capabilities of FPGAs to extract as much useful data as possible from each file during the *one* time the data is available before being stored. The idea is to extract the most useful but small meta-data from the incoming files so that the files can be "searched" later without having to take the time to retrieve all the files.

## 2 Text Analysis

The first area of investigation was analyzing files that contain recognizable ASCII text (Word documents and pdf documents, for example). The FPGA has one part of the processor to parse out words as blocks of letters from incoming documents. The next stage of the virtual processor can the analyze the words to store meta-data about the text in the document. One task is to make a short hash of every word in the file and keep a record of all hashes of all words. Multiple files in a collection have their hash lists sorted and correlated so that a hash that occurs for words in multiple files will only be listed once. Such a correlated hash list becomes a shorthand list for all the words in all the files in a group of files, say, a physical volume (a tape, a disk). When searching for files that contain specific words, words that don't appear in that archive can be immediately eliminated if the hash of that words does not appear in the hash list. By creating hash lists that correspond to different groups of files, groups that do not contain the search terms can be again quickly eliminated. The only files that have to be pulled from the archive and actually searched are the files that have matching hashes.

A new text analysis area will be looking at combinations of words in documents to see if there is any way to characterize the "style" of a writer quantitatively. By being able extract a digital writing style fingerprint, documents could be automatically analyzed to characterize which documents were more like another. This could be used to find documents of related style, related subject matter, or the the same or similar authors without having to do tight matches on subject matter.

## 3 Image Analysis

The ISL is also looking at analyzing images files using parallel processing on FPGAs. FPGA processing is particularly suited to image processing because

streams of pixels, and blocks of pixels that encompass a "stencil" can have all analysis done at once, only requiring each piece of the image to be loaded once. This maximizes the use of (typically limited) memory bandwidth.

One goal of image analysis is to collect a small amount of generalized meta-data to be able to determine with just the meta-data whether or not an image could be what is being searched for or not. For instance, if one of the general criteria for image searching is level of color, any image with a high color can be eliminated if you're searching for scans of medical X-rays. The goal of this study is to find and define a very small number of characteristics that will provide great searching utility.

Another thrust in image analysis is to analyze the image for very specific object recognition and record those results as the meta-data about the image. Some criteria would be very general, like is there sky in the picture? Are there human faces? Are the vehicles? Like the generalized questions, these specific object determinations could make searching for images much more efficient in a very large data storage volume. Choosing the criteria (objects to be searched for) will be determined in collaboration with NARA and with an overview of a large cross section of image files from their archives.