Virtualizing the CIC Floppy Disk Project: An Experiment in Preservation Using Emulation

Geoffrey Brown
Indiana University
Department of Computer Science

CIC Floppy Disk Project

GPO ACCESS

FDLP Desktop

University of Chicago

University of Illinois

Indiana University

University of Iowa

University of Michigan

Michigan State University

University of Minnesota

Northwestern University

Ohio State University

Pennsylvania State University

Purdue University

University of Wisconsin-Madison

CIC Floppy Disk Project

The CIC Floppy Disk Project is a partnership between the U.S. Government Printing Office and the Indiana University, Bloomington Libraries on behalf of the Committee on Institutional Cooperation (CIC), making publications that were distributed to federal depository libraries on floppy disk available via FTP over the Internet. The site is designed to provide a central location through which Federal Data, made available on floppy diskettes, can be identified, located and downloaded. While the project provides the information as it was originally presented on the floppy disks, it cannot provide software which might be needed to fully operate the disk. Records provided within the FDP inform the user about the specifications and may refer users to "read me" files. Consultation with local computer administrators may be required. With the ever-increasing number of products being released by the Government Printing Office, **FDP** enables libraries to either fill in gaps to their collections, or provide an immediate access point for patrons. While the collection represents over 200 entries, our collection is not complete. We would welcome contributions from your library's holdings or suggested titles.

A search will retrieve a title list from which individual titles can be "clicked" and down-loaded (click on the file name next to "link.")

1. **When clicking on the file link, you will be prompted to open or save the file.**
2. **Select SAVE and an appropriate location (e.g. floppy disk or hard drive). All files are compressed.**
3. **Once the file is downloaded, double click on it to "explode" the compressed files and directories (the files will be extracted within the directory where the compressed file resides).**

Every effort has been made to maintain the original file/directory structure of the diskettes. Please let us know if you have any problems or suggestions.

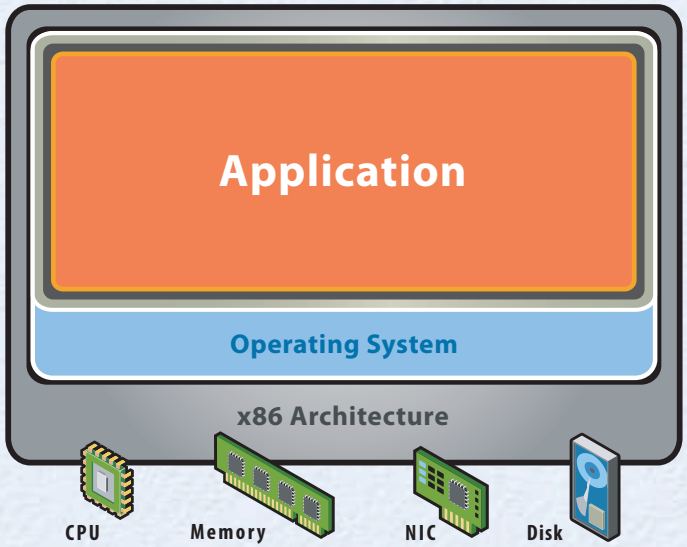
Click here to display titles with full bibliographic detail

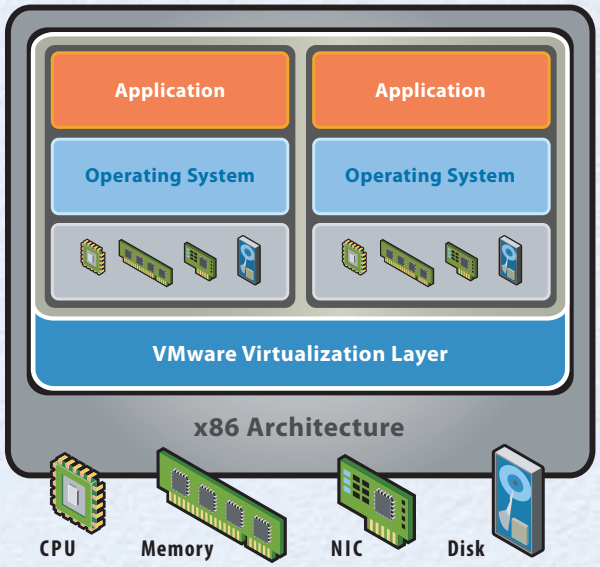
Last updated: Sat Apr 22
Questions and comments: libgpd@indiana.edu

ISSUES WE ARE TRYING TO SOLVE

- Documents in FDP (for example) require obsolete applications and operating systems
- Installing documents to access them requires specialized expertise
- These problems generalize to SUDOC documents on CD-ROM

VIRTUALIZATION





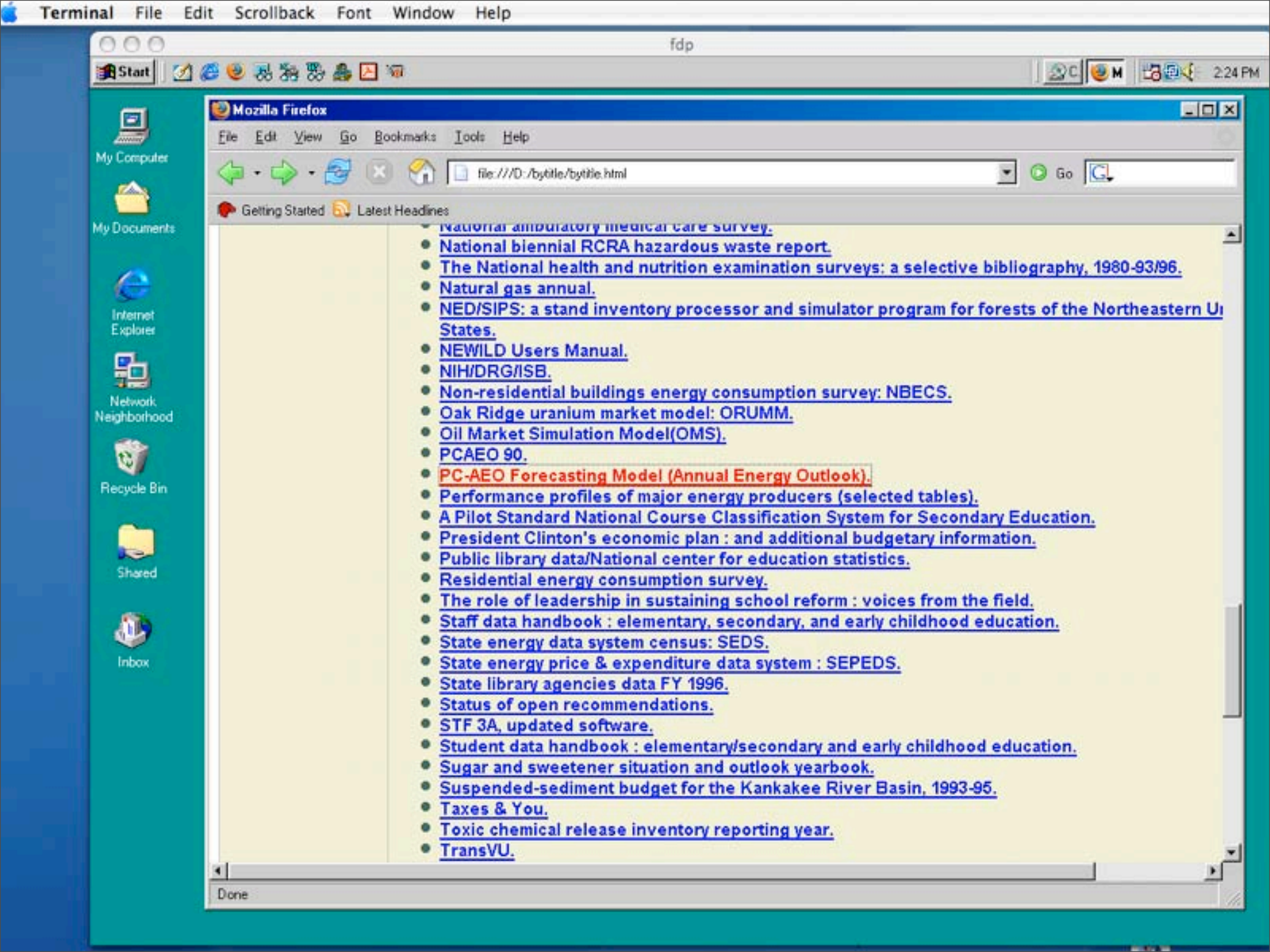
Before Virtualization:

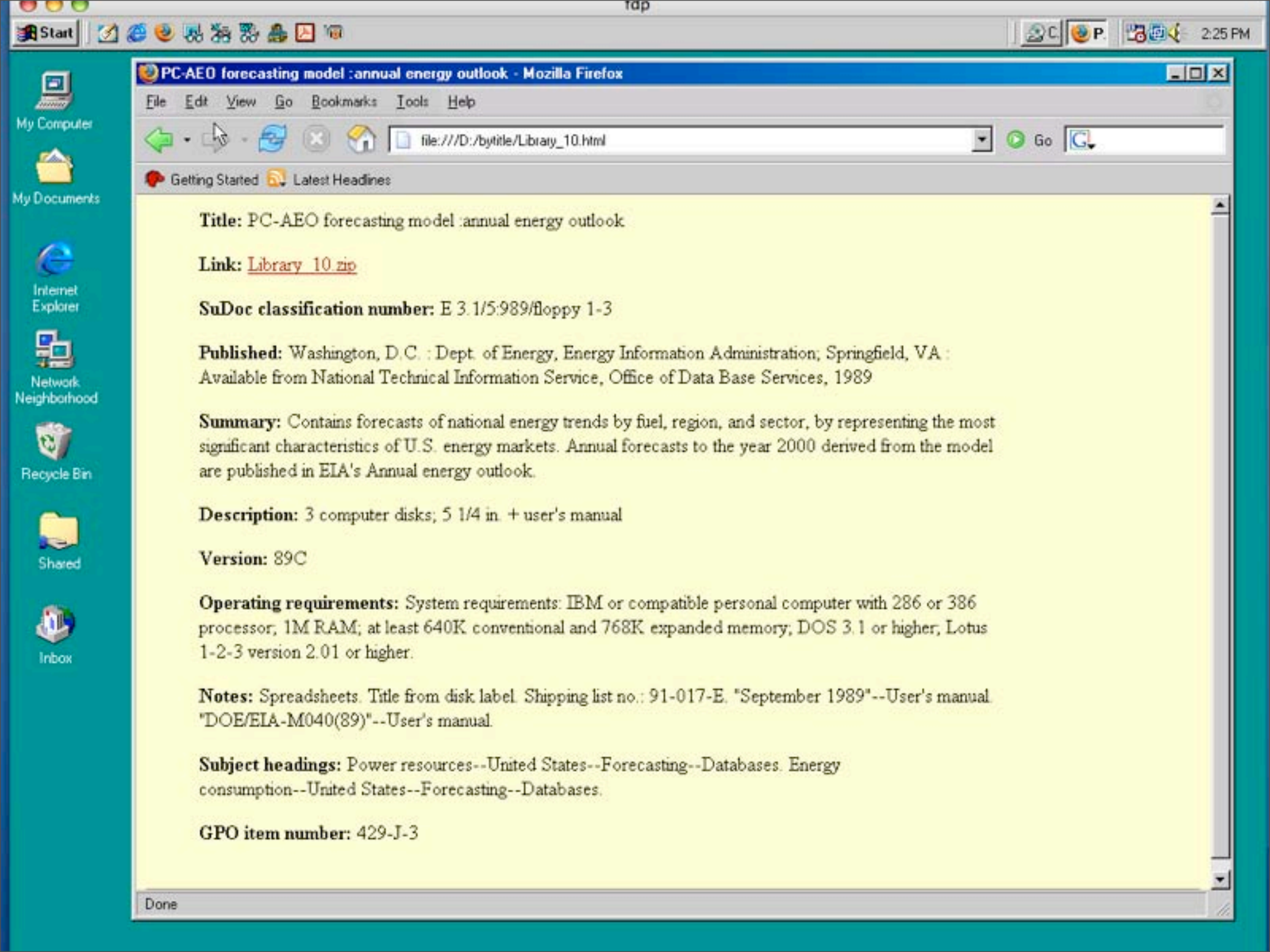
- Single OS image per machine
- Software and hardware tightly coupled
- Running multiple applications on same machine often creates conflict
- Underutilized resources
- Inflexible and costly infrastructure

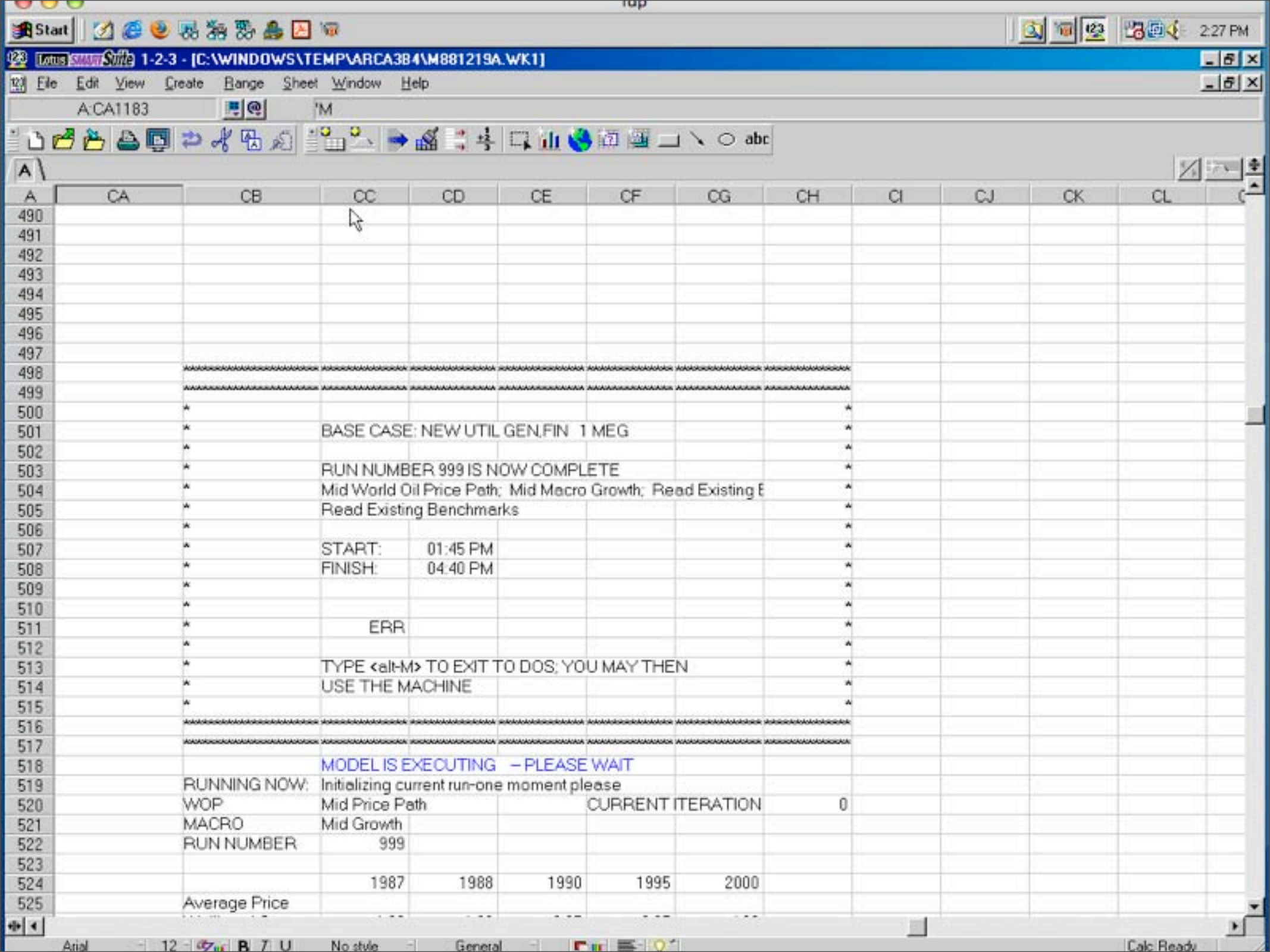
After Virtualization:

- Hardware-independence of operating system and applications
- Virtual machines can be provisioned to any system
- Can manage OS and application as a single unit by encapsulating them into virtual machines

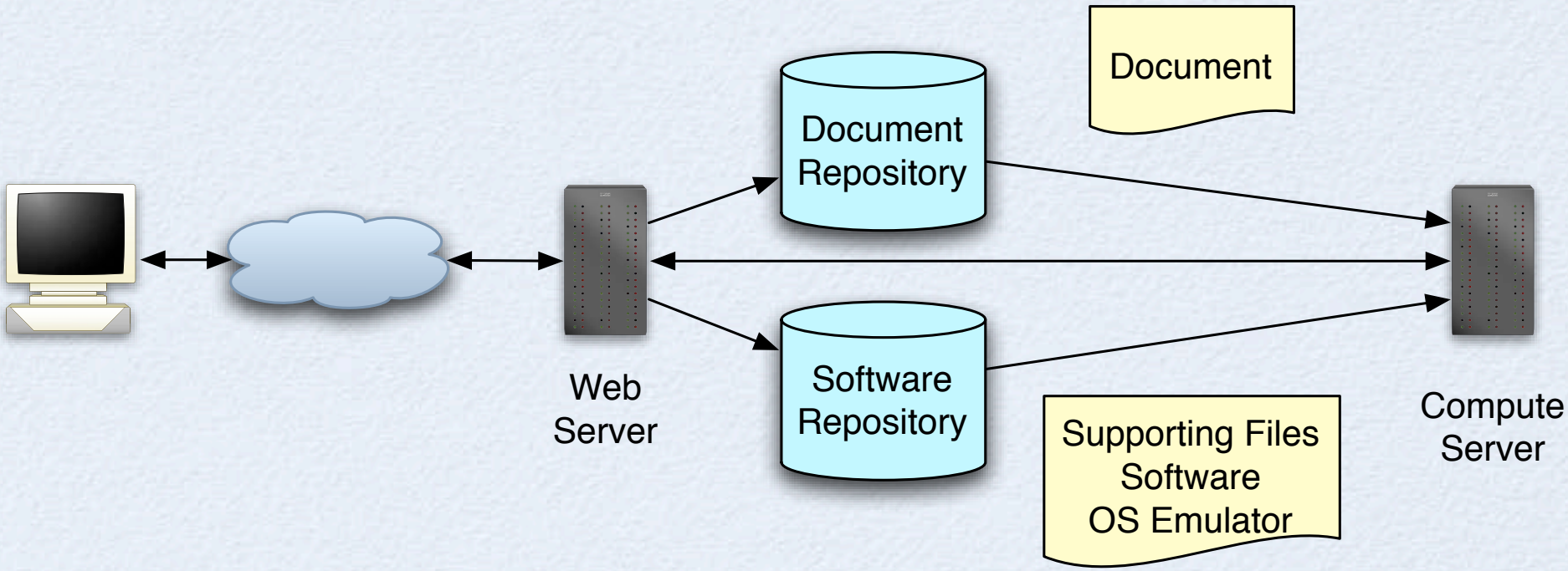
Source: Virtualization Overview, Copyright VMware


Terminal File Edit Scrollback Font Window Help
fdp
Start
2:24 PM
My Computer
My Documents
Internet Explorer
Network Neighborhood
Recycle Bin
Shared
Inbox
Mozilla Firefox
File Edit View Go Bookmarks Tools Help
file:///D:/bytitle/bytitle.html
Go
Getting Started
Latest Headlines
National ambulatory medical care survey.
National biennial RCRA hazardous waste report.
The National health and nutrition examination surveys: a selective bibliography, 1980-93/96.
Natural gas annual.
NED/SIPS: a stand inventory processor and simulator program for forests of the Northeastern U
States.
NEWILD Users Manual.
NIH/DRG/ISB.
Non-residential buildings energy consumption survey: NBECS.
Oak Ridge uranium market model: ORUMM.
Oil Market Simulation Model(OMS).
PCAEO 90.
PC-AEO Forecasting Model (Annual Energy Outlook).
Performance profiles of major energy producers (selected tables).
A Pilot Standard National Course Classification System for Secondary Education.
President Clinton's economic plan : and additional budgetary information.
Public library data/National center for education statistics.
Residential energy consumption survey.
The role of leadership in sustaining school reform : voices from the field.
Staff data handbook : elementary, secondary, and early childhood education.
State energy data system census: SEDS.
State energy price & expenditure data system : SEPEDS.
State library agencies data FY 1996.
Status of open recommendations.
STF 3A, updated software.
Student data handbook : elementary/secondary and early childhood education.
Sugar and sweetener situation and outlook yearbook.
Suspended-sediment budget for the Kankakee River Basin, 1993-95.
Taxes & You.
Toxic chemical release inventory reporting year.
TransVU.
Done


fdp
Start
2:25 PM
My Computer
My Documents
Internet Explorer
Network Neighborhood
Recycle Bin
Shared
Inbox
PC-AEO forecasting model :annual energy outlook - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
file:///D:/bytitle/Library_10.html
Go
Getting Started
Latest Headlines
Title: PC-AEO forecasting model :annual energy outlook
Link: Library_10.zip
SuDoc classification number: E 3.1/5:989/floppy 1-3
Published: Washington, D.C. : Dept. of Energy, Energy Information Administration; Springfield, VA : Available from National Technical Information Service, Office of Data Base Services, 1989
Summary: Contains forecasts of national energy trends by fuel, region, and sector, by representing the most significant characteristics of U.S. energy markets. Annual forecasts to the year 2000 derived from the model are published in EIA's Annual energy outlook.
Description: 3 computer disks; 5 1/4 in. + user's manual
Version: 89C
Operating requirements: System requirements: IBM or compatible personal computer with 286 or 386 processor; 1M RAM; at least 640K conventional and 768K expanded memory; DOS 3.1 or higher; Lotus 1-2-3 version 2.01 or higher.
Notes: Spreadsheets. Title from disk label. Shipping list no.: 91-017-E. "September 1989"--User's manual. "DOE/EIA-M040(89)"--User's manual.
Subject headings: Power resources--United States--Forecasting--Databases. Energy consumption--United States--Forecasting--Databases.
GPO item number: 429-J-3
Done


Start
2:27 PM
1-2-3 - [C:\WINDOWS\TEMP\ARCA3B4\M881219A.WK1]
File Edit View Create Range Sheet Window Help
A:CA1183
'M
A:
CA CB CC CD CE CF CG CH CI CJ CK CL
490
491
492
493
494
495
496
497
498 ********** ********** ********** ********** ********** ********** **********
499 ********** ********** ********** ********** ********** ********** **********
500 * *
501 * BASE CASE: NEW UTIL GEN,FIN 1 MEG *
502 * *
503 * RUN NUMBER 999 IS NOW COMPLETE *
504 * Mid World Oil Price Path; Mid Macro Growth; Read Existing E *
505 * Read Existing Benchmarks *
506 * *
507 * START: 01:45 PM *
508 * FINISH: 04:40 PM *
509 * *
510 * *
511 * ERR *
512 * *
513 * TYPE <alt-M> TO EXIT TO DOS; YOU MAY THEN *
514 * USE THE MACHINE *
515 * *
516 ********** ********** ********** ********** ********** ********** **********
517 ********** ********** ********** ********** ********** ********** **********
518 MODEL IS EXECUTING – PLEASE WAIT
519 RUNNING NOW: Initializing current run-one moment please
520 WOP Mid Price Path CURRENT ITERATION 0
521 MACRO Mid Growth
522 RUN NUMBER 999
523
524 1987 1988 1990 1995 2000
525 Average Price
Arial 12 B I U No style General
Calc Ready

MODEL ARCHITECTURE



Actions in response to patron request

- A pre-configured emulator is allocated
- Emulator is customized
  - Document file system mounted
  - Document specific installation executed
  - Shared file directories created for patron use
- Link to emulator and web accessible file system provided through patron browser
- Emulator executes remotely under patron control

Preparation for Virtualization

- Analyze software requirements of document collections
- Build software images (OS, applications)
- Build and test customization scripts

EXAMPLE -- FLOPPY DISK REQUIREMENTS

```
Library_2.zip: Zip archive data, at least v2.0 to extract
        INSTALL.EXE: MS-DOS executable (EXE)
        INSTALL.DAT: ASCII C program text, with CRLF line terminators
        DISK.ID: ASCII text, with CRLF line terminators
        DDB.001: data
Library_4.zip: Zip archive data, at least v2.0 to extract
        UMINSTR.DOS: (Corel/WP)
        UMINSTR.WIN: (Corel/WP)
        UMWP51.DOS: (Corel/WP)
        UMWP61.WIN: (Corel/WP)
        VOCED.EXE: MS-DOS executable (EXE), PKLITE compressed
                vl_help.dbf: DBase 3 data file (209 records)
                vl.exe: MS-DOS executable (EXE)
                vl_descr.ndx: DBase 3 index file
                vl_keycd.ndx: DBase 3 data file (12 records)
                vl_keywd.dbf: DBase 3 data file with memo(s) (219 records)
                vl_keywd.dbt: data
                vl_keywd.ndx: data
                vl_opmen.dbf: DBase 3 data file (3 records)
                vl_projc.dbf: DBase 3 data file with memo(s) (22 records)
                vl_projc.dbt: data
                vl_rcstr.dbf: DBase 3 data file with memo(s) (no records)
                vl_rcstr.dbt: data
                vl_rwcol.dbf: DBase 3 data file with memo(s) (57 records)
                vl_rwcol.dbt: data
                vl_tdesc.dbf: DBase 3 data file with memo(s) (21 records)
                vl_tdesc.dbt: data
                p016n.dbf: DBase 3 data file (900 records)
                p016r.dbf: DBase 3 data file (115 records)
                p017c.dbf: DBase 3 data file (80 records)
                p017n.dbf: DBase 3 data file (270 records)
                p017r.dbf: DBase 3 data file (30 records)
                p018c.dbf: DBase 3 data file (34 records)
                p018n.dbf: DBase 3 data file (45 records)
                p018r.dbf: DBase 3 data file (43 records)
                p019c.dbf: DBase 3 data file (78 records)
                p019n.dbf: DBase 3 data file (232 records)
                p019r.dbf: DBase 3 data file (51 records)
```

UNIX FILE

FILE(1) FILE(1)

NAME

file – determine file type

SYNOPSIS

file [**–bciknsvzL**] [**–f** *namefile*] [**–m** *magicfiles*] *file*
file -C [**–m** magicfile]

DESCRIPTION

This manual page documents version 3.39 of the **file** command.

File tests each argument in an attempt to classify it. There are three sets of tests, performed in this order: filesystem tests, magic number tests, and language tests. The *first* test that succeeds causes the file type to be printed.

UNIX FILE --EXAMPLE

```
-bash-2.05b$ file ddd.JPG  Debug.pdf  lab2.tex print.c


ddd.JPG:   JPEG image data, JFIF standard 1.01,
resolution (DPI), 96 x 96
Debug.pdf: PDF document, version 1.2
lab2.tex:  LaTeX 2e document text
print.c:   ASCII C program text
```

PERL SCRIPT TO AUTOMATE

```
sub fileinfo()
{
    my @allfiles;
    my $file;
    my $f = $_[0];
    my $d = cwd();
    my $tmpdir;

    printf "@tabs";
    open FH, "file \"$f\"|";
    my $info = <FH>;
    close FH;
    printf "$info";


    if (-d $f)  # recurse
    {
        $tmpdir = $f;
    }
    elsif ($info =~ m/Zip/)
    {
        $tmpdir = tempdir( CLEANUP => 1 );
        system("unzip $f -d $tmpdir > /dev/null");
    }
    elsif ($info =~ m/ARC archive/)
    {
        $tmpdir = tempdir( CLEANUP => 1 );
        pushdir($tmpdir);
        system("arc x \"$d/$f\" > /dev/null");
```

INTERESTING RULE

```
elsif ($info =~ m/SFX|PKLITE/)
{
    $tmpdir = tempdir( CLEANUP => 1 );
    pushdir($tmpdir);
    system("cp \"$d/$f\" .");
    system("dosemu –dumb $f >> /nfs/troy/home/digarchive/doserrs");
    system("rm $f");
    popdir();
}
```

TOP 20 FILE TYPES

```
2323  data (ASCII)
 858  Lotus 1-2-3 wk1 document data
 851  ASCII text, with CRLF line terminators
 551  data (binary)
 532  ASCII English text, with CRLF line terminators
 508  DBase 3 data file
 361  Zip archive data, at least v2.0 to extract
 228  binary Computer Graphics Metafile
 206  MS-DOS executable, MZ for MS-DOS
 166  Zip archive data, at least v1.0 to extract
 166  MS Compress archive data
 119  VMS Alpha executable
 118  directory
 101  MS-DOS executable, NE for MS Windows 3.x (driver)
  86  Lotus 1-2-3 wk3 document data
  82  MS-DOS executable, NE for MS Windows 3.x
  78  ARC archive data, dynamic LZW
  64  Lotus 1-2-3
  57  (Corel/WP)
  47  Non-ISO extended-ASCII English text, with CRLF line terminators
```

SOFTWARE REQUIRED FOR FDP

- Windows 98 (most disks were for msdos, win 3.1)
- DBase III -- (dbfviewer2000)
- WordPerfect
- Lotus 1-2-3 (smartsuite, smartsuite viewer)
- Microsoft Word (we use free msword viewer)
- Various Archive Tools
- Browser (we use Firefox)
- Generic Postscript Printer Driver
- Software we didn't install -- pascal, fortran, sas, arcinfo

PRACTICAL ISSUES

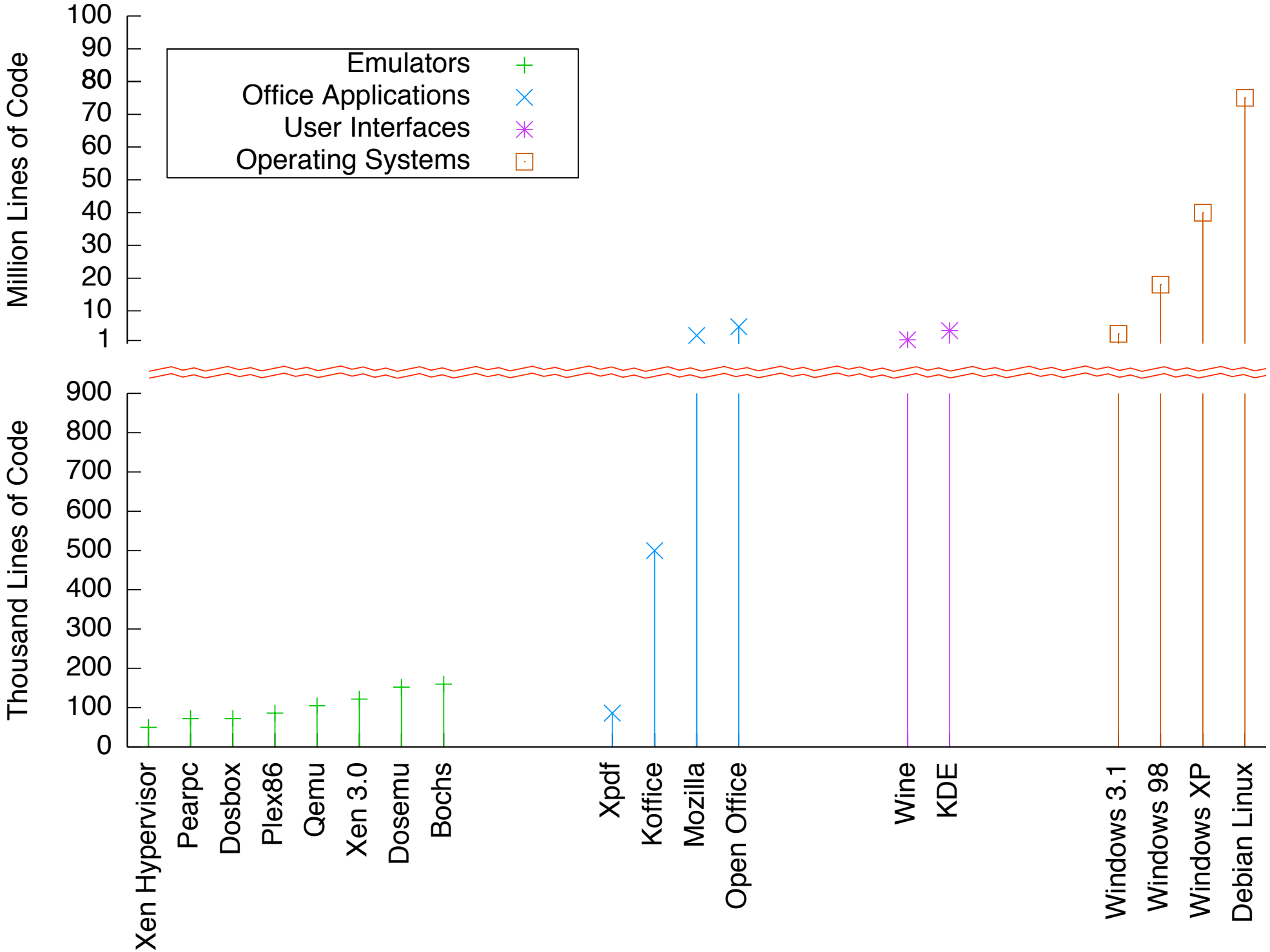
- Printing -- we print postscript to files
- File transfer -- windows sharing from guest to host
- Security

What we haven't done

- Adequate Testing
- Customization Scripts
- Web Delivery

PHILOSOPHICAL ISSUES

- Who Preserves the Emulator ?
- Why Not Just Migration ?


Million Lines of Code
100
90
80
70
60
50
40
30
20
10
1
Emulators
Office Applications
User Interfaces
Operating Systems
Thousand Lines of Code
900
800
700
600
500
400
300
200
100
0
Xen Hypervisor
Pearpc
Dosbox
Plex86
Qemu
Xen 3.0
Dosemu
Bochs
Xpdf
Koffice
Mozilla
Open Office
Wine
KDE
Windows 3.1
Windows 98
Windows XP
Debian Linux

WHY NOT JUST MIGRATION

- Loss of information -- e.g. word edits
- Loss of fidelity -- e.g. WordPerfect to Word isn't very good
- Loss of authenticity -- users of migrated document need access to original to verify authenticity
- Not always possible -- closed proprietary formats
- Not always feasible -- costs may be too high
- Emulation may necessary to enable migration

News Front Page
World
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Education
Science/Nature
Technology
Entertainment
Have Your Say
Magazine
In Pictures

Last Updated: Monday, 2 May, 2005, 17:18 GMT 18:18 UK

E-mail this to a friend Printable version

Readers 'declassify' US document

When news started circulating in Italy that a heavily censored Pentagon report into the death of secret agent Nicola Calipari had been decrypted, many thought it must be the work of some top-notch hacker.


Someone found a simple cut-and-paste job could restore the text

In fact, it turned out that the classified document, containing top-secret details - such as the name of the soldier who fired the deadly rounds of ammunition - could be made readable with two simple clicks of your computer mouse.

A few hours after the Pentagon published the report on its website, a few Italian readers found they could make the blacked-out paragraphs reappear by cutting and pasting them from the site into a Word document.

Salvatore Schifani, a 30-year-old IT worker, spotted the document at about 0300 local time (0100 GMT) on Saturday night.

BBC NEWS:VIDEO AND AUDIO
How the censored parts of the report were made public
VIDEO

THE STRUGGLE FOR IRAQ
KEY STORIES
- New Iraqi plan to curb violence
- Baghdad authorities lift curfew
- Baghdad mosque attack kills 10
- Iraq war fuels terror - US report
- Saddam thrown out for third time

BACKGROUND AND ANALYSIS

Tide of violence
Complete US victory in Iraq may be out of the question, argues John Simpson.

- Saddam Hussein on trial
- Iraq's federal fracas
- War justifications laid bare
- War dead figures
- Saddam trial one timeline
- Saddam trial two timeline

Digital Media

When Words Come Back From The Dead

David M. Ewalt, 12.13.05, 5:00 PM ET

By This Author

David M. Ewalt

- The Hartford CIO's Wireless LCD Television
- Trust

NEW YORK -In the year since pharmaceutical giant **Merck** withdrew its arthritis drug Vioxx from shelves, the company has been hit with 7,000 personal injury lawsuits--one of which already cost the company $253 million. But it could be a frequently misused feature of **Microsoft** Word that turns out to be the straw that broke Merck's back.

SUDOC VIRTUALIZATION PROJECT

- Approximately 2500 SUDOC CD-ROMs in IU Library
- We've done preliminary analysis on 150 CD-ROMs
- We are currently creating electronic copies of approximately 1000 CD-ROMs (skipping A, C3.278/2, C3.278/3, C21/5/6, E, HE, N, T, X, Y)

GOALS

- Deliver key SUDOC collections through virtualization
  - Develop web delivery techniques
  - Improve our software analysis tools
  - Develop image customization techniques (e.g. perform software install on the fly)

How this work might be used

- Libraries share pool of software images and licenses
- Libraries share expertise in supporting various document collections
- Libraries collaborate to provide redundancy
- Patrons access from anywhere without needing to obtain or install special software

How you can help

- Statistics on SUDOC usage
- Collaborate on building tools/infrastructure

geobrown @ cs . indiana . edu
(812) 855-4207

Acknowledgments

- Lou Malcomb (IU Head GIMSS)
- Julianne Bobay (IU Head SLIS Library)

INDIANA UNIVERSITY LIBRARIES BLOOMINGTON