Digital Vellum

Vint Cerf
Google

November 2017

Archiving Static Content

PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For *Anno* 1665, and 1666.

In the *SAVOY*,
Printed by *T. N.* for *John Martyn* at the Bell, a little without *Temple-Bar*, and *James Allestry* in *Duck-Lane*,
Printers to the *Royal Society*.

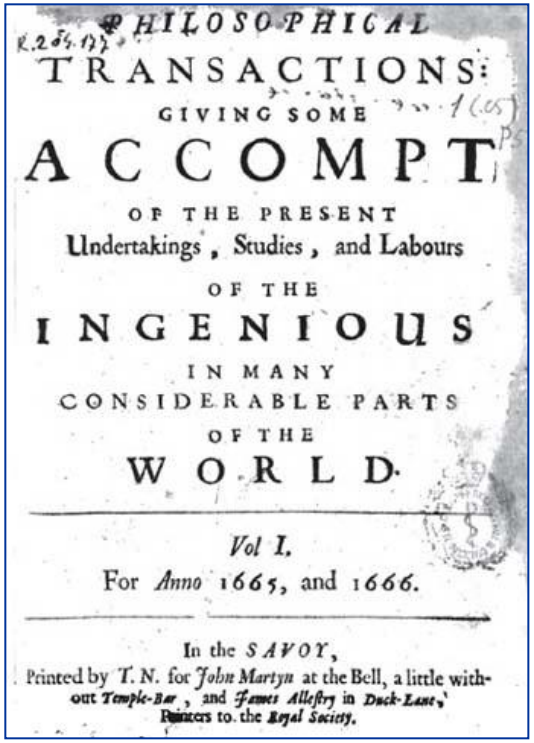

EC5274 5B/5D
ИЗОМ
СДЕЛАНО В БОЛГАРИИ
boedem
FLOPPY DISK
5,25" HD,DS
WRITE ENABLE
WRITE PROTECT


BUYDIG.COM


TDK


DVD
BluRayDisc

Archiving Static Text/Image Content


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

YouTube

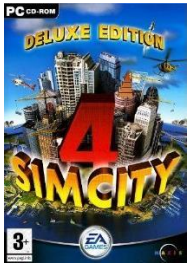
flickr

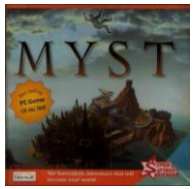
Picasa

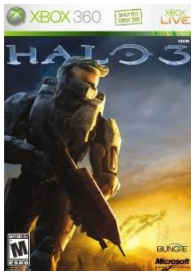
22nd Century
Doris Kearns Goodwin

- A Team of Rivals (Lincoln)
  - How did she reconstruct the dialog??
  - 100 Libraries and repositories w/physical correspondence
- What will the 22nd C. Doris Kearns Goodwin find?
- What will the National Archives be able to offer?
- What will our descendants know of our 21st Century?
  - Correspondence, entertainment, advertising, education, jobs, family life,…

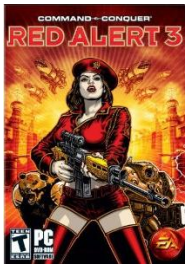
What About Executable Content?



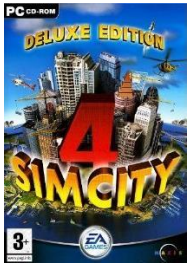


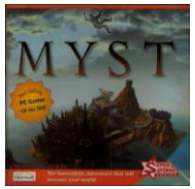


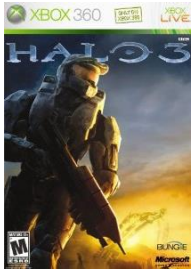
Games



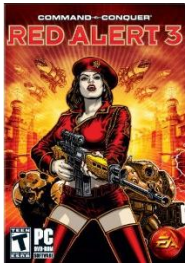
What About Executable Content?



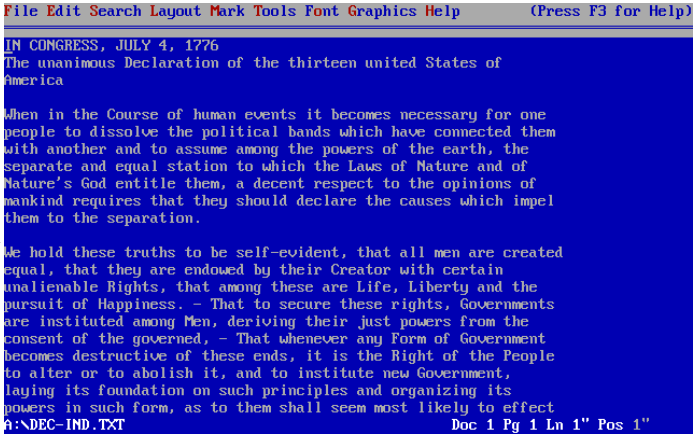


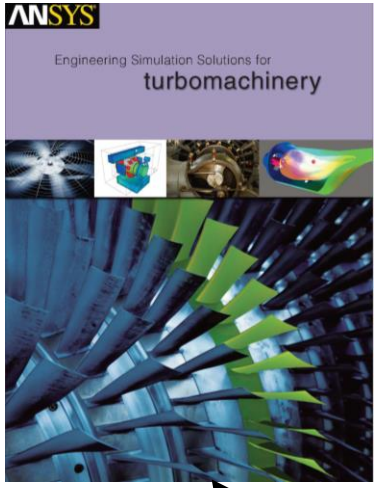


Games

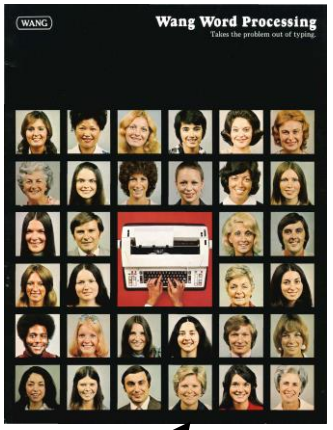


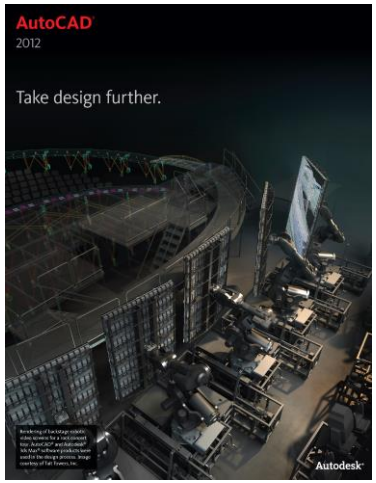
WordPerfect 1.0 doc
Can you read it today?
100 years from now?





Application-specific content





Simulation model
Can you re-run old model with new data?

Original Wang doc
Can you read it today?
100 years from now?

Challenges

- Interpretation of bits
- Metadata capture
- Source or executable code
- “Digital X-ray”
- Capacity for BIG DATA
- Bankruptcies, sunsetting of apps, OS, hardware
- Intellectual Property Rights
- Legal frameworks, exceptions for preservation

The OLIVE Project

- Carnegie-Mellon University
- Mahadev Satyanarayanan (“Satya”)
- NSF funded project on digital preservation

Execution Fidelity

Ability to precisely reproduce execution

Many moving parts

- **hardware**
- **operating system**
- **dynamically linked libraries**
- **configuration parameters**
- **language settings**
- **time zone settings**
- **…**

Inspiration: "Digital X-Ray" of the hardware and operating software

Very difficult to achieve and then maintain

Transform into a Scaling Problem

Pack up and carry the entire environment with you

- including the OS
- transitive closure of everything you need

Central idea of a (hardware) *virtual machine (VM)*

But VMs are huge

- many GB to tens of GB
- waiting to download → long launch delay
- inspiration from YouTube: *stream* instead of downloading

VM Streaming Not So Easy

Access to VM image is not linear

Reference pattern depends on many runtime factors

- **data dependencies**
- **human interaction**
- **spatial and temporal locality (program behavior)**

Our approach

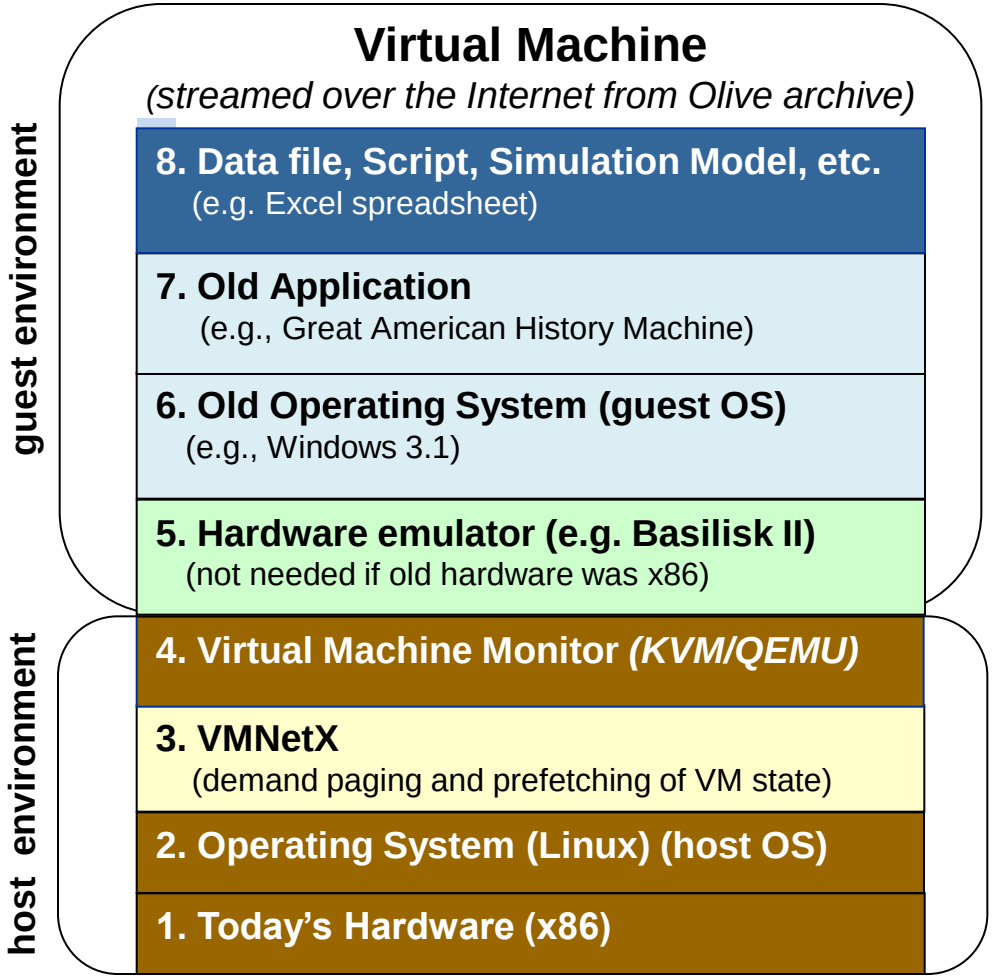
- ***demand paging***

  intercept missing VM pieces and fetch over Internet

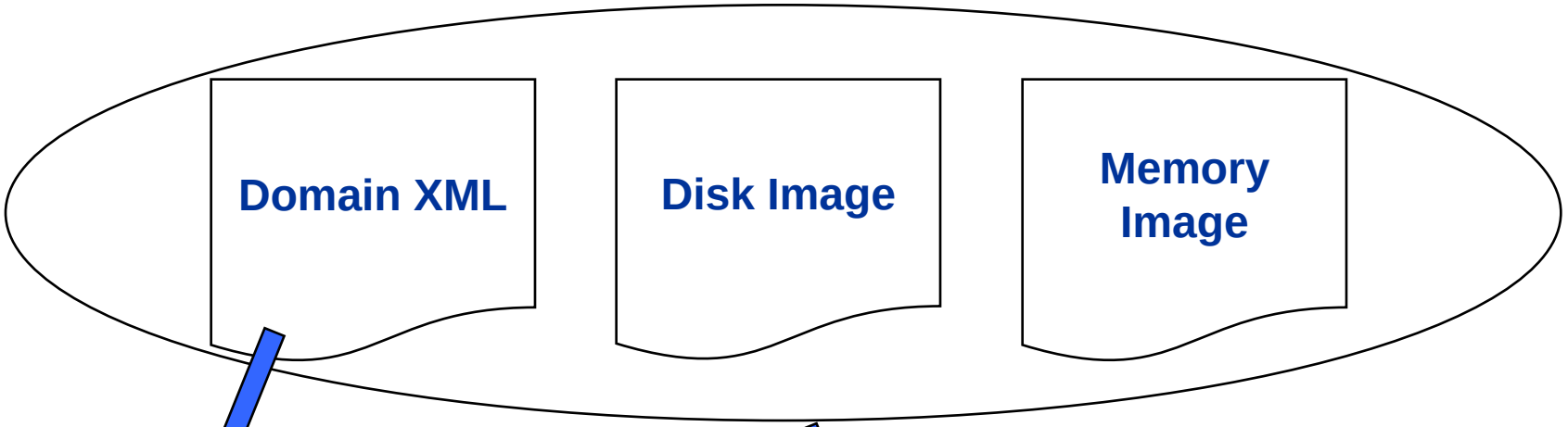
- ***prefetching***

  mask stalls due to demand misses
  (if hints are good)

Client Structure



VM Image Representation

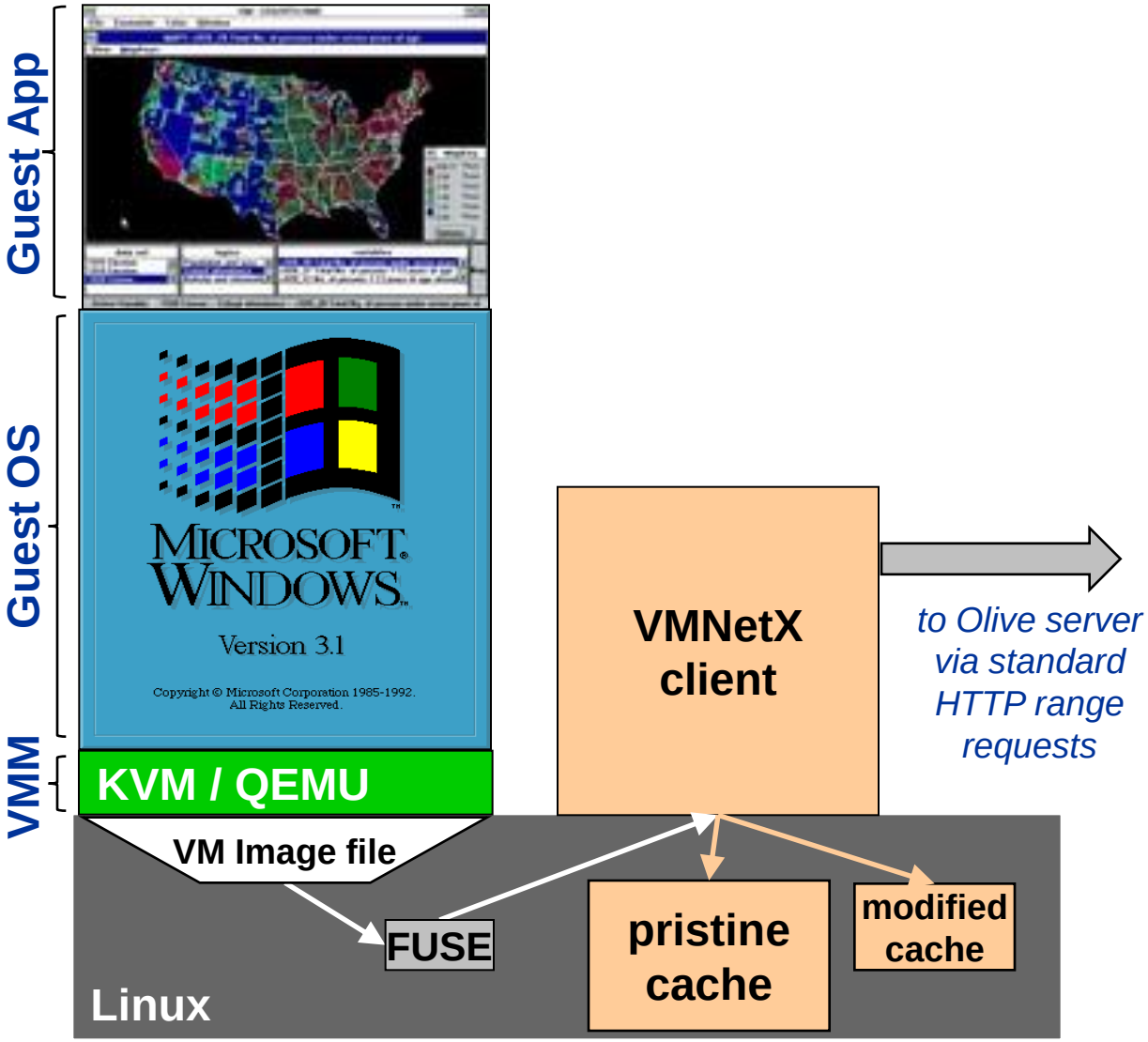


Machine details

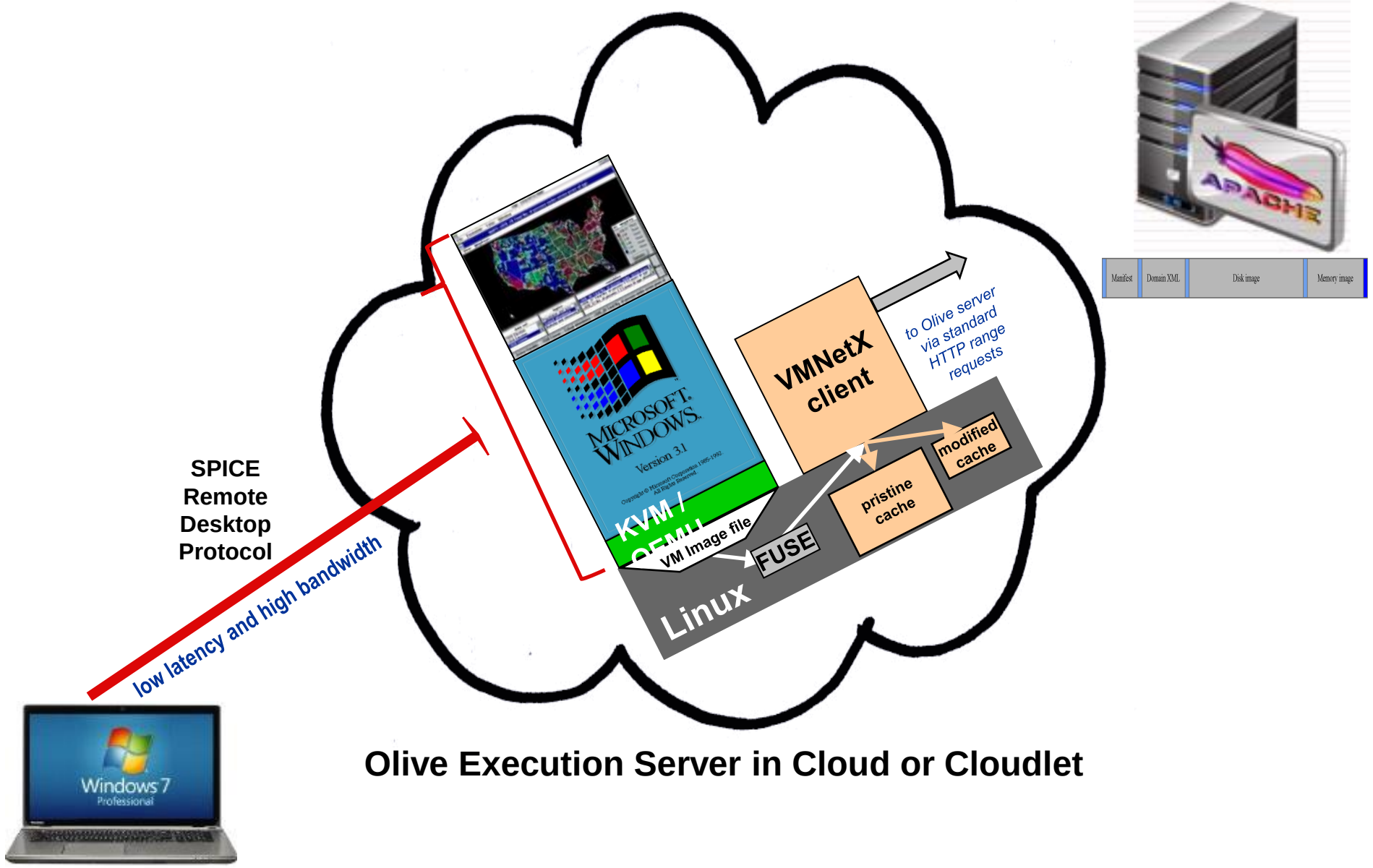
```
<domain type="kvm">
  <name>machine</name>
  <uuid>a7434757-631b-496d-a1ba-638014c74cc4</uuid>
  <memory>65536</memory>
  <currentMemory>65536</currentMemory>
  <vcpu>1</vcpu>
  <os>
    <type arch="i686" machine="pc">hvm</type>
    <boot dev="hd"/>
  </os>
  <features>
    <pae/>
  </features>
  <clock offset="utc"/>
  <devices>
    <emulator>/usr/libexec/qemu-kvm</emulator>
    <disk type="file" device="disk">
      <driver name="qemu" type="raw"/>
      <source file="/disk.img"/>
      <target dev="hda" bus="ide"/>
      <address type="drive" controller="0" bus="0" unit="0"/>
    </disk>
    <controller type="ide" index="0">
      <address type="pci" domain="0x0000" bus="0x00" slot="0x01" function="0x1"/>
    </controller>
    <interface type="user">
      <mac address="52:54:00:03:a0:11"/>
      <address type="pci" domain="0x0000" bus="0x00" slot="0x03" function="0x0"/>
    </interface>
    <input type="mouse" bus="ps2"/>
    <graphics type="vnc" autoport="yes"/>
    <video>
      <model type="vga" vram="9216" heads="1"/>
      <address type="pci" domain="0x0000" bus="0x00" slot="0x02" function="0x0"/>
    </video>
  </devices>
</domain>
```

Single file representation

Olive Implementation



Cloud Execution of Olive



Olive Execution Server in Cloud or Cloudlet

Many Future Technical Challenges

We are a long way from being "done"!

Scaling and performance issues

- VMs keep getting bigger, networks are never fast enough
- clever prefetching techniques

Precise emulation of hardware

- even x86 extended memory modes not quite right in QEMU (can't boot Windows 95 in KVM/QEMU)
- exotic hardware platforms
- host compatibility (e.g. CPU flags in x86) vs performance
- hardware performance accelerators (e.g. GPUs)

Multi-VM ensembles (e.g. HPC environments)

Tools for easy building of VMs (physical to virtual?)

Archiving entire cloud services

… *many others* …

Scope of Digital Preservation

- Digital object structures, representations, vocabulary and standard terminology (schema, OWL, …)
- Identifier spaces, registries, resolution mechanisms
  - The irony of WWW, URLs, DNS (TBL was at CERN)
  - Robert Kahn: Digital Object Architecture, CNRI
- Standard, rigorous ingestion processes
- Metadata (about the data, provenance, authenticity, calibration, ...)
- Legal frameworks for preservation (copyright, patents, licensing, special treatment for perserving bodies)
- Business Models for extended, long term operation

Milestones

- Technical means to capture and update digital storage media
- Capture and representation of relevant metadata
- Clearance of rights to share/execute digital objects
  - Possible legislation granting archives/libraries special “preservation” rights?
  - Might include both copyright and patent priviliges
- Provision for assuring integrity of digital objects
- Monitoring and management of changes to rights (e.g. expiration of copyright, patent)
- Development of business model(s) to sustain long-term preservation and access
  - Libraries, Archives, Universities, Museums
  - Long-lived institutions as vehicles or models?
    - E.g. Breweries, vineyards, Catholic (and other) Churches, Banks…. (!)
- Personalization of preservation options accessible to the general public

Other Projects

- The Internet Archive – Brewster Kahle et al
  - Library of Alexandria backup among others
  - Digital content, books, software
- The Computer History Museum
  - Software and computing artifacts
- Google Book Scans and Cultural Institute
- Digital Object Architecture and Identifiers (CNRI)

More Projects

- **RHIZOME, University of Freiburg**
- **Interplanetary File System (IPFS)**
- **International Internet Preservation Consortium**
- **UK Depositary Libraries Program**