Carnegie Mellon University

The Olive Executable Archive

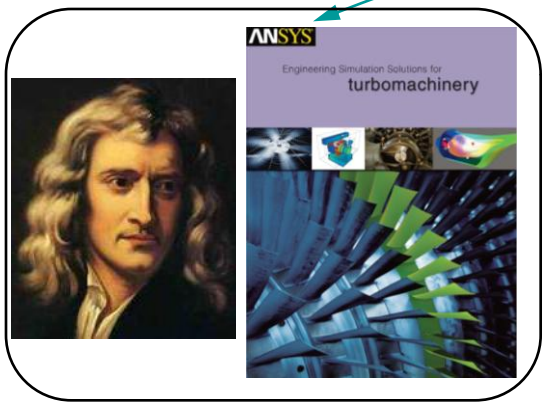
Presenter
Daniel Ryan
Curator of Executable Content

Transforming the Role of Libraries

Library of Alexandria

- Ability to archive static content transformed human history
- Ability to archive executable content will also transform human collaboration
- VM technology will play a starring role

reaching back in time



Isaac's archived VM image

I wonder what Isaac's model would say about this new data?

Olive: What and Why?

Olive seeks to close a gap in the ability of libraries and cultural memory institutions to meet the preservation requirements of constituents producing executable content.

- There is a high demand for access to and execution of early computer applications and games
- Preservation of data or code alone is insufficient – users must be able to actually manipulate this content in its original environment
- Offers precise replication of original executables (execution fidelity)

Think of Olive as YouTube for software. With the click of a button, users can interact with a piece of archived software as if it were new.

Funding

Sloan Foundation Grant – January 2013.

- Award: $400k over 2 years
- Goal: To develop the technical framework for Olive which supports the long term preservation of executable content; To plan for an effective organizational structure to sustain the archive and provide access to executable content

IMLS Leadership Grant – October 2012.

- Award: $497k over 2 years
- Goal: To understand what types of content can be ingested into Olive and to determine how an executable content archive can fit into existing trusted repository standards such as OAIS, OCLC, CRL and JISC

INSTITUTE of Museum and Library SERVICES

Software or Stone Tablet?

Unlike books, letters, and other traditional media, software is designed to be interpreted and executed by computers. It comes to life when it is executed, and this depends upon a number of things:

- Source code (provides combinations of possibilities, like DNA)
- Execution environment (operating system, runtime libraries, versions of libraries called during execution, hardware)
- User interaction (clicks, keystrokes, speed of input, etc)
- Data dependence (inputs to the software)

Each of these pieces is required to successfully reproduce software as experienced by its contemporary users. Without the correct execution environment, and without user interaction, the constituent bits of a program are as dead as stone tablets.

Legal Challenges

CMU's General Counsel Mary Jo Dively contracted with K&L Gates to assess the risks to CMU.

- In many circumstances, a carefully constructed and managed system will support legal use of legacy software (10 years old)
- We continue to study the opinion provided by K&L Gates

Sustainability Models

We commissioned Deanna Marcum at Ithaka S+R to provide a whitepaper on sustainability:

- Ithaka recommends a 3 year pilot/stress testing phase as the next step
- Partners in this pilot might become long term sustainers
- A recommended mechanism for sustainability would be a consortium of library related organizations

Demo

Execution Fidelity

Ability to precisely reproduce execution

- many moving parts

  hardware, operating system, dynamically linked libraries, configuration parameters, language settings, timezone settings, …

- very difficult to achieve execution fidelity

Solution: Take It With You

Transform the problem 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 **virtual machine (VM)**

But VMs are Huge!

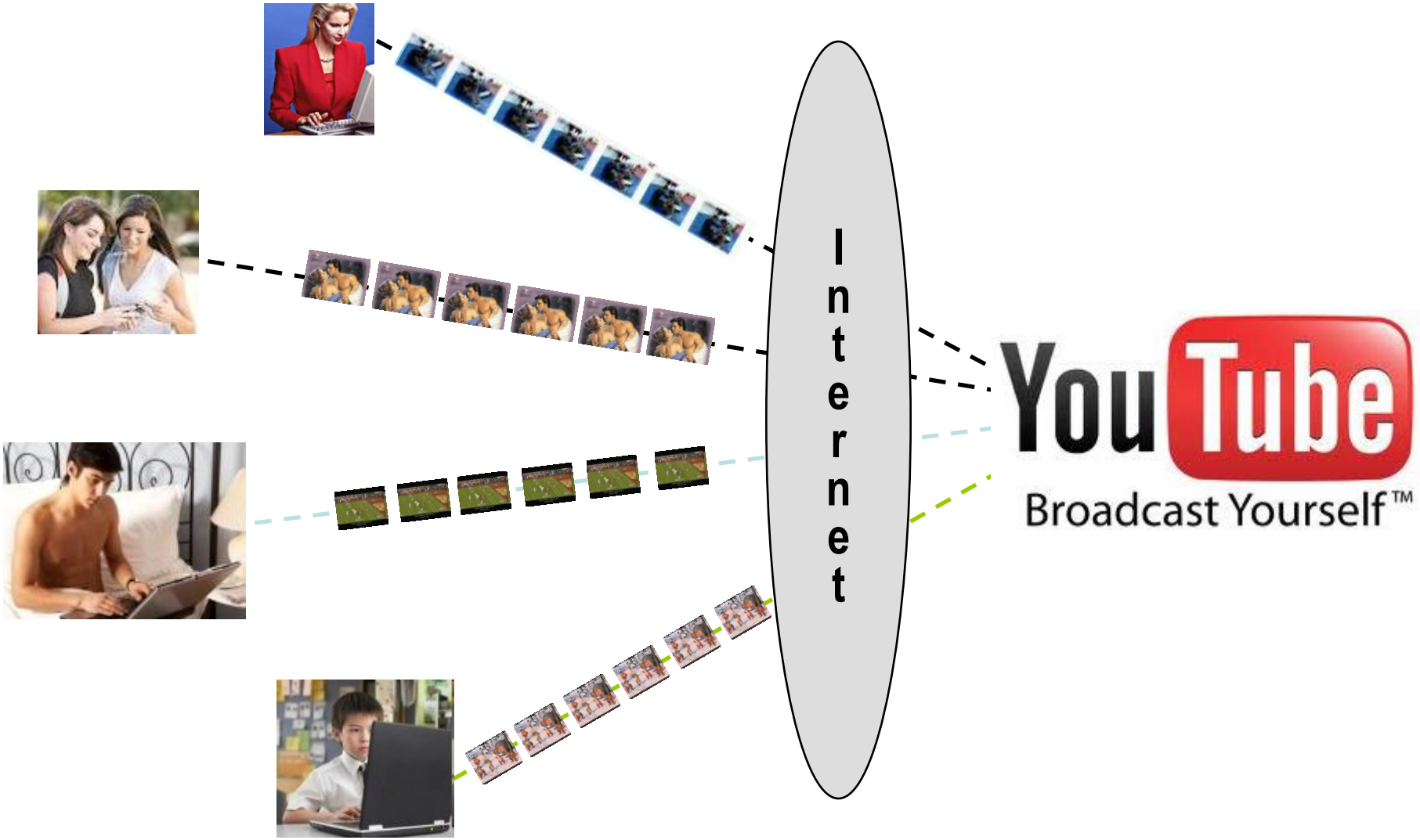
Imagine a 10 GB VM

- 100 Mbps → at least 800 seconds (13 minutes) download
- 10 Mbps → at least 8000 seconds (over two hours) download

No one will wait that long to look at something briefly!

How do we achieve quick launch?

Video Streaming



VM Streaming Not So Easy

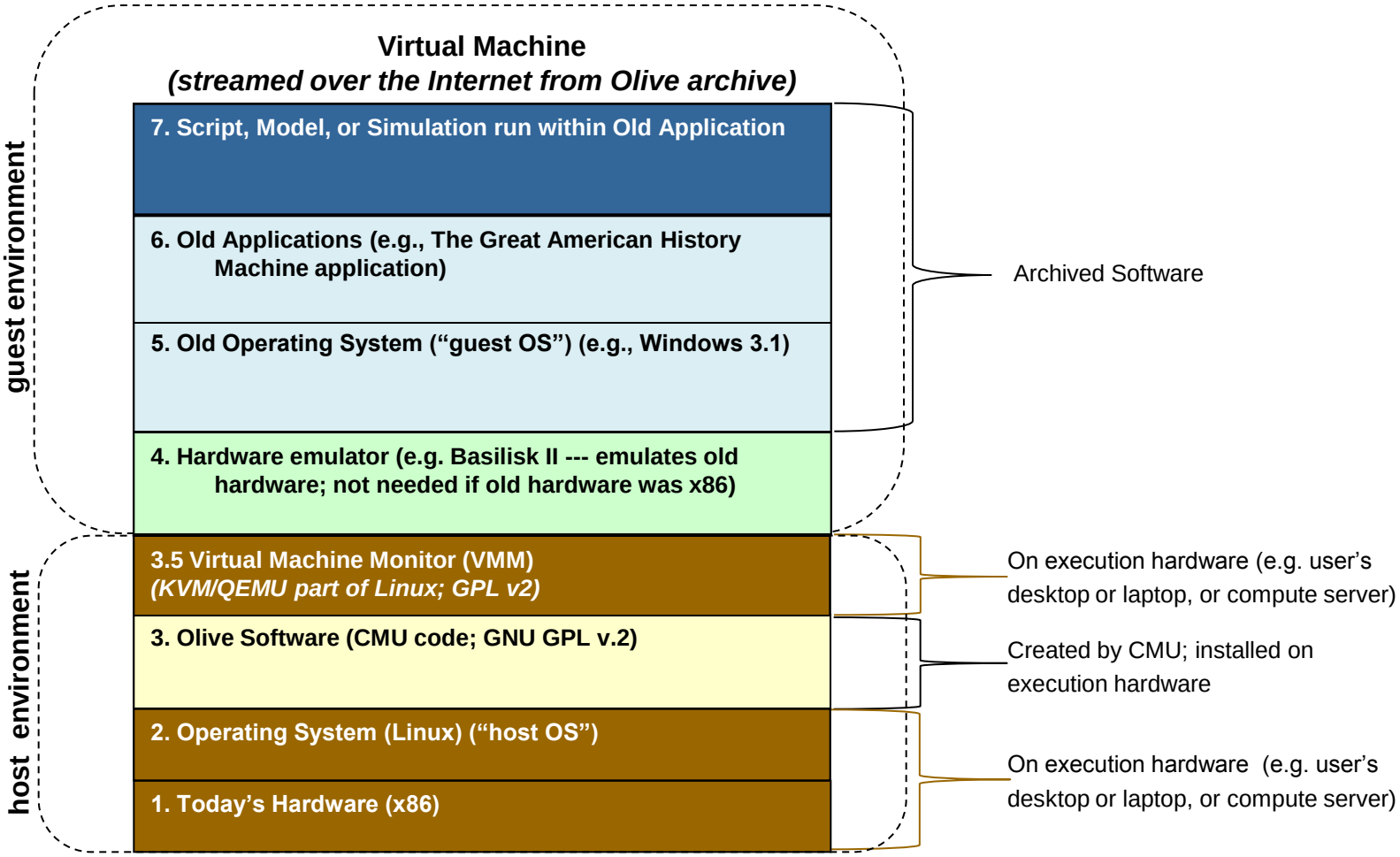
Access to VM image is not linear

- reference pattern depends on many runtime factors
- data dependencies
- human interaction
- spatial and temporal locality (program behavior)

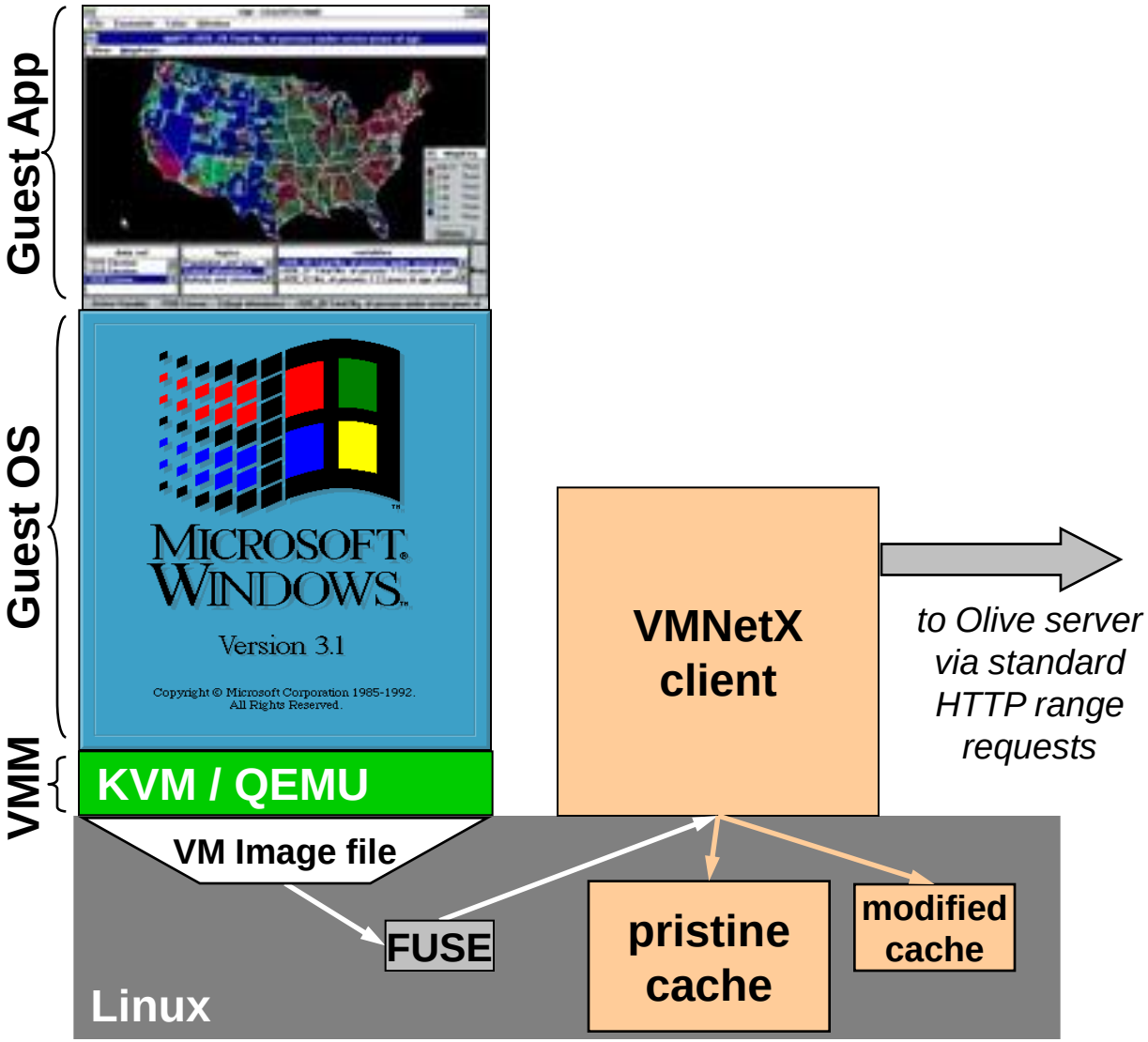
Borrow an old idea from operating systems: demand paging

- intercept missing VM pieces and fetch over Internet
- prefetching may speed up performance if predictions good

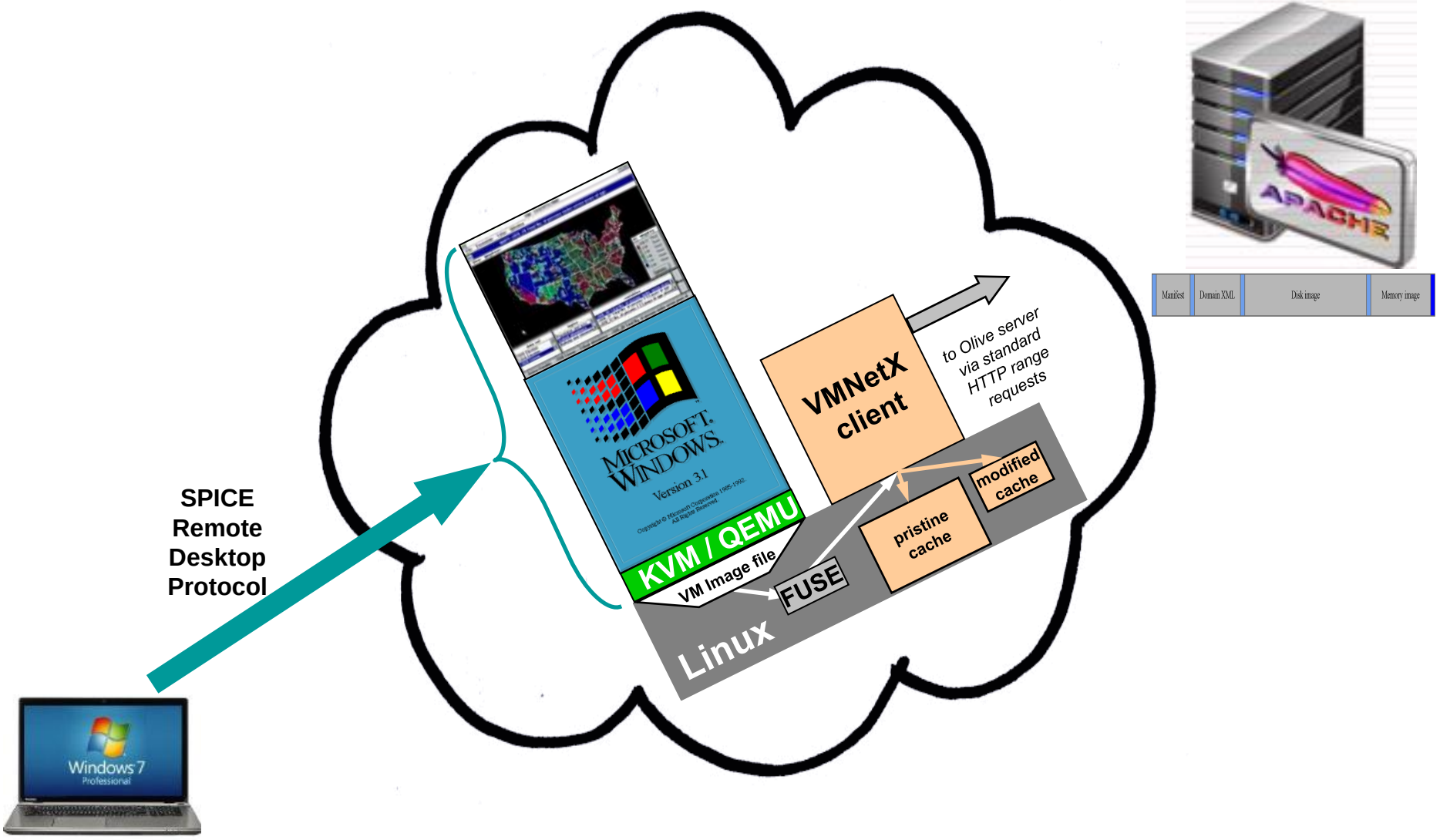
Olive Client Structure



Olive Implementation



Cloud Execution of Olive



Essence of Our Approach

The dialectic between an executing program and its environment generates the complex behaviors we observe with software. Olive captures this dialectic precisely.

Use Cases

Many pieces of software work only with the hardware and platform for which they were developed.

Educational Software

- The Great American History Machine
- ChemCollective

Games

- Doom
- Oregon Trail

Research Software

- Air Stripper Design and Costing
- Laboratory software for data processing and modeling

Metadata Challenges

Existing metadata standards may not be well-equipped to handle executable content.

Current State:

- Builds on work by Jerome McDonough at the University of Illinois
- Expanded MARC in OAI/OSS and METS

Future Considerations:

- **Curator's Workbench** (Carolina Digital Repository) – Desktop tool for metadata capture and organization across large collections
- Ithaka Suggestion: **TOTEM** (KEEP) – Records complex hardware and software relationships which apply to digital objects
- **XSEAD** (Carnegie Mellon University) – Crosses boundaries in Science, Engineering, Arts and Design to crowdsource metadata generation and curation

Curation Challenges

A nontrivial amount of technical work is required to successfully build a VM which achieves execution fidelity.

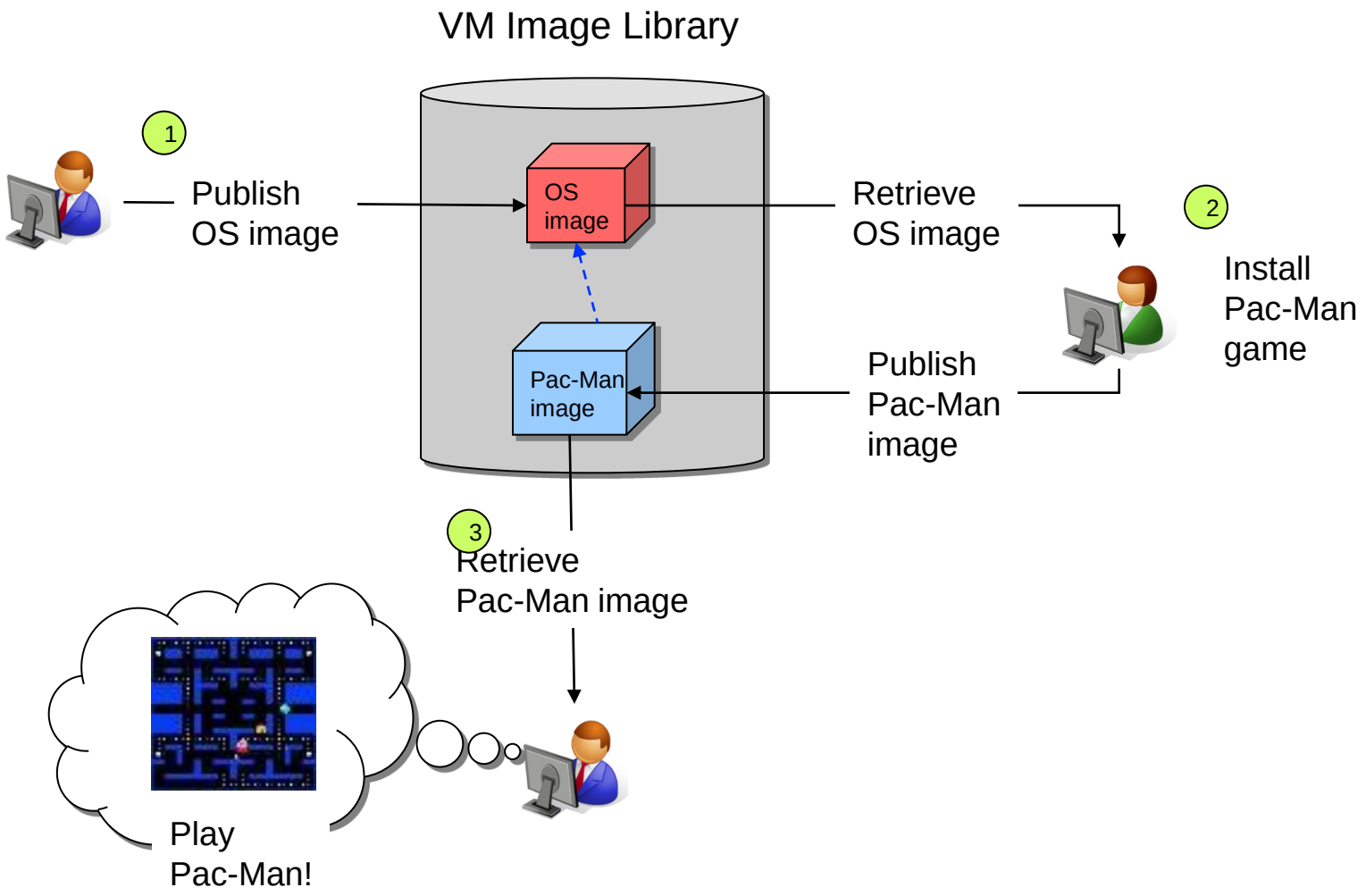
Under our current model, curators are responsible for:

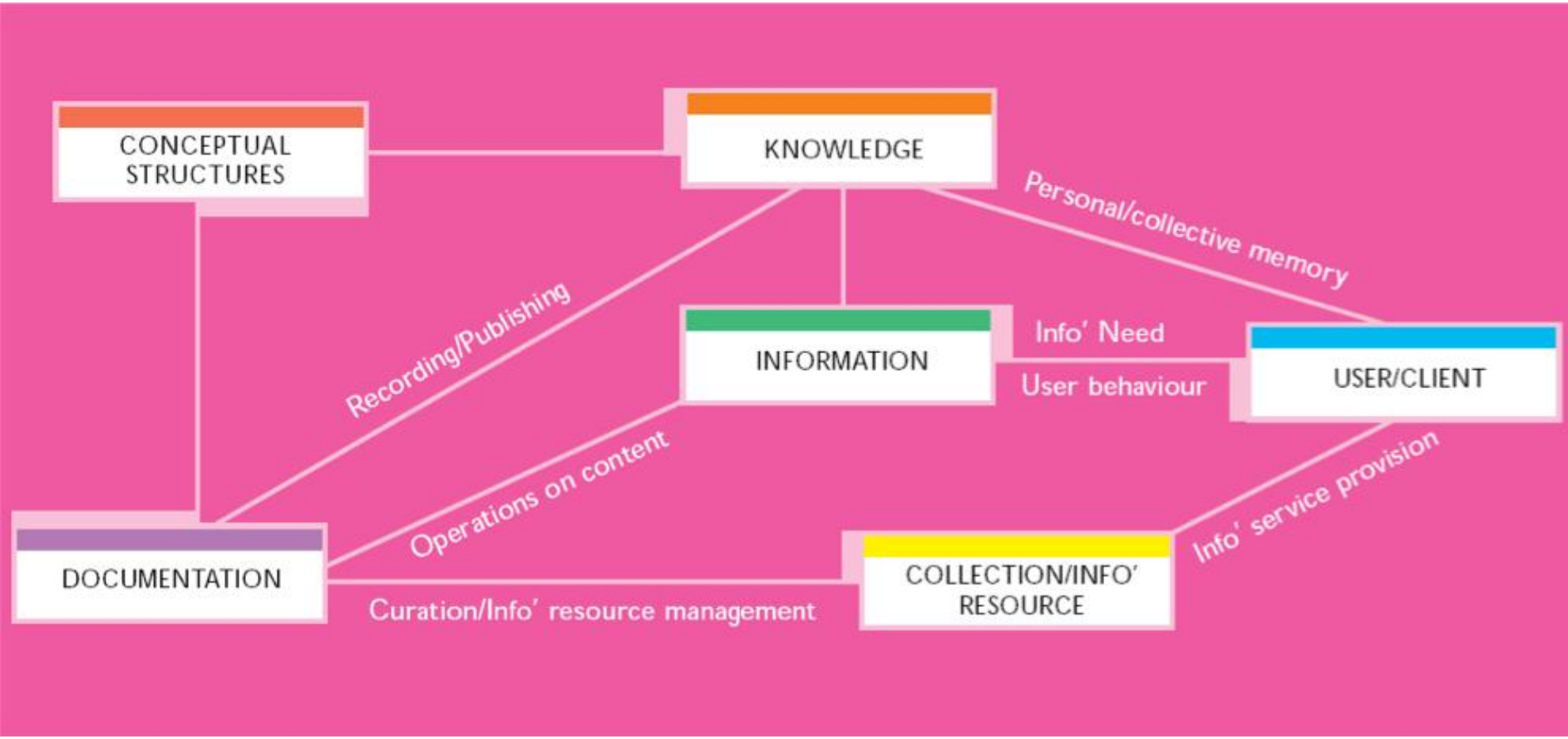
- Installing and configuring the platform
- Finding and including appropriate dependencies (libraries, etc.)
- Identifying and distinguishing bugs in the original implementation as compared with bugs in the virtualization stack

During the next phase, we are focusing on:

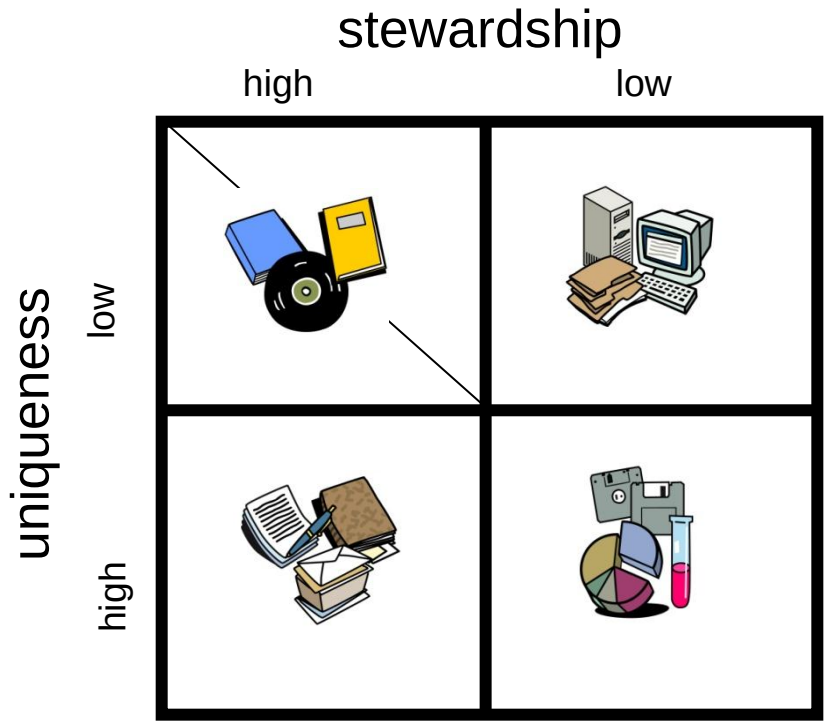
- **Spinoff VM's** – We are implementing functionality to save incremental changes in order to support collaborative curation and eliminate overlapping effort
- **Crowdsourcing** – Allowing any user who agrees to the appropriate terms & disclaimers to push new content back to Olive and be subject to a reputation based system a la Amazon

The Final Product




CONCEPTUAL STRUCTURES
KNOWLEDGE
Personal/collective memory
Recording/Publishing
INFORMATION
Info' Need
User behaviour
USER/CLIENT
Operations on content
Info' service provision
DOCUMENTATION
COLLECTION/INFO' RESOURCE
Curation/Info' resource management

Collections grid



Books
Journals
Newspapers
Gov. docs
CD, DVD
Maps
Scores

Special collections
Rare books
Local/Historical newspapers
Local history materials
Archives & Manuscripts,
Theses & dissertations

Freely-accessible web resources
Open source software
Newsgroup archives

Research, learning and administrative materials,
•ePrints/tech reports
•Learning objects
•Courseware
•E-portfolios
•Research data

•Institutional records
•Reports, newsletters, etc

Partners

INSTITUTE of Museum and Library SERVICES

Reed Elsevier

K&L GATES

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Thanks!

URL: https://olivearchive.org

GitHub: CMUSatyaLab/vmnetx

Twitter: @OliveArchive