A Future Emulation and Automation Research Agenda

Dirk von Suchodoletz

Albert-Ludwigs University Freiburg, Hermann-Herder Str. 10, 79104 Freiburg i. B., Germany

Abstract. After significant research and proven usefulness especially for complex, dynamic and interactive objects emulation is not widely adapted in digital preservation at all. While some significant building blocks of emulation based strategies are present a number of components are still unsatisfactory or missing. This paper proposes a research agenda for the future integration of emulation into preservation workflows. It discusses prerequisites and requirements for fully automated services operating in large scale environments. Those include the replacement of user interaction by using a standard interfacing protocol like Virtual Network Computing, proper system image and software components archiving and the "preservation aware" emulator. To achieve this additional channels to control the emulator and monitor its states are required. This paper analyses the state of the art in emulation and motivates the need for introducing additional control channels.

1 Introduction

Emulation is a extremely versatile and durable solution for retaining access to any kind of digital content. For some digital objects such as games, educational software or research applications, it is actually the only possible way as these objects usually can not be migrated. Nevertheless, emulation is not widely adopted in digital preservation. It has a status as a niche strategy handled by a few trained experts only. This is completely different to the wide adoption of emulators by hobbyists to run instances of their old computer systems or to sell old versions of electronic games for new platforms by entertainment studios, the wide spread deployment in the sub-field of virtualization and the use of emulation for future system software development. These scenarios imply an informed user already familiar with the technology: for example, deprecated systems and their user interfaces will be well understood by enthusiasts running emulators as part of the "game scene", as well as to system administrators dealing day to day with operating systems.

Emulation in digital preservation did not go much beyond the show-case scenario yet: A wide range of home computer and X86 emulators are available to demonstrate the feasibility of running old computer games, navigation of outdated web sites with the original tools of their times or displaying a complex object of a deprecated format. All this requires a certain effort including the

Dagstuhl Seminar Proceedings 10291
Automation in Digital Preservation
http://drops.dagstuhl.de/opus/volltexte/2010/2771

proper choice and configuration of the emulator, the recovery of the original or compatible software environments, the transfer of the object into this environment and finally its rendering or execution [5].

2 Unsteady Ground

While emulators and virtualization tools are taken for granted they have not proven their long-term availability. Much of the emulators and virtualization tools are younger than ten years, lots of emulators already vanished which might not completely replaced by new ones. The few approaches aiming at the long-term perspective like Dioscuri [11] or UVC [8] are often not as powerful or complete. Other tools like the VMware X86 virtualization like the *Workstation* changed the emulated system significantly deprecating old operating systems like Windows 3.X from Version 4.X on. Additionally the container formats of the virtual harddisk were updated regularily rendering actual *Workstation* Versions unable to access containers of earlier 3.X versions [15].

On the other side, if no mitigation strategies are taken, the knowledge about past computer systems will vanish. Important operational information, like the proper configuration of networking, graphical output in the correct resolution and color depth or configuration of the audio interfaces is lost. The same applies to the handling of once popular (graphical) user interfaces. How many owners of an iPhone or iPad would be comfortable with the DOS command line or the graphical user interface (GUI) of Windows 3.0? Hence, it is not sufficient to store the original computer systems or equivalents and its components, rather the knowledge about using these systems has to be kept and documented in order to safeguard usability for future usibility.

A major factor in the discussion of emulation strategies is widely missing: Up to now system images – the combination of software components running in a specific emulator to serve a certain purpose combined into to a runable original environments – were implicitly taken for granted. But results are not sufficiently reproducable as the required software components are implicitly used in todays experiments as they are not categorized and officially archived. Thus a component like a missing operating system for a specific X86 or Power PC machine or a firmware ROM of a home computer might render a digital object completely unusable, even with a perfectly running virtual substitute of the original machine. Beside the physical availability the legal issues of software are often ignored [12].

Further, up to now lots of knowledge and software is available on the net but will vanish with people loosing interest. A general software archiving – one of the building blocks of an emulation based preservation strategy – is not undertaken by any significant memory institution yet [10]. Despite the considerable efforts on digital preservation research, this essential groundwork has until now been largely neglected. This could lead to fatal gaps in the preservation workflows of future generations.

3 Emulation research 1.0

Previous research mainly focused on success criteria for the applicability of the emulation strategy in long-term preservation [14, 13, 16]. Most of the emulators taken into consideration like Dioscuri, QEMU, MESS or commercial virtualization tools are stand-alone applications not optimized for preservation services. During the PLANETS project [3] the prototype GRATE[1] was developed which allows the wrapping of various emulators with software environments within a single networked application. Designed as a general purpose remote access system it demonstrated a prototypical *create-view* service. Much of the involved procedures work in a black box model: Start the emulator with the object attached e.g. as virtual floppy or harddisk partition. Then wait for an uncertain amount of time until hopefully the expected action happens. Finally shut down the emulator and retrieve the altered object if required. While the traditional

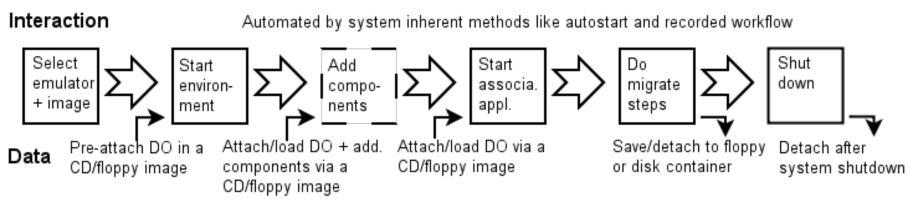

Fig. 1. Emulated environments could be used to migrate digital objects using the original tools they were created with.

human-interaction with the system strategy is acceptable for those scenarios it is not an option for the integration into large-scale production preservation frameworks requiring e.g. the un-attended migration of a batch of Word Perfect documents to ASCII text. Migration is yet another preservation service which might use emulation – at least if run in large scale – requires tools which could be deployed non-interactively (Fig. 1). A tedious issue from the view point of a digital archive manager is that for example, spread sheet, product design (CAD), audio/video or word processing programs cannot execute basic tasks in an unattended and fully automated manner. Thus a larger number of migration tools – the original applications the objects were created with – are not eligible restricting especially the handling of complex and proprietary file formats. Nevertheless research into the integration of emulation within automated archiving processes is still in its infancy [7, 2, 6]. Typically, current preservation frameworks in operation do not implement migration services using original applications within emulated environments as a backend for a number of reasons: Emulators offer a very limited measurements of functionalities. As the digital object of interest is

[1] GRATE – Global Remote Access To Emulation, *http://planets.ruf.uni-freiburg.de*, see [19] for further reference.

wrapped into an emulation environment which is a complete software ecosystem of its own no much observation from the outside is possible. The actual state of CPU, RAM and storage is often not available. This often gives a very limited idea on important states like application or operating system failures. In general a whole bunch of processes running within the environment is wrapped into an additional software layer and thus reduced often to a single process on the host machine generating file input/output on the system image.

4 Future Challenges in Emulation Research

The research and development of the last decade produced working preservation frameworks and emulators for a wide range of platforms. But significant building units connecting both spheres are unsatisfactory or missing. Emulation has to consider a range of additional challenges beside rebuilding a deprecated hardware or software environment in software executable on modern computer architectures. Thus the Keeping Emulation Environments Portable project[2] aims to develop an emulation wrapping system providing a generalized interface for emulation frameworks.

Previous efforts to integrate emulation services into long-term preservation frameworks such as proposed by PLANETS are not very well suited for large scale scenarios in their present shape. Neither a large number of concurrent users or the (parallel) processing of large collections in a specified timeframe is possible yet. After the demonstration of feasibility of integration [19, 2] the focus should be shifted towards the large-scale, production-system integration [18].

As long as the number of objects to be processed is manageable or just a few individual users interact with emulation environments the required computing power and wall clock time consumed for those processes is minor. If large scale preservation systems are to be run and preservation planning tools like PLATO [1] are to be used to evaluate runtimes and give reasonable cost estimates more information is needed.

A new generation of "digital preservation aware" emulators is desirable implementing a number of different interfaces. Beside the traditional screen output to the host system a VNC interface should be available. VNC offers an appropriate abstract layer to operate a standard computer interface providing screen output and keyboard and mouse input. The input activity and the resulted output could be observed and recorded. Such a recording could be used to replace the human user on the VNC client side by a machine agent sending events to the machine and interpreting the screen content [7].

The typical emulator of today is primarily focusing on direct human interaction by offering a GUI. The preservation perspective is often missing as the tools lacking certain capabilities which should be taken into account for basic preservation requirements. GUI enabled emulators are suitable for a range of certain create view requirements but sub-optimal for integration into large preservation

[2] KEEP, *http://www.keep-project.eu*

frameworks. For large scale migration scenarios requirements like predictability and accountability of actions taken play an important role. The time (calculated e.g. in wall time or CPU cycles consumed) should be predictable to give archive operators a base to calculate costs and amount of time consumed for a certain preservation action. A preservation ready emulation thus would include the availability of control interfaces to archiving systems to monitor and query the states the emulator at any time.

The aforementioned VNC interface has no access to other emulator controls like power and reset buttons or on removable devices. Thus beside the command line or configuration file interface for initial emulator setup and configuration an emulator should implement a so-called "monitor" like e.g. QEMU. It opens a channel to send commands during emulator operation for e.g. mounting removable devices, sending special keystrokes like CTRL-ALT-DEL and allows suspending and shutting down of the emulated system. It might implement the monitor interface allowing to request certain states of the running machine.

4.1 Software Archiving: Strengthening Weak Links

A future challenge is the reproduction of original environments from its single building blocks. View paths or pathways in other literature formalise the steps in an abstract way from the digital object into the actual rendering environment of the user. The number of involved steps might differ depending on layer emulated of the hardware-software stack [16]. View path deliver valuable hints about components e.g. operating systems and applications which are to be included. But they do not produce an exact installation order and dependency lists of software items to be taken into consideration and are not directly transferable into an object schema [4]. To automate software selection a tool registry describing software components and their dependencies like PRONOM [9] for file types is required.

In addition to storing and handling the digital objects themselves, it is essential that this complex set of software components is kept and managed. Archive (automatically) all relevant software components which a certain digital ecosystems consists of. Additionally be aware of all software components required by certain significant properties of the object. Those could be fontsets, codecs or specific decompression tools [17, 10].

4.2 Emulator Migration and Longevity

Emulation does not avoid migration, but moves it to a different level. When the host environment changes the emulators as applications made for this environment need to be changed too. The major challenge is to update the "outer" software layers of the emulator application without changing any inner components at all (Fig. 2). This has been accomplished very well with dead architectures like the old Apple Macintosh or home computers of the 1980th and early 1990th. A good example is the modular emulator MESS around for several years, updated from DOS up to todays modern Windows, Linux and Mac OS operating systems.

"Living" architectures like the X86 are more challenging: The typical problem observable e.g. with the virtual machine *VMware* available since end of the 1990th are the changing virtual hardware components and the updated virtual disk container formats. The audio, video, network and block device configuration

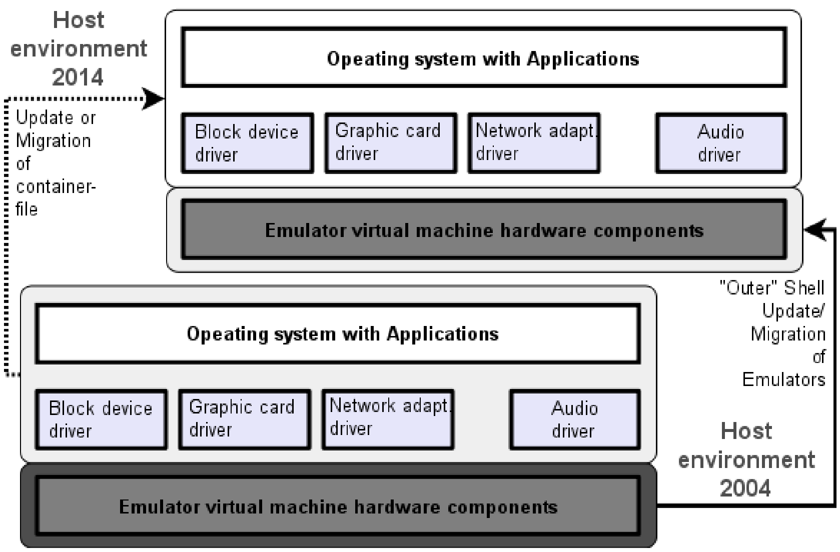

Fig. 2. Emulator updates e.g. required when changing from 2004 system to 2014 system shouldn't change anything of the virtual machine components.

of the standard X86 machine changed significantly requiring the installation of new hardware drivers within the emulated environment. During this period the format of the image file representing the harddisk changed several times. Newer VMware machines are not able to mount images of some versions earlier. The problem is comparable to the real hardware: If a computer is replaced by another a simple transfer of the old harddisk into the new machine or a blockwise copy of the old harddisk to the new one most probably will not work. Thus longevity of the virtual machines becomes a vital issue for the suitability of emulators in digital preservation. The optimal emulator adds new devices to the virtual machine, but keeps the old ones too. Plus, it does not change the format of the container files. This paradigm is partly fulfilled e.g. by QEMU. Nevertheless not all components were kept exactly the same, rendering some Microsoft operating systems like Windows 95 or 98 unusable on some versions.

5 Conclusion

Future research agendas in digital preservation need to bring the emulation strategy onto a next level. If emulators are not getting preservation ready and measurable to allow for comparison between each other and to be compared to other strategies they will not get out of their niche existence in digital preservation. An important sub-domain is an automated quality assurance for well defined test sets of standard environments for new versions especially of community and open source emulators. Defining emulation metrics to describe capabilities would help with testing and emulator comparisons. Additionally convenient methods for user feedback should be included into preservation frameworks making use of emulation. Nevertheless the most successful approach for the next level preservation aware emulator like Dioscuri would be a joint effort of national memory institutions [18]. The existing communities should be made aware of the needs of digital preservation, provided with feedback on the actual developments and for quality assurance. And they should be provided with a steady funding to keep up with the changing technology. Additionally emulation workflows need to get automated to be on equal terms with standard migration procedures. Manual emulation workflows are much too expensive regarding knowledge, personnel costs and time consumption to be taken into consideration for large-scale mass migration tasks.

References

1. Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber, and Hans Hofman. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, 10:133–157, 2009. 10.1007/s00799-009-0057-1.
2. Christoph Becker, Hannes Kulovits, Michael Kraxner, Riccardo Gottardi, Andreas Rauber, and Randolph Welte. Adding quality-awareness to evaluate migration web-services and remote emulation for digital preservation. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL09)*, 2009.
3. Adam Farquhar and Helen Hockx-Yu. Planets: Integrated services for digital preservation. *International Journal of Digital Curation*, 2(2), 2007.
4. Mario Philipps. Entwurf und implementierung eines softwarearchivs für die digitale langzeitarchivierung. Master's thesis, Albert-Ludwigs-Universität Freiburg, July 2010.
5. PLANETS. Planets - digital preservation research and technology. Online, *http://www.planets-project.eu*, 2010.
6. Klaus Rechert, Dirk von Suchodoletz, and Randolph Welte. Emulation based services in digital preservation. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 365–368, New York, NY, USA, 2010. ACM.
7. Klaus Rechert, Dirk von Suchodoletz, Randolph Welte, Maurice van den Dobbelsteen, Bill Roberts, Jeffrey van der Hoeven, and Jasper Schroder. Novel workflows for abstract handling of complex interaction processes in digital preservation. In *Proceedings of the Sixth International Conference on Preservation of Digital Objects (iPRES09)*, 2009.
8. Jasper Schroder and Raymond van Diessen. Digital asset preservation tool. Online, *http://www.alphaworks.ibm.com/tech/uvc*, September 2010.
9. The National Archives TNA. The technical registry pronom. Online, *http://www.nationalarchives.gov.uk/pronom*, 2010. Online resource.

10. Maurice van den Dobbelsteen, Dirk von Suchodoletz, and Klaus Rechert. Software archives as a vital base for digital preservation strategies. Online, *http://eprints.rclis.org/18764/*, July 2010.
11. Jeffrey van der Hoeven. Dioscuri: emulator for digital preservation. *D-Lib Magazine*, 13(11/12), 2007.
12. Jeffrey van der Hoeven and Jasper Schroder. Keep legal study. In Andreas Rauber, Max Kaiser, Rebecca Guenther, and Panos Constantopoulos, editors, *7th International Conference on Preservation of Digital Objects (iPRES2010) September 19 - 24, 2010, Vienna, Austria*, pages 373–381. Austrian Computer Society, 2010.
13. Jeffrey van der Hoeven and Dirk von Suchodoletz. Emulation: From digital artefact to remotely rendered environments. In *Proceedings of the Fifth International Conference on Preservation of Digital Objects (iPRES08)*, pages 93–98, The British Library, St. Pancras, London, 2008. The British Library.
14. Remco Verdegem and Jeffrey van der Hoeven. Emulation: To be or not to be. In *IS&T Conference on Archiving 2006, Ottawa, Canada, May 23-26*, pages 55–60, 2006.
15. Dirk von Suchodoletz. *Funktionale Langzeitarchivierung digitaler Objekte – Erfolgsbedingungen für den Einsatz von Emulationsstrategien.* Cuvillier Verlag Göttingen, 2009.
16. Dirk von Suchodoletz. Requirements for emulation as a long-term preservation strategy. Online, *http://eprints.rclis.org/18984/*, July 2009.
17. Dirk von Suchodoletz. Das softwarearchiv - eine erfolgsbedingung für die langzeitarchivierung digitaler objekte. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 63(3):38–55, 2010.
18. Dirk von Suchodoletz, Klaus Rechert, Jeffrey van der Hoeven, and Jasper Schroder. Seven steps for reliable emulation strategies – solved problems and open issues. In Andreas Rauber, Max Kaiser, Rebecca Guenther, and Panos Constantopoulos, editors, *7th International Conference on Preservation of Digital Objects (iPRES2010) September 19 - 24, 2010, Vienna, Austria*, pages 373–381. Austrian Computer Society, 2010.
19. Randolph Welte. *Funktionale Langzeitarchivierung digitaler Objekte – Entwicklung eines Demonstrators zur Internet-Nutzung emulierter Ablaufumgebungen.* Südwestdeutscher Verlag für Hochschulschriften, 2009.