Fixating on Fixity – Is Your Choice of Checksum in Conflict with Your Climate Goals?

Preliminary findings of research by:
George Blood, owner
Steve Burns, developer
Jeff Chestek, chief audio tech
George Blood Audio/Video/Film/Data

Premis

- Users need to *trust* archives
- Ttrust means objects are *authentic*
- When preservation of the *essence* of the object is complete and declared authentic, it stops changing and becomes *fixed*
- To confirm a digital object has not changed, we use *fixity*
- *Crytographic hashes*, in the form of checksums, are the mechanism for confirming digital objects are not changed
- Unchanged digital objects are truthworthy

How hard can it be?

- Select a range of checksum types to compare
  - MD5, SHA256, SHA1, Okum, CRC32
- Set up some computers to compare
  - New, old, Mac, PC, Linux, cloud
- Try different storage
  - External spinning drive, SSD, NAS, SAN, cloud
- Gather sample files (does file size have an impact?)
  - Lots of small files, some medium-sized files, big files
- Write some scripts
- Find a partner to run the tests in a different environment
  - WGBH/AAPB is our partner
- Report results

Early challenges

- How to measure the power consumption?
  - We purchased inexpensive plug-in meters
  - Do we include the external drive?
  - If so, how do we allocate power consumption from remote storage (NAS)
- Low repeatability of tests
  - Wildly varying results when test were run several times
- Quickly discover how many tests are needed

How hard can it be?

- Select a range of checksum types to compare
  - MD5, SHA256, SHA1, Okum, CRC32
- Set up some computers to compare
  - New, old, Mac, PC, Linux, cloud
- Try different storage
  - External spinning drive, SSD, NAS, SAN, cloud
- Gather sample files (does file size have an impact?)
  - Lots of small files, some medium-sized files, big files
- Write some scripts
- Find a partner to run the tests in a different environment
  - WGBH/AAPB is our partner
- Report results

5 (checksum types)

x5 (computer types)

x5 (external storage)

x3 (file sizes)

x3 (multiple passes)

x2 (two partners)

= 2,250 test runs!

Simplification

- Limit tests to MD5s and SHA256
- Use new, out of the box, iMac M1
  - Disconnected from network, nothing else installed
- Measure power consumption of computer only
- Expand the data size from 100GB to looping for 24 hours
  - Averages many test passes
  - Generates power consumption values big enough to compare

A peak at the data set

NDSA Research

File Edit View Insert Format Data Tools Extensions Help Last edit was made yesterday at 4:05 PM by Jeff Chestek

Share

39:46 | iMac M1 (Yellow)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		Storage Location	Data set size	Format	Checksum Type	Test Run	Start Time	End Time	Duration	Meter reading-start	Meter reading-end	Net Power Usage	Power consumption per hour	Cycle counter	Net Cycle count	Total number of files	Current file count	Fraction of current cycle	Net cycle count (decimal)	Data processed per measured period	Normalized result	Notes
2	computer specs	Storage type and connection	GB	100GB, XYZ audio files		ID of test run of 3	Time of day	Time of day	HH:MM:SS	kW/h	kW/h	Difference (W/h)	(W/h)	Integer	calculated	per cycle	integer	calculated	calculated	Gigabytes	W/h per TB	
36																					#DIV/0!	
37																					#DIV/0!	
38																					#DIV/0!	
39	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	1	7/11/2022 8:31:00 AM	7/11/2022 8:33:00 AM	0:02:00	3.01	3.02	10	300.00	1	1	190		0.00	1.00	256.00	40.00	256GB 1 cycle
40	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	2	7/11/2022 9:02 AM	7/11/2022 9:05 AM	0:03:00	3.02	3.03	10	200.00	1	1	190		0.00	1.00	256.00	40.00	256GB 1 cycle
41	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	3	7/11/2022 9:10 AM	7/11/2022 9:12 AM	0:02:00	3.03	3.03	0	0.00	1	1	190		0.00	1.00	256.00	0.00	256GB 1 cycle
42	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	4	7/11/2022 9:34 AM	7/11/2022 10:47 AM	1:13:00	3.04	3.08	40	32.88	34	34	190		0.00	34.00	8,704.00	4.71	256GB 34 cycles
43	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	5	7/11/2022 10:47 AM	7/11/2022 12:28 PM	1:41:00	3.08	3.15	70	41.58	75	41	190		0.00	41.00	10,496.00	6.83	256GB 75 cycles (net 41)
44	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	6	7/11/2022 12:28 PM	7/11/2022 1:54:00 PM	1:26:00	3.15	3.20	50	34.88	106	31	190		0.00	31.00	7,936.00	6.45	256GB 106 cycles (net 31
45	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	7	7/11/2022 2:10:00 PM	7/12/2022 14:24:00	24:14:00	3.69	4.42	730	30.12	686	686	190		0.00	686.00	175,616.00	4.26	256GB 686 cycles (new s
46	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	8	7/12/2022 14:24:00	7/13/2022 14:24:00	24:00:00	4.42	5.16	740	30.83	1361	675	190		0.00	675.00	172,800.00	4.39	256GB 1361 cycles (net 6
47	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	9	7/13/2022 14:24:00	7/14/2022 14:33:00	24:09:00	5.16	5.80	640	26.50	2040	679	190	166	0.87	679.87	174,047.66	3.77	256GB 2040.87 cycles (n
48	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	10	7/14/2022 14:33:00	7/15/2022 14:53:00	24:20:00	5.80	6.47	670	27.53	2725	685	190	172	0.91	685.91	175,591.75	3.91	256GB 2725.91 cycles (n
49	iMac M1 (Yellow)	Internal SSD	256	Audio	md5	11	7/15/2022 14:53:00	7/18/2022 14:58:00	72:05:00	6.47	7.97	1500	20.81	4070	1345	190		0.00	1,345.00	344,320.00	4.46	256GB 4070 cycles
50	iMac M1 (Yellow)	Internal SSD	256	Audio	sha256	12	7/18/2022 3:04 PM	7/19/2022 3:33:00 PM	24:29:00	7.97	8.75	780	31.86	1905	1905	190	181	0.95	1,905.95	487,923.87	1.64	256GB 1905 cycles
51	iMac M1 (Yellow)	Internal SSD	256	Audio	sha256	13	7/19/2022 3:33:00 PM	7/20/2022 15:39:00	24:06:00	8.75	9.55	800	33.20	3709	1804	190	6	0.03	1,804.03	461,832.08	1.77	256GB 3709 cycles
52	iMac M1 (Yellow)	Internal SSD	256	Audio	sha256	14	7/20/2022 3:39 PM	7/21/2022 15:43:00	24:04:00	9.55	10.36	810	33.66	5532	1823	190	0	0.00	1,823.00	466,688.00	1.78	256GB
53	iMac M1 (Yellow)	Internal SSD	256	Audio	sha256	15	7/21/2022 3:43 PM	7/22/2022 15:43:00	24:00:00	10.36	11.16	800	33.33	7348	1816	190	0	0.00	1,816.00	464,896.00	1.76	256GB
54	iMac M1 (Yellow)	Internal SSD	256	Audio	sha256	16	7/22/2022 3:43 PM	7/25/2022 15:44:00	72:01:00	11.16	13.74	2580	35.83	12470	5122	190	170	0.89	5,122.89	1,311,461.05	2.01	256GB
55	iMac M1 (Yellow)	Internal SSD	256	Audio	md5deep	17	7/25/2022 3:57 PM	7/26/2022 15:12:00	23:15:00	13.75	14.63	880	37.85	417	417	190	95	0.50	417.50	106,880.00	8.43	256GB
56	iMac M1 (Yellow)	Internal SSD	256	Audio	md5deep	18	7/27/2022 9:39 AM	7/28/2022 10:02:00	24:23:00	15.20	16.17	970	39.78	566	566	190	95	0.50	566.50	145,024.00	6.85	256GB
57	iMac M1 (Yellow)	Internal SSD	256	Audio	md5deep	19	7/28/2022 10:02 AM	7/29/2022 10:01:00	23:59:00	16.17	17.05	880	36.69	1122	556	190	184	0.97	556.97	142,583.92	6.32	256GB
58	iMac M1 (Yellow)	Internal SSD	256	Audio	md5deep	20	7/29/2022 10:01 AM		-1074514:01:00	17.05		-17050	0.02		-1122	190		0.00	-1,122.00		#DIV/0!	256GB
59	Cheesegrater	USB Drive, USB powered	1024	Audio	md5	1	7/12/2022 8:18 AM	7/13/2022 8:18:00	24:00:00	1.42	3.75	2330	97.08	3.72	3.72	1395			3.72	3,809.28	626.34	1TB, 4TB drive 3.72 cycle
60	Cheesegrater	USB Drive, USB powered	1024	Audio	md5	2	7/13/2022 8:18:00	7/14/2022 8:18:00	24:00:00	3.75	6.09	2340	97.50	7.28	3.56	1395			3.56	3,645.44	657.30	1TB, 4TB drive 7.28 cycle
61	Cheesegrater	USB Drive, USB powered	1024	Audio	md5	3	7/14/2022 8:18:00	7/15/2022 8:18:00	24:00:00	6.09	8.42	2330	97.08	10.93	3.65	1395			3.65	3,737.60	638.36	1TB, 4TB drive 10.93 (ne
62	Cheesegrater	USB Drive, USB powered	1024	Audio	md5	4	7/15/2022 8:18:00	7/16/2022 15:07:00	30:49:00	8.42	11.41	2990	97.03	15	4.07	1395	671	0.48	4.55	4,660.23	657.00	1TB, 4TB drive 15 (net 4.
63	Cheesegrater	USB Drive, USB powered	1024	Audio	sha256	5	7/18/2022 8:41:00 AM	7/19/2022 8:41:00	24:00:00	15.47	18.30	2830	117.92	2	2.26	1395	363	0.26	2.52	2,580.70	1,122.92	1TB, 4TB drive 2.26 cycle
64	Cheesegrater	USB Drive, USB powered	1024	Audio	sha256	6	7/19/2022 8:41:00	7/20/2022 8:41:00 AM	24:00:00	18.30	21.15	2850	118.75	4	2	1395	619	0.44	2.44	2,502.38	1,166.25	1TB, 4TB drive
65	Cheesegrater	USB Drive, USB powered	1024	Audio	sha256	7	7/20/2022 8:41:00	7/21/2022 8:41:00	24:00:00	21.15	23.99	2840	118.33	6	2	1395	845	0.61	2.61	2,668.27	1,089.90	1TB, 4TB drive
66	Cheesegrater	USB Drive, USB powered	1024	Audio	md5	8	7/21/2022 8:41:00	7/22/2022 10:23:00	25:42:00	23.99	26.50	2510	97.67	3	3	1395	1308	0.94	3.94	4,032.14	637.44	1TB, 4TB drive, md5 loop
67	Cheesegrater	USB Drive, USB powered	1024	Audio	md5	9	7/22/2022 10:23:00	7/25/2022 8:34:00	70:11:00	26.50	33.39	6890	98.17	14	11	1395	565	0.41	11.41	11,678.74	604.12	1TB, 4TB drive, md5 loop

Simplification – early results

- Repeatable test results!
- Confirmation of scripts and process
- Clarification of questions and next steps

…

Tests have expanded to

- External HDD vs SSD
- New computer vs old computer (aka “free hardware”)
- Using different CLI apps and scripts to test speed of code

Preliminary findings

- Do not use an old computer
  - Carbon footprint is 100—200x greater than current power efficient CPUs
- SSDs leave spinning discs in the dust
  - May be 25x less carbon intensive
- File size does not meaningfully impact results
  - 1 video file of 100GB has the same results as thousands of smaller files of 100GB
- Surprise: SHA256 *may* have lower carbon footprint than MD5
  - The opposite of a starting assumption
  - Building tests to control for possible variables to confirm finding

Planning testing and goals - 1

- #1 Quantify and provide data to support recommendations for best strategy in different use cases
  - Small archive with one computer and external drives
  - Small institutions or depart with large, local RAID (Drobo, Synology, etc)
  - Large institutional repository
  - Cloud
- Expand tests to other OS platforms
- Package process to share
  - First to WGBH for peer review
  - For anyone to use

Planning testing and goals - 2

- Run tests on AVPreserve's *Fixity Tool*
  - Engage AVPreserve in discussion – to test our methods and know their tool
- Quiescent tests
  - How much of the power is from a computer just being one vs. verifying checksums?
- Is it time and carbon efficient to download files from remote storage to a local SSD to verify the checksum?
- Publish results and share raw data
- Full presentation will be at the CLIR/NDSA/DLF in October

Additional information:

- george.blood@georgeblood.com
- On Facebook: George Blood LP
- On Instagram: @georgebloodlp
- Reports and PPTs at www.georgeblood.com
- CLIR/NDSA/DLF
  - Presentation in person at the 2022 Digital Preservation Conference in Baltimore, Maryland. The conference will be held October 12-13