Our setup:

- Package 1 used: see images-test-single-folder
- Package 2 used: see images-test-multi-folder
- Archivematica onfiguration - automated
    - Scan for viruses: yes
    - Assign UUIDs to directories: yes
    - Generate transfer structure report: no
    - Perform file format identification (transfer): yes
    - Extract packages: yes
    - Delete packages after extraction: yes
    - Perform policy checks on originals: no
    - Examine contents: skip examine contents
    - Create SIP(s): create single SIP and continue processing
    - Perform file format identification (ingest): yes
    - Normalize: normalize for preservation
    - Approve normalization: yes
    - Choose thumbnail mode: no
    - Perform policy checks on preservation derivatives: no
    - Perform policy checks on access derivatives: no
    - Bind PIDs: no
    - Document empty directories: no
    - Reminder add metadata if desired: continue
    - Transcribe SIP contents: no
    - Perform file format identification (submission documentation & metadata): yes
    - Select compression algorithm: 7z using bzip2
    - Select compression level: 5 - normal compression
    - Store AIP: yes
    - Store AIP location: default location
    - Upload DIP: do not upload DIP
    - Store DIP: store DIP
    - Store DIP location: default location

What we're testing:

- Relative consumption of energy for running one large package versus several smaller ones for the same content
- Relative consumption of energy for different stages of processing content (maybe)

How we're doing that (methodology):

- We are running PowerStat to measure the power drawn by Steve's computer (see specs above) for processing the same content as a single package versus several smaller ones (see Package 1 and Package 2, above, for details)
- We used the following processing configuration (identical for both) run automated
- For images-test-single-folder, we ran the package through 3 times and measured the power drawn at half-second intervals

- For images-test-multi-folder, we ran each folder through twice and measured the power drawn at half-second intervals
- We then ran a test run with manual intervention (later excluded from the data because it was less controlled than others)
- We subtracted the baseline for Steve's computer operating, which we calculated based on powerstat outputs when all applications that ran throughout the test were running, except for Archivematica
- The resulting numbers gave us an approximation of the power drawn as a result of running Archivematica with the different packages

Notes and times (note that these times were later calculated on a more precise basis in the "data processing" stage outlined below, using timestamps from the log files)

Date: September 7, 2023
9:25:30-9:26:00 – waiting with Chrome open, resting
9:56:59 – SingleFolder1 Automated transfer
9:57:20, done
9:58:07 - SingleFolder2 Automated transfer
9:58:42, done
9:59:25 - SingleFolder3 Automated transfer
9:59:59 - done
10:01:27 - first folder, first time, multi-transfer-1
10:02:05 - done
10:02:45 - second folder, first time, multi-transfer-2
10:03:15 - done
10:04:09 - third folder, first time, multi-transfer-3
10:04:40 - done
10:06:40 - first folder, second time, multi-transfer-1b
10:07:08 - done
10:07:48 - multi-transfer-2b, second folder, second time
10:08:24 - done
10:09:16 - Multi-transfer-3b, third folder, second time
10:09:54 - done
10:11:26 transfer-manual-decisions
10:11:54 - identify file format - yes
10:12:22 - examine contents - skip examine contents
10:12:50 - create SIP and continue processing
10:13:06 - create SIP and continue processing (for that bug)
10:13:54 - refresh page
10:14:16 - normalize for preservation
10:15:07 - continue without adding metadata
10:15:19 - no transcription
10:15:39 - store AIP
10:15:49 - store in default location

10:16:05 done

Data processing notes

1. Calculate baseline
    a. Identify when started processing based on start time of first package in mcp-server.log
    b. Pull time and watts from results.csv
    c. Result: average watts per half second from the time range 09:56:02 to 09:56:49 for running same computer with all applications constant, except for Archivematica
2. Pull data for each package
    a. Identify the two UUIDs associated with the content (pre-ingest and post-SIP creation, which remains as the AIP UUID) and then identify the start/end timestamps for each package
    b. Pull powerstat data (watts and timestamps) for the relevant time period
    c. Convert to joules (watts times seconds)
    d. Calculate total watts and joules, net draw (total minus baseline), average net draw per half second
    e. Calculate standard deviation, highlight all watts that are more than 1 SD above average
3. Comparison of approaches
    a. For the single folder: average total net watts and joules for all three times, average time to process
    b. For the multiple folders: average total net watts and joules for each folder (twice), add these to get average total net watts and joules for each

Findings

- Splitting content into separate packages drew more than twice as much power (1030 joules versus 2590 joules)
    - Time to process was also more than twice as high (24 seconds versus 1 minute)
    - Cannot generalize out to all types of content at all volumes, but within reason, consolidating packages is likely to be a lower draw when using the application..
- Started to look at processes that drew more power relative to others, defined as more than 1 SD above the average
    - Found that running ClamAV for virus checks was a big relative draw
    - Identify file format, extract metadata also relatively large
    - Normalization
- Still need to do more analysis of what can be modified but in interim, turning off virus check will likely reduce draw without significantly reducing performance, especially as most users have separate in-house processes
- The joules drawn per half-second were relatively constant across all packages.

Limitations

- Small dataset, means that findings may not be generalizable
- Findings are not linked to our particular power mix or any hardware, so they do not reflect actual or embedded emissions
    - However, this does allow us to make relative comparisons between different configurations and types of content
- The overall power draw is quite low, and all findings should be viewed with that context.

Future steps

- Investigate more closely which aspects of running Archivematica draw the most power, and whether there are any opportunities to add efficiencies, advise about this, or remove/disable unnecessary processes
- Test different types of content to see if this shifts the net joules per half-second, or if this number is relatively constant and it is processing time that determines power usage
- Add in data for the power mix we draw on and the hardware that we used in order to achieve an output more in line with the SCI specification.