ESTIMATION OF MEDIA STORAGE REQUIREMENT

GAGARINE YAIKHOM

1. Storage requirement for image tiling

The following equation estimates for a single image file the total storage requirement for maintaining a thumbnail and pre-generated image tiles, in addition to the original image file.

$$T = O + t + \lambda \times \sum_{i=0}^{|Z|} \left\lceil \frac{w \times z_i}{s} \right\rceil \times \left\lceil \frac{h \times z_i}{s} \right\rceil$$

where,

T total space (in bytes) for storing original image, thumbnail and tiles,
O size of the original image (in bytes),
t size of a thumbnail (in bytes),
w width of the original image (in pixels),
h height of the original image (in pixels),
s width, or height, of a square tile (in pixels),
λ size of a square tile (in bytes),
Z set of zoom levels (in percentages) where $Z = \{z_i : z_i > 0 \text{ and } z_i \leq 1\}$

2. Estimation from sample data

Based on sample image files that are representative of the data submitted under TIFF, DICOM and BMP image formats, we have the following sizes in bytes.

Type	Count	Min	1st Quar.	Median	Mean	3rd Quar.	Max
DCM	662	4196872	4196882	4196978	4196950	4196980	4196990
BMP	416	2159674	2180154	2180154	2178776	2180154	2180154
TIFF	86	954558	1196091	1622273	1435718	1665202	1712686
	1164	954558	2180154	4196876	3271669	4196980	4196990

Table 1. File sizes by image type (requires thumbnails and tiles)

Type	Count	Min	1st Quar.	Median	Mean	3rd Quar.	Max
PDF	313	3748070	3749141	3749369	3749832	3749625	3869153

Table 2. File sizes for other media (no thumbnails or tiles)

Min	1st Quar.	Median	Mean	3rd Quar.	Max
5058	8132	8902	9377	9393	27274

Table 3. Thumbnail sizes

Tile width	Min	1st Quar.	Median	Mean	3rd Quar.	Max
256	177	5246	8029	7361	9774	29081
128	165	1586	2250	2101	2705	8243

Table 4. Tile sizes by tile width

Based on the above statistics, and assuming that all of the submitted image files do not deviate far off, we can estimate storage requirements for thumbnails and 256 × 256 image tiles using the median values: $t = 8902$ and $\lambda = 8029$.

The following are representative storage requirements where we pre-generate tiles for each of the zoom levels $Z = \{0.1, 0.25, 0.5, 0.75, 1\}$. Since all of the thumbnails and tiles are in the same common image format, JPEG in our case, we can use the same values of t and λ for all image types.

	DCM	BMP	TIFF
T	4687620	2767144	2177147
O	4196978	2180154	1622273
w	2048	2048	1920
h	1024	1064	1168
e	11%	22%	26%

The value e gives the percentage of total storage required for storing the thumbnails and tiles. This value is directly proportional to the number of tiles generated for each of the zoom levels, and therefore, depends on Z.

3. Recommendation

Based on the conditions highlighted above, it would be safe to allocate approximately 30% extra space for tiles and thumbnails, assuming image files do not deviate too far from the median. Nonetheless, if the allocated space is insufficient, the tiling infrastructure allows horizontal scaling using multiple image servers.

As to estimating the total storage required for storing all of the original media data and the corresponding image tiles and thumbnails, we have limited information to make a firm safe estimate. Nonetheless, a rough estimate would be:

$$R = 1.3 \times 4196876 \times l \times (b + m) \times p \times i$$

where,

R	total space (in bytes) for storing original images, thumbnails and tiles,
l	number of lines,
b	number of baseline specimens per line,
m	number of mutant specimens per line,
p	number of media parameters in IMPReSS, and
i	number of images per parameter.

Hence, assuming that we receive media data for 5000 lines ($l = 5000$), and that each line requires 7 males and 7 females ($m = 14$) with twice as many baselines as mutants ($b = 28$), we would require approximately **26 terabytes** of storage for all of the 25 media parameters ($p = 25$) in IMPReSS, assuming each media parameter is associated with only one image file ($i = 1$).

Caveat: Note here that since not all media parameters are image files (e.g., PDF files), the actual storage requirement is likely to be less than the estimated 26 terabytes. On the other hand, we are not considering other image modalities such as 3D embryo data, segmented or temporal image sets where $i > 1$.