Errata for
‘Potential benefits of delta encoding and data compression for HTTP’

Jeffrey C. Mogul (Digital Equipment Corporation Western Research Laboratory)
250 University Avenue, Palo Alto, CA 94301; mogul@wrl.dec.com

Fred Douglis, **Anja Feldmann**, **Balachander Krishnamurthy** (AT&T Labs - Research)
180 Park Avenue, Florham Park, NJ 07932-0971; {douglis,anja,bala}@research.att.com

The quantitative results presented in our SIGCOMM ’97 paper [1] include numerous minor errors. These errors were caused by programming bugs that led to faulty analyses and simulations, and by inaccurate transcriptions during the preparation of the paper. Here we present corrected figures and tables, as well as corrections to values that appeared in the text of the original paper. The effect of correcting the errors is to reduce the differences between the results based on the proxy trace and those based on the packet-level trace. Our overall conclusions are not significantly altered.

Readers interested in an expanded treatment of the same material may wish to refer to our technical report [2], which now includes the corrections presented here. This report is available online at

http://www.research.digital.com/wrl/techreports/abstracts/97.4.html

In the third paragraph of section 4.2, ‘‘Packet-level trace analysis software,’’ the phrase

> In our traces we saw 1,366,401 requests, of which 26,501 (1.9%) had gaps

should have read:

> In our traces we saw 1,322,463 requests, of which 25,591 (1.9%) had gaps

the sentence

> Another 38,589 (2.8%) of the requests were detected as duplicates created by artifacts of the processing techniques.

should have read:

> Another 43,938 (3.3%) of the requests were detected as duplicates created by artifacts of the processing techniques.

and the sentence

> This left us with 1,080,143 usable responses (79% of the total).

should have read:

> This left us with 1,075,209 usable responses (81% of the total).

In section 5.2, ‘‘Overall response statistics for the packet-level trace,’’ the first two paragraphs should have read:

> The 1,075,209 usable records in the packet-level trace represent the activity of 465 clients, accessing 20,956 servers, referencing 499,608 distinct URLs. Of these references, 77,112 instances (39,628 distinct URLs) contained ‘‘?’’ and are classified as query URLs; these had 8,054 unique prefixes (up to the first ‘‘?’’ character). 52,670 of the instances (28,872 distinct URLs) contained ‘‘cgi’’, and so are probably references to CGI scripts.
>
> The mean request and response header sizes were 281 bytes and 173 bytes, respectively. 818,142 of the responses carried a full body, for a total of 6104 MB of response bodies (mean = 7881 bytes for full-body responses). 145,139 (13.5%) of the responses carried a status code of 304 (Not Modified). We omitted from our subsequent analyses 1,144 full-body responses for which we did not have trustworthy timing data, leaving a total of 816,998 fully-analyzed responses.

In section 5.3, ‘‘Characteristics of responses,’’ the sentence:

> We logged at least two full-body responses for more than half (57%) of the URLs in the trace, but only did so for 18% of the instances.

should have read:

> We logged at least two full-body responses for more than half (57%) of the URLs in the trace, but only did so for 30% of the instances.

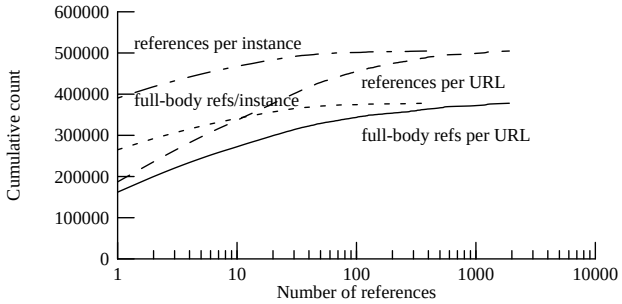

Figure 5-2: Cumulative distributions of reference counts (proxy trace)

In section 5.4, ‘‘Calculation of savings,’’ the sentences:

> In the proxy trace, 142913 of the 377962 status-200 responses (37.8%) were delta-eligible. In the packet-level trace, 83991 of the 819998 status-200 responses (10.2%) were delta-eligible.

should have read:

> In the proxy trace, 113356 of the 377962 status-200 responses (30.0%) were delta-eligible. In the packet-level trace, 83905 of the 816998 status-200 responses (10.3%) were delta-eligible.

The sentence

> In the proxy trace, only 18% of the status-200 responses were excluded from consideration for being identical, compared to 32% for the packet-level trace.

should have read:

> In the proxy trace, only 30% of the status-200 responses were excluded from consideration for being identical, compared to 32% for the packet-level trace.

The sentence

In the packet-level trace, 66% of the status-200 responses were GIF or JPEG images, but only 3.5% of those responses were delta-eligible; in contrast, 25% of the status-200 HTML responses were delta-eligible.

should have read:

In the packet-level trace, 66% of the status-200 responses were GIF or JPEG images, but only 3.0% of those responses were delta-eligible; in contrast, 19% of the status-200 HTML responses were delta-eligible.

The sentence

71446 (50%) of the delta-eligible responses in the proxy trace were text-format responses, as were 54856 (65%) of the delta-eligible responses in the packet-level trace.

should have read:

66413 (59%) of the delta-eligible responses in the proxy trace were text-format responses, as were 52361 (62%) of the delta-eligible responses in the packet-level trace.

In section 5.5, ‘‘Net savings due to deltas and compression,’’ the (garbled) sentence:

(These ‘‘unchanged’’ responses are delta-eligible because their last-modified time has changed, but their

should have read:

(An ‘‘unchanged’’ response is delta-eligible because its last-modified time has changed, although its body has not.)

The third and fourth paragraphs of section 5.5 should have read:

It is encouraging that, out of all of the full-body responses, table 5-1 shows over 30% of the response-body bytes could be saved by using *vdelta* to do delta encoding. This implies that the use of delta encoding would provide significant benefits for textual content-types. It is remarkable that over 83% of the response-body bytes could be saved for delta-eligible responses; that is, in those cases where the recipient already has a cached copy of a prior instance. And while it appears that the potential savings in transmission time is smaller than the savings in response bytes, the response-time calculation is quite conservative (as noted earlier).

For the 88017 delta-eligible responses where the delta was not zero-length, *vdelta* gave the best result 92% of the time. *diff -e* without compression and with compression each was best for about 2% of the cases, and simply compressing the response with *gzip* worked best in 2% of the cases. Just over 1% of the delta-eligible responses were best left alone. The *vdelta* approach clearly works best, but just using *diff -e* would save 52% of the response-body bytes for delta-eligible responses. That is, more than half of the bytes in ‘‘new’’ responses are easily shown to be the same as in their predecessors.

In section 5.5.1, ‘‘Analysis assuming client-applied deltas,’’ the sentence:

However, a much smaller fraction of the responses are delta-eligible at the individual clients (19% instead of 37.8%), and so the overall improvement from delta encoding is also much smaller.

should have read:

However, a much smaller fraction of the responses are delta-eligible at the individual clients (19% instead of 30%), and so the overall improvement from delta encoding is also much smaller.

In section 5.6, ‘‘Distribution of savings,’’ the sentences:

In fact, for delta-eligible responses in the proxy trace, the median number of bytes saved per response by delta encoding using *vdelta* is 3051 bytes (compared to a mean of 5517 bytes). For half of the delta-eligible responses, *vdelta* saved at least 98.5% of the response-body bytes (this includes cases where the size of the delta is zero, because the response value was unchanged).

should have read:

In fact, for delta-eligible responses in the proxy trace, the median number of bytes saved per response by delta encoding using *vdelta* is 2177 bytes (compared to a mean of 4994 bytes). For half of the delta-eligible responses, *vdelta* saved at least 96% of the response-body bytes (this includes cases where the size of the delta is zero, because the response value was unchanged).

In section 5.7, ‘‘Influence of content-type on coding effectiveness,’’ the phrase

for example, delta encoding of changed responses seems to be more effective for ‘‘application/octet-stream’’ resources than for ‘‘text/html’’ resources.

should have read:

for example, delta encoding of changed responses seems to be more effective for ‘‘text/html’’ resources than for ‘‘application/octet-stream’’ resources.

In section 5.8, ‘‘Effect of clustering query URLs,’’, the sentence:

Further, of the 86191 status-200 responses for query URLs, only 28395 (33%) were delta-eligible if the entire URL was used, but 77314 (90%) were delta-eligible if only the prefix had to match.

should have read:

Further, of the 86191 status-200 responses for query URLs, only 28186 (33%) were delta-eligible if the entire URL was used, but 76298 (89%) were delta-eligible if only the prefix had to match.

In section 9, ‘‘Summary and conclusions,’’ the sentence:

We found that, using the best known delta algorithm, for the proxy trace 83% of the delta-eligible response-body bytes and 31% of all response-body bytes could have been saved; at least 39% of the transfer time for delta-eligible responses and 12% of the total transfer time could have been avoided. For the packet-level trace, we showed even more savings for delta-eligible responses (85% of response-body bytes), although the overall improvement (9% of response-body bytes) was much less impressive.

should have read:

We found that, using the best known delta algorithm, for the proxy trace 77% of the delta-eligible response-body bytes and 22% of all response-body bytes could have been saved; at least 37% of the transfer time for delta-eligible responses and 11% of the total transfer time could have been avoided. For the packet-level trace, we showed even more savings for delta-eligible responses (82% of response-body bytes), although the overall improvement (8% of response-body bytes) was much less impressive.

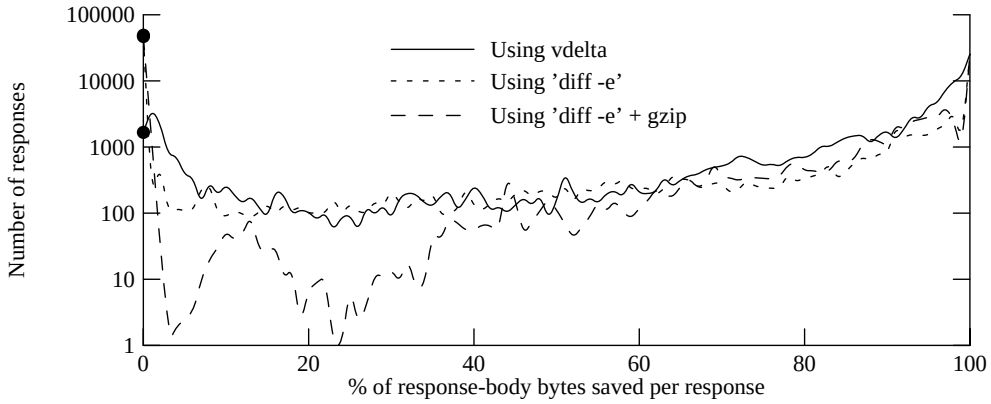

Figure 5-5: Distribution of response-body bytes saved for delta-eligible responses (proxy trace)

	Relative to delta-eligible responses *N* = 113356, 701 MBytes, 160551 seconds						**Relative to all status-200 responses *N* = 377962, 2462 MBytes, 557373 seconds**					
Computation	**Improved references**		**MBytes saved**		**Retrieval time saved**		**Improved references**		**MBytes saved**		**Retrieval time saved**	
unchanged	25339	(22.4%)	145	(20.8%)	11697	(7.3%)	25339	(6.7%)	145	(6.0%)	11697	(2.1%)
diff -e	37806	(33.4%)	215	(30.8%)	23400	(14.6%)	37806	(10.0%)	215	(8.8%)	23400	(4.2%)
diff -e (inc. *unchanged*)	63145	(55.7%)	361	(51.6%)	35098	(21.9%)	63145	(16.7%)	361	(14.8%)	35098	(6.3%)
diff -e | gzip	39800	(35.1%)	264	(37.7%)	32331	(20.1%)	39800	(10.5%)	264	(10.8%)	32331	(5.8%)
vdelta	86825	(76.6%)	394	(56.2%)	47647	(29.7%)	86825	(23.0%)	394	(16.1%)	47647	(8.5%)
vdelta (inc. *unchanged*)	112164	(98.9%)	539	(77.0%)	59344	(37.0%)	112164	(29.7%)	539	(22.0%)	59344	(10.6%)
vdelta compress	75414	(66.5%)	207	(29.6%)	27285	(17.0%)	302739	(80.1%)	832	(34.0%)	104092	(18.7%)
gzip compress	73142	(64.5%)	237	(33.8%)	31567	(19.7%)	289914	(76.7%)	965	(39.4%)	124045	(22.3%)
best algorithm above	112198	(99.0%)	541	(77.2%)	59490	(37.1%)	340845	(90.2%)	1270	(51.9%)	152086	(27.3%)

Table 5-1: Improvements assuming deltas are applied at proxy (proxy trace)

	Relative to delta-eligible responses *N* = 83905, 633 MBytes, 187303 seconds						**Relative to all status-200 responses *N* = 816998, 6193 MBytes, 2053027 seconds**					
Computation	**Improved References**		**MBytes saved**		**Retrieval time saved**		**Improved References**		**MBytes saved**		**Retrieval time saved**	
unchanged	6332	(7.5%)	8	(1.2%)	1459	(0.8%)	6332	(0.8%)	8	(0.1%)	1459	(0.1%)
diff -e	49681	(59.2%)	242	(38.2%)	56485	(30.2%)	49681	(6.1%)	242	(3.9%)	56485	(2.8%)
diff -e (inc. unchanged)	59744	(71.2%)	292	(46.2%)	57943	(30.9%)	59744	(7.3%)	292	(4.7%)	57943	(2.8%)
diff -e | gzip	50467	(60.1%)	280	(44.2%)	70487	(37.6%)	50467	(6.2%)	280	(4.5%)	70487	(3.4%)
vdelta	73483	(87.6%)	467	(73.8%)	100073	(53.4%)	73483	(9.0%)	467	(7.5%)	100073	(4.9%)
vdelta (inc. unchanged)	83546	(99.6%)	517	(81.7%)	101532	(54.2%)	83546	(10.2%)	517	(8.4%)	101532	(4.9%)
vdelta compress	76257	(90.9%)	250	(39.5%)	52424	(28.0%)	597469	(73.1%)	1099	(17.8%)	250822	(12.2%)
gzip compress	72819	(86.8%)	277	(43.8%)	59402	(31.7%)	604797	(74.0%)	1274	(20.6%)	294036	(14.3%)

Table 5-2: Improvements assuming deltas are applied at a proxy (packet-level trace)

References

[1] Jeffrey C. Mogul, Fred Douglis, Anja Feldmann, and Balachander Krishnamurthy. Potential benefits of delta encoding and data compression for HTTP. In *Proc. SIGCOMM '97 Conference*, pp. 181-194. ACM SIGCOMM, Cannes, France, September, 1997.

[2] Jeffrey C. Mogul, Fred Douglis, Anja Feldmann, and Balachander Krishnamurthy. *Potential benefits of delta encoding and data compression for HTTP*. Research Report 97/4, Digital Equipment Corporation Western Research Laboratory, July, 1997.

	Relative to delta-eligible responses *N* = 59550, 296 Mbytes, 105020 seconds						Relative to all status-200 responses *N* = 377962, 2450 MBytes, 557373 seconds					
Computation	**Improved references**		**MBytes saved**		**Retrieval time saved**		**Improved references**		**MBytes saved**		**Retrieval time saved**	
unchanged	16417	(27.6%)	67	(22.8%)	6175	(5.9%)	16417	(4.3%)	67	(2.8%)	6175	(1.1%)
diff -e	23072	(38.7%)	126	(42.9%)	15475	(14.7%)	23072	(6.1%)	126	(5.2%)	15475	(2.8%)
diff -e (inc. *unchanged*)	39489	(66.3%)	194	(65.7%)	21650	(20.6%)	39489	(10.4%)	194	(7.9%)	21650	(3.9%)
diff -e | gzip	24424	(41.0%)	157	(53.3%)	22326	(21.3%)	24424	(6.5%)	157	(6.4%)	22326	(4.0%)
vdelta	42223	(70.9%)	195	(66.0%)	31047	(29.6%)	42223	(11.2%)	195	(8.0%)	31047	(5.6%)
vdelta (inc. *unchanged*)	58640	(98.5%)	262	(88.8%)	37223	(35.4%)	58640	(15.5%)	262	(10.7%)	37223	(6.7%)

Table 5-3: Improvements assuming deltas are applied at individual clients (proxy trace)

Content-type	Delta-eligible Refs	MBytes	Total time	Refs unchanged	Bytes unchanged	Time wasted
All delta-eligible	83905	633	187303	12.0%	7.9%	0.9%
text/html	35066	282	96128	4.2%	3.2%	1.2%
application/octet-stream	31536	238	61743	0.1%	0.0%	0.0%
image/gif	14162	81	22415	52.1%	36.4%	1.6%
image/jpeg	2058	25	4844	50.9%	41.5%	1.8%
text/plain	479	2	422	19.8%	17.9%	6.9%
application/*other*	143	2	473	25.2%	20.3%	3.3%
image/*other*	83	0	115	3.6%	11.5%	0.1%
other or unknown	122	4	1099	3.3%	16.1%	0.6%

Table 5-4: Summary of unchanged response bodies by content-type (packet-level trace)

	All status-200		All delta-eligible			Not including unchanged			All delta-eligible		
Content-type	**Refs**	**MBytes**	**Refs**	**MBytes**	**Total time**	**Refs improved**	**Bytes saved**	**Time saved**	**Refs improved**	**Bytes saved**	**Time saved**
All content-types	816998	6193	83905	633	187303	87.6%	73.8%	53.4%	95.1%	75.0%	54.2%
text/html	184634	1271	35066	282	96128	95.8%	91.8%	60.6%	100.0%	93.6%	61.6%
application/octet-stream	75780	803	31536	238	61743	99.9%	83.6%	63.0%	100.0%	83.6%	63.0%
image/gif	434277	2221	14162	81	22415	45.5%	5.9%	7.1%	71.3%	7.8%	8.6%
image/jpeg	106022	1513	2058	25	4844	49.1%	7.0%	7.2%	99.8%	8.9%	8.9%
text/plain	6988	67	479	2	422	80.0%	73.4%	26.3%	99.6%	84.1%	32.2%
application/*other*	3789	146	143	2	473	64.3%	8.0%	8.2%	89.5%	15.8%	11.0%
application/x-msnwebqt	401	0	256	0	65	100.0%	80.2%	0.5%	100.0%	80.2%	0.5%
image/*other*	2328	9	83	0	115	96.4%	80.0%	12.9%	100.0%	84.0%	13.0%
other or unknown	2509	75	122	4	1099	95.9%	48.6%	72.9%	99.2%	50.9%	73.1%

Table 5-5: Summary of savings by content-type, for *vdelta* (packet-level trace)

Content-type	Refs	MBytes	Total time	Refs improved	Bytes saved	Time saved
All status-200	816998	6193	2053027	72.8%	19.8%	14.3%
image/gif	434277	2221	823214	55.7%	4.6%	3.0%
text/html	184634	1271	520817	99.7%	68.8%	41.5%
image/jpeg	106022	1513	434607	99.1%	2.8%	2.5%
application/octet-stream	75780	803	191968	66.0%	10.3%	12.2%
text/plain	6988	67	19561	95.2%	55.6%	30.1%
application/*other*	3789	146	31665	59.9%	28.8%	13.1%
image/*other*	2328	9	2713	98.6%	47.1%	27.4%
application/x-msnwebqt	401	0	94	99.5%	56.3%	0.4%
video/*	225	88	13083	93.3%	12.6%	11.0%
text/*other*	45	0	68	100.0%	71.7%	38.1%
other or unknown	2509	75	15236	77.0%	35.0%	33.3%

Table 5-6: Summary of *gzip* compression savings by content-type (all status-200 responses in packet-level trace)

	50 Mhz 80486 BSD/OS 2.1				90 MHz Pentium Linux 2.0.0				400 MHz AlphaStn 500/DUNIX 3.2G			
	Text		Non-text		Text		Non-text		Text		Non-text	
Computation	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**	**Mean**	**Std. dev.**
diff -e	72.2	57.2	∅	∅	136.5	134.6	∅	∅	406.9	305.2	∅	∅
gzip	72.9	43.4	56.8	31.5	100.5	78.6	106	78.4	252	151.6	189	139.4
gunzip	145.4	124.2	139.3	110.1	199.6	220.6	218.6	216.4	412.9	563.5	374.9	407.4
both steps above	47.6	31.2	39.6	24.3	64.2	54	70.1	56.8	147.9	103.2	121.8	100.3

Values are in Kbytes/sec., based on elapsed times ∅: not applicable

Corrected rows of **Table 6-1:** Overheads for compression and delta encoding

Computation	Improved References		MBytes saved		Retrieval time saved	
unchanged	9285	(10.8%)	12	(3.2%)	1575	(1.1%)
diff -e	4925	(5.7%)	27	(6.8%)	3437	(2.4%)
diff -e | gzip	5112	(5.9%)	34	(8.8%)	5226	(3.7%)
vdelta	18876	(21.9%)	61	(15.3%)	12217	(8.7%)

N = 86191, 419 MBytes, 141076 seconds

Table 5-7: Improvements relative to all status-200 responses to queries (no clustering)

Comput.	Improved References		MBytes saved		Retrieval time saved	
unchanged	14044	(16.3%)	6	(1.6%)	1145	(0.8%)
diff -e	38890	(45.1%)	97	(24.4%)	9800	(6.9%)
diff -e|gzip	40438	(46.9%)	226	(56.6%)	18015	(12.8%)
vdelta	60711	(70.4%)	262	(65.6%)	24817	(17.6%)
diff -e♣	52934	(61.4%)	103	(25.9%)	10946	(7.8%)
vdelta ♣	74755	(86.7%)	268	(67.2%)	25962	(18.4%)

♣: including unchanged responses

N = 86191, 419 MBytes, 141076 seconds

Table 5-8: Improvements when clustering queries (all status-200 responses to queries)