

Contents

Part I Traditional Deduplication Techniques and Solutions

1	Introduction	3
1.1	Data Explosion	3
1.2	Redundancies	4
1.3	Existing Deduplication Solutions to Remove Redundancies	5
1.4	Issues Related to Existing Solutions	7
1.5	Deduplication Framework	7
1.6	Redundant Array of Inexpensive Disks	8
1.7	Direct-Attached Storage	9
1.8	Storage Area Network	10
1.9	Network-Attached Storage	12
1.10	Comparison of DAS, NAS and SAN	13
1.11	Storage Virtualization	13
1.12	In-Memory Storage	15
1.13	Object-Oriented Storage	16
1.14	Standards and Efforts to Develop Data Storage Systems	16
1.15	Summary and Organization	20
	References	21
2	Existing Deduplication Techniques	23
2.1	Deduplication Techniques Classification	23
2.2	Common Modules	25
2.2.1	Chunk Index Cache	25
2.2.2	Bloom Filter	30
2.3	Deduplication Techniques by Granularity	34
2.3.1	File-Level Deduplication	34
2.3.2	Fixed-Size Block Deduplication	38
2.3.3	Variable-Sized Block Deduplication	44
2.3.4	Hybrid Deduplication	54
2.3.5	Object-Level Deduplication	55
2.3.6	Comparison of Deduplications by Granularity	55

2.4	Deduplication Techniques by Place	56
2.4.1	Server-Based Deduplication	56
2.4.2	Client-Based Deduplication	57
2.4.3	End-to-End Redundancy Elimination	58
2.4.4	Network-Wide Redundancy Elimination	60
2.5	Deduplication Techniques by Time	71
2.5.1	Inline Deduplication	71
2.5.2	Offline Deduplication	73
2.6	Summary	74
	References	75

Part II Storage Data Deduplication

3	HEDS: Hybrid Email Deduplication System	79
3.1	Large Redundancies in Emails	79
3.2	Hybrid System Design	80
3.3	EDMilter	80
3.4	Metadata Server	82
3.5	Bloom Filter	82
3.6	Chunk Index Cache	82
3.7	Storage Server	83
3.8	EDA	83
3.9	Evaluation	85
3.9.1	Metrics	85
3.9.2	Data Sets	86
3.9.3	Deduplication Performance	89
3.9.4	Memory Overhead	92
3.9.5	CPU Overhead	94
3.10	Summary	94
	References	94
4	SAFE: Structure-Aware File and Email Deduplication for Cloud-Based Storage Systems	97
4.1	Large Redundancies in Cloud Storage Systems	97
4.2	SAFE Modules	98
4.3	Email Parser	99
4.4	File Parser	100
4.5	Object-Level Deduplication and Store Manager	103
4.6	SAFE in Dropbox	104
4.7	Evaluation	106
4.7.1	Metrics	107
4.7.2	Data Sets	107
4.7.3	Storage Data Reduction Performance	109
4.7.4	Data Traffic Reduction Performance	109
4.7.5	CPU Overhead	110
4.7.6	Memory Overhead	113

4.8	Summary	113
	References	115

Part III Network Deduplication

5	SoftDance: Software-Defined Deduplication as a Network and Storage Service	119
5.1	Large Redundancies in Network	119
5.2	Software-Defined Network	121
5.3	Control and Data Flow	121
5.4	Encoding Algorithms in Middlebox (SDMB)	124
5.5	Index Distribution Algorithms	125
	5.5.1 SoftDANCE-Full (SD-Full)	125
	5.5.2 SoftDance-Uniform (SD-Uniform)	126
	5.5.3 SoftDANCE-Merge (SD-Merge)	127
	5.5.4 SoftDANCE-Optimize (SD-opt)	128
5.6	Implementation	130
	5.6.1 Floodlight, REST, JSON	130
	5.6.2 CPLEX Optimizer: Installation	130
	5.6.3 CPLEX Optimizer: Run Simple CPLEX Using Interactive Optimizer	135
	5.6.4 CPLEX Optimizer: Run Simple CPLEX Using Java Application (with CPLEX API)	137
5.7	Setup	139
	5.7.1 Experiment	139
	5.7.2 Emulation	140
5.8	Evaluation	140
	5.8.1 Metrics	140
	5.8.2 Data Sets	142
	5.8.3 Storage Space and Network Bandwidth Saving	142
	5.8.4 CPU and Memory Overhead	143
	5.8.5 Performance and Overhead per Topology	145
	5.8.6 SoftDance vs. Combined Existing Deduplication Techniques	147
5.9	Summary	150
	References	151

Part IV Future Directions

6	Mobile De-Duplication	155
6.1	Large Redundancies in Mobile Devices	155
6.2	Approaches and Observations	156
6.3	JPEG and MPEG4	156
6.4	Evaluation	156
	6.4.1 Setup	157

6.4.2	Throughput and Running Time per File Type	158
6.4.3	Throughput and Running Time per File Size	161
6.5	Summary	161
	References	164
7	Conclusions	165
 Part V Appendixes		
	Appendices	169
A	Index Creation with SHA1	171
A.1	sha1Wrapper.h	171
A.2	sha1Wrapper.cc	172
A.3	sha1.h	173
A.4	sha1.cc	177
B	Index Table Implementation using Unordered Map	193
B.1	cacheInterface.h	193
B.2	cache.h	195
B.3	cache.cc	198
C	Bloom Filter Implementation	201
C.1	bf.h	201
C.2	bf.c	202
D	Rabin Fingerprinting Implementation	209
D.1	rabinpoly.h	209
D.2	rabinpoly.cc	211
D.3	rabinpoly_main.cc	216
E	Chunking Core Implementation	219
E.1	chunk.h	219
E.2	chunk_main.cc	221
E.3	chunk_sub.cc	223
E.4	common.h	226
E.5	util.cc	227
F	Chunking Wrapper Implementation	231
F.1	chunkInterface.h	231
F.2	chunkWrapper.h	233
F.3	chunkWrapper.cc	233
F.4	chunkWrapperTest	237

G Sample Programs Using libnetfilter_queue Library 239

 G.1 ndedup.h..... 239

 G.2 ndedup.cc 243

 G.3 ndedup_main.cc..... 255

 References 260

Glossary 261

Data Deduplication for Data Optimization for Storage
and Network Systems

Kim, D.; Song, S.; Choi, B.-Y.

2017, XIII, 262 p. 89 illus., 61 illus. in color., Hardcover

ISBN: 978-3-319-42278-7