

---

# Contents

---

<b>List of Figures</b>	<b>xii</b>
<b>List of Tables</b>	<b>xiii</b>
<b>Glossary</b>	<b>xv</b>
<b>Preface to this edition</b>	<b>xvii</b>
<b>Preface</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 What is a network storage service? . . . . .	1
1.2 Research Motivation . . . . .	2
1.3 Research Statement . . . . .	2
1.4 Dissertation Outline . . . . .	2
<b>2 Background</b>	<b>5</b>
2.1 Introduction . . . . .	5
2.2 The Development of Distributed File Systems . . . . .	5
2.3 Limitations of Today's Distributed File Systems . . . . .	7
2.4 Support of Continuous-Medium Data . . . . .	8
2.5 Support of Structured Data . . . . .	10
2.6 Extensions to the Primary Storage Function . . . . .	11
2.6.1 File Indexing . . . . .	11
2.6.2 Persistent Programming Languages . . . . .	12
2.7 Better Data Placement Strategies . . . . .	13
2.8 Summary . . . . .	15

<b>3</b>	<b>Architectural Framework</b>	<b>17</b>
3.1	Goals	17
3.2	MSSA Entities	17
3.3	Storage Layers	19
3.3.1	Rationale	20
3.3.2	PS Layer	21
3.3.3	LS Layer	22
3.4	Custodes	23
3.5	Containers	24
3.6	Mapping	26
3.7	The Byte Segment Abstraction	27
3.7.1	Data Abstraction	27
3.7.2	Operations	27
3.8	BSC Sessions and Tickets	29
3.8.1	Rationale	29
3.8.2	Session Interactions	29
3.9	Related Work	30
3.9.1	Modular File System Design	31
3.9.2	HLSS and LLSS	31
3.9.3	DataMesh	32
3.10	Summary	33
<b>4</b>	<b>Access control</b>	<b>35</b>
4.1	Protection Requirements	35
4.2	Authorisation	37
4.3	Authentication	38
4.4	Access control in MSSA	39
4.5	The Use of Access Control Lists in MSSA	39
4.6	The Use of Capabilities in MSSA	41
4.7	Summary	47
<b>5</b>	<b>Naming and Related Issues</b>	<b>49</b>
5.1	Textual Names vs Identifiers	49
5.2	Naming	51
5.2.1	Considerations	51
5.2.2	Container and Object Identifiers	52
5.2.3	Generating Object Identifiers	53
5.2.4	Locating Containers and Objects	55
5.2.5	Naming and Value-Adding Clients	55
5.3	Existence Control	56
5.4	Summary	57
<b>6</b>	<b>The Design of a Byte Segment Custode</b>	<b>59</b>
6.1	Introduction	59
6.2	Design Considerations	59
6.2.1	Failure Recovery	59
6.2.2	Failure Recovery in MSSA	60

6.2.3	NVRAM and Atomic Updates . . . . .	60
6.3	NVRAM Transactions . . . . .	62
6.3.1	Overview . . . . .	62
6.3.2	NVRAM Buffer Blocks . . . . .	63
6.3.3	Intention Lists . . . . .	65
6.3.4	Committing Transactions . . . . .	66
6.3.5	Recovery . . . . .	66
6.3.6	A Loose End . . . . .	67
6.4	Metadata . . . . .	67
6.5	Disk Block Allocation . . . . .	74
6.6	Buffering and Disk I/O . . . . .	74
6.7	Other Implementation Details . . . . .	76
6.8	Summary . . . . .	77
<b>7</b>	<b>The Performance of the BSC</b>	<b>79</b>
7.1	Performance . . . . .	79
7.1.1	Best-case Performance . . . . .	79
7.1.2	Performance Cost of Atomic Writes . . . . .	80
7.1.3	Recovery Time . . . . .	83
7.1.4	I/O Throughput . . . . .	83
7.2	Related Work . . . . .	85
7.2.1	Existing systems . . . . .	86
7.2.2	Performance Studies . . . . .	86
7.2.3	LLSS . . . . .	87
7.3	Summary . . . . .	88
<b>8</b>	<b>Rate-Based Sessions: Concept &amp; Interface</b>	<b>89</b>
8.1	Introduction . . . . .	89
8.2	The CFC and the Translator . . . . .	90
8.2.1	The CFC . . . . .	90
8.2.2	The Translator . . . . .	91
8.3	Resource Reservation and Scheduling . . . . .	93
8.3.1	The Need to Reserve Resources . . . . .	93
8.3.2	Rate-based Sessions . . . . .	94
8.3.3	The Difficulties in Resource Reservation . . . . .	94
8.3.4	The Semantics of Rate-based Sessions . . . . .	95
8.4	Rate-Based Sliding Window . . . . .	95
8.4.1	Definition . . . . .	96
8.4.2	Relation with read-ahead scheduling . . . . .	96
8.4.3	Relation with byte segment accesses . . . . .	98
8.4.4	Variable rate sessions . . . . .	98
8.5	Rate-Based Session Interface . . . . .	98
8.5.1	Session Set-up and Shut-down . . . . .	99
8.5.2	Dynamic Window Adjustments . . . . .	100
8.6	Summary . . . . .	102

<b>9</b>	<b>Rate-based Sessions: Prototype Implementation &amp; Evaluation</b>	<b>103</b>
9.1	Prototype Implementation . . . . .	103
9.1.1	Progress Monitoring . . . . .	105
9.1.2	Storage Allocation . . . . .	105
9.1.3	Read-ahead and Write-behind Scheduling . . . . .	105
9.2	Evaluation . . . . .	107
9.2.1	Measured Parameters . . . . .	107
9.2.2	Experimental Setup . . . . .	108
9.2.3	Single Session Performance . . . . .	109
9.2.4	Multiple Session Performance . . . . .	112
<b>10</b>	<b>Conclusion</b>	<b>117</b>
10.1	Summary . . . . .	117
10.2	Further Work . . . . .	119
	<b>Bibliography</b>	<b>120</b>
	<b>Index</b>	<b>131</b>