Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Stages of Workload Model Design 2

1.3 Thesis Direction 4

1.3.1 Original Thesis Goals 4

1.3.2 Trials and Tribulations of Real Data 5

1.3.3 Revised Thesis Direction 6

1.4 Contributions of the Work 8

1.5 Thesis Organization 9

A1 Appendix for Chapter 1 **11**

A1.1 Data Analyst's Survival Guide 11

A1.1.1 Understand all Command Options 11

A1.1.2 Functionality of Hardware Components 12

A1.1.3 Granularity of Measurement 12

A1.1.4 I/O Scheduling 13

A1.1.5 Background Processes 13

A1.1.6 Impact of Data Collection 13

A1.1.7 Limitations of Collection Tools 14

A1.1.8 Resource Interaction 14

A1.1.9 Network Transfers 15

A1.1.10 Plan Carefully 16

2 Background and History **17**

2.1 General Concepts 17

2.2 Interactive Systems 26
2.2.1 Static Studies 27
2.2.2 Dynamic Studies 28
2.3 Distributed Systems 29
2.3.1 File System Studies 31
2.3.2 Network Studies 33
2.4 Classification of Users 36
2.5 Cluster Analysis 39
2.5.1 General Introduction 39
2.5.2 Workload Characterization and Modelling 43
2.6 Focus of this Thesis 44

3 Description and Measurement 47
3.1 Overview of the Study System 47
3.2 System Configuration 48
3.2.1 Processors and Operating Systems 50
3.2.2 Memory and Displays 50
3.2.3 Prestoserve 51
3.2.4 Window environment and login sessions 51
3.2.5 Disks 53
3.2.6 Network 55
3.3 Workload Overview 55
3.3.1 User Workload 56
3.3.2 System Workload 58
3.4 Data Collection Techniques 60
3.4.1 Process Accounting Data Collection 60
3.4.2 Static Data Collection 62
3.4.3 Dynamic Data Collection 62
3.5 Summary 64

A3 Appendix for Chapter 3 67
A3.1 Hardware and Software 67
A3.2 Disk Drives 68

A3.3 CDF Course Information 69
A3.4 Data Collection 70

4 Workload Characterization **75**
4.1 Data Reduction 76
4.2 General Analysis 77
4.2.1 UNIX Load Averages 77
4.2.2 CPU Utilization 79
4.2.3 Memory Usage 80
4.3 Interval Selection 81
4.3.1 Selecting A Day 81
4.3.2 Selecting A Time Interval 82
4.4 Analysis of Selected Interval 84
4.4.1 Disk 84
4.4.2 Memory 85
4.4.3 CPU 85
4.4.4 Network 85
4.5 Feature Set Extraction 86
4.6 Characterization for Model Design 87
4.6.1 Process Behaviour 87
4.6.2 Command Usage 93
4.6.3 User Behaviour 94
4.7 Summary 97

A4 Appendix for Chapter 4 **99**
A4.1 General Analysis 99
A4.1.1 UNIX Load Averages 99
A4.1.2 CPU Utilization 102
A4.1.3 Memory Usage 104
A4.2 Selecting A Time Interval 106
A4.2.1 Resource Intensive Jobs 106
A4.2.2 CPU Usage 107
A4.2.3 Disk Usage 109

A4.2.4 Number of Processes 110
A4.3 Analysis of Selected Interval 111
A4.3.1 Disk 112
A4.3.2 Memory 115
A4.3.3 CPU 118
A4.3.4 Network 121
A4.3.5 NFS 125
A4.4 Command Usage 128

5 Cluster Analysis for Workload Characterization 133
5.1 Scope of the Dissection Study 134
5.2 Clustering Method 134
5.2.1 Scaling 135
5.2.2 SAS FASTCLUS Procedure 135
5.2.3 SAS CLUSTER Procedure 136
5.2.4 Selecting a Clustering Method 137
5.2.5 Number of Clusters 139
5.3 Cluster Variability for Eddie 140
5.3.1 Command Variability 141
5.3.2 User Variability 144
5.4 Cluster Variability for Client Workstations 146
5.4.1 Host Variability 146
5.4.2 Comparison with Eddie 149
5.5 Summary 150

A5 Appendix for Chapter 5 153
A5.1 Clustering Methods 153
A5.1.1 Methods Examined 153
A5.1.2 Choosing a Clustering Method 154
A5.1.3 Reclustering the Large-Resource Usage Cluster 156
A5.2 Number of Clusters 157
A5.3 Workstations 159
A5.4 Marvin 160

6 Model Design **163**

6.1 Model Outline 164

6.1.1 Heterogeneity 164

6.1.2 Scope 165

6.1.3 Outliers 165

6.1.4 Periodic Elements 165

6.2 Model Overview 166

6.2.1 Timing of Components 167

6.2.2 Model Components 168

6.2.3 Identification of Components 170

6.3 Defining User Sessions 170

6.3.1 General Knee Criterion 171

6.3.2 Applying the Knee Heuristic 172

6.4 Classification of Users 174

6.4.1 Clustering Method 175

6.4.2 Workstations 175

6.4.3 Eddie 178

6.5 Classification of Commands 179

6.5.1 Clustering Method 179

6.5.2 Workstations 180

6.5.3 Eddie 181

6.5.4 Handling the Large Resource Command Class 181

6.6 Timing of Components 182

6.6.1 Command Interarrival Distributions 182

6.6.2 User Session Distributions 183

6.7 Summary 183

A6 Appendix for Chapter 6 **185**

A6.1 Periodic Workload Element Identification 185

A6.2 Number of User Classes 191

A6.2.1 Workstations 192

A6.2.2 Eddie 193

A6.3 User Class Dendrograms 195
A6.4 Number of Command Classes 196
A6.4.1 Workstations 200
A6.4.2 Eddie 201
A6.4.3 Distributions for Resource Usage 202
A6.5 Timing of Components 203
A6.5.1 Command Interarrival Distributions 203
A6.5.2 Session Interarrival Distributions 207
A6.5.3 Session Duration Distributions 210

7 Conclusions and Future Work **213**
7.1 Summary 213
7.2 Recommendations 216
7.2.1 General Observations 217
7.2.2 Clustering Generalities 218
7.3 Future Work 219
7.3.1 Applications of the Model 219
7.3.2 Alternative Systems and Workloads 220
7.3.3 Alternative Validation Techniques 221
7.3.4 Examine Dynamic Models 222
7.3.5 Parallel Simulation 222

A7 Appendix for Chapter 7 **225**
A7.1 Dynamic Models 225

Bibliography **229**