

Dirk Heger, Franz Saenger (Hrsg.),  
Karlsruhe

# **Optimale Kopplung von Hochleistungsrechnern**

Reihe **10**: Informatik/  
Kommunikationstechnik Nr. **526**

# Inhalt

## Teil A: Projektübersicht, Grundlagen, Methoden und Werkzeuge

<b>1 Projektübersicht: Ziele und Durchführung .....</b>	<b>1</b>
1.1 Einführung .....	1
1.2 Projektziele .....	9
1.3 Projektdurchführung.....	9
1.3.1 Selektion geeigneter Kommunikationsstrukturen von Hochleistungsrechnern .....	10
1.3.2 Bewertungsmethoden und Werkzeuge .....	16
1.3.3 Untersuchte Rechnerkonfigurationen.....	20
<b>2 Leistungsbewertung mittels Simulation .....</b>	<b>27</b>
2.1 Die Methode .....	27
2.1.1 Was ist Simulation?.....	27
2.1.2 Warum simuliert man? .....	28
2.1.3 Nutzen der Simulation .....	29
2.1.4 Das Vorgehen .....	29
2.1.5 Der Simulationskreislauf .....	30
2.1.6 Die Modellierung .....	31
2.2 Das Simulationswerkzeug OOST.....	37
2.2.1 Werkzeugkonzept.....	37
2.2.2 Der Simulationskern .....	38
2.2.3 Die Modellierung in OOST .....	40
2.2.4 Integrierte Datenbank .....	44
2.2.5 Die Ausführung der Experimente .....	45
2.2.6 Genetische Algorithmen zur Feinabstimmung .....	46
2.2.7 Auswertung .....	49
<b>3 Datenbanken in Hochleistungsrechensystemen .....</b>	<b>51</b>
3.1 Hochleistungsrechensysteme .....	51
3.2 Datenbanken .....	53
3.3 Beispiel .....	57
3.4 Einsatzgebiete .....	57
3.5 SMILE .....	58
<b>4 Gemeinsame Komponenten der untersuchten Rechensysteme .....</b>	<b>60</b>
4.1 Einleitung.....	60
4.1.1 Systemmodell.....	60

4.1.2 Lastmodell.....	61
4.1.3 Bewertung der Leistungsfähigkeit.....	62
4.1.4 Leistungsmaße für Datenbank-Applikationen .....	62
4.1.5 Meßumgebung .....	63
4.2 Verwendete Lasten.....	63
4.2.1 Einsatz realer Lasten (Applikationen).....	63
4.2.2 Vorteile synthetischer Lasten (Benchmarks) .....	64
4.2.3 Transaction Processing Performance Council (TPC).....	64
4.2.4 Benchmarks TPC-A und TPC-B .....	65
4.2.5 Bewertung der Benchmarks TPC-A und TPC-B .....	67
4.2.6 Weitere Benchmarks .....	68
4.2.7 Lasten für spezielle Untersuchungen.....	69
4.3 ORACLE Datenbank-Managementsysteme .....	69
4.4 Der Betriebssystem-Modul Lock Management .....	70
4.4.1 Aufgaben der Sperrverwaltung.....	70
4.4.2 Klassifikation von Lock Management-Varianten.....	71
4.4.3 Distributed Lock Management .....	72
4.4.4 DLM des HIPLEX-Systems .....	73
4.4.5 DLM für Workstation Cluster.....	81
4.4.6 DLM für RM1000 - Konfigurationen .....	84
4.5 Organisation der Messungen.....	85
4.5.1 Meßkonzept.....	85
4.5.2 Qualität der Meßergebnisse .....	88
4.5.3 Organisation der einzelnen Messung.....	90
4.5.4 Organisation des Meßablaufs .....	94

## **Teil B: Leistungsbewertung und -optimierung ausgewählter Rechensysteme**

### **BI: Multiprozessor-/Mehrclustersystem HIPLEX**

<b>5 Messungen an Experimentalkonfigurationen .....</b>	<b>99</b>
5.1 Einleitung und Zielsetzungen.....	99
5.2 HIPLEX-Konzept und Experimentalkonfigurationen .....	100
5.2.1 Hardware der Experimentalkonfigurationen .....	100
5.2.2 Betriebssystem und Distributed Lock Manager .....	101
5.2.3 Datenbank-Managementsystem.....	101
5.2.4 Anwendungs-, Steuerungs- und Meß-Software.....	101
5.2.5 Datenbank .....	102

5.3 Bewertung der HIPLEX-Architektur durch Messungen .....	102
5.3.1 Messungen zur Ermittlung optimaler Konfigurationen .....	102
5.3.2 Messungen mit Oracle Parallel Server .....	104
5.3.3 Messungen an einer Global-Store-Konfigurationen .....	106
5.3.4 Messungen mit dem OMEK-SQL-Benchmark .....	107
5.4 Zusammenfassung und Ausblick.....	109
<b>6 Simulationsexperimente mit DLM-/GS-Varianten .....</b>	<b>110</b>
6.1 Einleitung.....	110
6.2 Modellbeschreibung .....	110
6.2.1 Der synthetische Auftraggeber SAG .....	111
6.2.2 Die Komponente Cluster .....	112
6.2.3 Die Komponente Festplatten .....	113
6.2.4 Die Komponente Global Store.....	113
6.2.5 Die Komponente Inter-Cluster-Kommunikation .....	113
6.2.6 Mögliche Systemengpässe der HIPLEX-Architektur.....	114
6.3 Simulationsexperimente .....	115
6.3.1 Modellverifikation und Modellvalidierung.....	115
6.3.2 Bewertung eines Mehr-Cluster-Systems.....	121
6.3.3 Optimierung eines Mehr-Cluster-Systems .....	128
6.4 Fazit .....	132

## **BII: Workstation Cluster**

<b>7 Vergleich unterschiedlicher Netzkonfigurationen/Simulation eines Shared Database Systems.....</b>	<b>133</b>
7.1 Einleitung.....	133
7.2 Vergleich unterschiedlicher Netzkonfigurationen.....	134
7.2.1 Netzwerktechnologien .....	134
7.2.2 Meßkonfigurationen und Simulationsmodelle .....	136
7.2.3 Workstation-Architektur und Treiberprotokolle.....	136
7.2.4 Benchmarks.....	137
7.2.5 Ergebnisse der Leistungsmessungen und Simulationsexperimente.....	138
7.3 Engpaßanalyse für ATM - Netz.....	144
7.3.1 Fazit.....	149
7.4 Leistungsbewertung eines <i>Shared Database Systems</i> auf einem Workstation Cluster durch Simulation.....	150
7.4.1 Motivation.....	150
7.4.2 Modellstruktur und -komponenten .....	151

7.4.3 Distributed Lock Management .....	153
7.4.4 Simulationsexperimente und -ergebnisse .....	153
7.4.5 Fazit.....	157

### **BIII: Massiv-Parallelrechner RM1000/SMILE**

<b>8 Modellierung und Simulation des MESH-Kommunikationsnetzes.....</b>	<b>158</b>
8.1 Modellsicht .....	158
8.2 Simulationsergebnisse zur Leistungsbewertung .....	163
8.2.1 Wahl der Kommunikationslast.....	163
8.2.2 Einfluß des Kommunikationsnetzes .....	164
8.2.3 Einfluß der Kommunikationslast.....	167
8.2.4 Einfluß der Knotenlage.....	169
8.2.5 Auslastung der Ressourcen .....	171
8.3 Validierung des Simulationsmodells .....	171
8.4 Untersuchung von Entwurfsalternativen.....	173
8.4.1 MESH-Ressourcen.....	174
8.4.2 Bypass-Routing.....	174
8.4.3 Deadlock-beherrschendes-Routing / Torus-Routing .....	175
8.4.4 Durchsatzvergleich 2D Mesh - 3D Mesh .....	179
8.4.5 Wormhole-Routing-Algorithmen.....	180
8.5 MESH-Modell im RM1000/SMILE-Gesamtmodell.....	183
8.6 Messungen an der Experimentalkonfiguration .....	184
<b>9 Messungen an Konfigurationen mit unterschiedlichen Netzknoten.....</b>	<b>185</b>
9.1 Einleitung und Zielsetzungen.....	185
9.2 SMILE-Konzept und Meß-Konfigurationen.....	186
9.2.1 Hardware der RM1000-Konfigurationen (MPP) .....	186
9.2.2 Hardware der SMILE-Konfigurationen (MPP/SMP).....	187
9.2.3 Betriebssystem und Distributed Lock Manager .....	189
9.2.4 Datenbank-Managementsystem.....	189
9.2.5 Anwendungs-, Steuerungs- und Meß-Software.....	190
9.2.6 Datenbank .....	190
9.2.7 Die Anwendungsklasse SAP R/3 .....	191
9.3 Leistungsbewertung durch Messungen.....	193
9.3.1 Messungen an RM1000-Konfigurationen (MPP) .....	194
9.3.2 Messungen an SMILE-Konfigurationen (MPP/SMP) .....	196
9.3.3 Messungen an RM600E-Konfigurationen (SMP).....	197
9.3.4 Leistungsvergleich MPP - MPP/SMP - SMP .....	199

9.3.5 Messungen zum Vergleich OSS - OPS .....	201
9.3.6 Messungen mit Multi-Threaded Server.....	201
9.4 Rückgekoppelte Langzeitmessungen .....	203
9.5 Zusammenfassung und Ausblick.....	206
<b>10 Leistungsprognose für skalierbare MPP-/SMP-Systeme durch</b>	
<b>Simulation .....</b>	<b>207</b>
10.1 Modellbeschreibung .....	207
10.1.1 Die Komponente Disk-Configuration.....	208
10.1.2 Die Komponente Computing Node .....	209
10.2 Simulationsexperimente RM1000.....	210
10.2.1 Ermittlung einer geeigneten Systembelastung .....	210
10.2.2 Ermittlung der notwendigen Simulationsdauer.....	212
10.2.3 Leistungsuntersuchungen bei höherer Rechenkapazität .....	215
10.2.4 Engpaßanalyse und Optimierung bei weniger gut partitionierten Datenbanken .....	217
10.2.5 Skalierbarkeit der RM1000-Architektur bei weiterer Erhöhung der Knotenanzahl.....	220
10.3 Experimente an SMILE-Modellen .....	222
10.3.1 Untersuchung der SMILE-Konfiguration mit lokaler Lasterzeugung.....	223
10.3.2 Untersuchung der SMILE-Konfiguration mit der Lasterzeugung auf den RM1000 Knoten .....	225
 <b>Teil C: Fazit und Anhang</b>	
 <b>11 Fazit .....</b>	<b>227</b>
<b>12 Verwendete Abkürzungen .....</b>	<b>232</b>
<b>13 Literaturverzeichnis.....</b>	<b>234</b>