Spur: Optimizing Resource Allocation for Determined AI

T2000

for the

Bachelor of Science

from the Course of Studies Computer Science

at the Cooperative State University Baden-Württemberg Stuttgart

by

Max Mustermann

12. September 2022

Due Date	June - September
Student ID, Course	1946566, STG-TINF20A
Company	Hewlett Packard Enterprise, Böblingen
Supervisor in the Company	Max Mustermann

Confidentiality Statement

The T2000 on hand

Spur: Optimizing Resource Allocation for Determined AI

contains internal resp. confidential data of Hewlett Packard Enterprise. It is intended solely for inspection by the assigned examiner, the head of the Computer Science department and, if necessary, the Audit Committee at the Cooperative State University Baden-Württemberg Stuttgart. It is strictly forbidden

- to distribute the content of this paper (including data, figures, tables, charts etc.) as a whole or in extracts,
- to make copies or transcripts of this paper or of parts of it,
- to display this paper or make it available in digital, electronic or virtual form.

Exceptional cases may be considered through permission granted in written form by the author and Hewlett Packard Enterprise.

Stuttgart, 12. September 2022

Max Mustermann

Author's declaration

Hereby I solemnly declare:

1. that this T2000, titled *Spur: Optimizing Resource Allocation for Determined AI* is entirely the product of my own scholarly work, unless otherwise indicated in the text or references, or acknowledged below;
2. I have indicated the thoughts adopted directly or indirectly from other sources at the appropriate places within the document;
3. this T2000 has not been submitted either in whole or part, for a degree at this or any other university or institution;
4. I have not published this T2000 in the past;
5. the printed version is equivalent to the submitted electronic one.

I am aware that a dishonest declaration will entail legal consequences.

Stuttgart, 12. September 2022

_____

Max Mustermann

Abstract

Building Machine Learning (ML) models requires significant domain knowledge, and training such models is a time-consuming and resource-intensive process. Determined AI is a deep learning training platform that simplifies and automates the model building process and accelerates the training process by seamlessly applying distributed training with multiple Graphics Processing Units (GPUs). In this project, we propose Spur to optimize the resource allocation (GPU and CPU) for different distributed training strategies (e.g., AllReduce and ParameterServer) to further accelerate the training performance and improve resource efficiency.

Contents

Acronyms

ML	Machine Learning
GPU	Graphics Processing Unit
CPU	central processing unit

List of Figures

List of Tables

1 Introduction

(**Alexander.2007**)

Alexander.2007

2 Materials and Methods

3 Results

4 Discussion

5 Conclusion and Implication

Appendix

A. Figures

B. Tabels

A Figures

B Tabels