1. Record Nr. UNINA9910633930803321

Titolo	Network and parallel computing : 19th IFIP WG 10.3 International Conference, NPC 2022, Jinan, China, September 24-25, 2022 proceedings. / / Shaoshan Liu, Xiaohui Wei, editors
Pubbl/distr/stampa	Cham, Switzerland : , : Springer, , [2022] ©2022
ISBN	3-031-21395-5
Descrizione fisica	1 online resource (360 pages)
Collana	Lecture notes in computer science ; ; 13615
Disciplina	004.6
Soggetti	Computer networks Parallel computers Parallel processing (Electronic computers)
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di bibliografia	Includes bibliographical references and index.
Nota di contenuto	Intro -- Preface -- Organization -- Contents -- Architecture -- A Routing-Aware Mapping Method for Dataflow Architectures -- 1 Introduction -- 2 Background and Related Works -- 2.1 Dataflow Architecture -- 2.2 Related Works -- 3 Motivation -- 4 Our Method -- 5 Evaluation -- 5.1 Methodology -- 5.2 Performance Improvement -- 5.3 Energy Saving -- 5.4 Scalability -- 5.5 Compilation Time -- 6 Conclusion -- References -- Optimizing Winograd Convolution on GPUs via Partial Kernel Fusion -- 1 Introduction -- 2 Background -- 2.1 Implementations of Convolution -- 2.2 Winograd Convolution -- 2.3 NVIDIA GPU Architecture and Tensor Cores -- 3 Related Work -- 4 Methodology -- 4.1 Optimizing EWMM Stage -- 4.2 PKF (Partial Kernel Fusion) -- 5 Implementation and Experiment -- 5.1 Implementation PKF on TVM -- 5.2 Experiment -- 6 Conclusion -- References -- Adaptive Low-Cost Loop Expansion for Modulo Scheduling -- 1 Introduction -- 2 Expanded Modulo Scheduling -- 2.1 Data Dependence Graph -- 2.2 Expansion Count and Iteration Interval -- 2.3 Scheduling -- 2.4 Resolving Expansion Faults -- 2.5 Completing the MRT -- 3 Performance Evaluation -- 3.1 Target Architecture -- 3.2 Adaptation of EMS -- 3.3 Experiment Setup -- 3.4 Experiment Results -- 4 Conclusion -- References -- SADD: A Novel Systolic Array

Accelerator with Dynamic Dataflow for Sparse GEMM in Deep Learning -- 1 Introduction -- 2 Background -- 2.1 Dataflows in the Systolic Array -- 2.2 Sparsity -- 3 SADD Architecture -- 3.1 Group-Structure-Maintained Compression -- 3.2 The SIS and SWS -- 3.3 The Performance of SIS and SWS with Different GEMM Sizes -- 3.4 The SDD and SADD -- 4 Experimental Results -- 4.1 Experimental Setup -- 4.2 Performance Comparison of Different Dataflows -- 4.3 Comparison of the SADD and the TPU -- 4.4 Scalability Analysis -- 4.5 Hardware Cost Analysis -- 5 Related Work -- 6 Conclusion.
References -- CSR&amp -- RV: An Efficient Value Compression Format for Sparse Matrix-Vector Multiplication -- 1 Introduction -- 2 The Compressed Sparse Row and Repetition Value Format -- 2.1 CSR&amp -- RV Representation -- 2.2 SpMV Algorithm -- 3 Experimental Results -- 3.1 Performance Comparison -- 3.2 Memory Overhead -- 3.3 Pre-processing -- 4 Conclusion -- References -- Rgs-SpMM: Accelerate Sparse Matrix-Matrix Multiplication by Row Group Splitting Strategy on the GPU -- 1 Introduction -- 2 Related Work and Motivation -- 3 Rgs-SpMM Design -- 3.1 Data Organization in Rgs-SpMM -- 3.2 Row Group Splitting -- 4 Experiment Evaluations -- 4.1 Overall Performance -- 4.2 Analysis of Results -- 5 Conclusion -- References -- Cloud Computing -- Interference-aware Workload Scheduling in Co-located Data Centers -- 1 Introduction -- 2 Related Work -- 3 Interference-aware Solution -- 3.1 Performance Interference Metric -- 3.2 Performance Interference Model Based on Linear Regression -- 3.3 Interference-aware Workload Scheduling -- 4 Experiment and Evaluation -- 4.1 Prediction Accuracy of Performance Interference Models -- 4.2 Evaluation of Scheduling Strategies on Throughput -- 5 Conclusion -- References -- FaaSPipe: Fast Serverless Workflows on Distributed Shared Memory -- 1 Introduction -- 2 Related Work -- 3 Design and Implementation -- 3.1 The PipeFunc Programming Model -- 3.2 System Architechture of FaaSPipe -- 3.3 Intra-workflow Memory Sharing -- 3.4 Full-Duplex Memory Transfer -- 4 Evaluation -- 4.1 Distributed Word Count -- 4.2 LightGBM -- 4.3 Efficency of FaaSPipe vs. Faasm -- 5 Conclusion -- References -- TopKmer: Parallel High Frequency K-mer Counting on Distributed Memory -- 1 Introduction -- 2 Background -- 2.1 Parallel K-mer Counting -- 2.2 Heavy Hitters -- 3 TopKmer Counter -- 3.1 Multi-layer Hash Table -- 3.2 Insert -- 3.3 Query.
4 Parallel K-Mer Counting Framework -- 5 Results -- 5.1 Experiment Setup -- 5.2 Quality of Counting -- 5.3 Performance Comparison -- 5.4 Scaling Capability -- 5.5 Time Consumption Analysis -- 6 Conclusion -- References -- Flexible Supervision System: A Fast Fault-Tolerance Strategy for Cloud Applications in Cloud-Edge Collaborative Environments -- 1 Introduction -- 2 Flexible Supervision System Architecture -- 3 Fault Detection and Fault-Tolerance Strategy -- 4 Experimental Evaluation -- 5 Related Work -- 6 Conclusions and Future Work -- References -- Adjust: An Online Resource Adjustment Framework for Microservice Programs -- 1 Introduction -- 2 A QoS Awareness Framework for Microservices -- 2.1 Microservice Analyzer (MSA) -- 2.2 Microservice Prediction Model (MSPM) -- 2.3 Microservice Performance Guarantor (MSPG) -- 3 Evaluation -- 3.1 Performance Guarantee -- 3.2 Resource Re-collection -- 4 Conclusion -- References -- Cloud-Native Server Consolidation for Energy-Efficient FaaS Deployment -- 1 Introduction -- 2 Key Design Considerations -- 3 DAC Design -- 3.1 System Overview -- 3.2 Function Classifier -- 3.3 Consolidation Controller -- 4 Evaluation -- 4.1 Methodologies -- 4.2 Evaluation Results -- 5 Conclusion -- References -- Deep Learning -- NeuProMa: A Toolchain for Mapping Large-Scale Spiking Convolutional Neural Networks onto Neuromorphic Processor -- 1 Introduction -- 2

Background -- 2.1 Neuromorphic Processor -- 2.2 Spiking Convolutional Neural Network -- 3 Related Work -- 4 NeuProMa -- 4.1 Splitting -- 4.2 Partitioning -- 4.3 Mapping -- 5 Experiment Setup -- 5.1 Experiment Platform -- 5.2 Evaluated SCNNs -- 6 Experiment Results -- 6.1 Splitting Performance -- 6.2 Partitioning and Mapping Performance -- 7 Conclusion -- References -- Multi-clusters: An Efficient Design Paradigm of NN Accelerator Architecture Based on FPGA -- 1 Introduction.
2 Background -- 2.1 Design Patterns of Accelerator -- 2.2 Related Work -- 3 Overall Method -- 3.1 Division Method -- 3.2 Architecture Design -- 3.3 Design Space Exploration -- 3.4 Scheduling Strategy -- 4 Experiment -- 4.1 Experiment Setup -- 4.2 Comparison with CPU and GPU -- 4.3 Comparison with Previous FPGA Accelerators -- 5 Conclusion -- References -- TrainFlow: A Lightweight, Programmable ML Training Framework via Serverless Paradigm -- 1 Introduction -- 2 Background and Challenges -- 2.1 Distributed ML Training -- 2.2 Challenges -- 3 TrainFlow Design -- 3.1 Overview -- 3.2 Serverless Process Model and Training Basics -- 3.3 Programmability Extension with Event-Driven Hook -- 4 Implementation -- 5 Evaluation -- 5.1 Availability -- 5.2 Programmability -- 6 Related Work -- 6.1 Classic ML Training -- 6.2 Serverless ML Training -- 7 Conclusion -- References -- DRP:Discrete Rank Pruning for Neural Network -- 1 Introduction -- 2 Related Work -- 2.1 Compression Techniques -- 2.2 Sparse Method -- 2.3 Structured Pruning -- 3 Consideration Bias Sparsity -- 4 Discrete Rank Pruning -- 5 Experiment and Evaluation -- 5.1 Datasets and Network Models -- 5.2 Implementation -- 5.3 Results and Analysis on CBS -- 5.4 Results and Analysis on DRP -- 6 Conclusion -- References -- TransMigrator: A Transformer-Based Predictive Page Migration Mechanism for Heterogeneous Memory -- 1 Introduction -- 2 TransMigrator -- 2.1 Design of Neural Network -- 2.2 Page Migration -- 3 Evaluation and Analysis -- 3.1 Trace Collection -- 3.2 Network Training -- 3.3 Migration Simulation -- 3.4 Access Time -- 3.5 Energy Consumption -- 3.6 Network Overhead -- 4 Related Work -- 5 Conclusion -- References -- Hardware Acceleration for 1D-CNN Based Real-Time Edge Computing -- 1 Introduction -- 2 Background -- 2.1 State-of-the-Art CNN Accelerators -- 2.2 CNN in Real-Time Computing.
3 Proposed Architecture for 1D-CNN -- 3.1 Data Reuse -- 3.2 Accelerated 1D-CNN Architecture -- 3.3 Compiler for 1D-CNN Architecture Generation -- 4 Results -- 4.1 Setup -- 4.2 Evaluations of Power, Latency and Bandwidth -- 4.3 Comparative Analysis -- 5 Conclusion -- References -- Emerging Applications -- DLC:
An Optimization Framework for Full-State Quantum Simulation -- 1 Introduction -- 2 Background and Related Work -- 2.1 Quantum States and Quantum Circuits -- 2.2 Full-State Quantum Simulator -- 2.3 Related Work -- 3 Framework Overview -- 4 CPU-GPU Locality Enhancement -- 4.1 Data Dependency Analysis -- 4.2 CPU-GPU Locality Enhancement -- 5 Communication Optimization Among Multi-GPU -- 5.1 Challenges of Multi-GPU -- 5.2 Communication Scheme -- 5.3 Optimization of Communication -- 6 Performance Evaluation -- 6.1 Environment Setup -- 6.2 Performance on Single Node -- 6.3 Performance on Multiple Nodes -- 7 Conclusion -- References -- Approximation Algorithms for Reliability-Aware Maximum VoI on AUV-Aided Data Collections -- 1 Introduction -- 2 Related Works -- 2.1 AUV-Aided Data Collection -- 2.2 Orienteering Problem and Variants -- 3 System Model and Problem Definition -- 3.1 System Model -- 3.2 Problem Definition -- 4 Approximation Algorithm for the Path Finding Problem -- 4.1 Approximation Algorithm for the Path Finding Problem

Without Real-Time VoI Decay -- 4.2 Approximation Algorithm for the Path Finding Problem with Real-Time VoI Decay -- 5 Simulation and Performance Evaluation -- 6 Conclusion -- References -- CCSBD: A Cost Control System Based on Blockchain and DRG Mechanism -- 1 Introduction -- 2 Related Work -- 3 System Design -- 3.1 System Overview -- 3.2 Medical Evidence-Based Classification Model -- 3.3 Contract Strategy for Clinical Data Sharing -- 4 Evaluation -- 5 Discussion -- 6 Conclusion -- References.
Number of UAVs and Mission Completion Time Minimization in Multi-UAV-Enabled IoT Networks.