Overview

The following describes the Huawei AI full-stack solution and the position of MindSpore in the solution. Developers who are interested in MindSpore can visit the MindSpore community and click Watch, Star, and Fork.

Introduction to MindSpore

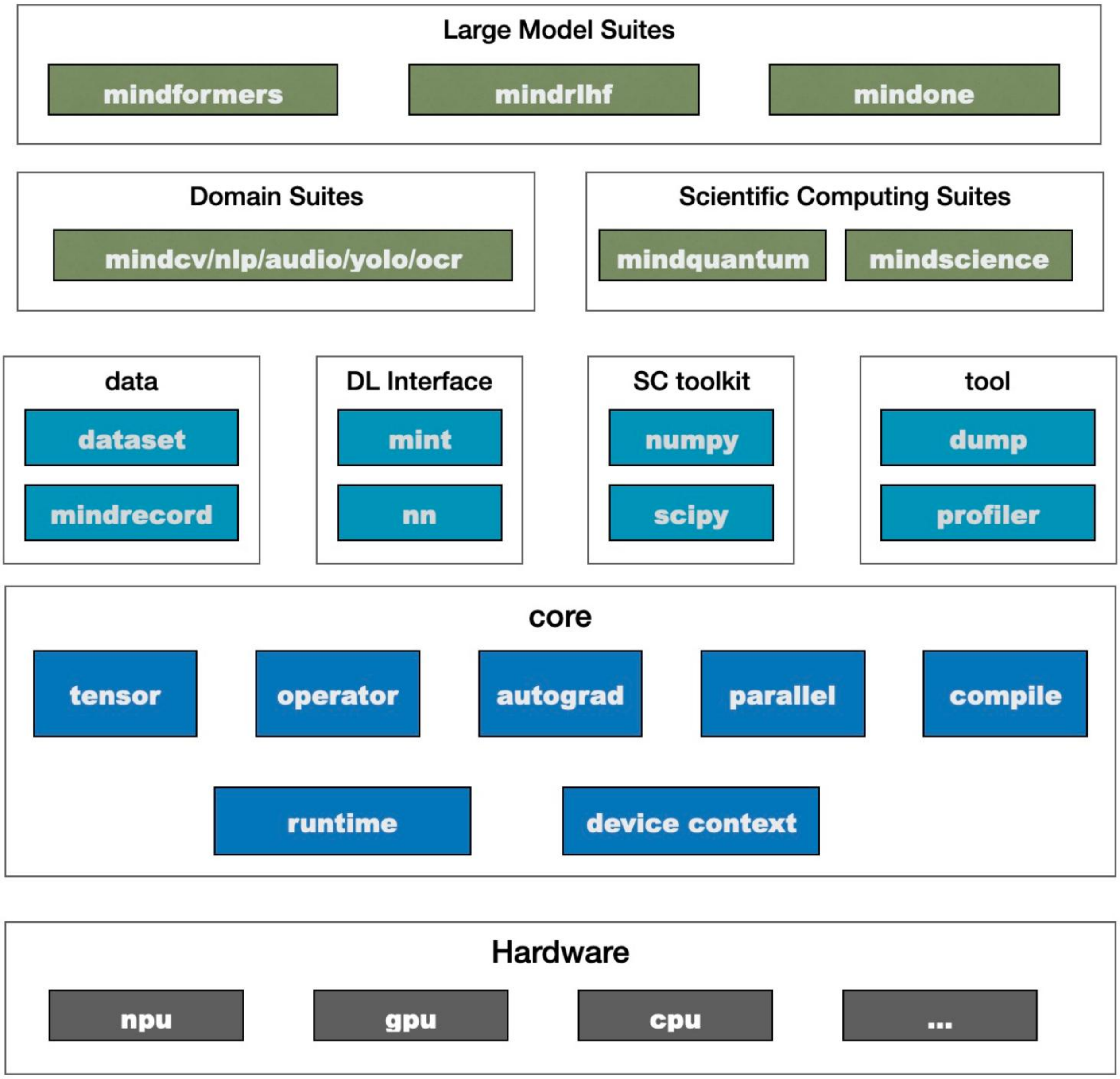
Overall Architecture

The overall architecture of MindSpore is as follows:

MindSpore Model Suite Layer: Provides developers with ready-to-use models and development kits, such as the large model suite MindSpore Transformers, MindSpore ONE, and scientific computing libraries for hot research areas;
•Domain Suites: Computer Vision/Natural Language Processing/Speech/Detection/OCR and other model suites
•Scientific Computing Suites: Quantum Computing/Fluid Dynamics/Chemistry and other model suites
• Large Model Suites: LLM/Diffusion/DPO and other large language generative model suites

MindSpore DL+Scientific Computing Interface Layer: Provides developers with various Python interfaces required for AI model development, maximizing compatibility with developers' habits in the Python ecosystem;
•Data: High-performance data processing engine
Deep Learning Interface: Common neural network encapsulation layer
•Scientific Computing Interface: Encapsulation layer for common NumPy/SciPy functions
•Tools: Debugging and optimization dump data/performance analysis profiler interface

MindSpore Core: As the core of the AI framework, it builds the Tensor data structure, basic operation operators, autograd module for automatic differentiation, Parallel module for parallel computing, compile capabilities, and runtime management module.
•tensor: Basic high-dimensional tensor data structure
•operator: Basic computational unit operator
•autograd: Automatic differentiation
•parallel: Parallel basic module, including operator/optimizer/pipeline capabilities
compile: containing AST/bytecode graph construction/computation optimization capabilities
•runtime: Runtime management module
•device context: Device management module



Design Philosophy

MindSpore is a full-scenario deep learning framework designed to achieve three major goals: easy development, efficient execution, and unified deployment across all scenarios. Easy development is reflected in API friendliness and low debugging difficulty; efficient execution includes computational efficiency, data preprocessing efficiency, and distributed training efficiency; full-scenario means the framework simultaneously supports cloud, edge, and device-side scenarios.