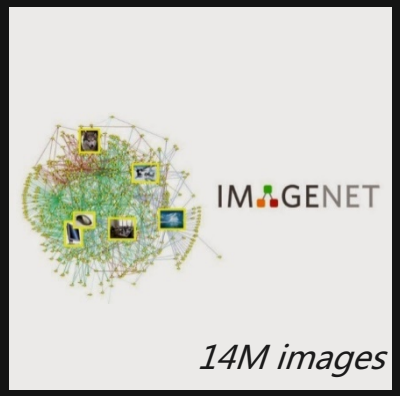
# Open Platform and Tools for AI

Tingting Qin | System Research Group, MSRA

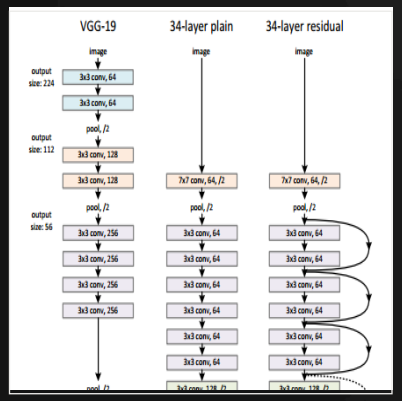
01/28/2021

# The bottlenecks of DL systems



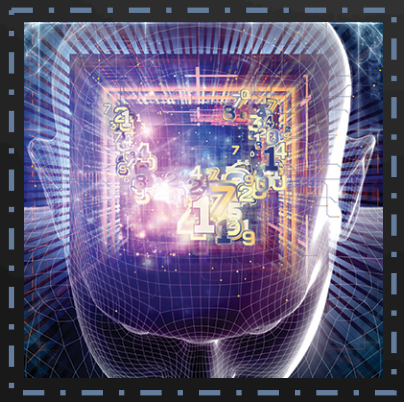

Big Data

- Large size
- Various in type
- Load & Share



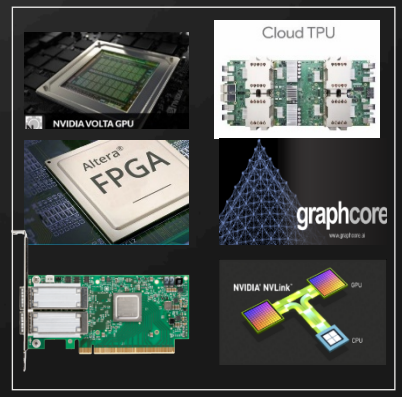

Advanced Model

- Easy to develop
- Fast model iteration
- Fast training & inference



Deep Learning systems

- Customizable
- Scalable
- Intelligent in System Level
- Performance Optimization and Validation




Hardware acceleration and high-speed network

- Powerful computing power
- Fast network communication

# Microsoft Management and Intelligence Platform

## 深度学习智能探索

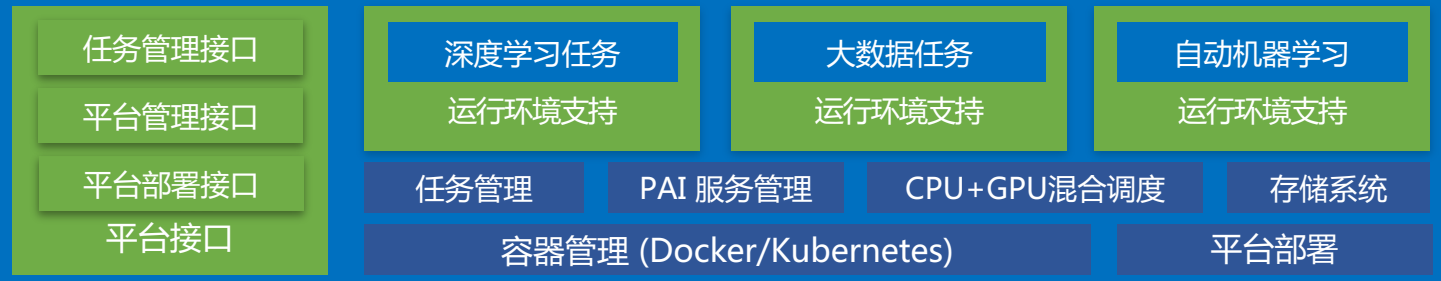
- 全新的定义搜索空间的语言
- 面向开发者的更友好的接口
- 可扩展的最新调参、NAS算法的支持

Neural Network Intelligence


## 异构集群管理调度

- 隐藏各种深度学习框架的复杂性，
- 界面友好，支持各种工具集成
- 减少资源冗余

Open Platform for AI

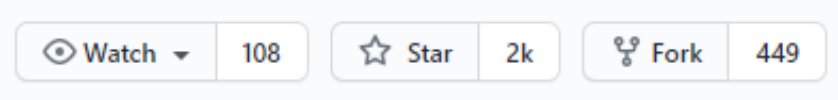



## 系统级智能优化

- 针对深度学习模型训练的全栈式优化
- 高效的资源调度和管理系统
- 提高系统利用率

# Open Platform for AI (OpenPAI)

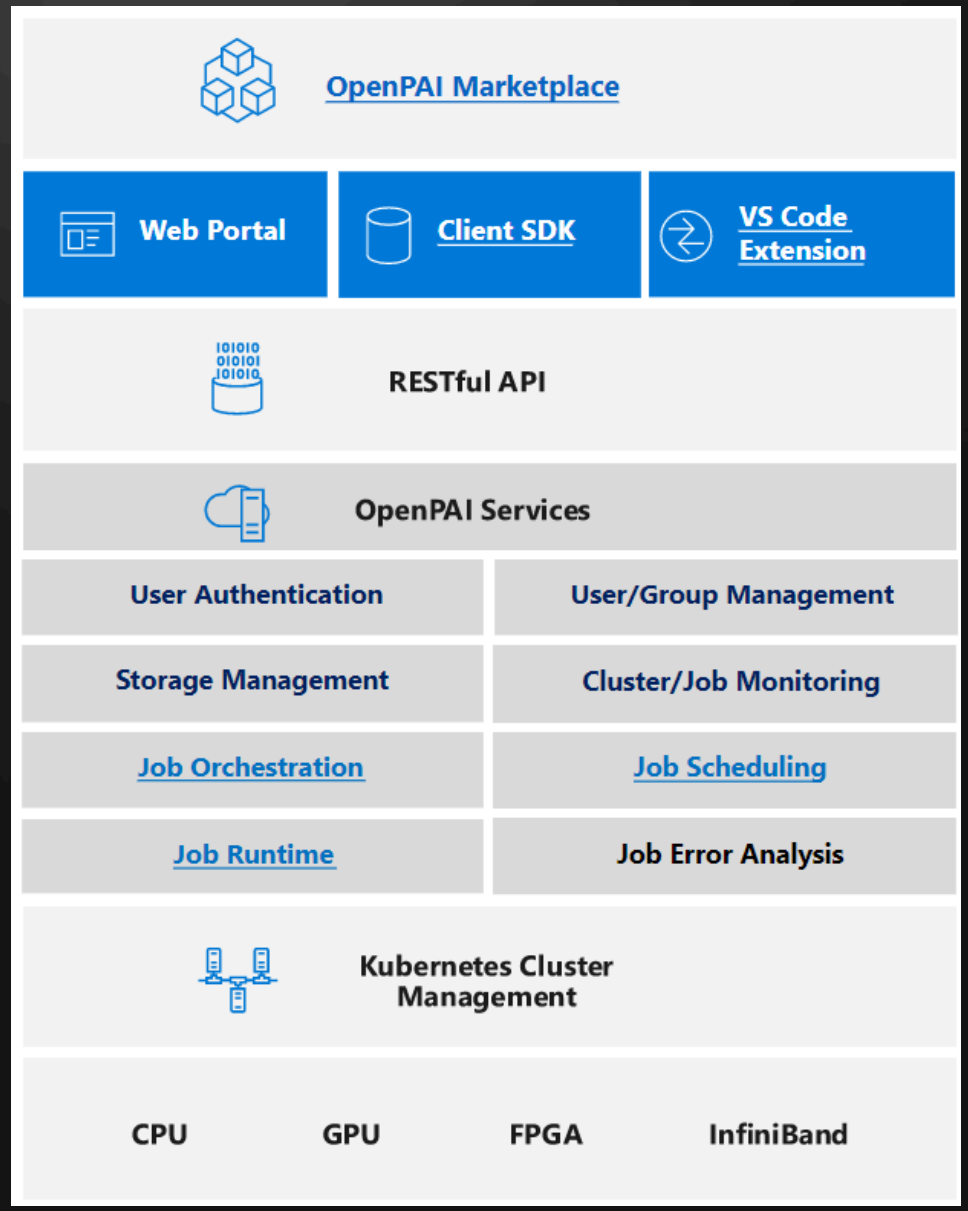
- GitHub Repo:
  https://github.com/Microsoft/pai
- Docs:
  https://openpai.readthedocs.io/
- Snapshot on 1/26/2021




build passing chat on gitter release v1.4.1

OpenPAI v1.5.0 has been released!

With the release of v1.0, OpenPAI is switching to a more robust, more powerful and lightweight architecture. OpenPAI is also becoming more and more modular so that the platform can be easily customized and expanded to suit new needs. OpenPAI also provides many AI user-friendly features, making it easier for end users and administrators to complete daily AI tasks.



# When to Consider OpenPAI

- Need to share powerful AI computing resources (GPU/FPGA farm, etc.) among teams
- Needs to share and reuse common AI assets like Model, Data, Environment, etc
- Needs an easy IT ops platform for AI
- Want to run a complete training pipeline in one place

# Why Choose OpenPAI

## Easy to Deploy

- Full stack solution
- Deployable in heterogeneous environments
  - On-premises
  - Hybrid
  - Public cloud
  - Single-box
- Simple deployment scripts

## Easy to Use

- Containerized AI Platform
- Unified AI job training Standard
- Pre-built docker for popular AI frameworks
- Marketplace for learning & asset sharing

## Easy to Monitor

- Friendly Web portal
- Server monitoring, load , log management
- Job status track
- Multiple-dimension’ s resource usage view
  - Cluster level
  - Node level
  - Job level
  - … …

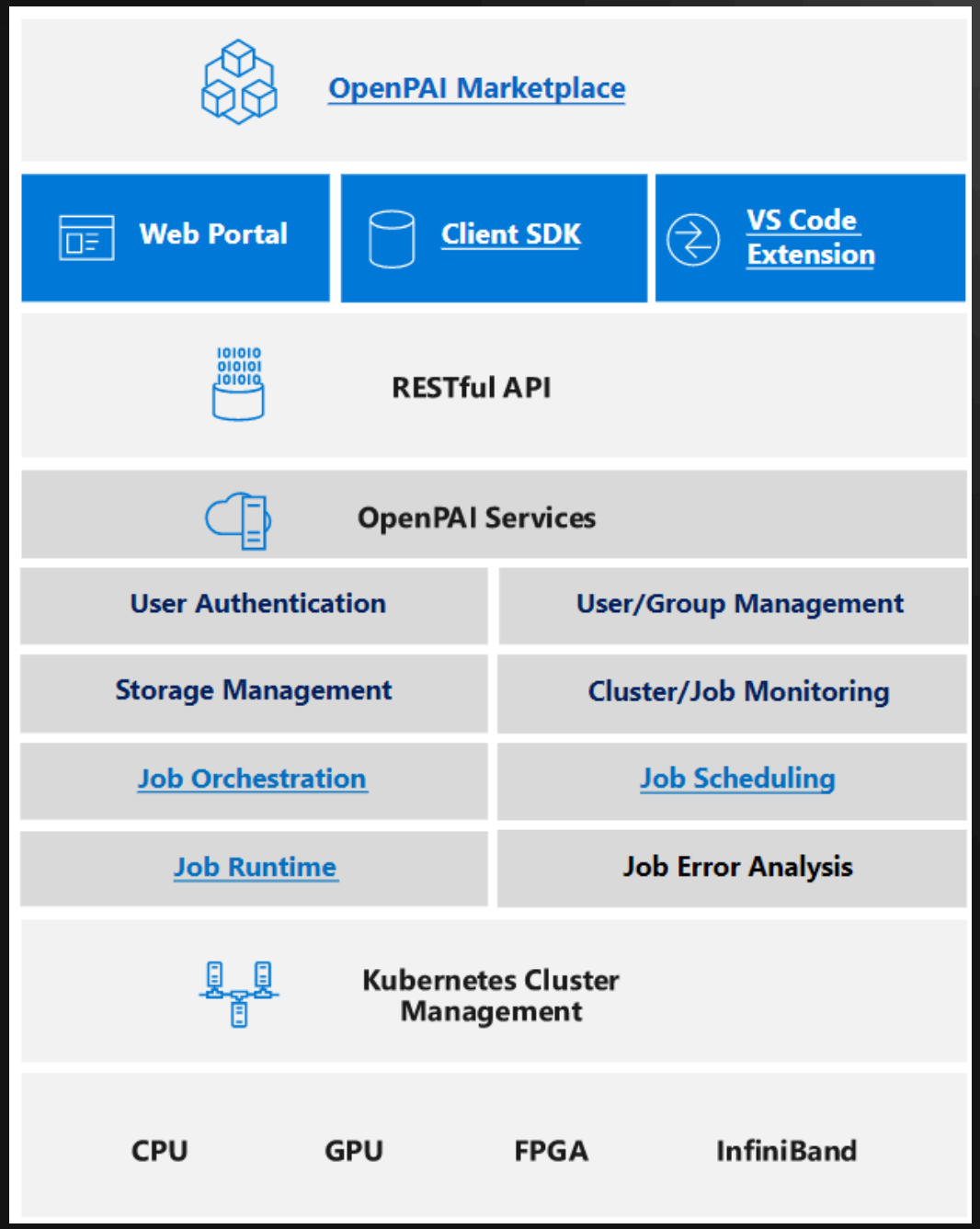
## Easy to Extend

- Most complete solution for DL
- Modular architecture
  - Different module can be plugged in as appropriate
- Compatible with Kubernetes eco-system

OPEN Compute Project®

Microsoft

# OpenPAI Architecture




Evolving on top of reproducible state-of-the-art results

Experiments optimized for researchers, and data scientists

High performance, flexible and easy to extend, best use of GPU resources.

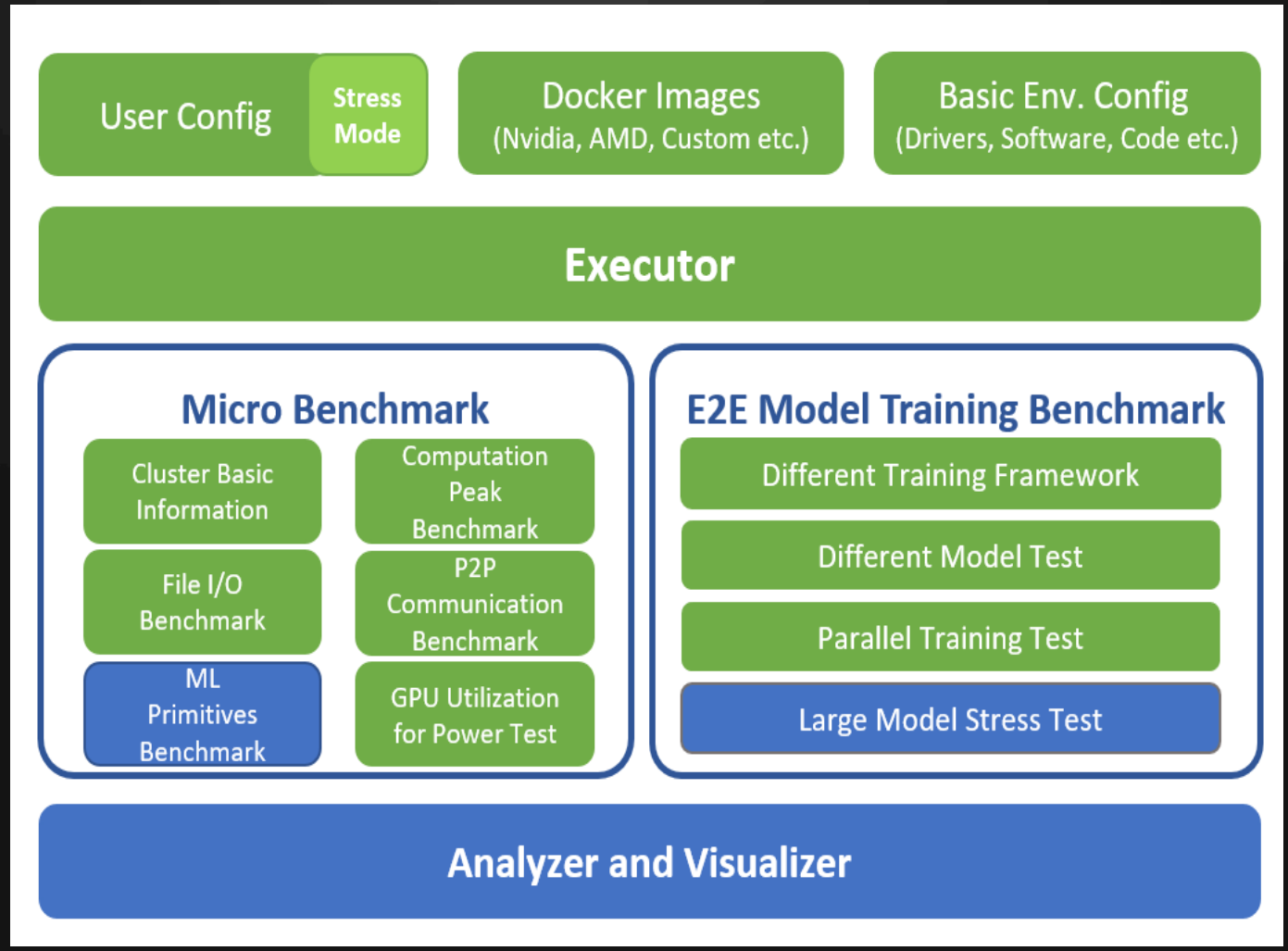
- *general workload supporting*
- *batch training*
- *elastic offline inference*
- *extensible plugins*

Cloud, on-premise and hybrid

Heterogeneous computing resources

# Hardware Validation and Benchmarking

- A Benchmark framework to compare & better understand DL software & hardware
  - Micro-benchmark for computation & communication
  - Domain-aware E2E DL workloads
- Provide guidance for better resource allocation & utilization
  - Checking hardware performance
  - Deployment decisions
  - Arranging AI workloads with best-fit hardware
- An open source & extensible benchmark
  - Accept new DL hardware and software



# Collaboration

Microsoft

Microsoft

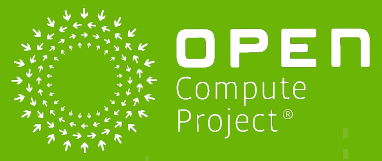
INESA 雲賽 上海仪电

ADVANTECH

Enflame 燧原科技

Contact Us:

https://github.com/Microsoft/pai




Microsoft

# Thanks!

## Q & A