Оценочное тестирование эффективности вычислительных узлов суперкомпьютера cHARISMa для задач глубокого обучения

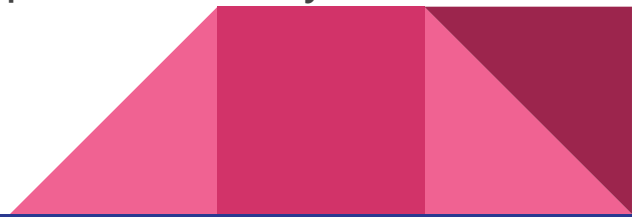
Выполнили: Писарев В.В., Промыслов Г.В.,Тимофеев А.В.

Национальный исследовательский университет «Высшая школа экономики»,
Объединенный институт высоких температур РАН

Введение

1. Важность суперкомпьютерных систем в решении задач машинного обучения.
2. Ключевой показателей производительности - вычислительная эффективность выполнения поставленных задач.
3. Стандартные подходы неэффективны при расчетах конкретных вычислительных задач.

В связи с этим появляются альтернативные оценки производительности, в том числе при помощи бенчмарков приближенных к реальным научным приложениям.

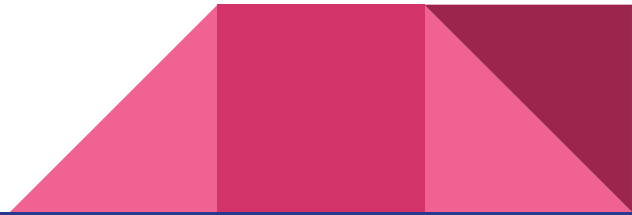


Задачи

Была подготовлена методика оценивания эффективности суперкомпьютеров и их компонент при работе с задачами машинного обучения.

Будет проверено, что принадлежность к определенному классу научных задач приведет к похожим результатам производительности на различных системах

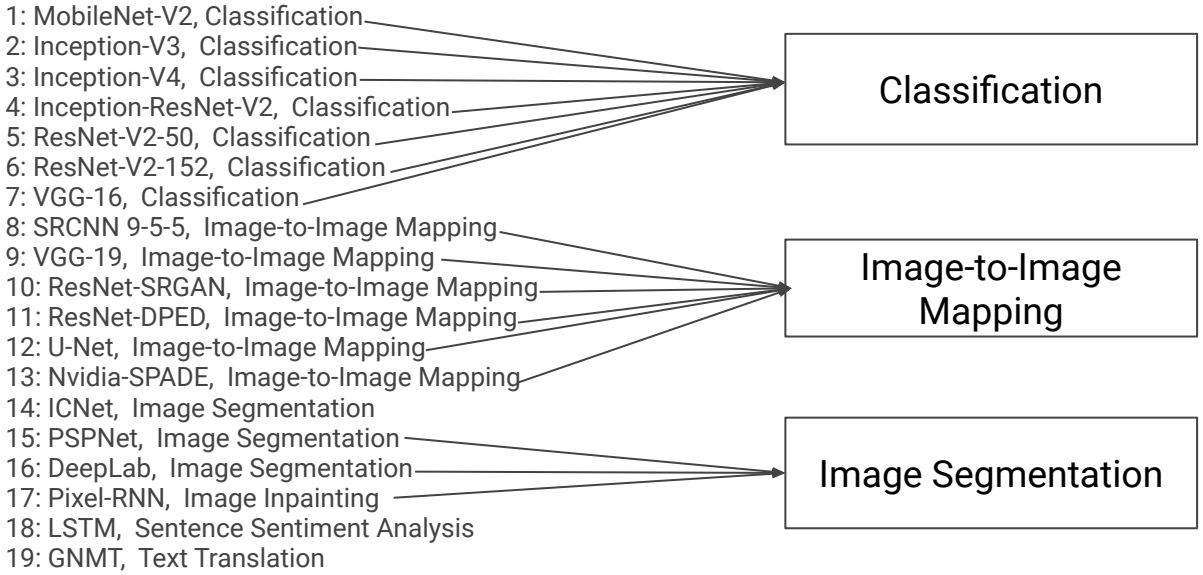
Будут определены наилучшие системы для работы с задачами Machine Learning



Описание метода

На текущем этапе нами используется база бенчмарка ai-benchmark, позволяющая оценить эффективность расчетов на примеры работы 19 нейронных сетей на стадиях обучения и работы, что покрывает большинство архитектур глубокого обучения, используемых в современных задачах.

Используемые модели



Описание архитектуры

На текущем этапе расчет проводился на графических ускорителях узлов суперкомпьютере cHARISMa обладающего следующими спецификациями:

Количество узлов/CPU/Ядер/GPU/Ядер GPU	47/94/2616/166/954880
Модель процессора	Intel Xeon Gold / AMD EPYC
Модель GPU	116 x NVIDIA V100 32 ГБ SXM (NVLink) 48 x NVIDIA A100 80 ГБ SXM (NVLink) 2 x NVIDIA H100 80 ГБ PCIe
Оперативная память	52.9 ТБ
Параллельная система хранения данных	Lustre 848 ТБ
Тип вычислительной сети	InfiniBand EDR (2x100 Гбит/с)
Тип управляющей сети	Gigabit Ethernet
Пиковая производительность (FP64)	2.12 Петафлопс
Производительность в LINPACK	927.4 Терафлопс
Пиковая производительность для ИИ (FP16)	33.3 AI-Петафлопс