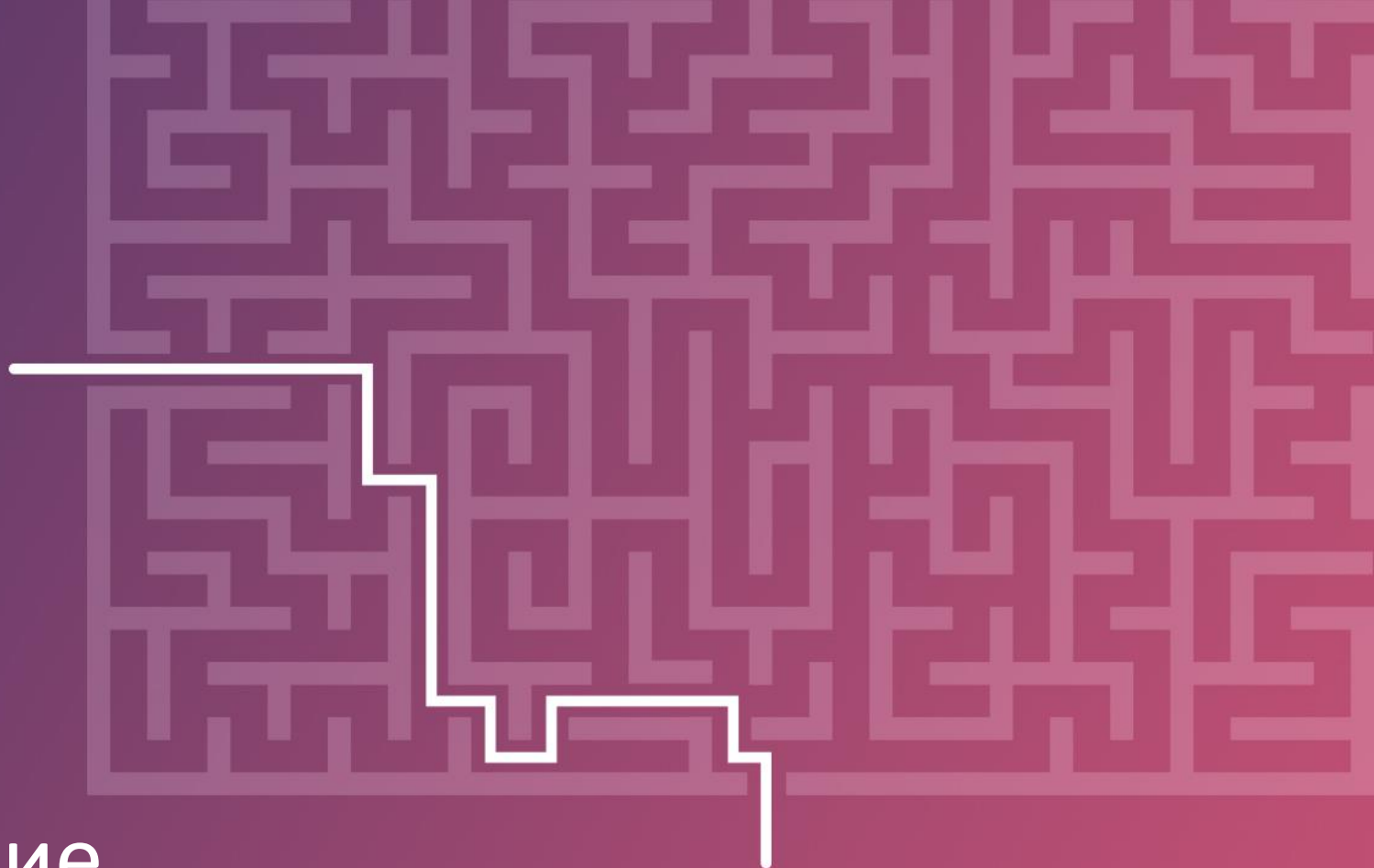




Машинное обучение с использованием решений Lenovo

Антон Болдин
Андрей Сысоев
Lenovo

Почему сейчас самое время врендрять технологии машинного обучения и ИИ?



Много данных

90%

компаний из списка
500 крупнейших
используют
Большие Данные

Источник: University of Maryland, US

Высокая производительность

в 1.000.000 раз

быстрее; современный
смартфон, чем все
компьютеры лунной миссии
NASA в 1969 г.

Источник: ZME Science Journal

Денежная выгода

\$1.2 трлд.

Отвоюют компании,
использующие ИИ, у своих менее
«информированных»
конкурентов к 2020 году

Источник: Forrester



Примеры задач, решаемых с помощью ИИ

Предиктивная аналитика

- Принятие решений и прогнозирование
- Реклама и персонализация предложений
- Рекомендательные системы
- Банковский скоринг
- Персональная медицина
- Предсказание отказа оборудования

Компьютерное зрение:

- Распознавание лиц и других объектов
- Видеоаналитика
- Описание содержания изображений и видео

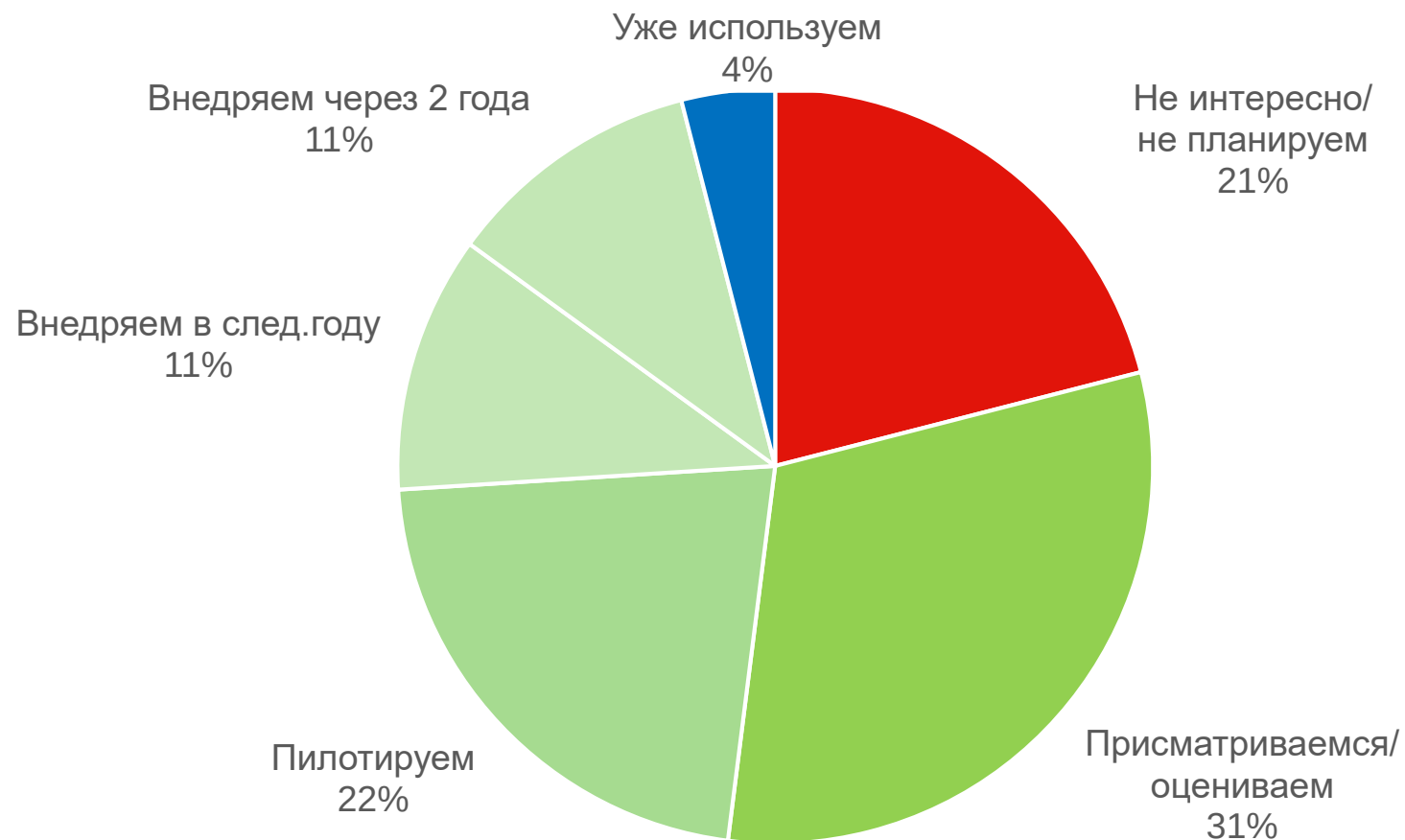
Обработка естественного языка:

- Чтение и анализ неструктурированного текста
- Анализ запросов пользователя
- Машинный перевод
- Антифрод и антиспам
- Генерация текстов на заданную тему

Биометрия:

- Голос
- Отпечатки пальцев
- Сетчатка глаза
- ДНК

Что говорят заказчики в мире об инициативах ИИ?

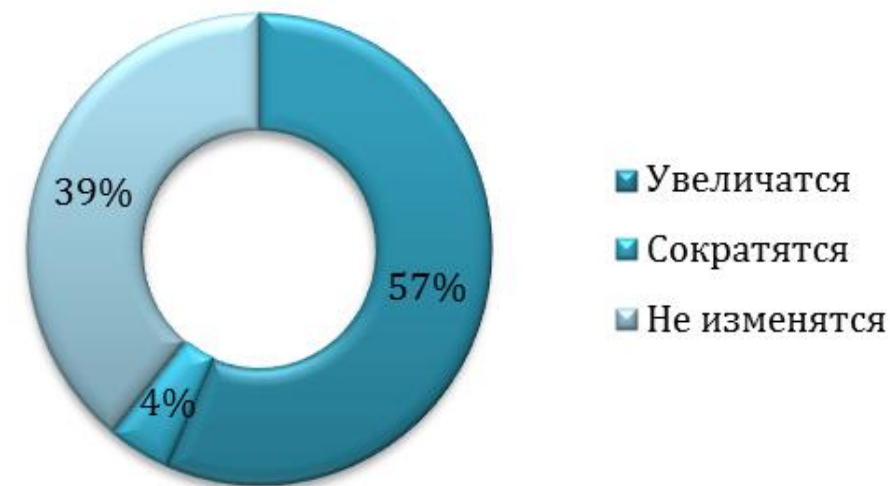
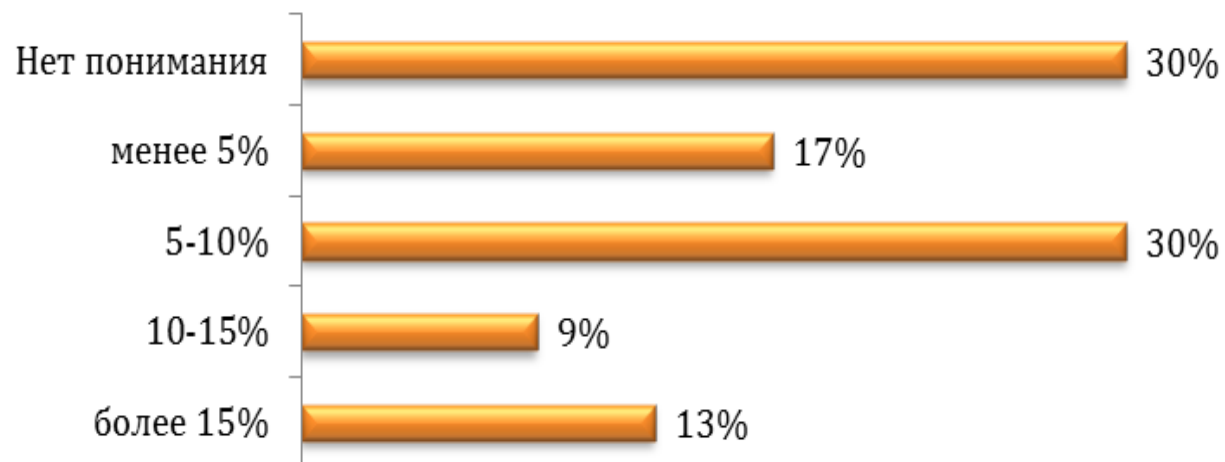


n = 400

Вопрос: Используете ли вы решения ИИ в своей компании?

Источник: Cloud and Artificial Intelligence Perceptions Survey 2018, IDC, January, 2018

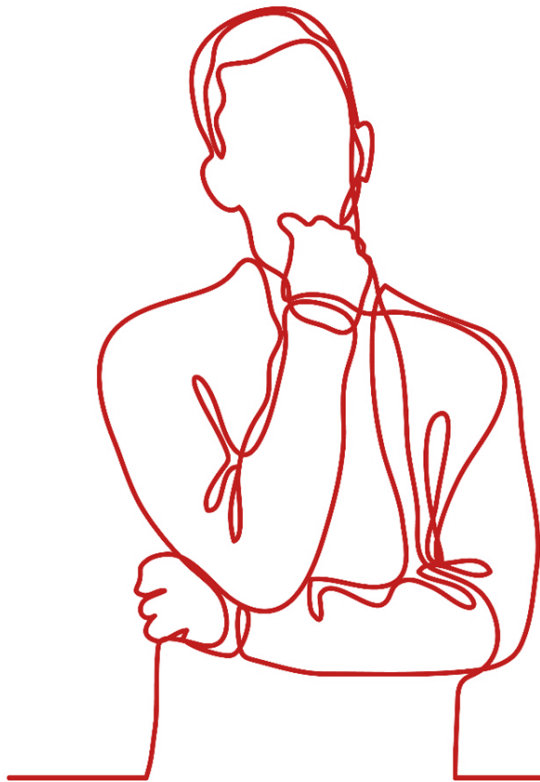
Доля расходов на ИИ от ИТ-бюджета в России в ближайшие 3-5 лет



n = 100

Источник: Tadviser, Актуальные тенденции рынка искусственного интеллекта и машинного обучения, октябрь 2017

Что сдерживает темпы роста рынка ИИ в России?



1. Низкий уровень осведомленности заказчиков о технологиях и реализованных на их базе проектах
2. Высокая ресурсозатратность для заказчика
3. Сомнения в практической пользе или в ощутимой экономической отдаче от внедрения

Практический подход Lenovo к ИИ



Понять

Нужен ли мне ИИ?

Экосистема из
экспертов и партнеров



Разработать

Инструменты для
быстрого старта проекта

Простые и удобные
в пользовании



Развернуть

Сервисные услуги:
ввод и сопровождение

Быстрое развертывание
«под ключ»



Инновационные
центры ИИ



HW и SW платформы
для разработки
приложений



Валидированные
решения и
профессиональные
услуги

Мы предоставляем доступ к инновационным центрам ИИ в трех частях света



ЦЕЛИ

Познакомиться с ИИ
Принять решение для себя

ИНФРАСТРУКТУРА

ThinkSystem 

ДЕМОНСТРАЦИИ

ИИ в медицине
ИИ в производстве

ПАРТНЕРЫ

Исследовательские
институты

Технологические
партнеры

Экосистемные
партнеры



Моррисвилль



Штутгарт



Пекин

Инфраструктура Lenovo для ИИ и маш. обучения



Платформы
для обучения
моделей



ThinkSystem SR650
2U, 2 Intel Xeon,
Big Data Platform



ThinkSystem SD530
2U, 2 x 2 Intel Xeon,
2 x 2 NVIDIA Tesla V100



NEW!

ThinkSystem SR670
2U, 2 x Intel Xeon,
4 x 2 NVIDIA Tesla V100



LiCO v5.1



NEW!

Производительность



ThinkSystem SD530
2U, 4 x 2 Intel Xeon



ThinkSystem SR650
2U, 2 x Intel Xeon,
2 NVIDIA Tesla P100



ThinkSystem SD530
2U, 2 x 2 Intel Xeon,
2 x 2 NVIDIA Tesla V100

**Lenovo intelligent
Compute Orchestrator**
ПО для управления нагрузками
машинного обучения
и HPC

ThinkSystem



Lenovo ThinkSystem SR650



24x 2.5" + 2x 3.5" HDDs/SSDs



12x + 2x 3.5" HDDs/SSDs

- 2U в стойке
- 2x процессора до 205W
- 24x DIMM слота
- 2.5" или 3.5" HDDs/SSDs
- до 12x NVMe (до 24x в январе '19)
- 7x PCIe слотов + LOM
- 2x GPU

Lenovo ThinkSystem SD530

- 4x 2-проц. сервера (ноды) в 2U
- процессоры до 205W
- 16 DIMM слотов на ноду
- 6x 2.5" HDDs/SSDs на ноду,
2x из которых NVMe
- 4x GPU в 2U

На каждую ноду на выбор:

- 2x PCIe x8 «холодной замены»
- 1x PCIe x16 «теплой замены»



SD530 и GPU модуль



4x SD530 в шасси D2

Lenovo ThinkSystem SR670



вид спереди



вид сзади

- 2U в стойке
- 2x процессора до 205W
- 24x DIMM слота
- 7x PCIe 3.0 слотов
- 4x GPU
- 8x 2.5" HDDs/SSDs
- 2x 2000W БП

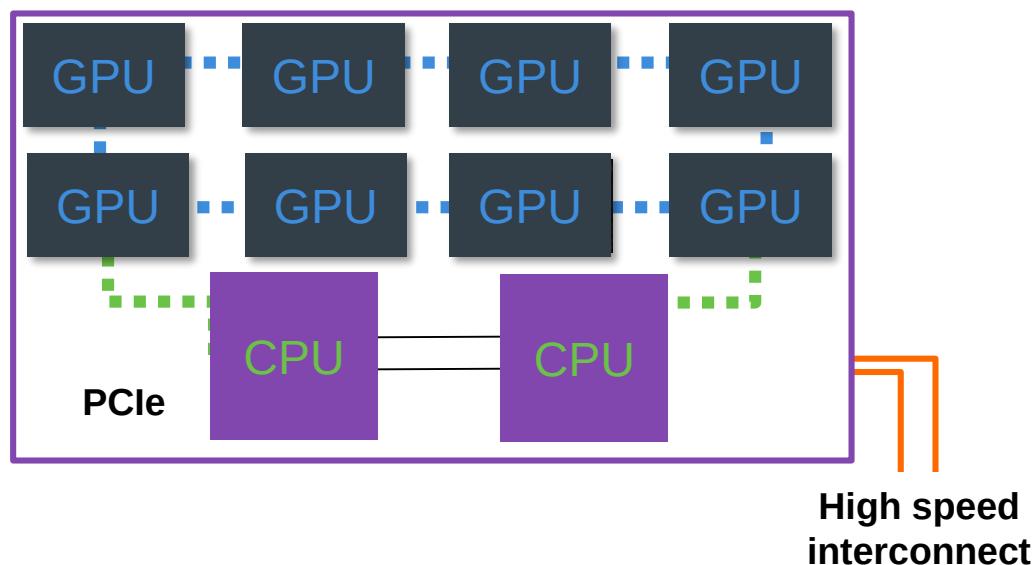
Почему 2U, а не 1U?

	2U ThinkSystem SR670	1U GPU Server	Почему важно?
CPU	До 205W (48 ядер)	До 150W (40 ядер)	Выше производительность
Диски	До 8 шт.	До 2 шт.	Больше хранимых данных для загрузки GPU
Вентиляторы	согласно OSHA	Выше скорости вращения -> выше энергопотребление -> выше шум	Энергоэффективность, шум, амортизация
Сеть	HDR ready (x32 Gen3 PCIe)	One x16 Gen3, capped at EDR	Готовность к высоким скоростям передачи данных по сети
GPU	Сбалансированный дизайн между CPU, GPU и I/O – все компоненты на одном CPU или разных	I/O разведен только от одного CPU, меньше гибкости	Возможность установки GPU следующего поколения (400W)
Плотность в стойке	До 35-40kW/стойку	До 80kW/стойку	Демократичнее к питанию на уровне стойки

Scale-Up vs Scale-Out

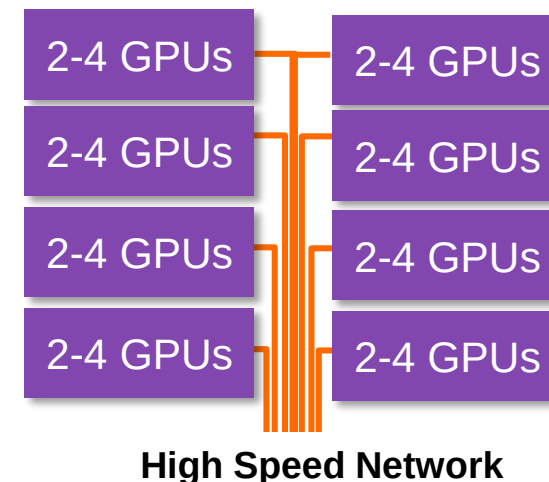
Scale-Up (8x GPU на ноду)

- Монолитный дизайн – один пользователь на систему
- Под узкие задачи
- Сложности с питанием/охлаждением/дизайном



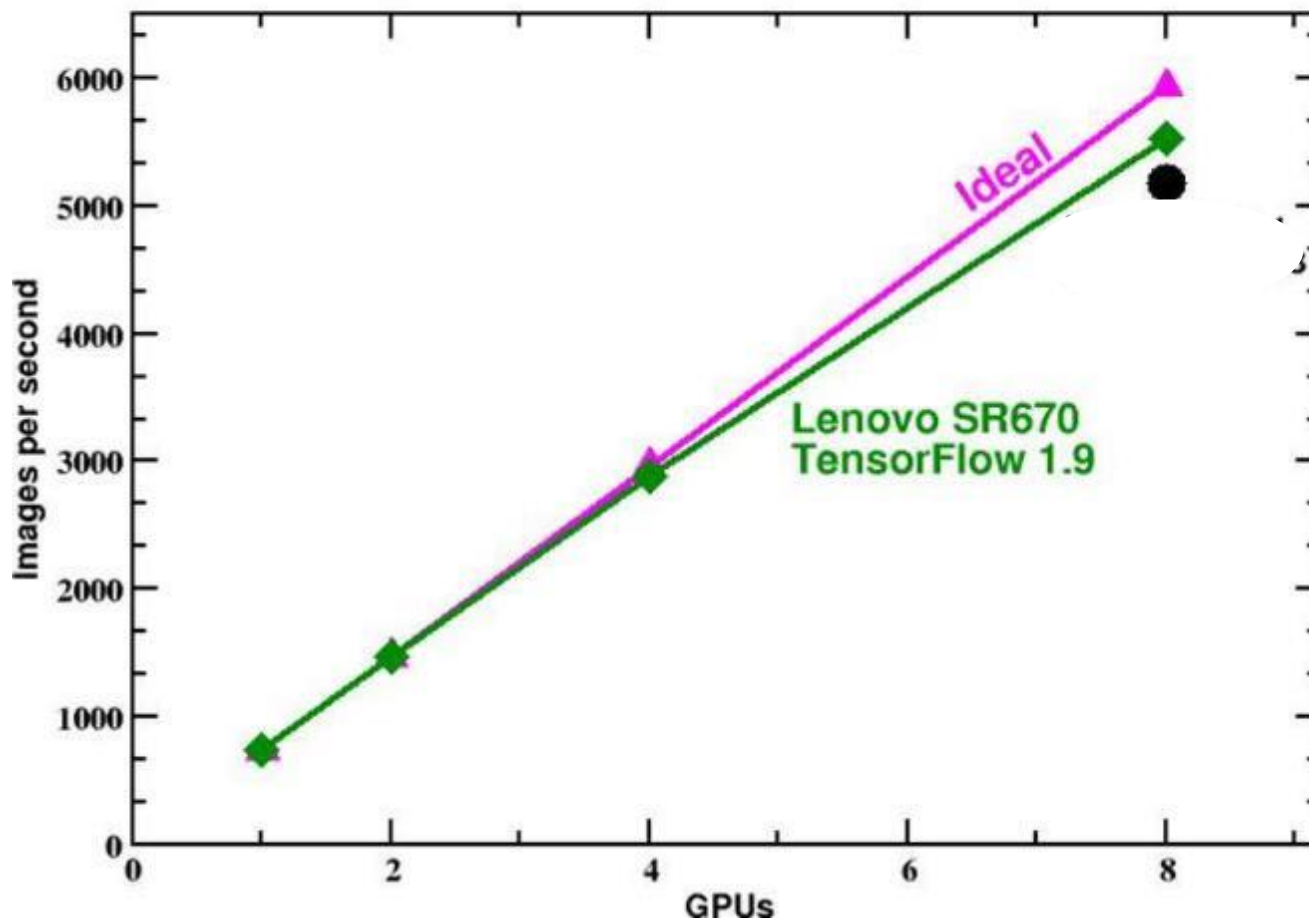
Scale-Out (2-4x GPU на ноду)

- Универсальные строительные блоки, соединенные по высокоскоростной сети
- Гранулярность – поддержка нескольких пользователей; параллельные нагрузки
- Мультизадачность (HPC / AI / VDI)



Scale-Out в массы!

V100 Benchmark, resnet50, 256 batch, fp16



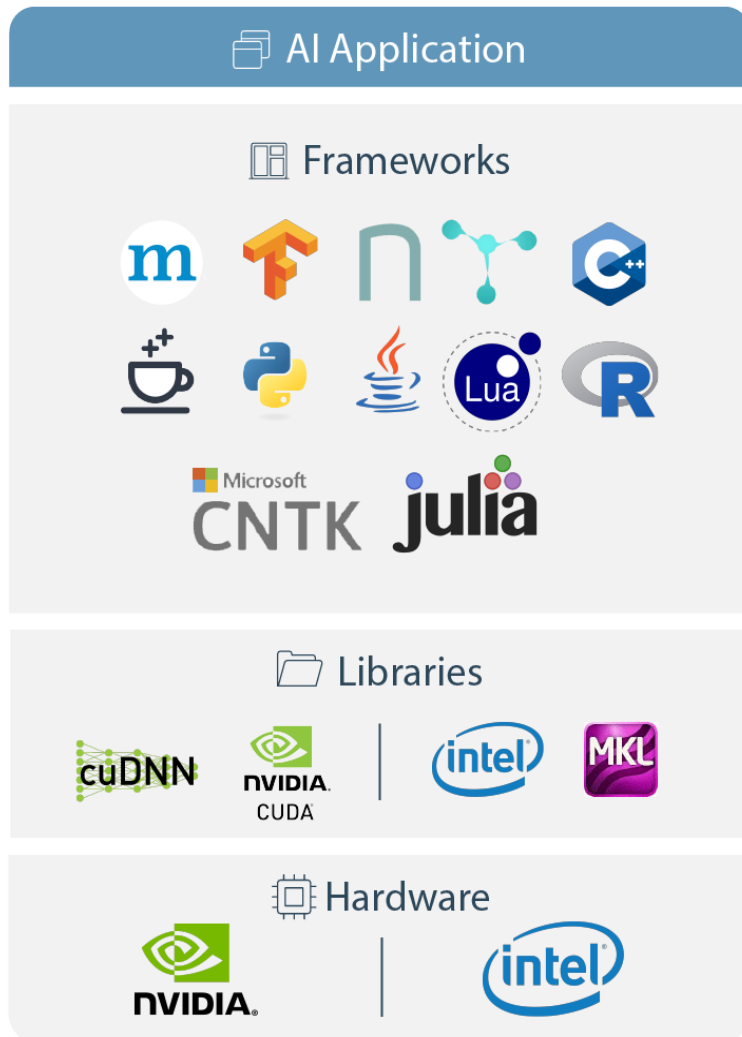
Scale-Up системы обеспечивают БОЛЬШУЮ производительность при условии:

- Большой дата-сет
- Постоянная/непрерывная нагрузка

Scale-Out системы обеспечивают лучшее цена/производительность

- Удобнее подобрать конфигурацию к дата-сету
- Обучение или инференс
- Не только ИИ (+HPC, +VDI)
- **95% производительности при 50% ниже стоимости, чем scale-up система**

Построить ИИ решение: так ли это просто?



Сложность рабочего процесса

- Предварительная оценка входных данных, подготовка и валидация моделей обучения

Сложность разработки ИИ приложений

- Дата-ученые тратят время на написание кода, вместо того, чтобы непосредственно работать с моделями обучения

Сложность интеграции стэка Open Source технологий

- Успешный ввод в эксплуатацию всего программного стэка для ИИ – нетривиальная задача

Различные фреймворки – различные стэки SW

Различное HW – различные библиотеки

Сложность масштабирования

- Применение распределенных архитектур для маш. обучения требует спец. навыков и может повлечь за собой изменение процессов для заказчиков

LiCO – чтобы легко управлять машинным обучением

Упростить рабочий процесс при обучении ИИ

- Простой и понятный GUI и шаблоны для управления
- Набор предтренированных моделей, мониторинг обучения, история обучений
- Поддержка множества типов пользователей с разграничением прав

Упростить развертывание технологий open source

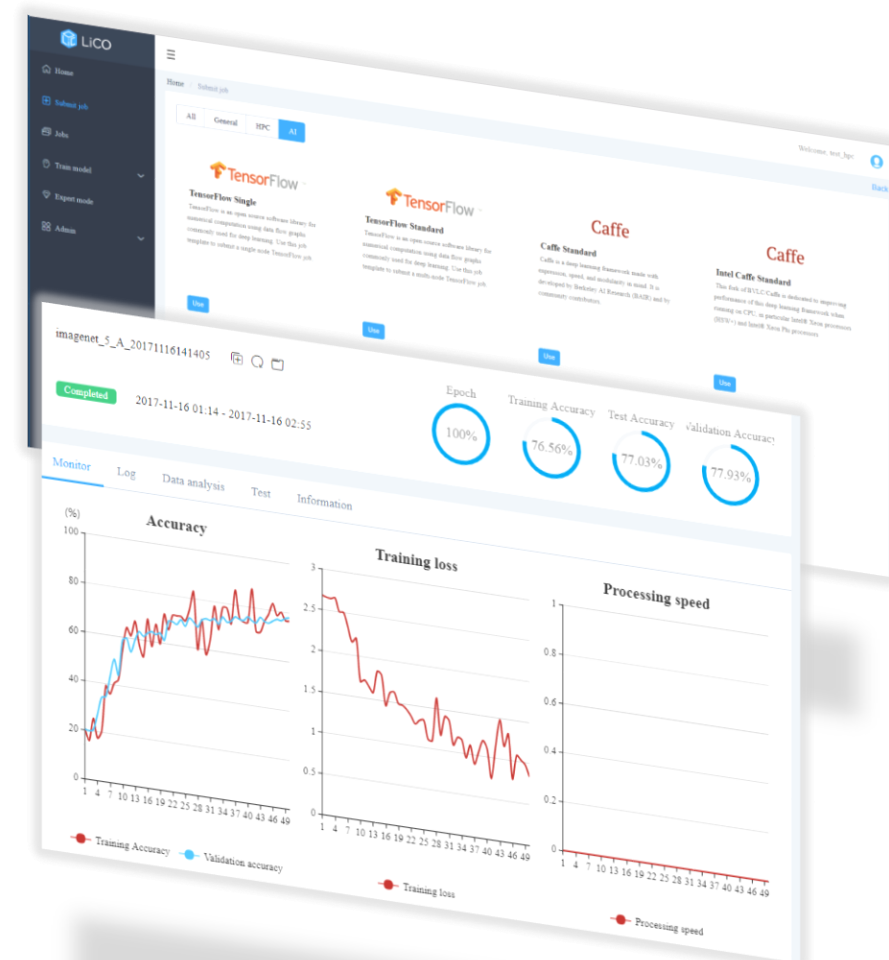
- Валидированные стэки для сокращения времени на развертывание и ввод решения в эксплуатацию
- Поддержка различных ИИ фреймворков

Поддержка необходимых HW компонентов

- Различные NVIDIA GPUs и процессоры Intel для разных типов нагрузок

Оптимальное использование HW для улучшения TCO

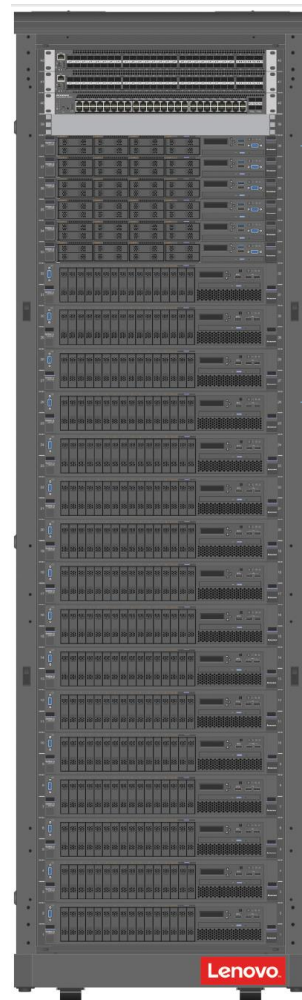
- Поддержка распределенных архитектур
- Управление ресурсами кластеров при выполнении ИИ нагрузок



Валидированные решения и референсные архитектуры Lenovo для ИИ



- Оптимизированные архитектуры для всего жизненного цикла ИИ решения, включая управление данными, обучение модели и инференс
- Для различных нагрузок
- Модульная архитектура для удобства масштабирования
- Оптимизация TCO и упрощение развертывания



Сеть



Управляющая нода / Хранение данных



Ноды для инференсов



Ноды для обучения



Ноды Big Data

ИИ решения Lenovo

Не только «железо»



LiCO
ПО для управления



Профессиональные
услуги



Партнерские
решения



ПРИМЕРЫ

ИИ на базе Lenovo быстрее всех обнаруживает раковую опухоль печени



Задача:

Разработать автоматический алгоритм сегментации для выявления раковых аномалий на КТ снимках

Совместно с Мюнхенским Техническим Университетом, Университетом Тель-Авива и другими исследовательскими институтами

35 стран, включая США, Китай, Индию

ИИ решение на базе Lenovo (LeHealth):

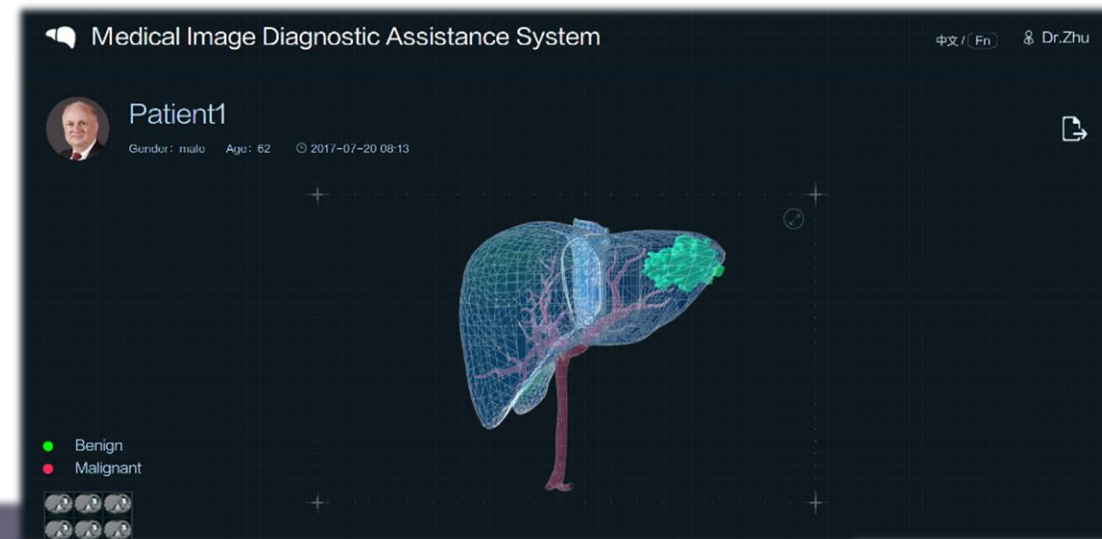
Анализ КТ снимков пациентов для раннего обнаружения и классификации опухолей

Облегчение диагностики заболевания посредством обнаружения зон распространения и текущего состояния опухолей

Предоставление возможностей 3D моделирования для врачей

Помощь в работе для врачей, увеличение эффективности лечения, составление отчетов

[Подробнее](#)



Lesion										
#	User	Dice per case ▲	Dice global ▲	VOE ▲	RVD ▲	ASSD ▲	MSD ▲	RMSD ▲	Precision at 50% overlap ▲	Recall at 50% overlap ▲
1	leHealth	0.7020 (1)	0.7940 (5)	0.394 (11)	5.921 (18)	1.189 (12)	6.682 (5)	1.726 (8)	0.156 (14)	0.437 (3)
2	hchen	0.6860 (2)	0.8290 (1)	0.356 (3)	5.164 (17)	1.073 (5)	6.055 (1)	1.562 (2)	0.409 (4)	0.408 (4)
3	hans.meine	0.6760 (3)	0.7960 (4)	0.383 (10)	0.464 (12)	1.143 (8)	7.322 (12)	1.728 (9)	0.496 (2)	0.397 (5)
4	medical	0.6610 (4)	0.7830 (2)	0.357 (12)	12.124 (1)	1.075 (7)	6.317 (3)	1.596 (1)	0.446 (3)	0.374 (6)

ИИ в производстве

Задача:

Производство алюминиевых банок

Контроль качества и выявление брака

Поиск косметических и структурных дефектов

Где ИИ может помочь:

Анализ выпускаемой продукции на конвейере

Мониторинг конвейера 24x7

Mark III Systems

Принятие решения на уровне одной банки или целой партии

Lenovo предоставляет доступ к ресурсам Artificial Intelligence Innovation Center в Моррисвилле, а также поддержку специалистов и инженеров

[Подробнее](#)



ИИ в сельском хозяйстве

Предпосылка:

70% всей пресной воды в мире расходуется на сельское хозяйство

Рост населения Земли (10 млрд. к 2050)

Проблема:

Сохранение природных ресурсов

Как обеспечить растущее население пищей?

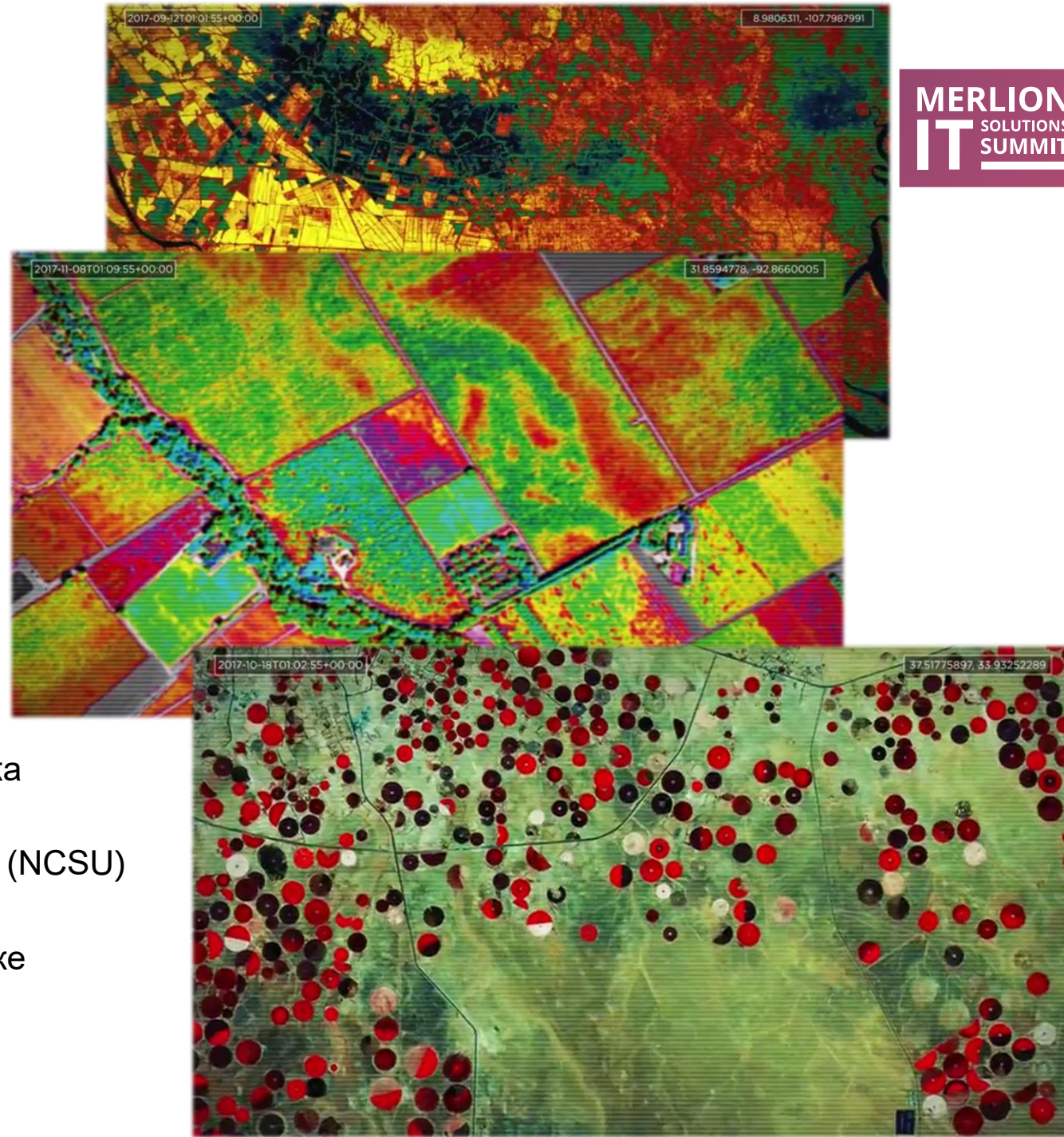
Где ИИ может помочь:

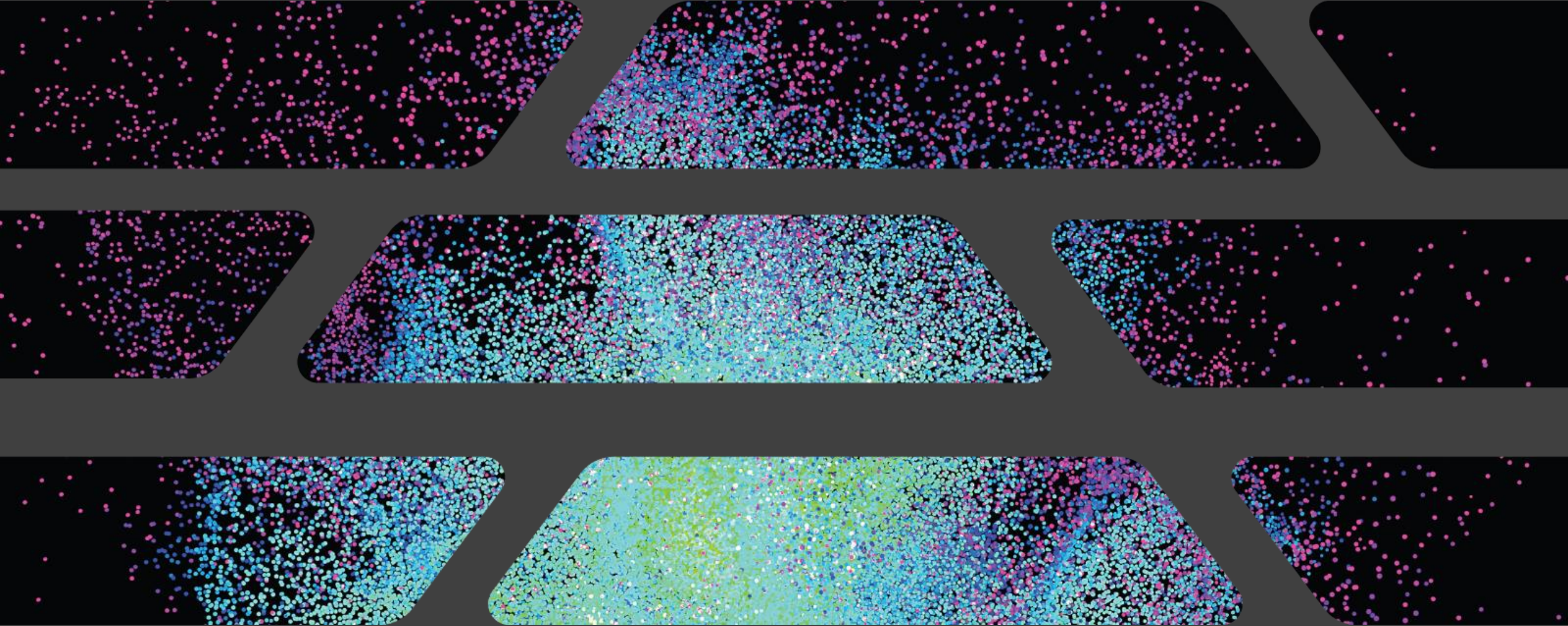
Работа со спутниковыми изображениями для поиска оптимальной территории с учетом всех факторов

Государственный университет Северной Каролины (NCSU)

Lenovo предоставляет доступ к ресурсам Artificial Intelligence Innovation Center в Моррисвилле, а также поддержку специалистов и инженеров

[Подробнее](#)

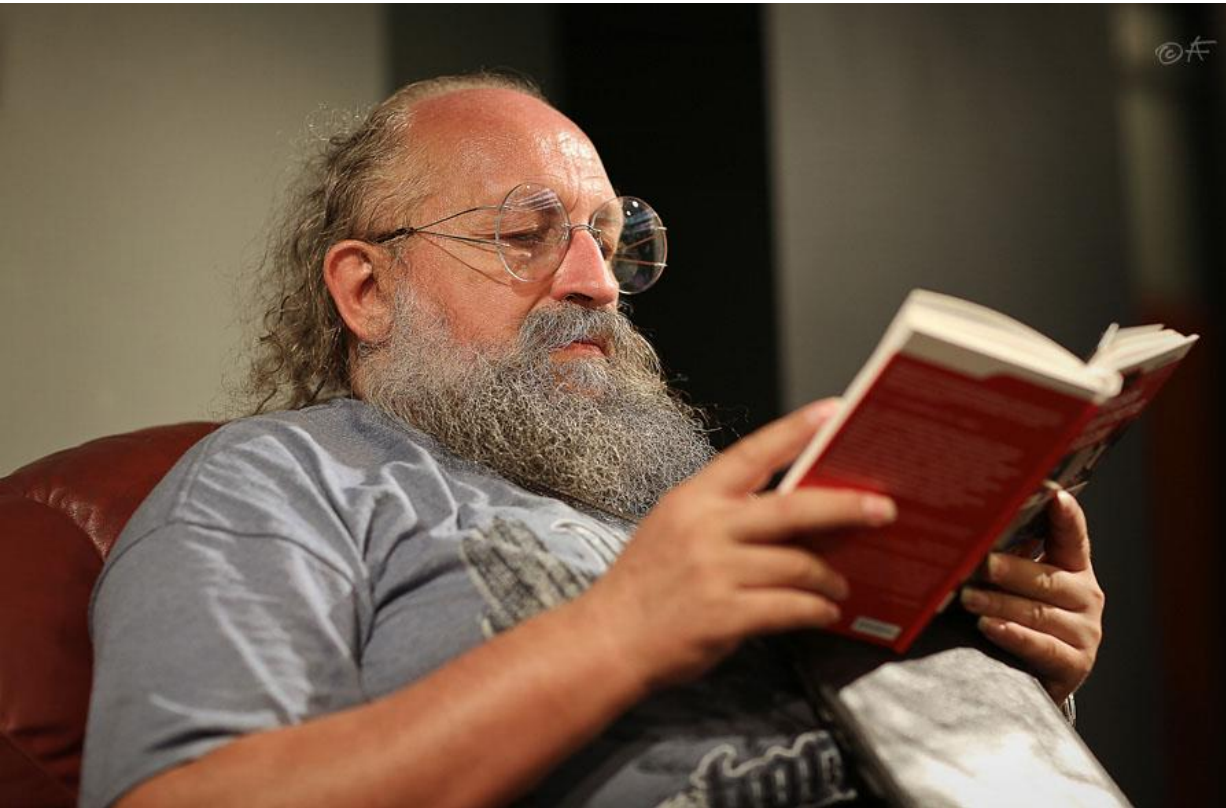




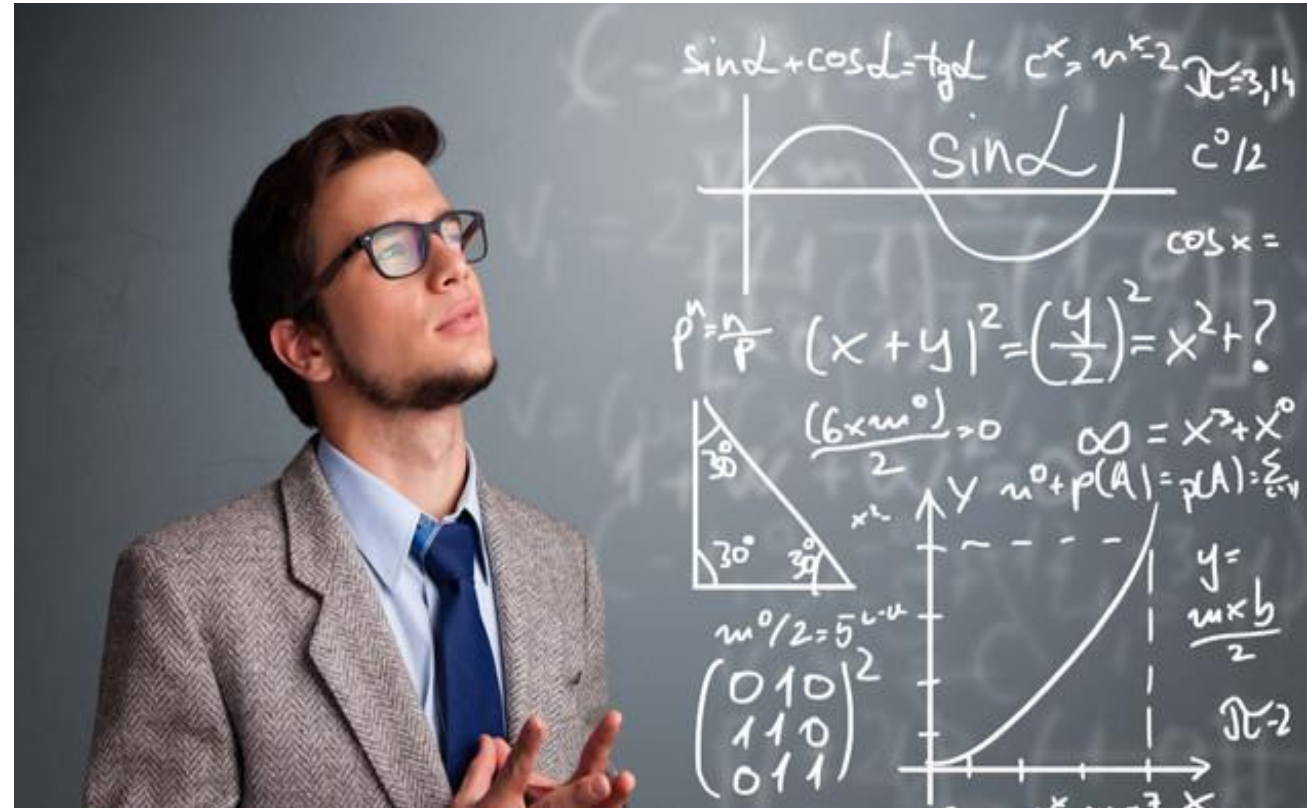
→ LiCO

Приминение опыта Lenovo HPC в ИИ

HPC Guy



Data Scientist



- Командная строка, планировщики, скрипты, мониторинг и т.д.
- Требуются специальные знания и навыки, чтобы работать с HPC кластером

- Нет нужды в знаниях HPC
- Но нужны навыки работы с фреймворками, библиотеками, языками программирования
- Оптимизация всего стека для задач ИИ

+ LiCO simplifies managing and using AI

Configures the hardware and software resources for the data scientist, so they can get to work

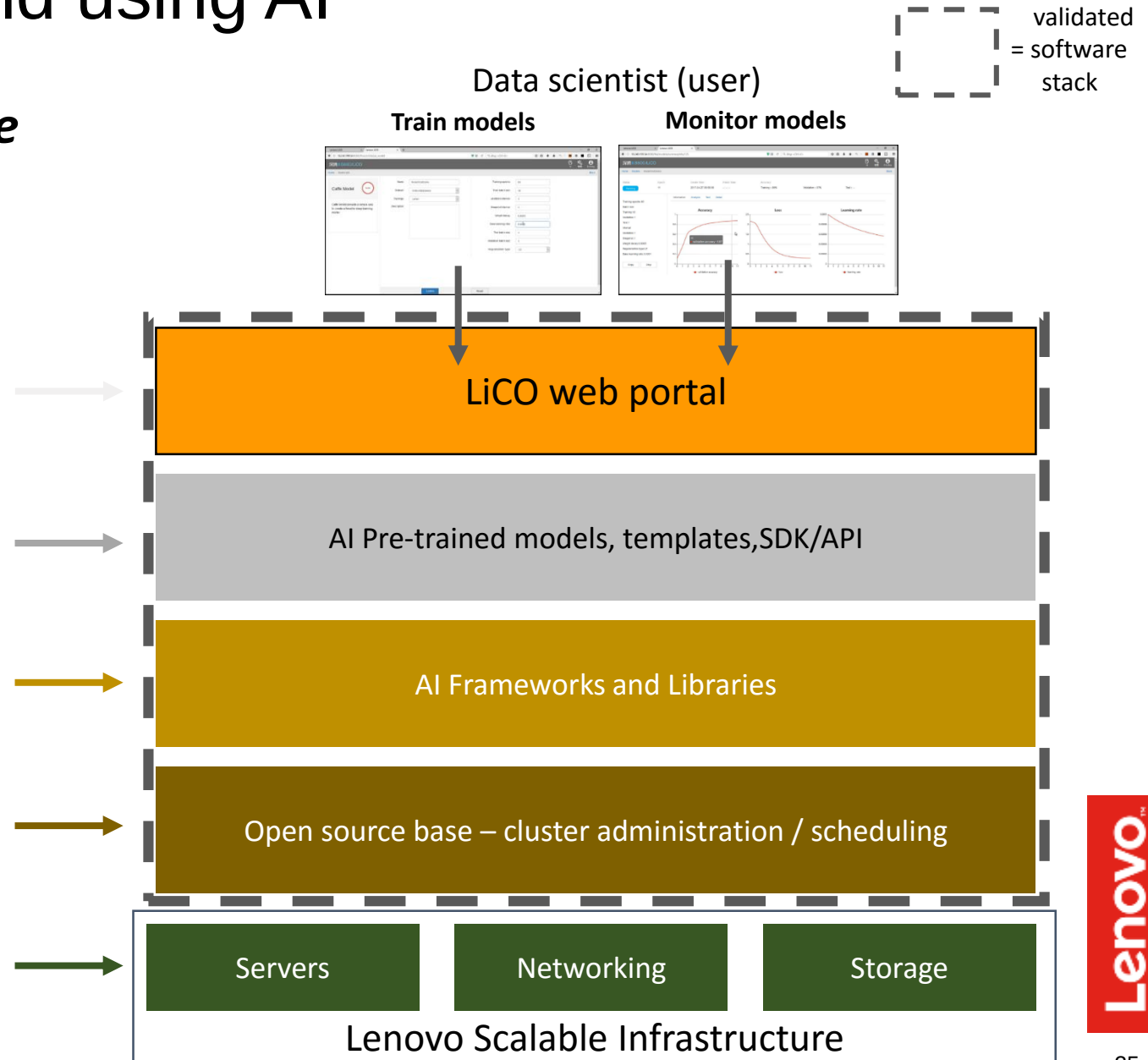
Administrative and user GUI access to the cluster, to deploy, monitor and manage AI training jobs

Pre-trained models as a foundation to extend upon and reduce data & time-to-train new models

Open source AI frameworks and libraries for training (TensorFlow, Caffe, Intel-Caffe, MxNet, Containers)

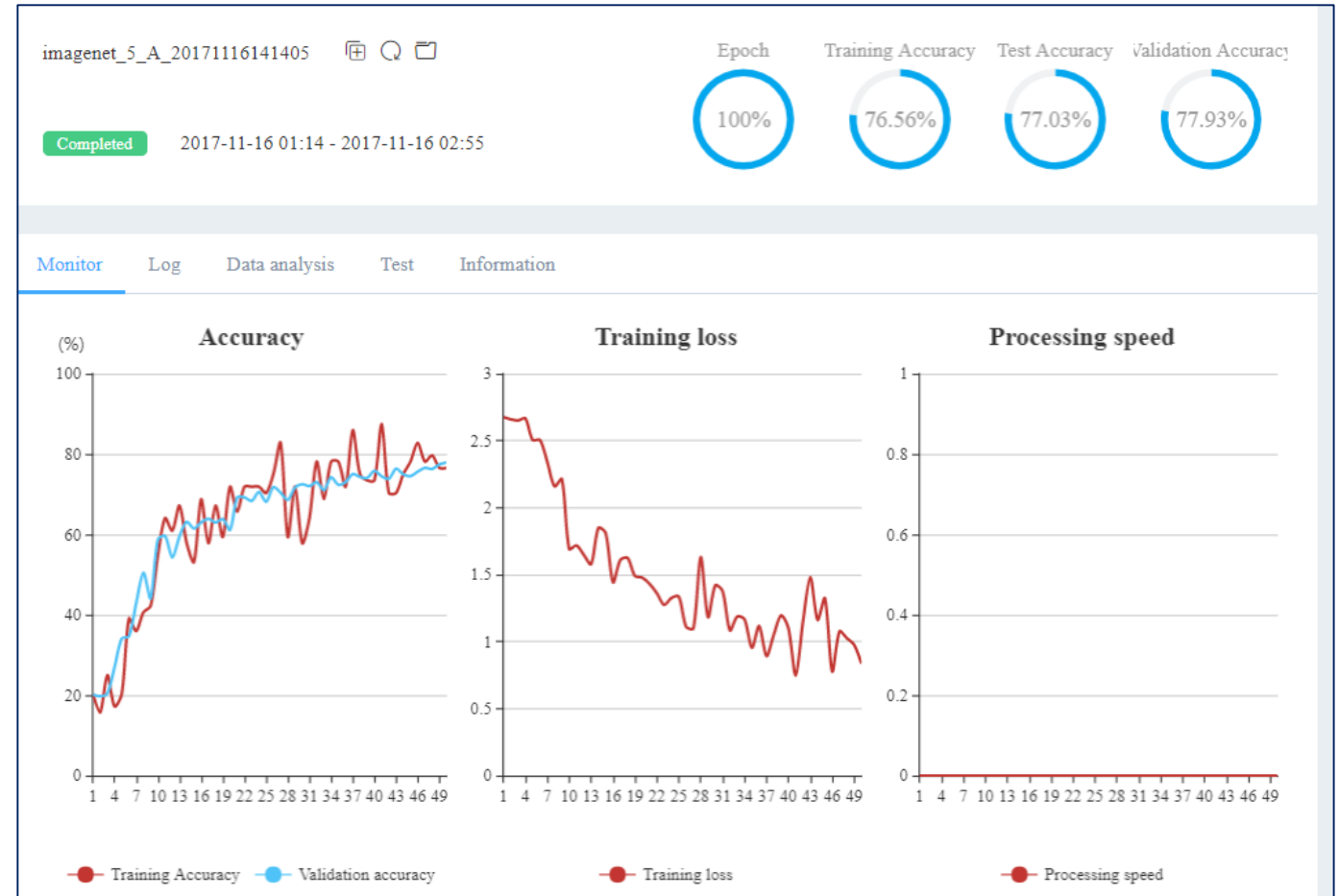
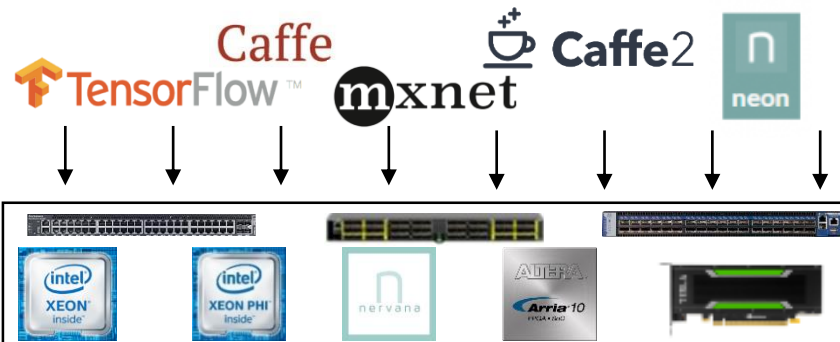
Open tools to manage / schedule resources for deploying AI training jobs to a scale-out cluster

Lenovo Infrastructure – mix of server types, accelerators, processors, networking, storage, deployed as a single solution

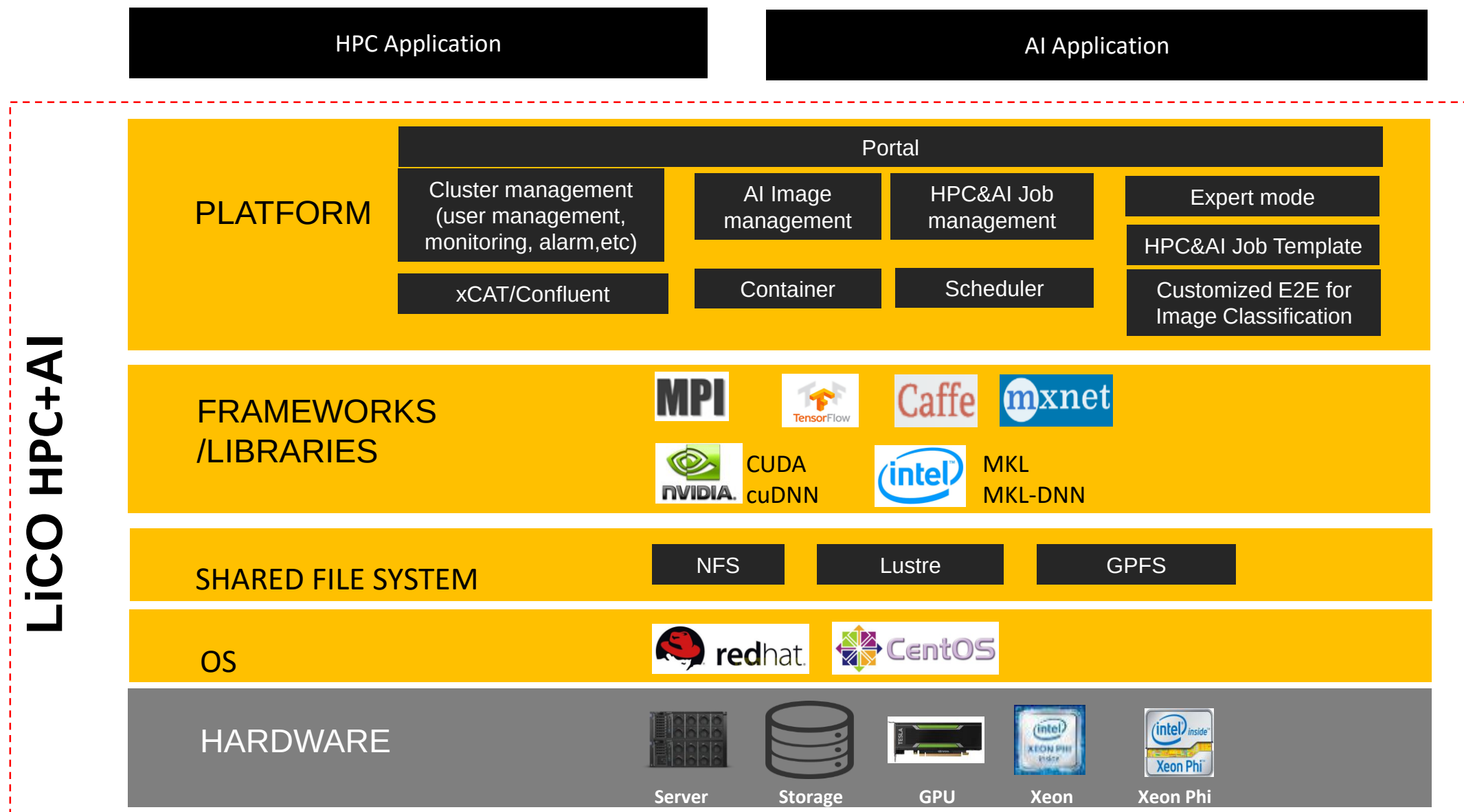


+ LiCO enables data scientists to optimize AI

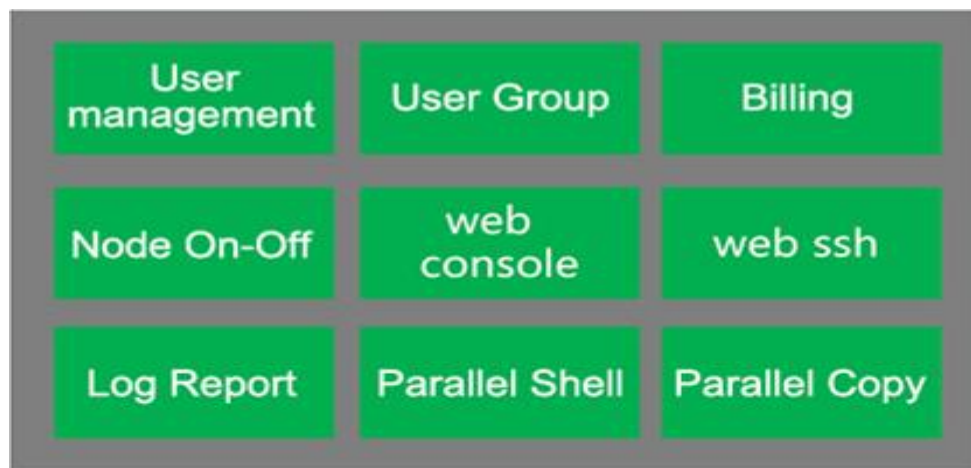
- Monitor the training in flight
 - Is the model achieving the expected accuracy?
 - Is it improving over time?
- Run training jobs with different AI frameworks and hardware resources to optimize models
 - Experiment to find the optimal combination for each AI model



+ LiCO Architecture



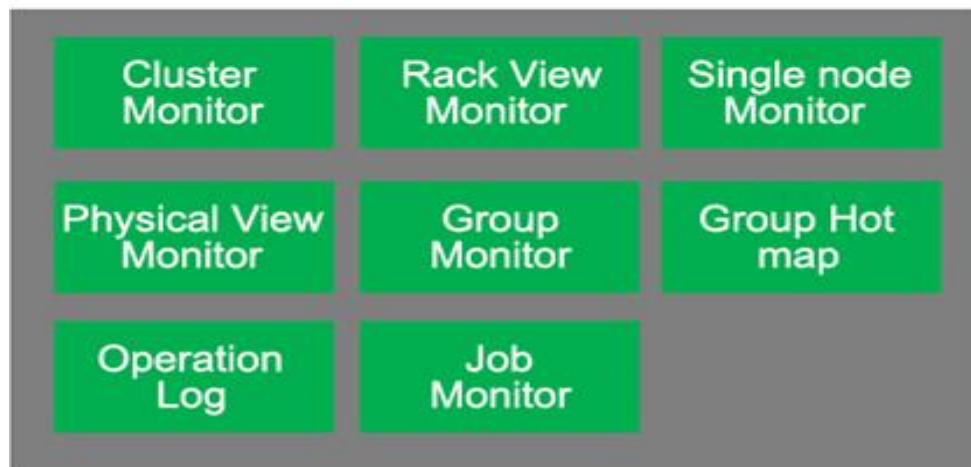
+ LiCO Features



Cluster management



Job Management



Monitor



Alarm

MERLION IT SOLUTIONS SUMMIT

**Спасибо
за внимание!**

