Флопс

Современный подход к созданию решений для задач HPC, ЦОД, облаков

07.07.2025

Ключевые технологические тренды HPC и Big Data

Инструменты искусственного интеллекта и машинного обучения (Commercial AI and machine learning)

Использование технологий искусственного интеллекта, машинного обучения и обработки естественного языка для содействия в подготовке данных, понимании и трактовке результатов анализа, то есть в качестве расширения возможностей человека и традиционных способов формирования и использования аналитического контента

Непрерывная интеллектуальная обработка данных (Continuous Intelligence)

Подход, при котором результаты аналитики в реальном времени интегрируются в бизнес-операции, происходит обработка потоковой контекстной информации, поступающей с датчиков IoT, и исторических данных, позволяющий моментально реагировать на изменения и предписывать поведение моделей

Генеративный искусственный интеллект (Generative AI)

Генеративный искусственный интеллект — тип системы искусственного интеллекта (ИИ), способной синтезировать текст, изображения или комбинированный медиаконтент
Известные системы: ChatGPT, YandexGPT, Bard, DeepSeek

«Расширенная» (дополненная) аналитика (Augmented analytics)

Совершенствование процесса анализа за счет автоматизации процесса поиска, обработки данных с использованием технологий машинного обучения (ML) и (AI)

ARM в HPC и Big Data

Архитектура ARM64 используется в HPC и Big Data благодаря высокой энергоэффективности, стабильной производительности, линейной масштабируемости, оптимизации для нагрузок обработки данных

Новые программные среды и инструменты управления кластерами (K8s, REST API, NCCL, NEXUS)

Широкое внедрение контейнерных технологий, конвергенция оркестраторов и классических планировщиков, новых программных интерфейсов взаимодействия и доступа, облачных объектных хранилищ данных

Общие тренды в создании ЦОД и HPC

Данные

- К 2025 г. – 160 Збайт, увеличение в 3 раза по сравнению с 2020 г.
- Разработка новых ИИ-моделей требует больших СХД, компания OpenAI построит ЦОД для хранения 5 Збайт данных
- Около 60% данных, которые хранят компании, остаются неиспользованными

Потребление электроэнергии

- ИИ ЦОД развиваются быстрее, чем электростанции и линии электропередач
- IEA обещает к 2030 г. увеличение энергопотребление ЦОД более 2 раз на 23% в год
- Стоимость солнечной энергии - $38–$78 за 1 МВт/ч, $107 для газовых электростанций
- Энергоэффективность является важной частью сокращения энергопотребления ЦОД
- Доступность электроэнергии определяет выбор места для центра обработки данных
- Стандартная стойка ЦОД 12 кВт, ЦОД гиперскейлеров 40-60 кВт, ИИ ЦОД более 200 кВт
- ЦОД переходят на локальную выработку электроэнергии (19% уже), 62% планируют

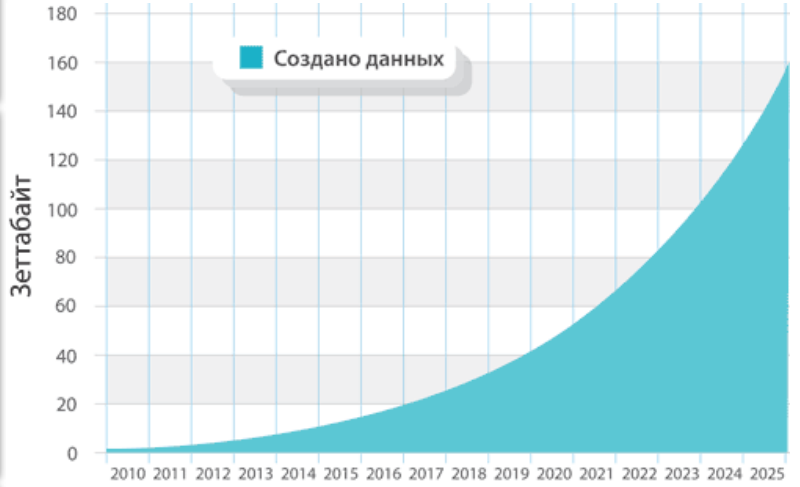
Выбросы, экология

- Выбросы углерода Google выросли на 48% за пять лет из-за ИИ
- Выбросы Alphabet составили 14,3 млн метрических тонн углерода в 2023 г. Это на 48% выше, чем в 2019 г.
- Выбросы углерода Microsoft выросли на 30% с 2020 г.

Рост числа и мощности ЦОД

- Отчет IDC показывает, что к 2028 г. расходы Европы на ИИ достигнут $133 млрд., среднегодовой рост 30,3% с 2024 г.
- К 2027 г. доход рынка жидкостного охлаждения вырастет до $2 млрд., рост 60% в год
- Рост числа ИИ ЦОД увеличиваться на 60–80% в год
- Тренд на создание так называемых «фабрик искусственного интеллекта»

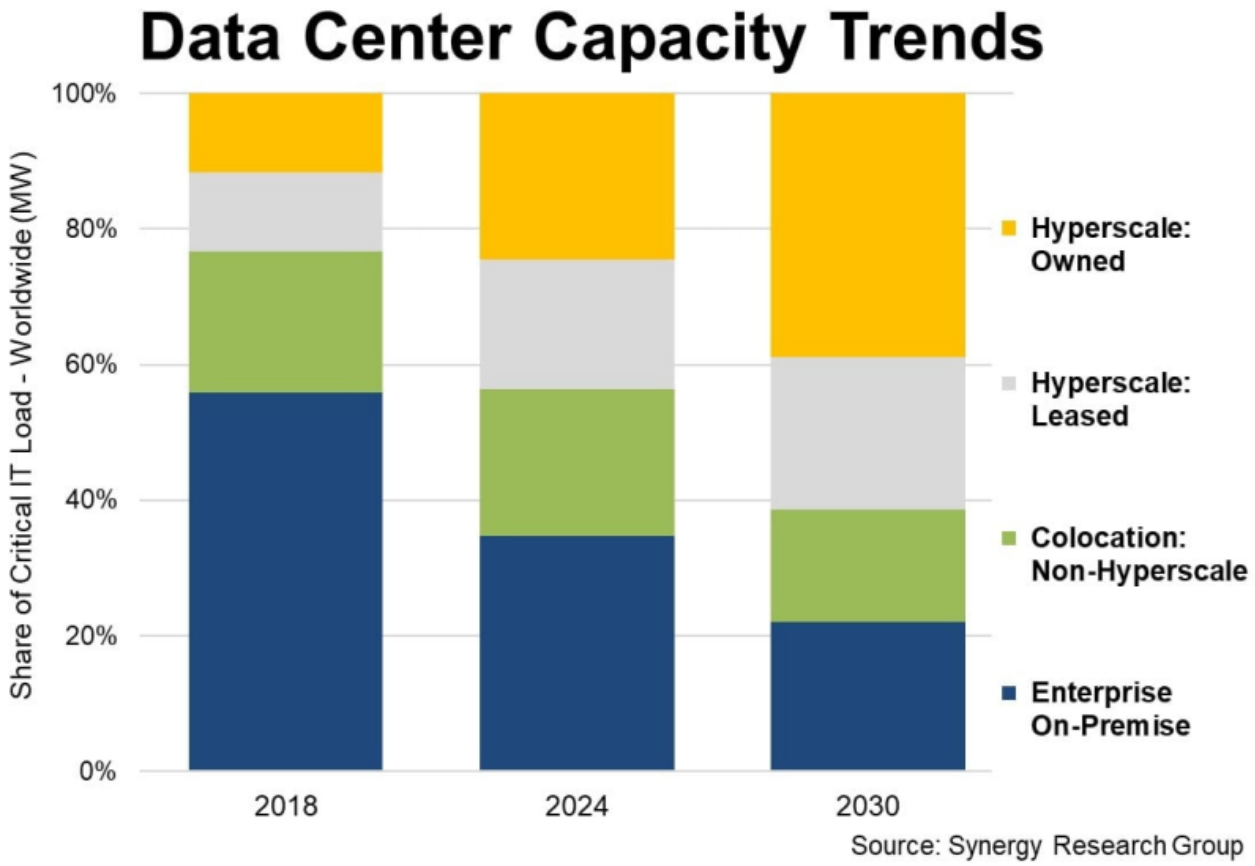
К 2025г. – 160 зеттабайт



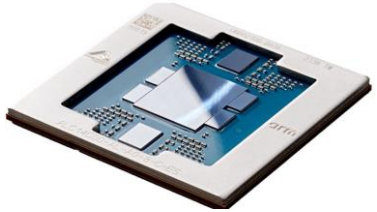
Источник: IDC, исследование Data Age 2025

Тренд на Фабрики ИИ

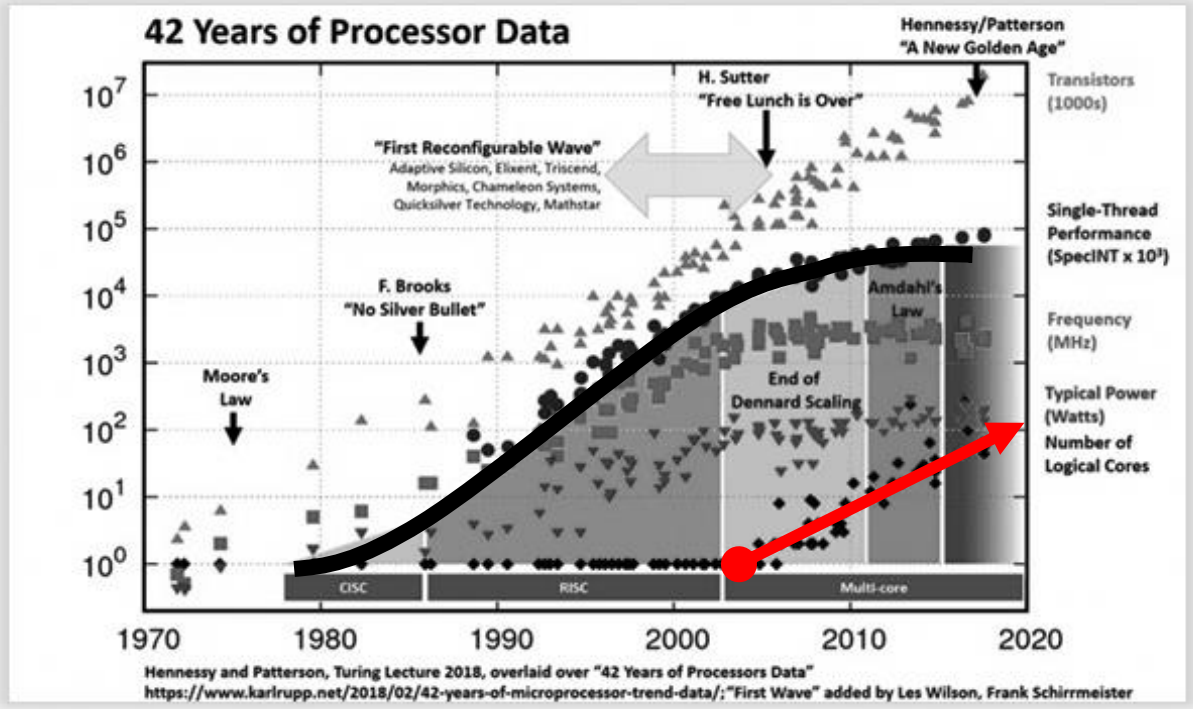
- К 2030 году на гиперскейлеров будет приходиться **61 %** всех мощностей ЦОД в мире, что обусловлено ростом увеличением спроса на ИИ-вычисления
- В распоряжении гиперскейлеров уже находится **1189** ЦОД. Вместе на долю IT-гигантов приходится **44 %** мировой мощности ЦОД. Из них более половины приходится на собственные ЦОД.
- На колокейшн-ЦОД, не связанных с гиперскейлерами, приходится **22 %** от общей мощности
- На долю корпоративных ЦОД остаётся всего **34 %** мощностей. 6 лет назад на них приходилось **56 %** ёмкости
- В качестве примера, в проект Stargate будет инвестировано **$500 млрд.**
- В ЦОД xAI Илона Маска размещено **100 000** графических процессоров Nvidia H100
- Nvidia выпустила чертежи для строительства фабрик ИИ, называемые эталонным дизайном предприятия
- Мощность стойки фабрик ИИ более **200 кВт**
- ИИ фабрики будут потреблять **более ГВт** электронергии

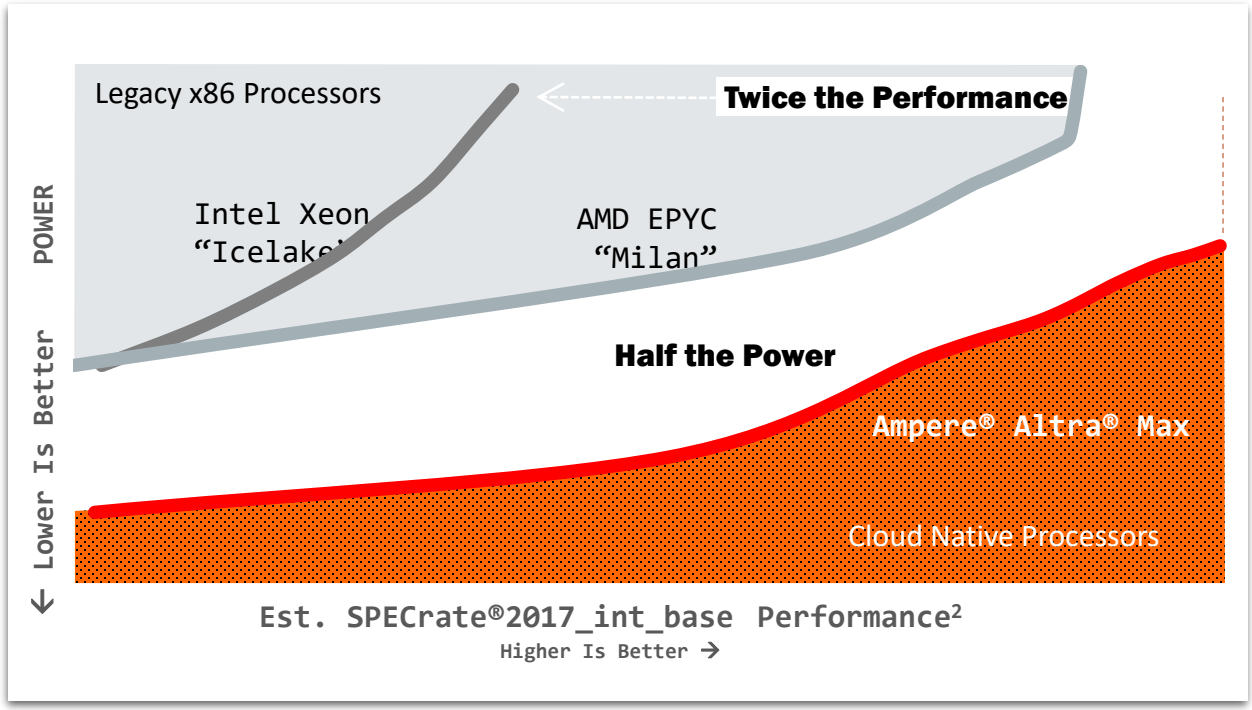


CPU ARM в HPC

Google		Arm-процессор **Axion** 72 ядра Neoverse V2 Titanium network and storage offloads
Amazon Web Services (AWS)		Arm-процессор **Graviton4** ИИ-ускоритель Trainium2 96 ядер Neoverse V2 12 каналов DDR5
Microsoft		Arm-процессор **Azure Cobalt 100** ИИ-ускоритель Azure Maia 100 128 ядер Armv9 Neoverse N2
Ampere Comptuing (Oracle)		Arm-процессор **Ampere One** 192 ядер собственной разработки 8 каналов DDR5

Архитектура ARM64 обеспечивает эффективность





Turbo Frequency

Hyperthreading

Scale Up Accelerators

Power Optimized Consistent Performance

Linear Core Scaling

High Performance General-Purpose Cores

Paradigm Shift →

Архитектура ARM64 оптимизированная для облачных нагрузок

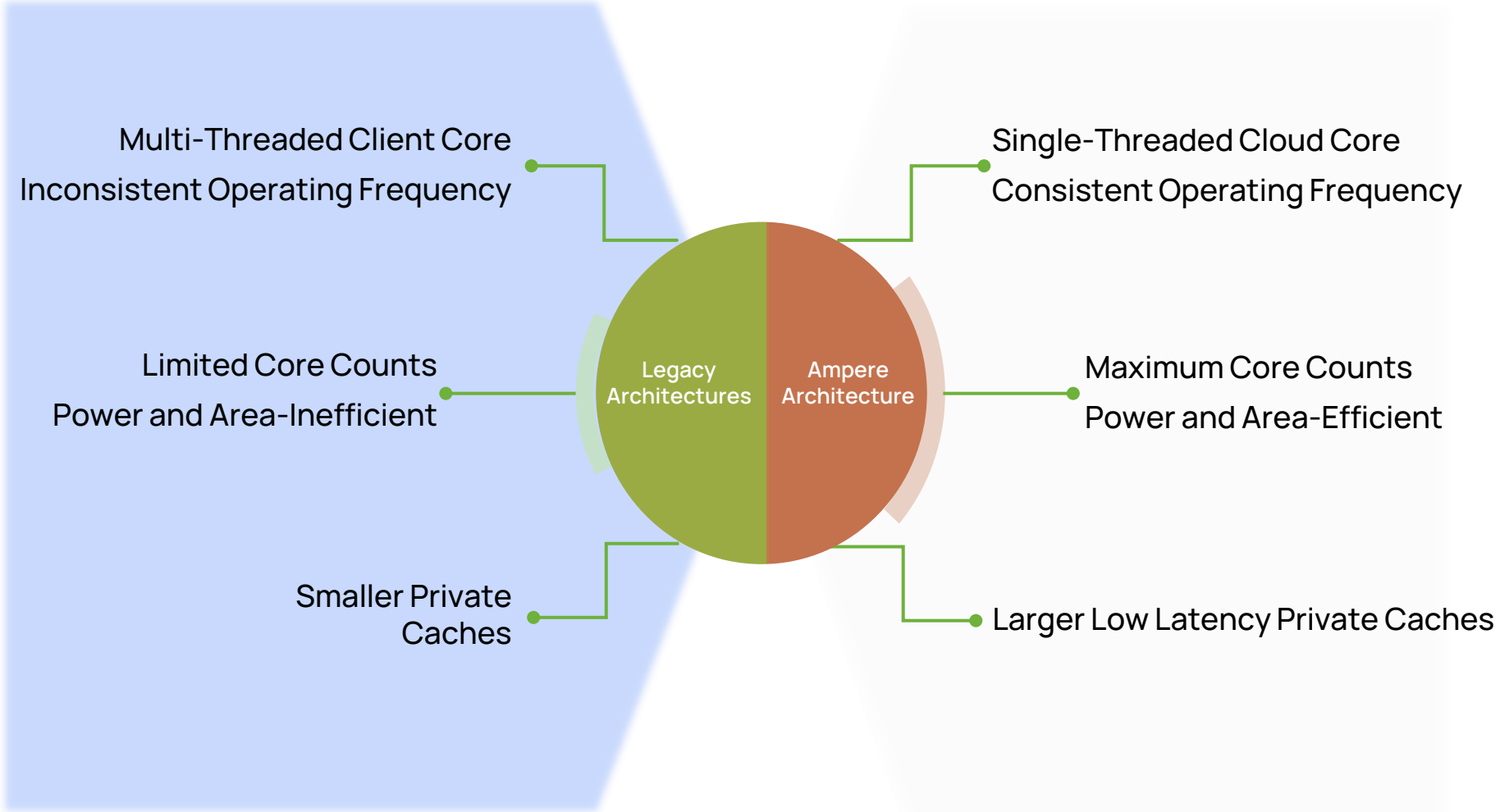
Predictable High Performance

+

Elastic and Scalable

+

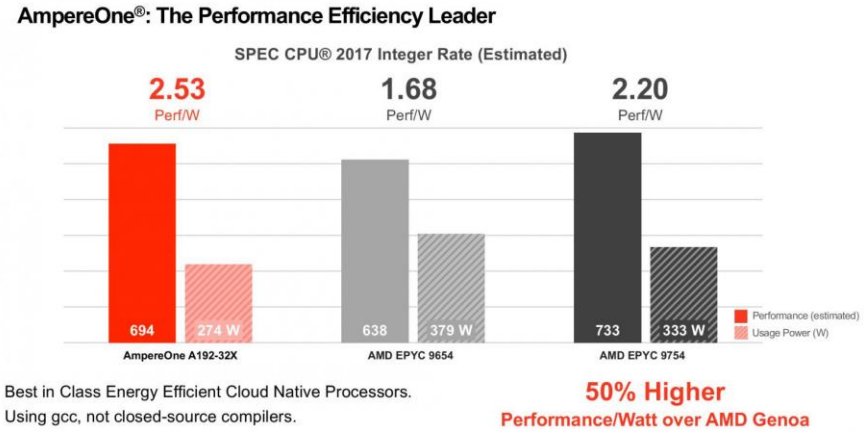
Power Efficient and Sustainable

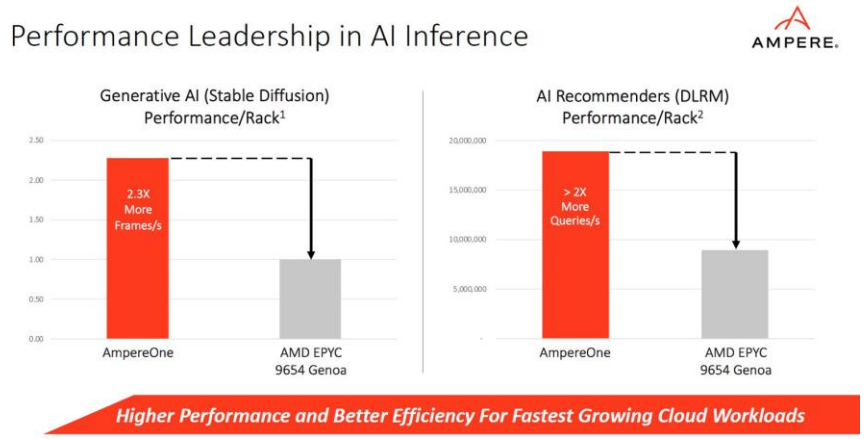


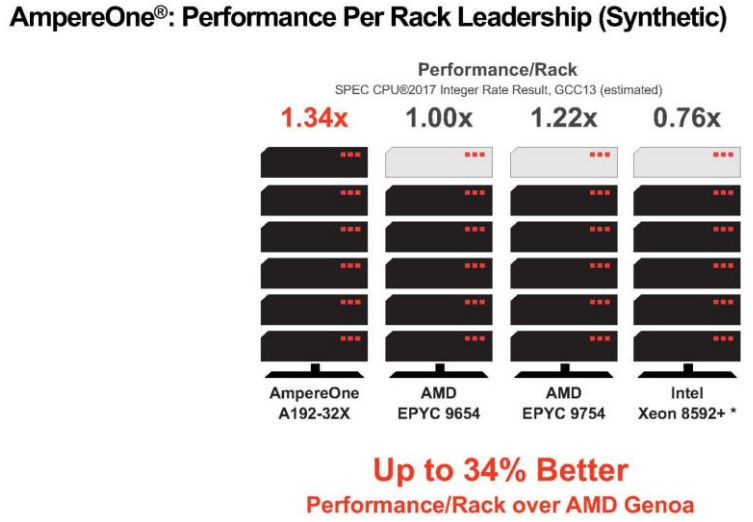
Пример Cloude Native процессоров

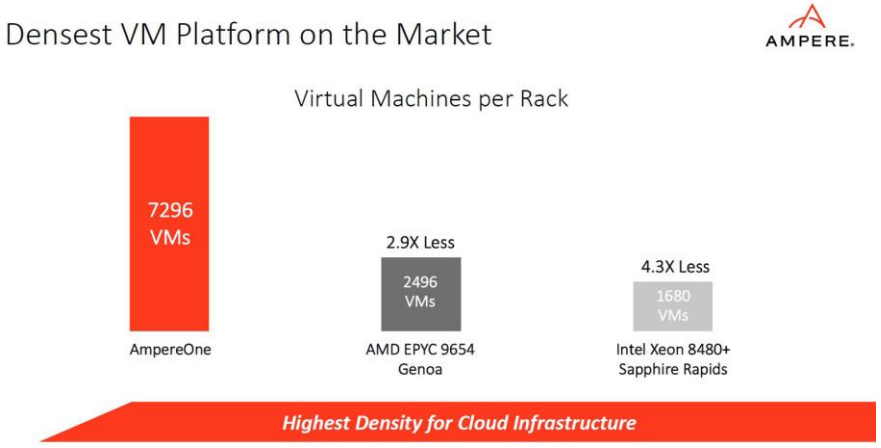
Превосходство Ampere One

- Энергоэффективность выше в 2,5 раза по сравнению с серверами на процессорах AMD EPYC
- На 34% больше производительности на стойку
- В 2,9 раза больше виртуальных машин по сравнению с серверами на процессорах AMD EPYC
- В 4,3 раза больше виртуальных машин по сравнению с серверами на процессорах Intel Xeon

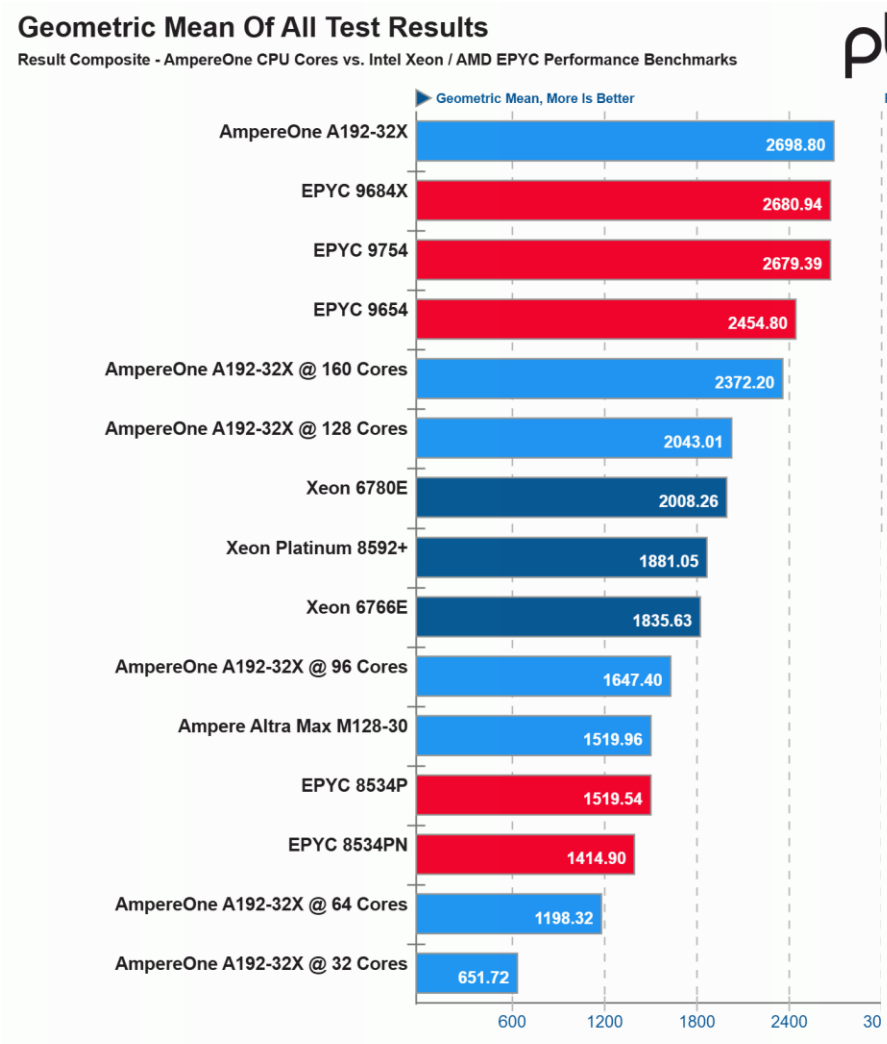


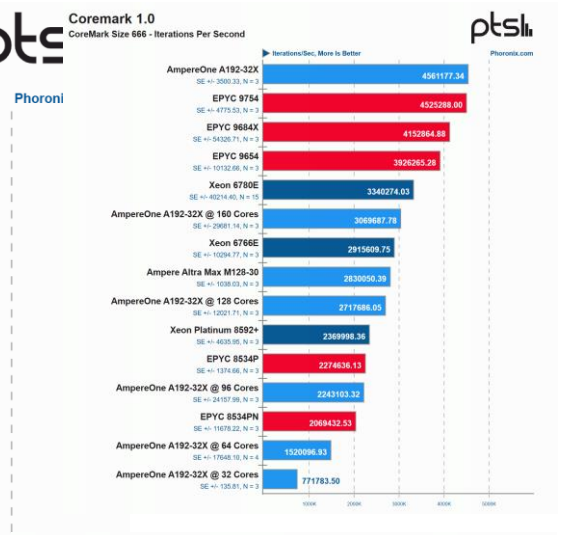


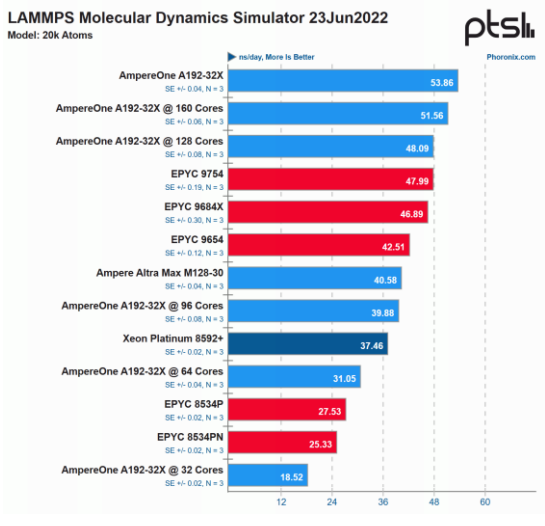


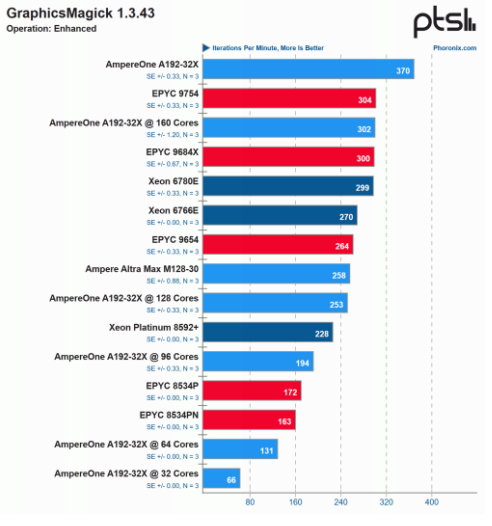


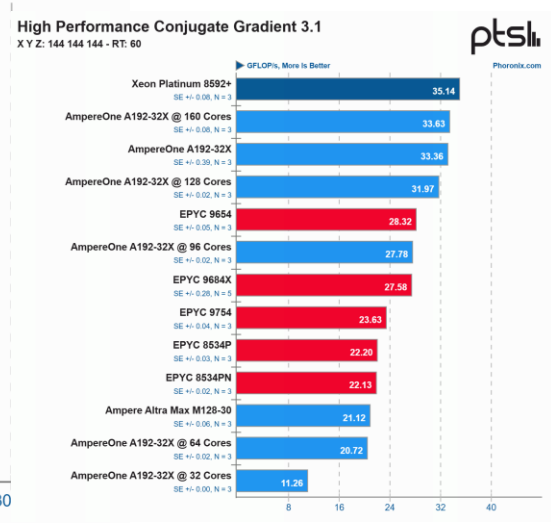
Независимые тесты производительности

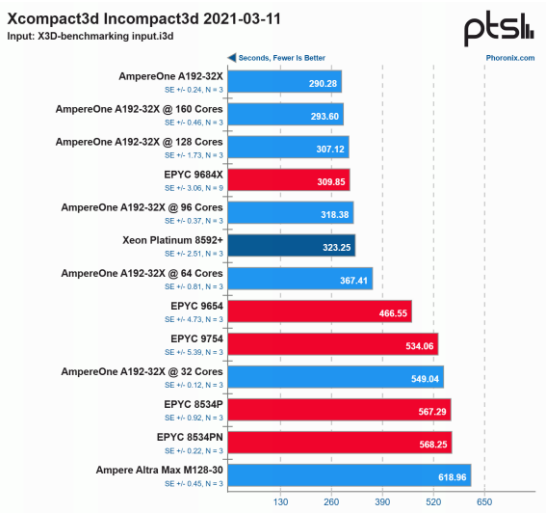


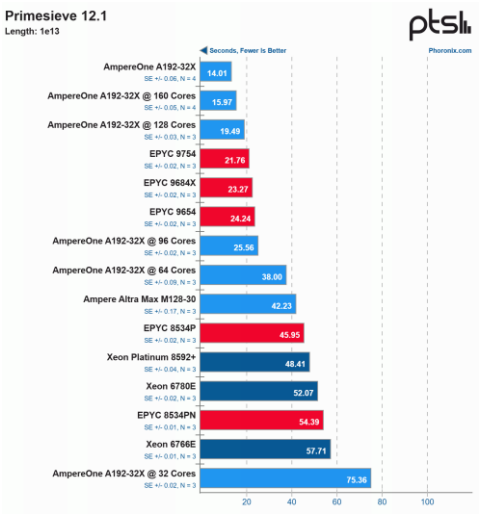












Ampere One в задачах AI Inference

1 MW Data Center with a PUE of 1.5 @ 15 Kw/Rack

Ampere® Altra® Family

40 Servers/Rack
60 Racks – 2400 Servers

Cost = $17.2M
AI Inference Performance = 901M tps

Nvidia DGX H100

1 Server/Rack
60 Racks – 60 Servers

Cost = $27.3M
AI Inference Performance = 190M tps

4.7x More AI Inference and More than $7M in Savings[3]

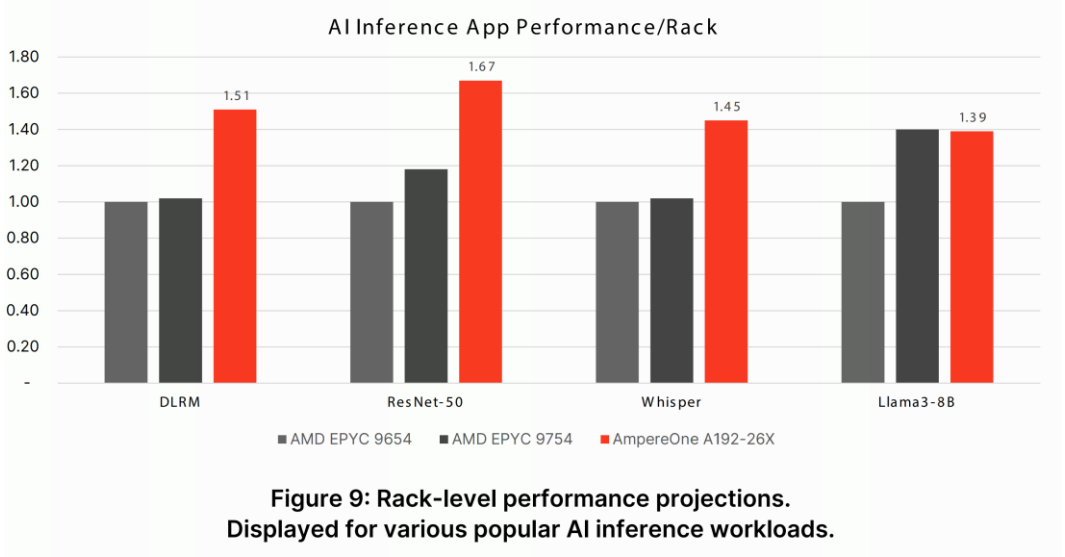
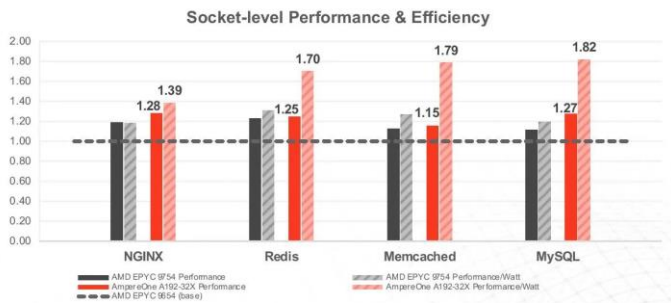


Figure 9: Rack-level performance projections.
Displayed for various popular AI inference workloads.

AmpereOne® for GPU-Free AI Inference

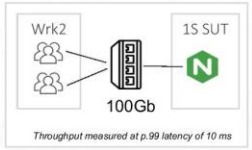
UP TO **15%** more performant and
2.2x more efficient than AMD Genoa

Performance Leadership on popular Cloud Native Apps

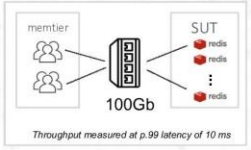


UP TO **28%** more performant and
82% more efficient than AMD Genoa

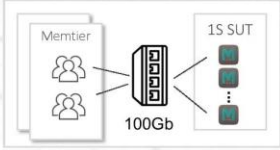
NGINX Test setup



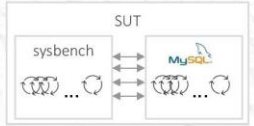
Redis Test setup



Memcached Test setup



MySQL Test setup



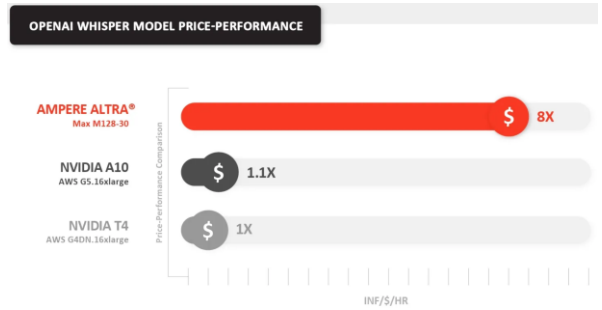
ИИ без использования GPU-ускорителей

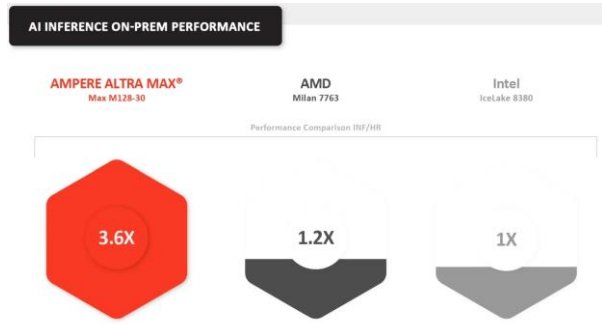
Правильный выбор архитектуры

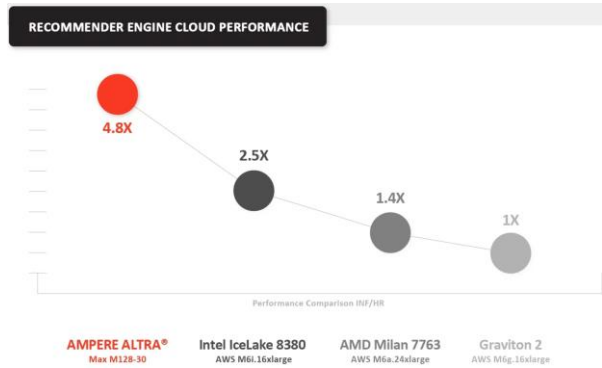
- Выбирайте энергоэффективные совместимые с облаками процессоры
- Развертывайте только тот объем вычислений, который вам нужен для удовлетворения требований производительности вашего приложения
- Объедините GPU с энергоэффективными процессорами для более тяжелых рабочих нагрузок обучения ИИ или вывода LLM

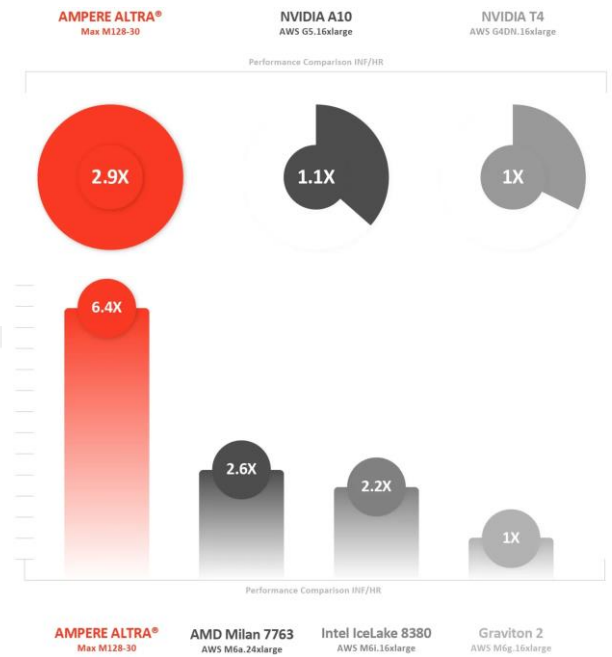
Основные преимущества

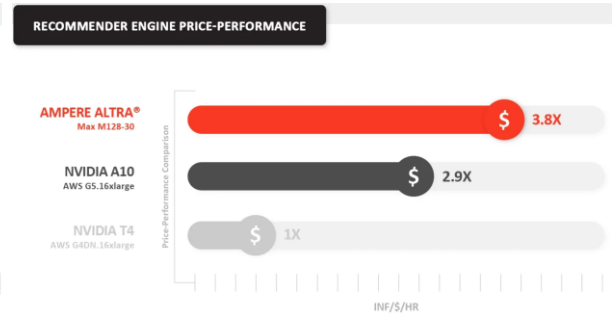
- До 8 раз лучшее соотношение цены и качества модели распознавания речи в облаке
- Производительность модели распознавания речи до 2,9 раз выше
- Повышение производительности вывода ИИ до 3,6 раз при локальном развертывании
- Повышение производительности вывода ИИ в облаке до 6,4 раз
- Повышение производительности рекомендательного механизма в облаке до 4,8 раз
- До 3,8 раз лучшее соотношение цены и производительности рекомендательного механизма в облаке











Аналитика данных

Совместимые программные решения

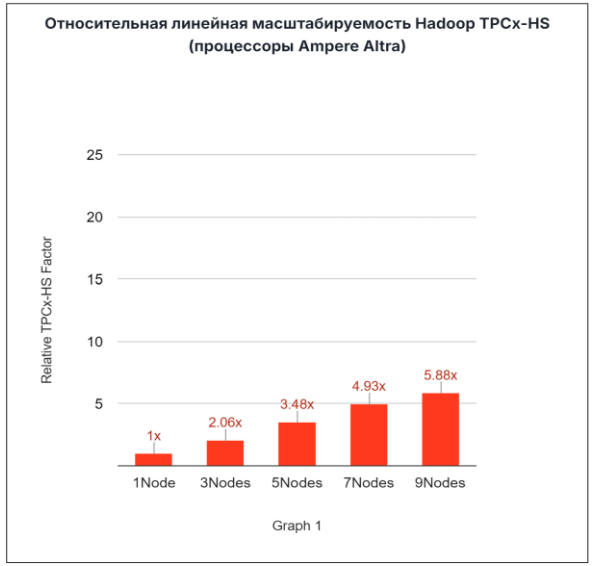
- **HDFS** - компонент уровня хранения больших данных
- **YARN** - управляет ресурсами для приложений
- **MapReduce** – алгоритм распределения заданий по всему кластеру
- **Hadoop** - программная среда с открытым исходным кодом для хранения данных и запуска приложений на кластерах стандартного оборудования
- **Apache Spark** - инженерия данных, машинное обучение
- **Hive** - распределенная система хранения данных
- **Hbase** - база данных, работающая поверх HDFS

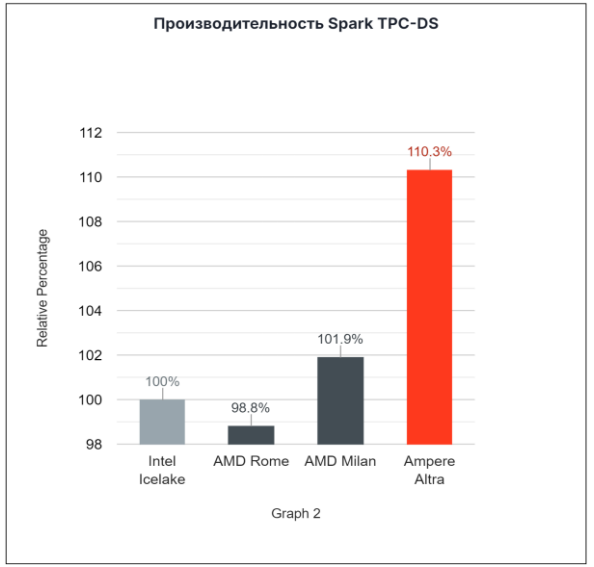
Основные преимущества

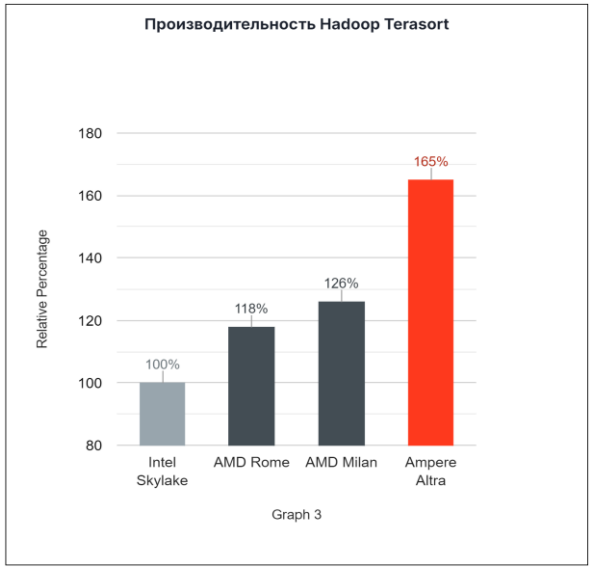
- Линейная масштабируемость
- Производительность облачных технологий
- Предсказуемая производительность при пиковых нагрузках
- Эффективное энергопотребление

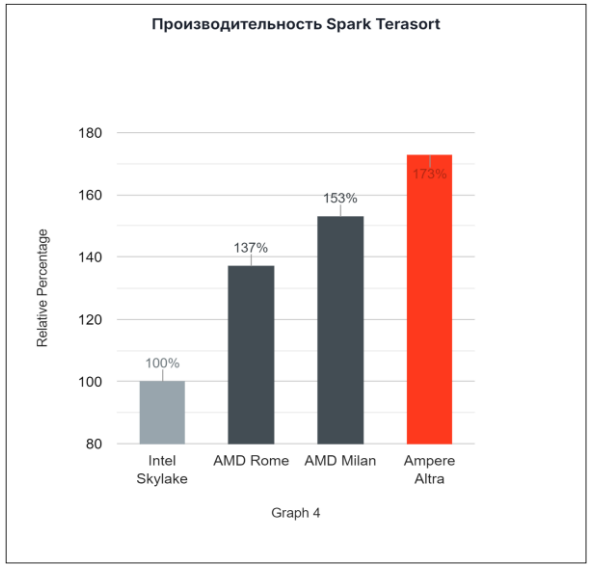
Тестируемая система

- Тесты масштабируемости - кластер из 9 узлов, голое железо
  - ✓ CPU - Single socket per server with Ampere Altra N1 processor with 80 cores @3.0GHz
  - ✓ Memory - Samsung DDR4 512GB DRAM 3200 MHz
  - ✓ Storage - 4 Micron 7300 NVME Drives each with 3.5 TB storage space
  - ✓ Network Cards - Mellanox CX-6, 2 x 100Gb/s with a bonded configuration
- Тесты Terasort Hadoop и Spark на виртуальных машинах:
  - ✓ Intel - Skylake 2000 GHz
  - ✓ AMD - Rome 2249 MHz
  - ✓ AMD - Milan 2449 MHz
  - ✓ Ampere - Altra processors 3000 MHz
  - ✓ vCPU's - 16 ARM Cores or 16 x86 threads.Memory - 64 GB
  - ✓ Storage - 1TB storage with 1000 MB/sec read/write performance









Программно-определяемые хранилища данных

Совместимые программные решения

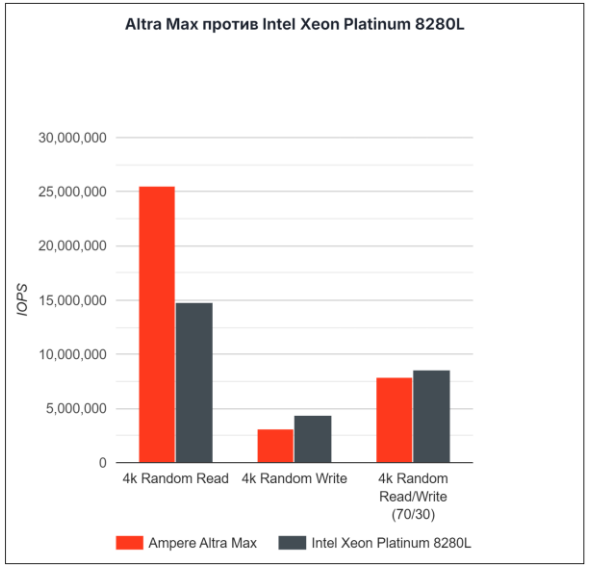
- **LINSTOR** - решение для управления
- **Ceph** - Программно-определяемое решение для хранения данных с открытым исходным кодом (блочное, файловое, объектное хранение)
- **MinIO** – объектное хранилище, доступное в Kibernetes
- **Linbit SDS** - программно-определяемое решение для хранения данных для платформ Linux (доступно для Kubernetes, Red Hat OpenShift и т. Д)

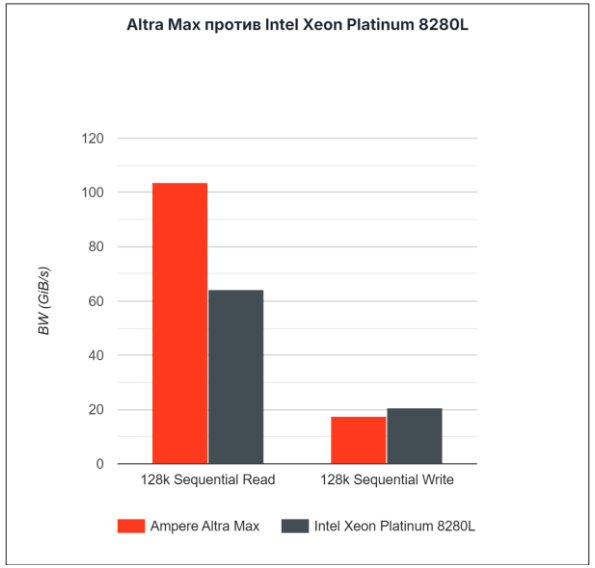
Основные преимущества

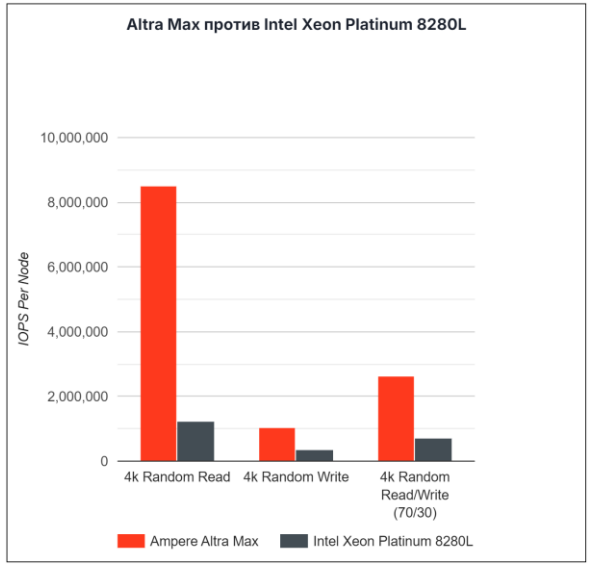
- Обеспечивает до 30 млн IOPS
- Большее количество ядер по сравнению с конкурентами
- Предсказуемая и линейная производительность рабочей нагрузки
- Низкое энергопотребление

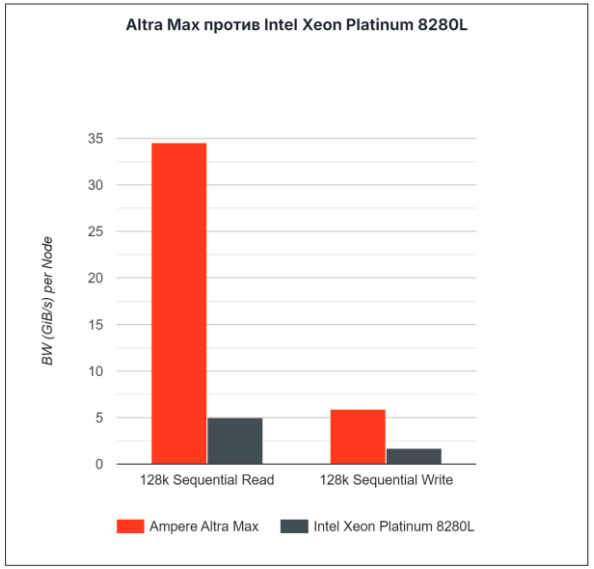
Тестируемая система

- Ampere Altra Max - 128 ядер до 3,0 ГГц на узел
  - ✓ 3-узловой кластер microk8s
  - ✓ Версия LINSTOR: 1.16.0
  - ✓ Версия DRDB: 9.2.0
- 2 процессора Intel Xeon Platinum 8280L на узел
  - ✓ 12-узловой кластер
  - ✓ Версия DRDB: 9.0.21-1









Особенности задач вычислительных кластеров

Классические задачи

- Продолжительные расчеты - могут длиться несколько дней или недель. Конечное время расчета
- Линейность - характеризуются линейными взаимосвязями между переменными, что делает их более простыми и предсказуемыми
- Низкая размерность - традиционные инженерные задачи часто включают в себя небольшое количество переменных и данных, что позволяет использовать простые и эффективные методы решения

Задачи AI/ML

- Краткосрочные итеративные расчеты, не ограниченные по времени задания
- Нелинейность - часто характеризуются нелинейными взаимосвязями между переменными, что делает их более сложными и трудными для решения
- Высокая размерность - часто включают в себя большое количество переменных и данных, что требует использования специальных алгоритмов и методов обработки данных
- Большая ресурсоемкость

Технологии и инструментарий вычислительных кластеров

Для классических задач

- MPI, openMP
- Управление - диспетчер заданий (Slurm)
- Хранение данных – posix ФС (NFS, Lustre), объектные СХД (S3)
- Контейнеризация (HPC) – singularity, enroot
- Grafana+Prometeus+InfluxDB, Nagios
- Точка входа – ssh login-server, инженерные приложения

Для задач AI/ML

- NCCL, CUDA, MPI
- Управление – оркестратор (K8s)
- Хранение данных – объектные СХД (S3), распределенные СХД (HDFS), POSIX ФС (NFS)
- Контейнеризация – docker, containerd …
- Grafana+Prometeus+InfluxDB / Grafana + VictoriaMetrics
- Точка входа – jupyter notebook, ssh login-server, инструменты для управления жизненным циклом машинного обучения
- Pipelines, CI/CD - VRay, MLFlow, Kubeflow, Helm
- Фреймворки - pytorch, tensorflow, keras

Возможен ли универсальный кластер

Различия (требований, инструментов)

- Современные нагрузки требуют большей динамичности, готовности к большей гибкости со стороны инфраструктуры
- Cloud-native application
- Задачи класса Inference (бесконечные)

Совпадения (требований, инструментов)

- Распределение нагрузки в многоузловой кластерной среде
- Общие коммуникационные библиотеки MPI, NCCL
- Способность эффективно задействовать ускорители GPU

Опыт создания универсального кластера на основе Kubernetes

- mpi-operator запускает распределенные приложения через механизмы ssh, а не современных библиотек семейства pmi, pmix
- В контейнерах mpi приложения запускаются под root - к этому могут возникать большие вопросы от специалистов по IB и не только
- В контейнерах, запускаемых в K8s видны все ядра и всё ОЗУ вычислительного узла
- Несоответствие в запущенном контейнере списков устройства и ресурсов IB в файловых системах /sys и /dev
- Современные NCCL-приложения работают без проблем, но при работе NCCL с множеством IB-устройств код приложения, связанный с NCCL, мог вызвать ошибки индексации IB-устройств
- Некоторые расширения для K8s (операторы nvidia-gpu, nvidia-network) реализованы не универсально, а только для определенного набора конкретных ОС, а не для семейств ОС Linux
- Не реализованы алгоритмы и функционал планировщиков заданий (bind/pinning, ranks mapping, backfill, и т.п.)
- Средства ресурсного биллинга и организация пространств команд (KubeSphere, KubeCost, OpenCost)

Возможен ли универсальный кластер

Различия (требований, инструментов)

- Современные нагрузки требуют большей динамичности, готовности к большей гибкости со стороны инфраструктуры
- Cloud-native application
- Задачи класса Inference (бесконечные)

Совпадения (требований, инструментов)

- Распределение нагрузки в многоузловой кластерной среде
- Общие коммуникационные библиотеки MPI, NCCL
- Способность эффективно задействовать ускорители GPU

Опыт создания универсального кластера на основе Kubernetes

- Mpi-operator запускает распределенные приложения через механизмы ssh, а не современных библиотек семейства pmi, pmix
- В контейнерах mpi приложения запускаются под root - к этому могут возникать большие вопросы от специалистов по IB и не только
- В контейнерах, запускаемых в K8s видны все ядра и всё ОЗУ вычислительного узла
- Несоответствие в запущенном контейнере списков устройства и ресурсов IB в файловых системах /sys и /dev
- Современные NCCL-приложения работают без проблем, но при работе NCCL с множеством IB-устройств код приложения, связанный с NCCL, мог вызвать ошибки индексации IB-устройств
- Некоторые расширения для K8s (операторы nvidia-gpu, nvidia-network) реализованы не универсально, а только для определенного набора конкретных ОС, а не для семейств ОС Linux
- Не реализованы алгоритмы и функционал планировщиков заданий (bind/pinning, ranks mapping, backfill, и т.п.)
- Средства ресурсного биллинга и организация пространств команд (KubeSphere, KubeCost, OpenCost)

Элементы модульного сервера «M1/M2»

Сервер

Модульный сервер «M1/M2»

Основные характеристики

- Форм-фактор – 6U
- 10 вычислительных модулей
- Тепловыделение от 1,3 кВт, до 8,5 кВт

Доп. характеристики

- 8 сдвоенных вентиляторов
- 4 блока / преобразователей питания
- Модуль управления сервером

Шасси модульного сервера

Модульное серверное шасси M1РШ

Особенности

- Форм-фактор – 6U
- Поддержка стандарта 19”
- 4 блока питания
- Коммутация сзади, обслуживание спереди

Модульное серверное шасси M1ОШ

Особенности

- Форм-фактор – 6 OU
- Поддержка стандарта OCP 2.0/3.0, 12В, 48В
- Отсутствие блоков питания
- Коммутация и обслуживание спереди

Вычислительные модули

Модуль MB1

Особенности

- 1 слот
- 2 ЦПУ
- 1 PCIe

Модуль MB2

Особенности

- 2 слота
- 2 ЦПУ
- 2 GPU

Модуль MB3

Особенности

- 2 слота
- 4 ЦПУ

Модуль MB4

Особенности

- 2 слота
- 2 ЦПУ
- 5 PCIe

Компоненты

Электронные и корпусные компоненты

Встроенное ПО и ПО управления

Варианты наполнения сервера «M1/M2»

M1/10x MB1

Характеристики

- **10 серверов в шасси**
- До 2560 ядер, до 40 ТБ ОЗУ
- До 80 SSD M.2 (22110) (raw – 640 ТБ)
- До 140x 1-40 Гб/с & 20x 100/200 Гб/с
- Тепловыделение типовое – от 1,8/2,8 кВт
- Тепловыделение пиковое – до 5,5/8,5 кВт
- Производительность до 7,7/61,4 Tflops

M1/5x MB2

Характеристики

- 5 серверов в шасси
- До 1280 ядер, до 20 ТБ ОЗУ
- **До 10 GPU, до 1x 500 Вт или 2x 300 Вт**
- До 40 SSD M.2 (22110) (raw – 320 ТБ)
- До 70x 1-40 Гб/с & 10x 100/200 Гб/с
- Тепловыделение типовое –до 2,2/2,7 кВт
- Тепловыделение пиковое - до 6,5/8,1 кВт
- **Производительность до 101/291 Tflops**

ДОСТУПНО ЛЮБОЕ СОЧЕТАНИЕ ВЫЧИСЛИТЕЛЬНЫХ МОДУЛЕЙ В ШАССИ

- 7 серверов в шасси
- 4x MB1 + 3x MB2

- 6 серверов в шасси
- 2x MB1 + 3x MB2 + 1x MB3

- 8 серверов в шасси
- 6x MB1 + 2x MB2

M1/5x MB3

Характеристики

- 5 серверов в шасси
- До 2560 ядер, до 40 ТБ ОЗУ
- До 80 SSD M.2 (22110) (raw – 640 ТБ)
- До 70x 1-40 Гб/с & 10x 100/200 Гб/с
- Тепловыделение типовое – от 1,7/2,7 кВт
- Тепловыделение пиковое – до 5,2/8,2 кВт
- Производительность до 7,7/61,4 Tflops

M1/5x MB4

Характеристики

- 5 серверов в шасси
- До 1280 ядер, до 20 ТБ ОЗУ
- До 40 SSD M.2 (22110) (raw – 320 ТБ)
- До 70x 1-40 Гб/с & 10x 100/200 Гб/с
- Тепловыделение типовое – от 1,3/1,8 кВт
- Тепловыделение пиковое – до 3,9/5,5 кВт
- Производительность до 3,8/30,7 Tflops

Применимость в ЦОД и HPC

ЦОД 19” и OCP

- Варианты моделей и конфигураций под любые задачи
- Эффективное использование воздушного охлаждения любого типа:
  - ✓ Фрикулинг
  - ✓ Чиллерное охлаждение
- Экономия ресурсов систем электропитания и охлаждения
- Развитая система мониторинга и управления
- Защита инвестиций в шасси
- Снижение расходов на поддержку

Типовая стойка HPC

Характеристики

- До 8 шасси и до 80 серверов в стойке
- До 48U в стойке
- До 7680/20480 ядер, до 120/320 ТБ ОЗУ
- Тепловыделение типовое – от 17,4/22,8 кВт
- Тепловыделение пиковое – до 52/68 кВт
- До 2,3 Pflops в стойке
- Возможность подключения как по Ethernet, так и по InfiniBand

Особенности

- Воздушное охлаждение
- Модуль управления одного из шасси управляет всей стойкой

Типовая стойка ЦОД

Характеристики

- До 4 шасси и до 40 серверов в стойке
- До 24U в стойке
- До 3840/10240 ядер, до 60/160 ТБ ОЗУ
- До 320 SSD M.2 (22110) (raw – 2,5 ПБ) в стойке
- Тепловыделение типовое – от 8,7/11,4 кВт
- Тепловыделение пиковое – до 26/34 кВт
- До 1,16 Pflops в стойке
- До 560x 1-40 GbE & 80x 100/200 GbE

Особенности

- Модуль управления одного из шасси управляет всей стойкой
- Совместимость как с 19”, так и с OCP 3.0 инфраструктурой

HPC

- Высокая плотность вычислительных ресурсов
- Высокая масштабируемость
- Развитая система мониторинга и управления
- Снижение сложности коммутации
- Унификация компонентов
- Длительный жизненный цикл шасси и модулей

Е·Флопс

Спасибо!

www.e-flops.ru

+7 495 795-33-93

info@e-flops.ru