

# OCP Global Summit решения для Computational Storage и компонуемых масштабируемых архитектур

На состоявшейся конференции OCP<sup>1</sup> (Open Compute Project) Global Summit 2021 в ноябре 2021 г. (San Jose, California, USA) был представлен ряд новых разработок в области Computational Storage и компонуемых инфраструктур. Среди компаний, представивших новые решения: Samsung, Inspur, ScaleFlux, Kioxia, SK hynix, Seagate Technology и др. Большая часть анонсированных решений (на середину 2022 г.) уже доступны.

## Samsung

### F2FS

Разработки Samsung с открытым исходным кодом начались еще в 2012 г. (рис. 1, [1]) с файловой системы для флэш-памяти – F2FS (Flash-Friendly File System), разработанной Samsung Electronics для ядра Linux [3]. Мотивом для F2FS было создание файловой системы, которая с самого начала учитывает характеристики устройств хранения данных на основе флэш-памяти NAND (таких как твердотельные диски, eMMC и SD-карты), которые широко используются в компьютерных системах от мобильных устройств до серверов. Поддержка F2FS включена в ядро Linux, начиная с версии 3.8. Параллельно развивается пакет f2fs-tools, содержащий набор утилит для обслуживания разделов F2FS (mkfs.f2fs, fsck.f2fs).

- 1) Open Compute Project (OCP) – это концепция, сообщество и организация, в рамках которых участники в форме открытого диалога делятся разработками в сфере программного, аппаратного и физического проектирования современных центров обработки данных (ЦОД) и оборудования для них. Дата основания OCP – 2011 г. Основной задачей проекта является снижение CAPEX и OPEX инфраструктуры крупномасштабных ЦОД. Объединение включает в себя такие компании, как Facebook, IBM, Intel, AMD, Nokia, Google, Huawei, Microsoft, Seagate Technology, Western Digital, Dell, Rackspace, Cisco, Goldman Sachs, Lenovo, Alibaba Group, Schneider Electric, Samsung и многие другие.

Разработчики, участвующие в деятельности сообщества, стремятся достичь универсальности и простоты масштабирования инфраструктуры вычислительных мощностей. При этом манипуляции, производимые над оборудованием, должны минимально влиять на его работоспособность, производиться в режиме горячей замены и задействовать минимальное количество обслуживающего персонала, а также автоматизировать мониторинг потребляемых ресурсов и наблюдать статистику сбоев ([https://ru.wikipedia.org/wiki/Open\\_Compute\\_Project](https://ru.wikipedia.org/wiki/Open_Compute_Project)).

F2FS разработана специально с учётом специфики флэш-памяти и учитывает такие особенности, как неизменное время доступа и ограниченный ресурс количества циклов перезаписи данных. F2FS была разработана на основе подхода к файловой системе с журнальной структурой (log-structured file system или log-structured merge-tree – журнально-структурное дерево со слиянием, <https://ru.wikipedia.org/wiki/LSM-дерево>), которая адаптирована к новым форматам хранения.

F2FS устраняет некоторые известные проблемы старых файловых систем с журнальной структурой, такие как эффект снежного кома блуждающих деревьев и высокие накладные расходы на очистку. Кроме того, поскольку запоминающее устройство на основе NAND демонстрирует различные характеристики в зависимости

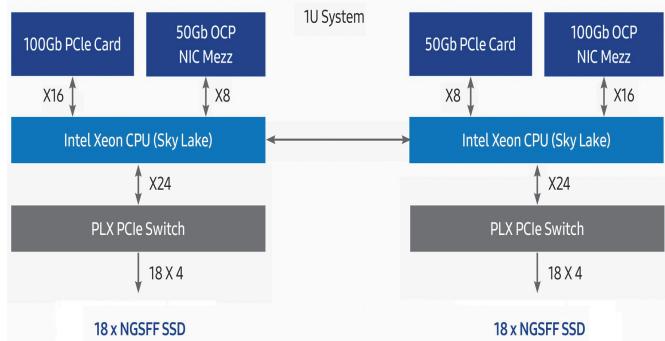


Рис. 2. Data-path архитектура Mission Peak.

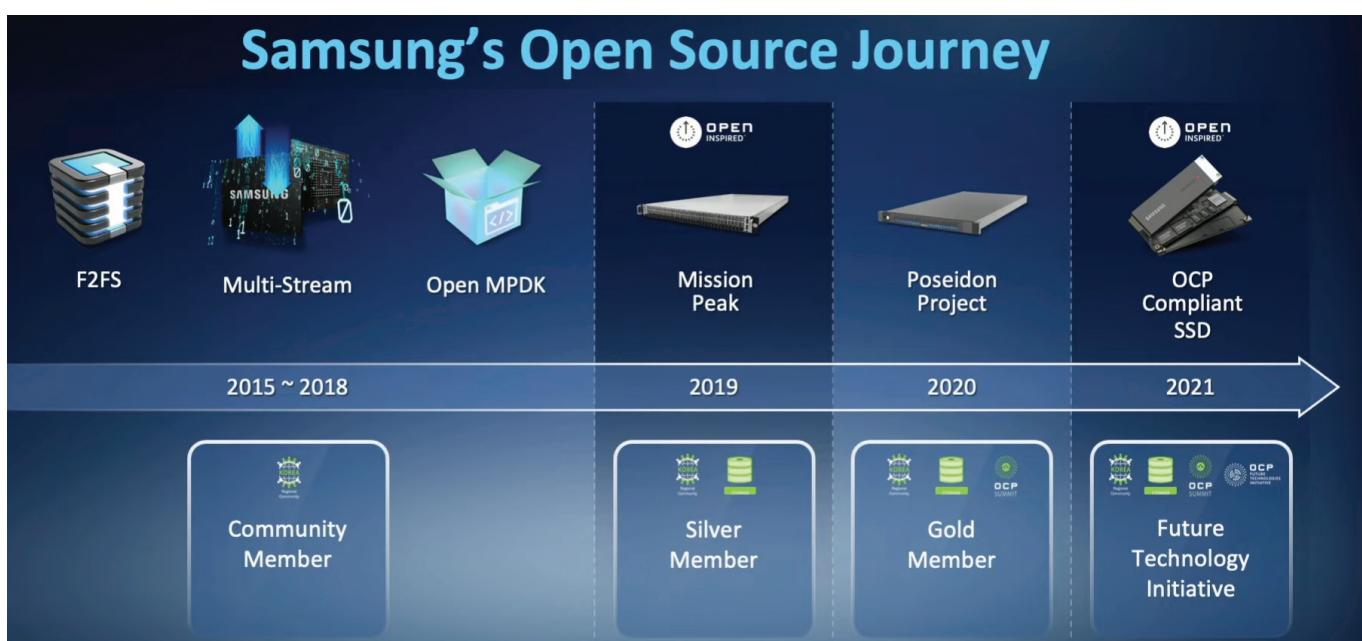


Рис. 1. Вклад Samsung в разработку ОСР-решений [1].

от своей внутренней геометрии или схемы управления флэш-памятью (например, уровня трансляции флэш-памяти или FTL), оно поддерживает различные параметры не только для настройки разметки на диске, но и для выбор алгоритмов выделения и очистки.

Хранение структур данных, организованных в форме LSM-дерева, используется во многих СУБД и предоставляет быстрый доступ по индексу в условиях частых запросов на вставку (например, при хранении журналов транзакций). LSM-деревья, как и другие деревья, хранят пары «ключ — значение». LSM-дерево поддерживает две или более различные структуры, каждая из которых оптимизирована под устройство, в котором она будет храниться. Синхронизация между этими структурами происходит блоками.

#### Mission Peak

Mission Peak — это монтируемый в стойку 1U-сервер хранения данных на базе флэш-памяти в форм-факторе 1U EIA-310D (19 дюймов) на базе твердотельных накопителей NGSFF и поставляемый по OCP-лицензии. Mission Peak был разработан совместно с AIC. Также в разработке принимали участие (по состоянию на 2019 г.) Mellanox, E8 Storage и Memory Solution.

Система обеспечивает производительность хранения 10 млн IOPs при произвольном чтении, сбалансированную с пропускной способностью Ethernet 300 Гбит/с. Эта производительность и баланс с пропускной способностью сети идеально подходят для приложений, передающих сохраненные данные по сети, таких как сети доставки контента и масштабируемые системы хранения. Благодаря двухпроцессорным процессорам Intel Skylake-SP и до 24 модулям памяти DDR4 DIMM система также хорошо подходит для локальных рабочих нагрузок с интенсивным вводом-выводом, таких как аналитика в реальном времени и серверы баз данных.

Система состоит из 36 NGSFF SSD, подключенных к материнской плате через коммутационную матрицу PLX PCIe (с фронтальным обслуживанием), двух процессоров Intel Xeon и двух сетевых карт 100GbE и 2x50GbE. Устройства NGSFF специально разработаны для all-flash серверов и оптимизированы для 1U-конструкций. Это позволяет Mission Peak предлагать в 5 раз большую емкость по сравнению с конструкциями U.2 предыдущего поколения.

#### Poseidon V2 E3.x

В мае 2020 г. (на OCP Virtual Global Summit) Samsung представила новый подход (проект Poseidon, основанный на мульвендорной коллоквиации) к сотрудничеству в области хранения данных с открытым исходным кодом в различных отраслях, в котором особое внимание уделялось разработке платформы с открытым исходным кодом (OSP — open-source platform), фундаментально привязанной к развертыванию облачной инфраструктуры. Одновременно также был представлен эталонный дизайн этой платформы, совместно разработанный с Inspur. Результатом этой программы стала не только новая инновационная аппаратная платформа и первый в мире твердотельный накопитель, признанный OSP, но и вклад в виде ПО с открытым исходным кодом как в существующие, так и в новые проекты.

На OCP Virtual Global Summit 2021 Samsung и Inspur совместно анонсировали уже вторую версию платформы Poseidon — Poseidon V2 E3.x — OCP-совместимое NVMe-oF-решение для дезагрегиро-

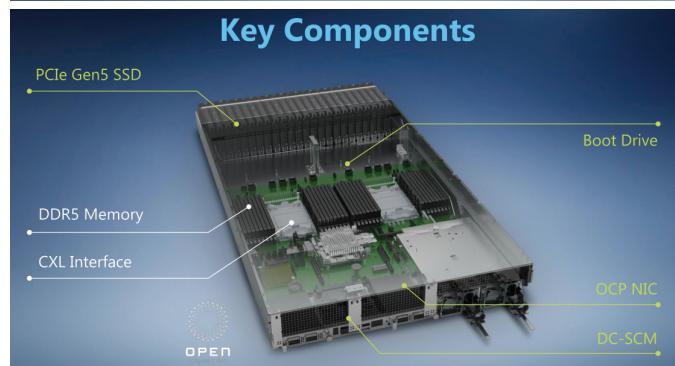


Рис. 3. Ключевые компоненты Poseidon V2 E3.x.

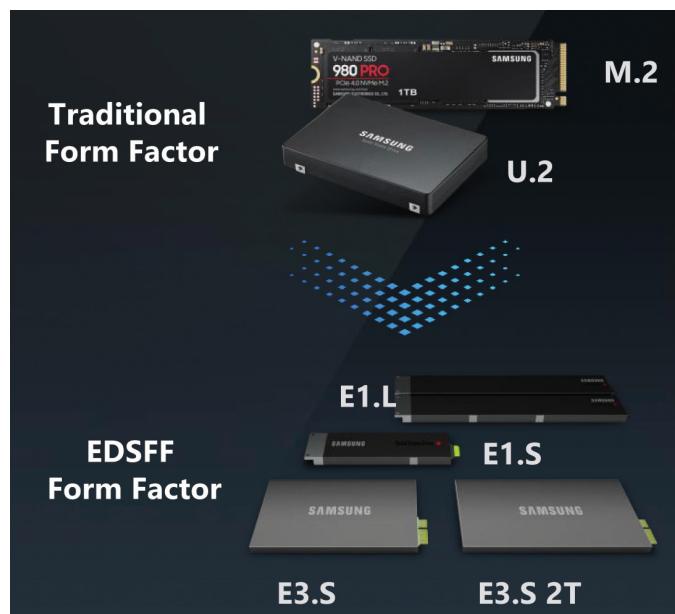


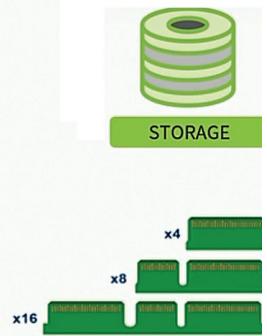
Рис. 4. Традиционные форм-факторы накопителей и форм-фактор EDSFF.



Рис. 5. Компоненты для хранения данных и расширения ОП в составе Poseidon V2 E3.x.

## EDSFF – Benefits

- Designed to overcome conventional device limitations
- Advantageous to capacity & performance scaling
  - ✓ Improves thermal characteristics, power capacity, and scalability
- Provides many flexibilities
  - ✓ Various power (12W, 16W, 20W, 25W, 40W, 70W) and PCB tab (x4, x8, x16) options
  - ✓ Various size: E1.S (5.9/8.01/9.5/15/25mm), E1.L (9.5/18mm), E3.S (1T/2T), E3.L(1T/2T)
- Built in LEDs, carrier-less design



	E1.L	E3	E3 Short, Thin	E3 Long, Thin	E3 Short, Thick	E3 Long, Thick
Size	38.4 x 318.75 x 9.5mm	38.4 x 318.75 x 18mm	76.2 x 112.75 x 7.5mm	76.2 x 142.2 x 7.5mm	76.2 x 112.75 x 16.8mm	76.2 x 142.2 x 16.8mm
Recommended Power(W)	25W	25W(low fan), 40W(high fan - 1.5x)	25W	40W	40W	70W
	E1.S					
Size	31.5 x 111.49 x 5.9mm	31.5 x 111.49 x 8.01mm	33.75 x 110.75 x 9.5mm	33.75 x 110.75 x 15mm	33.75 x 110.75 x 25mm	
Recommended Power(W)	12W	16W	20W	20W	25W	



Рис. 6. Преимущества форм-фактора EDSFF по поддерживаемым мощностям (от 12 до 70 Вт) и широкому диапазону размеров.

вванного и общего хранилища. Решение построено на базе эталонной системы Samsung Poseidon V2 E3.x на базе PCIe Gen5 и на сервере общего назначения Inspur NF5180M6. Решение предлагается для ведущих облачных провайдеров и гиперскэйлеров для построения компонуемых дезагрегированных архитектур (рис. 3, 4). В отличие от первой версии, V2 поддерживает более высокую производительность и плотность упаковки, интерфейсы DDR5 и CXL (DDR5 для near мемоги с низкой задержкой, интерфейсы расширения CXL для “далней” (far) памяти с более высокой задержкой), а также DC-SCM (Datacenter-ready Secure Control Module, [6]). Продукты Samsung SSD и DRAM для Poseidon V2 включают: SSD ZNS+QLC, SSD PCIe Gen5, SmartSSD, Z-SSD, CXL Memory Expander и память DDR5 DRAM.

Poseidon V2 E3.x использует компонуемую архитектуру, чтобы максимально использовать преимущества форм-фактора EDSFF E3.x (Enterprise & Data Center SSD Form Factor, [https://en.wikipedia.org/wiki/Enterprise\\_%26\\_Data\\_Center\\_SSD\\_Form\\_Factor](https://en.wikipedia.org/wiki/Enterprise_%26_Data_Center_SSD_Form_Factor)), смешивая различные типы устройств в зависимости от варианта использования (рис. 5, 6, 7, [4]). Система Poseidon V2 может предоставлять не только высоко-производительное разделяемое SSD-хранилище, но и расширять па-

мять за счет CXL Memory Expander и работать с ускорителями AI/ML (рис. 6). Пользователи центра обработки данных могут настроить систему в соответствии с потребностями приложения.

Помимо аппаратной платформы, Samsung разработала для нее операционную систему PoseidonOS (POS, [5]), которая представляет собой open-source NVMe-oF решение для дезагрегированных хранилищ данных и является облегченной ОС для хранения данных, поддерживающей высокую производительность и ценные функции в сети хранения данных (рис. 8):

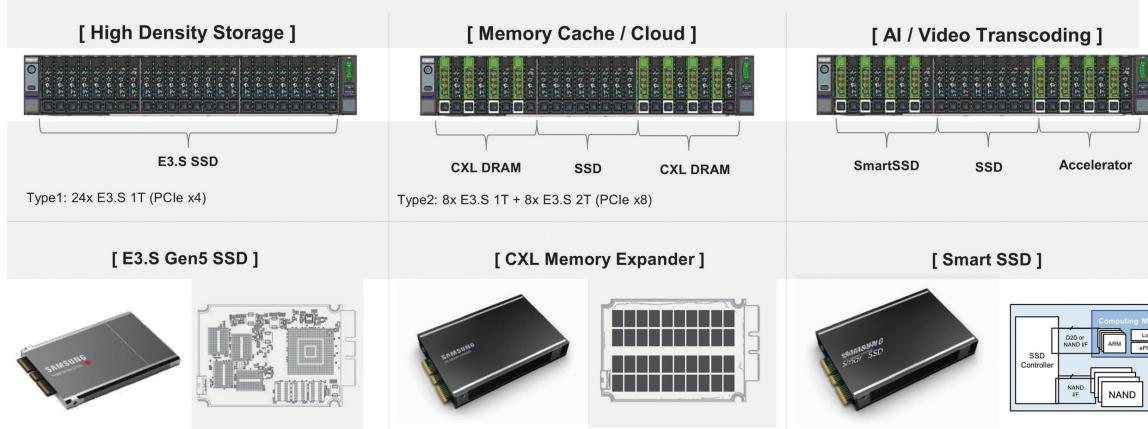
- функции хранения:
  - управление томами;
  - регулирование производительности;
  - провижининг;
- функции поддержания доступности:
  - программный RAID;
  - многоканальное соединение;
  - 2-узловая высокая доступность;
- оптимизация производительности:



STORAGE

## E3 Reference System – PSD V2

- Designed to maximize the benefits of E3.x form factor
- Can configure the system according to application's needs



OPEN POSSIBILITIES.

Рис. 7. Платформа Poseidon V2 E3.x дает возможность ее конфигурирования для различных применений.

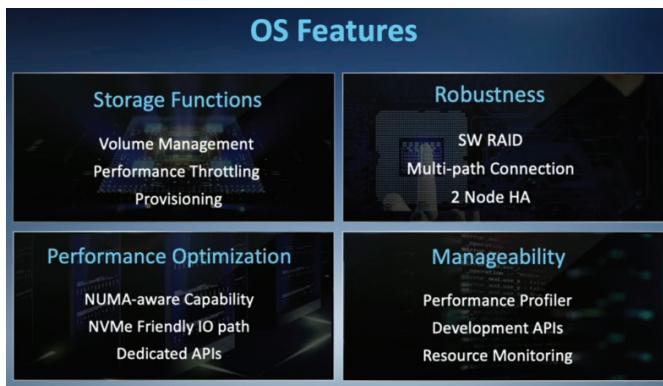


Рис. 8. Функции Poseidon OS.



Рис. 9. План развития платформы Poseidon V2.

- поддержка NUMA;
- дружественный NVMe путь ввода-вывода;
- выделенные API;
- управляемость:
  - профилировщик производительности;
  - API-интерфейсы разработки;
  - мониторинг ресурсов.

POS использует преимущества твердотельных накопителей (SSD) NVMe, оптимизируя стек хранения и используя современный высокоскоростной интерфейс.

План развития платформы Poseidon V2 E3.x представлен на рис. 9. Продукты Poseidon V2 должны быть доступны уже в 2022 году.

## Inspur

Inspur Information и Samsung разделяют общее видение продвижения платформы OCP для более широкого внедрения. Inspur Information входит в тройку крупнейших производителей серверов в мире и участвует в трех крупных организациях открытых вычислений — OCP, ODCC и Open19, что демонстрирует стремление компании взять на себя ведущую роль в продвижении коммерциализации открытых технологий [7].

«После разработки эталонной системы E1.S мы ожидаем, что этот тип решения для хранения данных станет одним из самых востребованных и экономичных решений для хранения данных на рынке для ведущих серверов облачных центров обработки данных и гипермасштабируемых компаний, которые управляют крупными центрами обработки данных», — говорит Джоньюл Ли (*Jongyoul Lee*), исполнительный вице-президент группы разработчиков программного обеспечения памяти Samsung.

«Благодаря нашему объединенному видению дизайна серверов общего назначения Inspur и платформы Poseidon от Samsung мы считаем, что E1.S и E3.x принесут революционный вариант использования, который удовлетворит потребность в эффективной высокопроизводительной системе хранения данных с высокой плотностью», — заявил Алан Чанг (*Alan Chang*), вице-президент по технической эксплуатации Inspur Information. — Клиенты, которые используют серверы общего назначения в качестве своих вычислений, могут плавно перейти на Poseidon, чья модульная конструкция сократит избыточное проектирование и проверку по всем направлениям. Мы ожидаем еще более широких моделей использования и приложений с новой спецификацией Poseidon v2».

С 2016 г. Inspur сотрудничает с Liqid в части построения компонуе-

мых дезагрегированных инфраструктур (CDI-решений, Composable Disaggregated Infrastructure). Результатом этого партнерства является стойка Inspur GPU CI (Composable Infrastructure) Rack (рис. 10, [8]).

Inspur GPU CI Rack, интегрированный с Liqid Composable, обеспечивает беспрецедентную гибкость в масштабе стойки, отражая следующую эволюцию в архитектуре центра обработки данных. Данное решение позволяет пользователям управлять, масштабировать и настраивать физические серверы без операционной системы за считанные секунды, а также имеет усовершенствованную систему управления, обеспечивающую подлинную дезагрегацию стандартных серверных компонентов.

Inspur GPU CI Rack позволяет изначально развертывать несколько элементов графической обработки (GPU) в фабрике PCI-Express (PCIe) и мгновенно динамически распределять их на любой узел Inspur. Назначение ресурсов GPU-to-CPU можно настраивать и перенастраивать с помощью автоматизации на основе политик в режиме реального времени по мере необходимости изменения инфраструктуры .

## Оборудование Inspur GPU CI Rack

### Шасси расширения: GX4 /9J

2U платформа расширения графического процессора (JBOG) с подключением по PCIe, вмещающая 4 графических процессора (double-wide), с гибкой топологией для различных применений и поддерживающая несколько типов графических процессоров (NVIDIA-certified; V100, P100, P40, M40, K80, M60, M10 и др.).

### Liquid Grid

Интеллектуальная управляемая коммутационная фабрика со сверхнизкой задержкой, которая электрически соединяет пулы разрозненных элементов системы.

### Вычислительные узлы: i24

2U с четырьмя 2-сокетными половинной ширины узлами на процессорах Intel Xeon Scalable с 16 модулями DIMM DDR4 и до 6 твердотельных накопителей и одним сетевым адаптером OCP на узел.

### ПО Liquid Command Center

ПО для управления компонуемой инфраструктурой, которое автоматизирует, оркестрирует и динамически компонует машины без ПО (bare-metal) из пулов разрозненных элементов без ПО (bare-metal).

## *Inspur Information AI серверы с GPU NVIDIA A100 с тензорными*

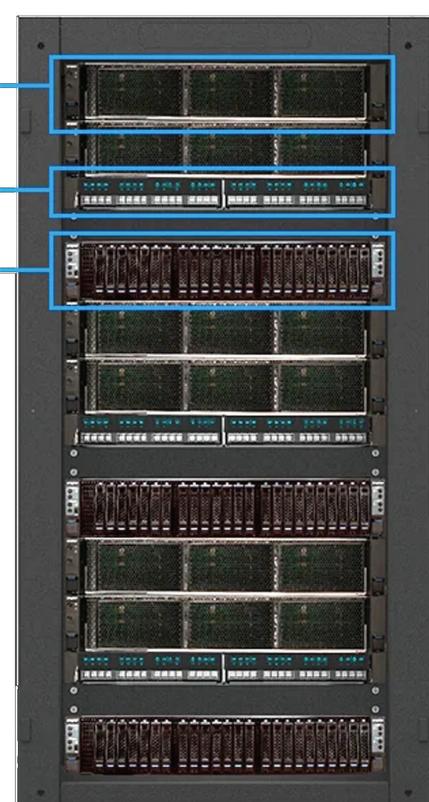


Рис. 10. Inspur GPU CI Rack.

## *ядрами сохраняют лидирующие позиции по производительности одного узла в тестах в MLPerf Training v2.0 [10]*

В начале июля 2022 г. открытый инженерный консорциум MLCommons™ опубликовал последние результаты MLPerf™ Training v2.0, в которых серверы Inspur AI лидируют по производительности одиночных узлов.

MLPerf — самый влиятельный в мире тест AI-производительности. Он управляет MLCommons, в который входят представители более 50 ведущих мировых компаний в области искусственного интеллекта и ведущих академических учреждений, включая Inspur Information, Google, Facebook, NVIDIA, Intel, Гарвардский университет, Стэнфордский университет и Калифорнийский университет в Беркли. Тесты MLPerf AI Training проводятся два раза в год для отслеживания улучшений в производительности вычислений и предоставления пользователям авторитетных данных.

В последней версии MLPerf Training v2.0 приняли участие 21 глобальных производителей и исследовательских учреждений, включая Inspur Information, Google, NVIDIA, Baidu, Intel-Habana и Graphcore.

Среди тестов закрытого подразделения (closed division benchmarks) для одноузловых систем серверы Inspur AI показала лучшие результаты в обработке естественного языка с помощью BERT, рекомендации с помощью DLMR и распознавании речи с помощью RNN-T. Среди основных высокопроизводительных AI-серверов, оснащенных восемью GPU NVIDIA A100 Tensor Core, серверы Inspur Information AI заняли первое место в пяти задачах (BERT, DLMR, RNN-T, ResNet и Mask R-CNN).

Серверы Inspur AI продолжают достигать прорывов в производительности искусственного интеллекта благодаря комплексной оптимизации программного и аппаратного обеспечения. По сравнению с результатами MLPerf v0.5 в 2018 году, серверы Inspur AI продемонстрировали значительное повышение производительности до 789% для типичных моделей серверов с 8 GPU.

Лучшая производительность серверов Inspur AI в MLPerf является результатом инноваций в дизайне и возможностей оптимизации полного стека для ИИ. Для высоконагруженного совместного планирования задач с несколькими GPU передача данных между узлами NUMA и GPU оптимизирована, чтобы гарантировать, что ввод-вывод данных в задачах обучения находится в состоянии максимальной производительности. Что касается рассеивания тепла, Inspur Information лидирует, развернув восемь NVIDIA Tensor Core A100 GPU (<https://www.nvidia.com/en-us/data-center/a100>) мощностью 500 Вт в пространстве 4U и поддерживая воздушное и жидкостное охлаждение. Серверы Inspur AI продолжают оптимизировать производительность обработки данных перед обучением и применяют комбинированные стратегии оптимизации, такие как гиперпараметр и NCCL-параметр, а также множество улучшений, предоставляемых программным стеком NVIDIA AI, чтобы максимизировать производительность обучения AI-модели.

### *Значительное улучшение результатов тренировок сети Transformer*

Предварительно обученные массивные модели, основанные на архитектуре нейронной сети Transformer, привели к разработке алгоритмов искусственного интеллекта нового поколения. Модель BERT в тестах MLPerf основана на архитектуре Transformer. Лаконичная и наращиваемая архитектура Transformer делает возможным обучение массивных моделей с огромным числом параметров. Это привело к значительному улучшению алгоритмов больших моделей, но требует более высоких требований к производительности обработки, взаимосвязи связи, производительности ввода-вывода, параллельных расширений, топологии и рассеиванию тепла для систем ИИ.

В тесте BERT серверы искусственного интеллекта Inspur еще больше повысили производительность обучения BERT за счет использования таких методов, как оптимизация предварительной обработки данных, улучшение плотной передачи параметров между графическими процессорами NVIDIA, автоматическая оптимизация гиперпараметров и т. д. Серверы информационного искусственного интеллекта Inspur могут выполнять обучение модели BERT примерно с 330 млн параметров всего за 15 869 минут с использованием 2 850 176 фрагментов данных из набора данных Википедии, повышение производительности на 309% по сравнению с максимальной производительностью 49,01 минуты в Training v0.7. На данный момент серверы Inspur AI выиграли тест MLPerf Training BERT в третий раз подряд.

Два сервера Inspur AI с лучшими результатами в MLPerf Training v2.0 — это NF5488A5 и NF5688M6. NF5488A5 — один из первых серверов в мире, который поддерживает восемь GPU NVIDIA A100 с тензорными ядрами с технологией NVIDIA NVLink и два процессора AMD Milan в пространстве 4U. Он поддерживает как жидкостное, так и воздушное охлаждение. Он выиграл в общей сложности 40 титулов MLPerf. NF5688M6 — это масштабируемый сервер искусственного интеллекта, предназначенный для оптимизации крупномасштабных центров обработки данных. Он поддерживает восемь GPU NVIDIA A100 с тензорными ядрами и два процессора Intel Ice Lake, до 13 интерфейсов ввода-вывода PCIe Gen4 и выиграл в общей сложности 25 титулов MLPerf.

## **ScaleFlux**

ScaleFlux<sup>2)</sup>, основанная в 2014 г. и разрабатывающая решения Computational Storage, является членом многих сообществ, продвигающих стандартизацию и унификацию этого класса решений. В частности, ScaleFlux — участник OCP Open Domain-Specific Architecture (ODSA) Subproject [11]. С момента своего создания в марте 2019 года в рамках OCP подпроект ODSA предпринял важные шаги в определении и разработке архитектуры на основе чиплетов сведением новых интерфейсов, уровней канала, а также ранней проверки концепции.

Десятилетия прогресса в области CPU общего назначения замедлились, в то время как требования к производительности рабочих нагрузок резко возросли, что привело к значительному спросу на ускорители для конкретных предметных областей. Конструкции на основе чипсетов, объединяющие несколько кристаллов в одном корпусе, могут сократить время разработки и стоимость производства ускорителей. Согласно предварительному исследовательскому отчету IHS/Informa, совокупный рынок чиплетов к 2024 году составит почти 3 млрд долларов, а к 2030 году вырастет до 10 млрд долларов.

Миссия подпроекта ODSA состоит в том, чтобы определить открытый интерфейс и архитектуру, которые позволяют смешивать и сопоставлять кремниевые чиплеты от разных поставщиков через открытый рынок на одной SoC. Для достижения этой цели в рамках ODSA было создано 9 рабочих групп [12, 13].

### **Семейство продуктов ScaleFlux 3000**

В ноябре 2021 г. на саммите OCP ScaleFlux продемонстрировала новое семейство продуктов ScaleFlux 3000 на базе своего нового чипа ScaleFlux SFX 3000. В семейство также входит ПО, оптимизированное для вычислительных хранилищ («CSware»), которое упрощает для клиентов внедрение вычислительных хранилищ в более широком диапазоне вариантов использования [14].

Благодаря набору продуктов следующего поколения ScaleFlux пользователи могут более легко развертывать Computational Storage, решая различные проблемы, такие как стоимость и плотность хранения данных, эффективность вычислений (производительность на сервер, на ватт), несогласованность задержек, надежность SSD и перемещение данных.

Отличительной особенностью накопителей ScaleFlux является то, что в мае 2021 г. в них была интегрирована флэш-память NAND с четырьмя ячейками (QLC) от Micron Technology. По словам компании, новые накопители QLC с интегрированным и прозрачным сжатием могут снизить стоимость флэш-памяти до уровня менее 0,01 доллара за гигабайт в год. При этом производительность обеспечивается на уровне более дорогих дисков TLC — до 6 раз быстрее, чем NVMe SSD QLC. Выносимость находится на одном уровне с дисками TLC, что до 4 раз выше, чем у других SSD QLC [16].

Кроме того, ScaleFlux представила CSware: сочетание готового кода и примера кода для расширения внедрения и углубления ценности и простоты использования функций вычислительного хранилища. Это упрощает для клиентов процесс внедрения вычислительных накопителей в дополнительных сегментах рынка, таких как традиционные твердотельные накопители NVMe, карты процессоров вычислительных хранилищ (CSP) и решения SoC.

Новый набор продуктов ScaleFlux, основанный на технологии Arm®, состоит из четырех инновационных предложений, в том числе:

2) ScaleFlux является пионером в развертывании масштабируемого вычислительного хранилища. Computational Storage — это основа современной инфраструктуры, управляемой данными, которая обеспечивает высокую производительность, доступное масштабирование и гибкие платформы для приложений, интенсивно использующих вычисления и операции ввода-вывода. Компания ScaleFlux, основанная в 2014 году, является хорошо финансируемым стартапом, лидером которого доказали свою способность развертывать сложные вычислительные решения и твердотельные хранилища в больших объемах (<https://www.scaleflux.com/intro/>).

- SoC серии SFX 3000 для блоков обработки данных позволяет производителям накопителей и оборудования разрабатывать собственные твердотельные накопители и карты-ускорители. Для поставщиков приводов включение встроенного ПО ScaleFlux под ключ сократит цикл разработки привода и снизит затраты на разработку;
- диски NVMe Computational Storage серии CSD 3000 позволяют пользователям сократить расходы на хранение данных в 3 раза, при этом производительность приложений удваивается, а срок службы флэш-памяти увеличивается в 9 раз по сравнению с обычными дисками. Кроме того, пользователи могут развертывать распределенные вычислительные функции для конкретных приложений с помощью ядер Arm;
- диски NVMe SSD серии NSD 3000 представляют пользователям более интеллектуальные SSD NVMe, обеспечивающие двукратное увеличение выносливости и двукратное увеличение производительности при произвольной записи и смешанном чтении/записи по сравнению с другими накопителями NVMe;
- серия CSP 3000 обеспечивает сжатие, шифрование и программируемые функции для пользователей, желающих использовать Computation Storage, но не желающих развертывать CSD;
- оптимизированное ПО CSware Computational Storage позволяет получить максимальную отдачу от использования CSD, NSD и CSP ScaleFlux.

#### ScaleFlux SFX 3000 Storage Processor

ScaleFlux SFX 3000 Storage Processor является основой семейства ScaleFlux 3000. Процессор хранения ScaleFlux SFX 3000 сочетает в себе хост-интерфейс PCIe Gen4 x8 с 16-канальным интерфейсом NAND, что позволяет создавать лучшие в своем классе конструкции твердотельных накопителей. Процессоры хранения серии SFX 3000 основаны на базовой архитектуре управления флэш-памятью и технологиях ускорения, разработанных ScaleFlux в поколениях CSS 1000 и CSD 2000. SFX 3000 состоит из 4 основных функциональных блоков (рис. 11):

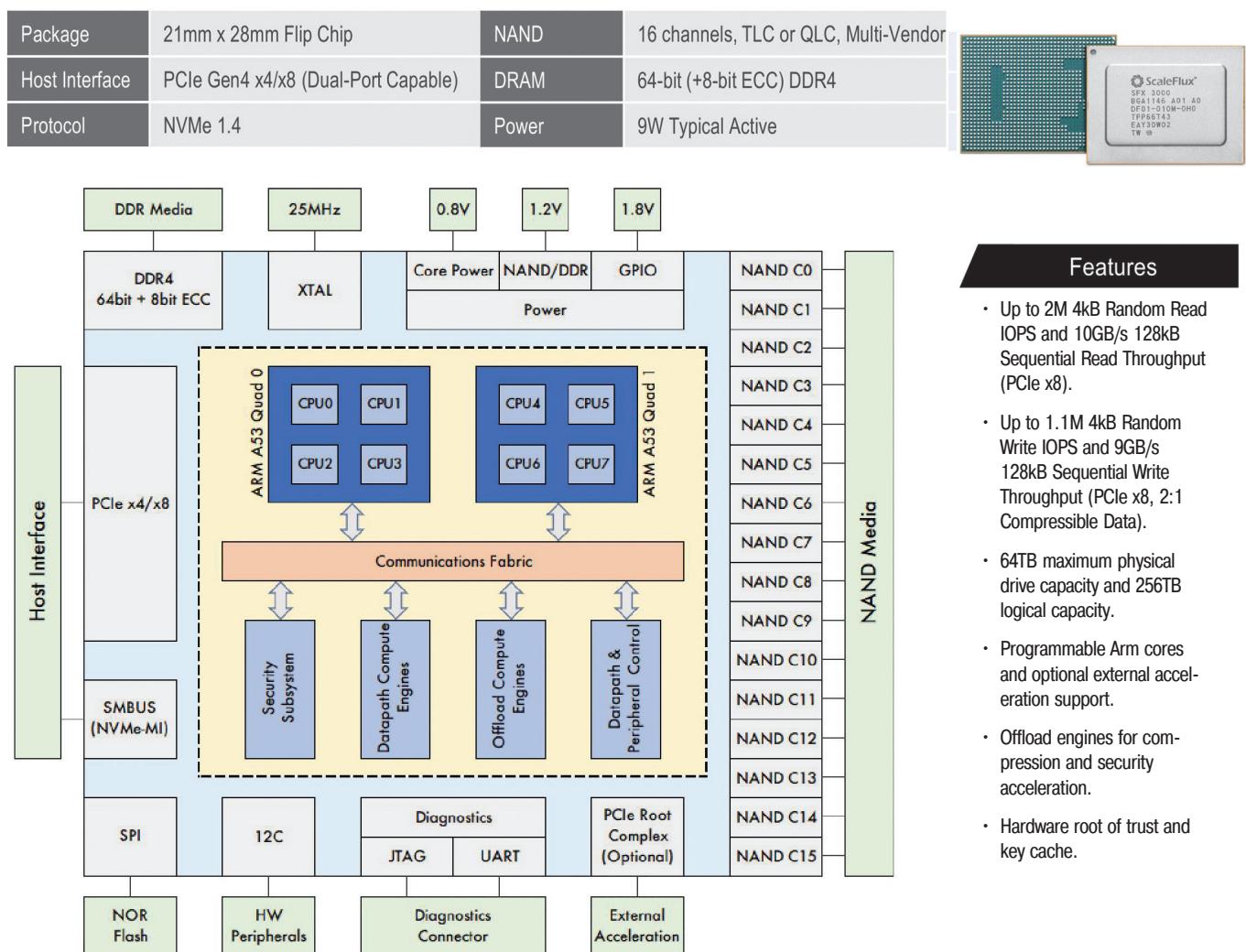


Рис. 11. Архитектура и особенности ScaleFlux SFX 3000 Storage Processor.

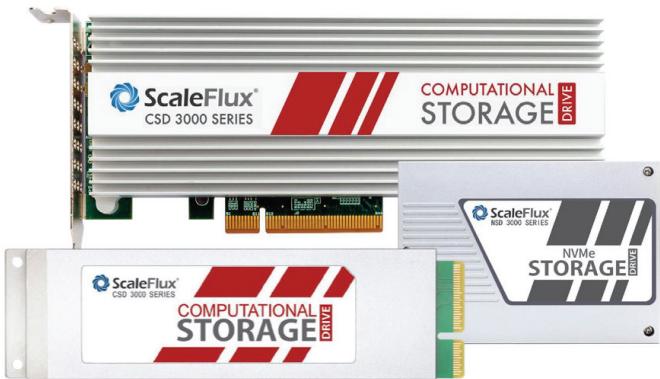


Рис. 12. SSD ScaleFlux CSD 3000 и NSD 3000.

Табл. 1. Отличительные особенности ScaleFlux CSD 3000 и NSD 3000.

	NSD 3000	CSD 3000
Transparent Compression & Decompression	✓	✓
Leading IO Performance & Low Latency	✓	✓
Enhanced Endurance	✓	✓
Available with TLC & QLC Flash	✓	✓
NVMe Compliant	✓	✓
Data-at-Rest Encryption with TCG Opal	✓	✓
Extendable Capacity		✓
Programmable ARM® Cores		✓
Hardware Acceleration Engines		✓

CSD 3000 сочетает функции высокопроизводительного NVMe SSD с аппаратными вычислительными модулями (Hardware Compute Engines) и программируемым ядром, обеспечивая гибкое и простое в использовании вычислительное решение для хранения данных. Функциональные особенности CSD 3000:

- высокопроизводительный Enterprise & Data Center NVMe SSD:
  - совместимость с NVMe 1.4 & OCP NVMe Cloud SSD Spec;
  - стандартный форм-фактор: U.2, U.3, E1.x, HHHL Add-in Card;
  - хост-интерфейс X8 PCIe Gen 4 поддерживает однопортовые от 1x1 до 1x8 и двухпортовые конфигурации;
  - TLC & QLC NAND;
  - до 64TB физической емкости;
  - конфигурируемый overprovisioning;
  - TCG Opal 2.0;
  - защита данных при потере питания;
  - атомная запись;
  - >=2.5 млн часов наработка на отказ (MTBF);

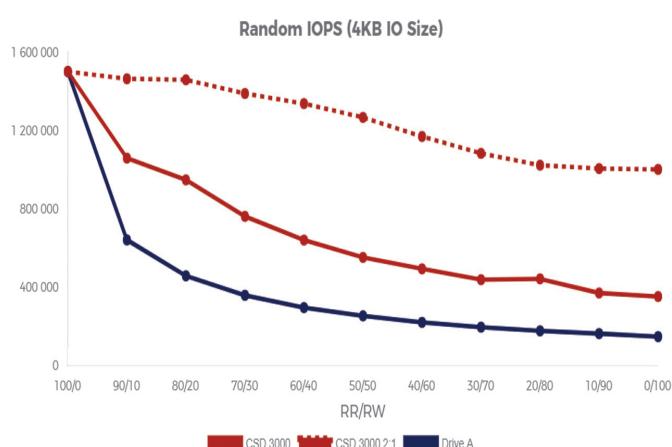


Рис. 13. Производительность CSD3000 при изменении соотношения операций read/write и коэффициента компрессии.

- функции computational storage:

- прозрачное без штрафных санкций (“Penalty-Free”) сжатие, что умножает срок службы, производительность (рис. 12) и емкость хранилища данных, а также повышает энергоэффективность и экономичность на несколько порядков, чем при сжатии в ядрах CPU;
- до 256TB емкость хранилища с помощью функции Capacity Multiplier;
- программируемые ядра для пользовательских функций.

IO-производительность CSD3000 при изменении соотношения операций read/write и коэффициента компрессии представлена на рис. 13.

#### ScaleFlux CSP 3000

Вычислительные процессоры хранения данных серии CSP 3000 обеспечивают платформу для распределения и распараллеливания вычислительных функций. CSP 3000 отделяет функции ускорения и разгрузки от устройства хранения, позволяя пользователям развертывать его с любыми дисками (не только с дисками ScaleFlux).

Пользователи CSP 3000 получают ускоритель со встроенными аппаратными вычислительными ядрами (HCE, hardware compute engines) и программируемыми ядрами, которые помогают им эффективно масштабировать производительность системы.

HCE отлично подходят для обработки задач с фиксированным алгоритмом, таких как сжатие, которые могут потреблять огромное количество циклов ЦП общего назначения. Благодаря специально разработанным HCE эти задачи выполняются быстрее, при этом потребляется на несколько порядков меньше энергии.

Помимо фиксированных функций HCE, CSP 3000 может помочь в выполнении функций подготовки данных, таких как фильтрация, с помощью своих программируемых ядер Arm®.

CSP 3000 будет поддерживать стандартные API и соответствовать новым стандартам рабочих групп Computational Storage в Ассоциации производителей сетей хранения данных (SNIA) и NVMe.org.

#### CSware

С помощью CSware ScaleFlux упрощает получение большей отдачи от ваших решений для хранения данных ScaleFlux. CSware включает в себя не только готовые приложения, но и улучшения существующих приложений, интеграцию в приложения, подключаемые модули и примеры кода.

Для первой волны CSware ScaleFlux разрабатывает 3 проекта:

- *KallaxDB* — хранилище KV, оптимизированное для вычислительных накопителей;
- *CSware RAID* — программный RAID, использующий все преимущества прозрачного сжатия на дисках;
- *оптимизация В-дерева* — модификации В-дерева для повышения производительности произвольной записи в реляционных базах данных.

Хранилища KV, такие как RocksDB, обычно используются для обработки больших объемов неструктурированных данных, особенно когда отдельные элементы данных имеют небольшой размер (например, <256 байт). Структуры данных в хранилищах KV могут быть довольно сложными, требующими больших объемов памяти и высокой вычислительной мощности процессора.

Используя прозрачное сжатие на диске, KallaxDB использует простую структуру данных, что значительно снижает нагрузку на CPU и память при индексировании. Результаты:

- более высокая эффективность;
- меньшая нагрузка на CPU на операцию;
- требуется меньше DRAM на ТБ данных в наборе данных;
- лучшая производительность;
- больше операций в секунду;
- меньшая задержка на операцию.

Реляционные базы данных, такие как MySQL, PostgreSQL, MariaDB, Oracle, SQL Server и MongoDB, используют структуры данных В-дерева. В-деревья отлично подходят для быстрого чтения. Однако они могут страдать от низкой производительности записи, поскольку случайные обновления/вставки могут вызвать очень сильное увеличение времени записи В-дерева.

ScaleFlux продемонстрировал, что В-дерево может использовать прозрачное сжатие на диске, чтобы значительно уменьшить увеличение времени записи, что приводит к гораздо более высокой производительности записи В-дерева. Результаты:

- 10-кратное уменьшение времени записи;
- улучшенная производительность записи и задержка;
- поддержка производительность чтения и эффективности хранения.

Благодаря широкому распространению NVMe SSD для высоко-производительных приложений пользователи должны сбалансировать затраты на обеспечение отказоустойчивости на своем уровне флэш-памяти NVMe с производительностью.

Конфигурация RAID 5 экономит общий объем хранилища, необходимый для хранения данных пользователя. Но это сопряжено со значительными накладными расходами на расчеты четности и требует больше времени для восстановления данных.

RAID 10 избавляет от накладных расходов на вычисление четности и повышает производительность восстановления. Но это значительно увеличивает затраты на хранение, так как требуется 2 ТБ дискового пространства на 1 ТБ пользовательских данных.

ScaleFlux CS RAID может предоставить пользователям более высокую производительность, чем у RAID 10, при более низких затратах на хранение, чем у RAID 5. ScaleFlux CS RAID отслеживает степень сжатия данных и физическое свободное пространство на дисках. Когда данные достаточно сжимаемы, ПО использует избыточное свободное пространство на дисках, чтобы работать как RAID 10 без увеличения физического количества дисков или физической емкости каждого диска. Результаты:

- снижение нагрузки на CPU при расчетах четности;
- стоимость хранилища и структура емкости как у RAID 5;
- производительность и время восстановления RAID 10.

## KIOXIA

### *Новые форм-факторы NVMe SSD выходят получают признание*

На 2021 OCP Global Summit KIOXIA America, Inc. продемонстрировала последние дополнения к своему обширному портфелю SSD для центров обработки данных. KIOXIA представила Enterprise and Data Center Standard Form Factor (EDSFF) E3.S, CD7 Series PCIe® 5.0 SSD как часть первой в отрасли “живой”<sup>3)</sup> демонстрации привода E3.S. Компания также представила свои Data Center SSD серии XD6, совместимые с EDSFF E1.S и Open Compute Platform (OCP) NVMe® Cloud SSD [17].

Провозглашенный форм-фактором NVMe SSD будущего, EDSFF позволяет использовать твердотельные накопители следующего поколения для будущих архитектур центров обработки данных, поддерживая при этом множество новых устройств и приложений. KIOXIA является активным участником отраслевой разработки решений EDSFF E1.S/L и E3.S/L и сотрудничает с ведущими разработчиками центров обработки данных, серверов и систем хранения, чтобы раскрыть весь потенциал флэш-памяти, NVMe и PCIe технологий.

*Серия KIOXIA CD7.* EDSFF E3.S SSD, разработанные с использованием технологии PCIe 5.0, увеличивают плотность флэш-памяти на диск для оптимизации энергоэффективности и консолидации стоек по сравнению с SSD форм-фактора 2,5 дюйма.

*Серия KIOXIA XD6.* EDSFF E1.S SSD для центров обработки данных были первыми<sup>4)</sup> SSD EDSFF E1.S, отвечающими конкретным требованиям гипермасштабируемых приложений, включая требования к производительности, мощности и температуре согласно спецификации OCP NVMe Cloud SSD.

Также было представлено *KumoScale™ Shared Accelerated Storage Software*. Данное ПО делает возможным масштабирование сетевой флэш-памяти NVMe в рамках центра обработки данных.

«Форм-фактор EDSFF E3 призван изменить подход к проектированию корпоративных серверов и хранилищ, — отметил Грег Вонг, основатель и главный аналитик Forward Insights. — Мы ожидаем, что рынок перейдет на EDSFF, начиная с 2022 года, с появлением систем на базе PCIe 5.0».

3) Источник: Корпорация KIOXIA, по состоянию на 9 ноября 2021 г.

4) На основе обзора общедоступной информации по состоянию на 3 ноября 2020 г.

5) По сравнению с накопителями KIOXIA серии CD6.

Основные характеристики накопителя серии CD7:

- форм-фактор EDSFF E3.S с емкостью до 7,68 ТБ;
- разработан в соответствии с последней спецификацией PCIe 5.0 и оптимизирован для производительности x2 PCIe lane;
- использование меньшего количества линий PCIe увеличивает количество поддерживаемых устройств PCIe;
- построен на флэш-памяти KIOXIA BiCS FLASH™ 3D TLC;
- скорость чтения до 6450 МБ/с и 1050 тыс. операций ввода-вывода в секунду при произвольном чтении;
- задержки чтения 75 мкс и записи 14 мкс, что на 17% и 60% ниже, чем у накопителей PCIe 4.0 SSD предыдущего поколения соответственно.

### *Серия EM6 обеспечивает новые варианты использования для приложений NVMe-oF*

Также KIOXIA America, Inc. объявила [18] о начале производства SSD серии EM6 Enterprise NVMe-oF™ для систем Ethernet Bunch of Flash (EBOF). Используя конвертерный контроллер Marvell® 88SN2400 NVMe-oF SSD для преобразования NVMe® SSD в двухпортовый 25 Гбит/с NVMe-oF SSD, накопители KIOXIA серии EM6 предоставляют всю пропускную способность SSD для сети.

Благодаря способности масштабировать производительность NVMe SSD, собственные архитектуры NVMe-oF хорошо подходят для таких приложений, как искусственный интеллект (AI)/машинное обучение (ML), высокопроизводительные вычисления (HPC) и расширение хранилища.

В случае HPC использование файловой системы Lustre® для обеспечения высокой пропускной способности и параллельного доступа к вычислительным кластерам, выгодно применение хранилищ на основе NVMe-oF таких, как системы EBOF с EM6 SSD, которые обеспечивают высокую доступность (HA) конфигурации. Пример конфигурации HPC HA состоит из нескольких резервных сетевых подключений между вычислительным хостом и EBOF с подключенными к 88SN2400 NVMe SSD для обеспечения масштабируемой пропускной способности в зависимости от количества SSD.

В случае использования расширения дезагрегированного блочного хранилища с высокой доступностью одна и та же конфигурация может вместить несколько хост-систем, обеспечивая общее и масштабируемое блочное хранилище с высокой утилизацией пропускной способности NVMe SSD.

Накопители KIOXIA серии EM6 успешно прошли сертификационные испытания NVIDIA® GPUDirect Storage® (GDS). GDS подключается к системам на базе NVIDIA A100 для ускорения приложений AI/ML, а платформы хранения EBOF позволяют масштабировать флэш-хранилище NVMe для этих ресурсоемких рабочих нагрузок.

Основные характеристики серии EM6:

- одиночное или двойное подключение к сети 25Gb Ethernet и RoCEv2;
- совместимость со спецификациями NVMe-oF 1.1 и NVMe 1.4;
- 2,5-дюймовый форм-фактор Z-высоты 15 мм;
- выносимость 1 DWPD с емкостями 3840 ГБ, 7680 ГБ.

Накопители KIOXIA EM6 Series уже доступны через Ingrasys — дочернюю компанию Foxconn — на платформе ES2000 EBOF (<https://www.ingrasys.com/es2000>). ES2000 — это система хранения данных высотой 2U, вмещающая до 24 накопителей форм-фактора 2,5 дюйма, и может быть сконфигурирована с несколькими сетевыми подключениями для повышения пропускной способности и резервирования.

### *Серия CD8 повышает производительность по сравнению с твердотельными накопителями PCIe 4.0<sup>5)</sup> на 135%.*

22 марта 2022 г. компания KIOXIA America, Inc. объявила о том, что пробует твердотельный накопитель PCIe® 5.0 второго поколения для корпоративных клиентов и центров обработки данных. Будучи первым поставщиком, предложившим накопитель с интерфейсом PCIe 5.0, KIOXIA представила новое семейство твердотельных накопителей NVMe™ для центров обработки данных серии CD8. Твердотельные накопители CD8 удваивают пропускную способность на линию по сравнению с твердотельными накопителями PCIe 4.0 с 16 гигатранзакций в секунду (ГТ/с) до

32 ГТ/с и оптимизированы для рабочих нагрузок гипермасштабируемых центров обработки данных и корпоративных серверов.

«Сегодня твердотельные накопители PCIe 4.0 считаются лидерами с точки зрения обеспечения высочайшего уровня производительности накопителей, — прокомментировал Грег Вонг, основатель и главный аналитик Forward Insights. — Твердотельные накопители PCIe 5.0 следующего поколения обеспечивают удвоенный уровень производительности и будут продолжать стимулировать рынок твердотельных накопителей PCIe/NVMe, который, как ожидается, будет расти со среднегодовым темпом роста более 20% до 2026 года».

Основанный на технологии флэш-памяти BiCS FLASH™ TLC 3D 5-го поколения KIOXIA, корпоративный твердотельный накопитель 7-го поколения серии CD8 использует фирменный контроллер и микропрограмму KIOXIA в форм-факторе 2,5 дюйма<sup>6)</sup>, 15 мм по оси Z. Новые накопители разработаны в соответствии со спецификациями PCIe 5.0®, Open Compute Project (OCP) Datacenter NVMe SSD и NVMe 1.4 и хорошо подходят для приложений и сценариев использования, включающих высокопроизводительные вычисления, искусственный интеллект, уровень кэширования, потоковую передачу контента, а также финансовый трейдинг и анализ.

#### Ключевые особенности CD8 SSD:

- выносливость 1 DWPD для интенсивного чтения, предназначенная для гипермасштабируемых и ориентированных на серверные нагрузки; емкость от 960 гигабайт до 15,36 терабайт;
- смешанное использование с ресурсом 3 DWPD, предназначенная для корпоративных рабочих нагрузок и рабочих нагрузок с интенсивным записью; емкость от 800 гигабайт до 12,8 терабайт;
- обеспечение скорости последовательного чтения 7,2 ГБ/с и скорость произвольного чтения 1,25 млн операций ввода-вывода в секунду;
- обеспечение скорости последовательной записи 6,0 ГБ/с и производительности произвольной записи 200 КБ на 135% больше<sup>5)</sup>, чем в версии предыдущего поколения;
- доступно несколько вариантов защиты устройства, включая очистку с мгновенным стиранием (SIE<sup>7)</sup>) и диск с самошифрованием (SED<sup>8)</sup>) в 2,5-дюймовом форм-факторе.

#### Источники, доп. ресурсы

- [1] 2021 OCP Global Summit – Open Innovation The Next Step in Memory Evolution, Jongyoul Lee, Executive Vice President Samsung – [https://www.youtube.com/watch?v=nsIvVm\\_e5ko](https://www.youtube.com/watch?v=nsIvVm_e5ko).
- [2] 2021 OCP Global Summit – <https://www.opencompute.org/events/past-events/2021-ocp-global-summit>.
- [3] F2FS – <https://en.wikipedia.org/wiki/F2FS>.
- [4] Challenges and Opportunities of EDSFF based Storage Solution – <https://www.youtube.com/watch?v=mzFCiO-UyKA>.
- [5] Poseidon OS – <https://github.com/poseidonos/poseidonos>.
- [6] Datacenter Secure Control Module Specification – <https://www.opencompute.org/documents/ocp-dc-scm-spec-rev-1-0-pdf>.
- [7] Samsung Advance Joint Open Storage Solution for OCP – <https://www.inspursystems.com/newsroom/inspur-samsung-advance-joint-open-storage-solution-for-ocp/>.
- [8] Inspur GPU CI (Composable Infrastructure) Rack – <https://www.inspursystems.com/composable-infrastructure-solutions/>.
- [9] GX4 2U Quad GPU Box. Flexible Expansion, Ultra High-Performance – <https://www.inspursystems.com/product/gx4>.
- [10] Inspur Information AI Servers with NVIDIA A100 Tensor Core GPUs Maintain Top Ranking in Single-Node Performance in MLPerf Training v2.0 – <https://www.inspursystems.com/newsroom/inspur-information-ai-servers-with-nvidia-a100-tensor-core-gpus-maintain-top-ranking-in-single-node-performance-in-mlperf-training-v2-0-global-ai-benchmarks/>.
- [11] The OCP Open Domain-Specific Architecture (ODSA) Subproject Makes Significant Gains in Chiplet-based Architecture, Design and Industry Collaboration – <https://www.globenewswire.com/news-release/2019/09/26/1921320/0/en/The-OCP-Open-Domain-Specific-Architecture-ODSA-Subproject-Makes-Significant-Gains-in-Chiplet-based-Architecture-Design-and-Industry-Collaboration.html>.
- [12] Server/ODSA – <https://www.opencompute.org/wiki/Server/ODSA>.
- [13] OCP ODSA Workshop - Prototyping Chiplet Based Open Data Accelerators, 2021-08-27 – <https://www.opencompute.org/events/past-events/ocp-odsaworkshop-prototyping-chiplet-based-open-data-accelerators>.
- [14] ScaleFlux Announces Next Generation Portfolio, Bringing Highly Efficient Computational Storage to the Masses, 04 nov. 2021 – <https://www.globenewswire.com/fr/news-release/2021/11/04/2327623/0/en/ScaleFlux-Announces-Next-Generation-Portfolio-Bringing-Highly-Efficient-Computational-Storage-to-the-Masses.html>.
- [15] ScaleFlux 3000 Family of Products – <https://www.scaleflux.com/landing/csd3000family>.
- [16] THE 10 COOLEST SSD AND FLASH STORAGE PRODUCTS OF 2021 – <https://www.crn.com/slide-shows/storage/the-10-coolest-ssd-and-flash-storage-products-of-2021/10>.
- [17] KIOXIA America Addresses the Future of Flash Storage in Hyperscale Data Centers at OCP Global Summit – <https://business.kioxia.com/en-us/news/2021/ssd-20211109-1.html>.
- [18] KIOXIA Announces Production Availability of Native Ethernet Flash-Based SSDs – <https://business.kioxia.com/en-us/news/2021/ssd-20211111-1.html>.
- [19] KIOXIA Introduces 2nd Generation SSDs for Enterprise and Hyperscale Data Centers Designed with PCIe 5.0 Technology – <https://business.kioxia.com/en-us/news/2022/ssd-20220322-1.html>.

6) «2,5 дюйма» указывает форм-фактор SSD. Он не указывает физический размер диска.

7) SIE: Опция Sanitize Instant Erase поддерживает Crypto Erase, стандартизированную функцию, определенную техническими комитетами (T10) INCITS (Международный комитет по стандартам информационных технологий).

8) SED: вариант диска с самошифрованием поддерживает TCG Enterprise SSC.