# СКС для GPU–кластеров. Первые шаги

13-й форум Data Centre Design & Engineering
Москва, 26 мая 2026 года

Семенов Андрей Борисович, д.т.н., профессор МГСУ и МТУСИ

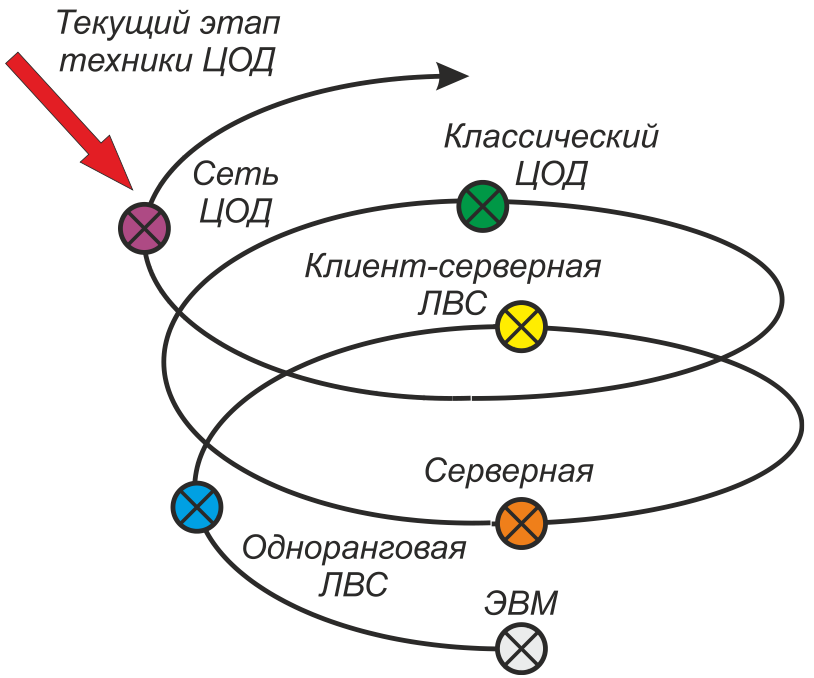
# Разновидности ЦОДов как технического объекта – 1(2)

ЦОДы по определению ГОСТ Р 58812—2020 представляют собой связанную систему ИТ-инфраструктуры, инженерной инфраструктуры, оборудование (серверное и сетевое) и части которых размещены в здании или помещении, подключенном к внешним сетям, как инженерным, так и телекоммуникационным. ЦОД пользуются большой популярностью при построении различных информационных систем. При необходимости здание ЦОД может иметь прилегающую территорию, но эта особенность в данном случае значения не имеет.

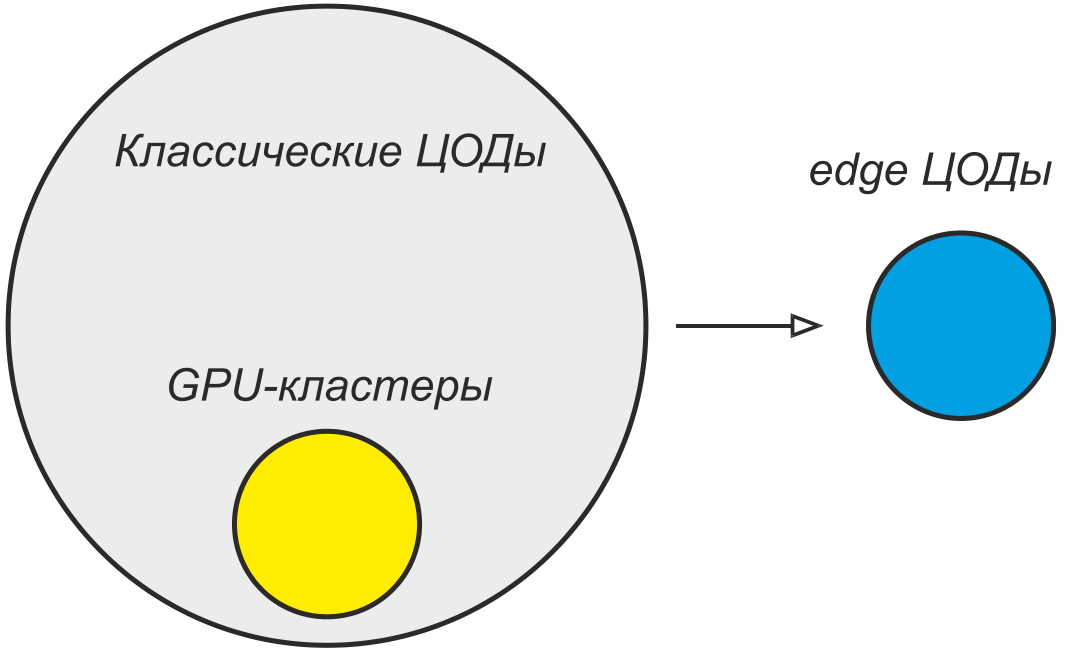
За три десятка лет своего существования как технического объекта создано великое множество таких объектов, которые классифицируются по различным критериям. Первоначально ЦОДы как развитие классических серверных реализовывали исключительно централизованную модель хранения и обработки данных, в настоящее время с учетом естественного эволюционного развития по спирали с технической точки зрения могут быть разбиты на три основные разновидности

- классические ЦОДы;
- edge (периферийные) ЦОДы;
- GPU-кластеры как структура внутри классического ЦОДа.

Появление двух последних разновидностей фактически означает, что отрасль в своем развитии начинает выходить на распределенную схему обработки данных на новом техническом уровне. В полном соответствии с воззрениями Гегеля и его принципом диалектики развитие конкретной области техники происходит не линейно, а по спирали с выходом каждый раз на новый уровень.

# Разновидности ЦОДов как технического объекта – 2(2)

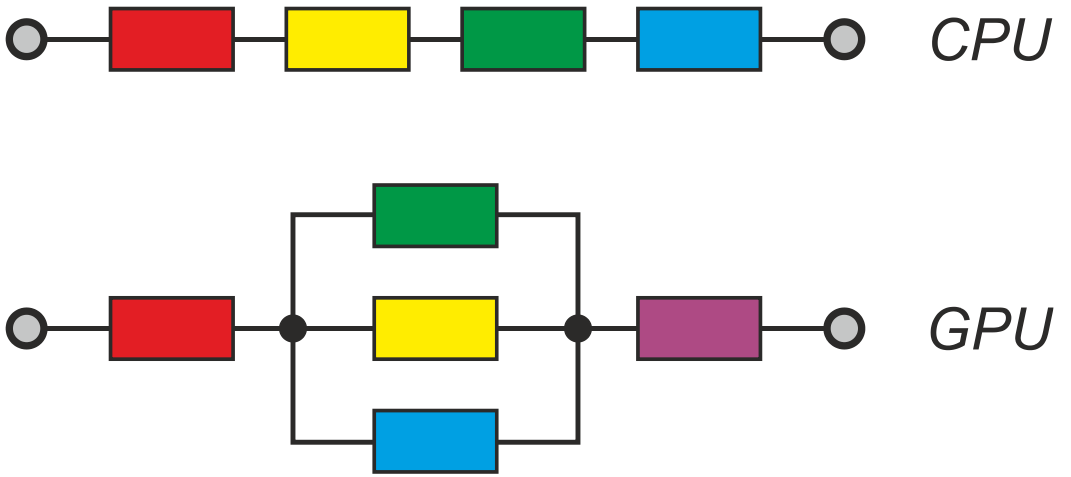



Соотношение между отдельными объектами в схематической форме можно представить себе так, как это показано на эскизе в правой верхней части слайда. При этом опять же по Гегелю каждый новый виток спирали включает в себя предыдущие достижения, т.е. при создании нового класса объектов д.б. в максимально полной степени использованы все сделанные ранее наработки.

Применительно к СКС данное положение означает, что на раннем этапе формирования технического направления кабельная система реализуется на уже имеющейся элементной базе по хорошо отработанным принципам. Может меняться только соотношение по объемам применения того или иного решения. В частности, в GPU-кластерах из-за их геометрической компактности (см. далее) заметно более широко применяются

- удлиненные коммутационные шнуры;
- активные кабельные сборки.

# Схема обработки запроса в GPU-кластере




GPU-кластер согласно одному из определений представляет собой сеть взаимосвязанных компьютеров (узлов), каждый из которых оснащён одним или несколькими графическими процессорами (GPU). Эти узлы работают совместно, распределяя вычислительные задачи и ускоряя обработку данных за счёт параллельной обработки.

Соответственно, содержит

- графические процессоры, которые, в отличие от CPU, изначально ориентированы на параллельную обработку (см. эскиз в правой части слайда);
- узлы кластера в виде компьютеров, содержащих
  - GPU;
  - центральные процессоры (CPU) для управления операциями и выполнения задач, не подходящих для GPU;
  - оперативную память (RAM).

Для обеспечения нормальной работоспособности такой структуры требуются

- высокоскоростные каналы связи (например, InfiniBand как типично кластерная технология, NVLink, Ethernet), обеспечивающие быструю передачу данных между узлами и GPU;
- минимальное время задержки передачи данных между узлами как одно из условий наращивания объема обрабатываемой информации.

# Структурные особенности GPU-кластера

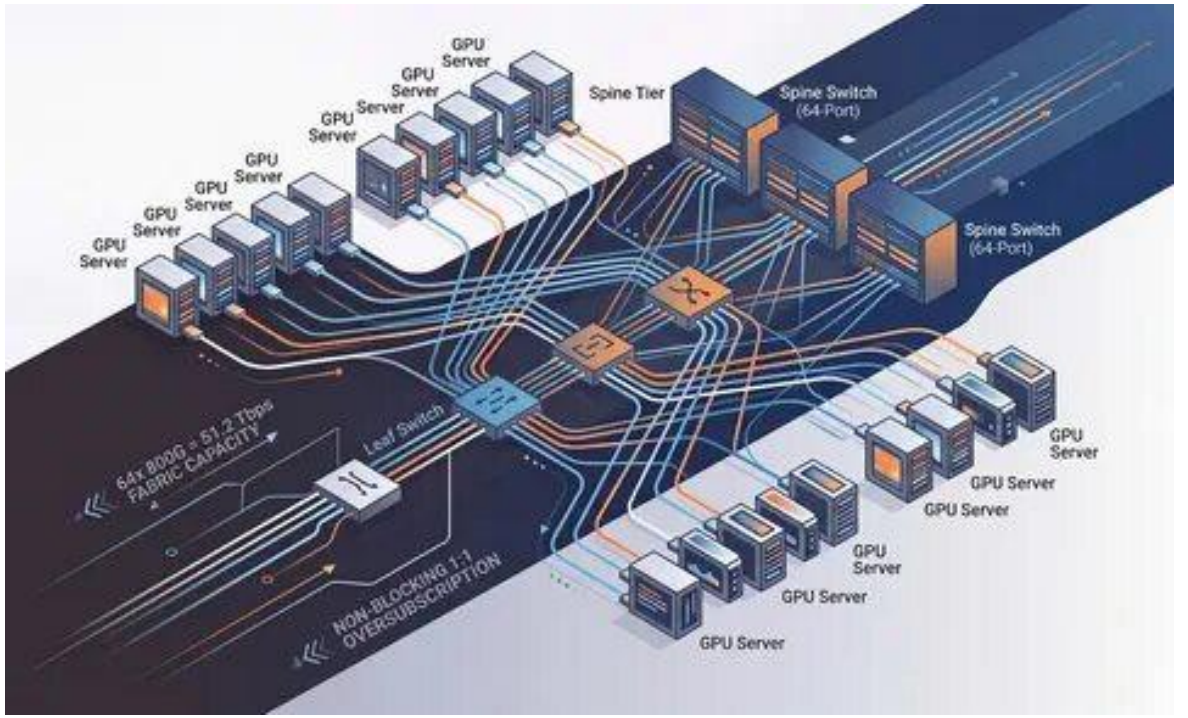
GPU-кластер характеризуется следующими отличительными особенностями:

1. геометрической компактность как прямое следствие необходимости уменьшения времени задержки передачи сигналов между узлами;
2. необходимостью применения высокоскоростного транспорта данных.

Минимизация времени задержки достигается проектными приемами и, в первую очередь,

- обращением к высокоскоростным протоколам с малой задержкой;
- ограничением максимальной протяженности линии значением примерно 50 м.

Следствием последнего становится то, что средняя длина линии GPU-кластера примерно на 20% меньше по сравнению с классическим ЦОД.



# Важность параметра delay (задержка сигнала) для GPU-кластера


Функционирование GPU-кластера предполагает активное применение параллельной обработки. При такой организации функционирования вычислительные процедуры обработки не могут начинаться до момента получения необходимых данных от других серверов.

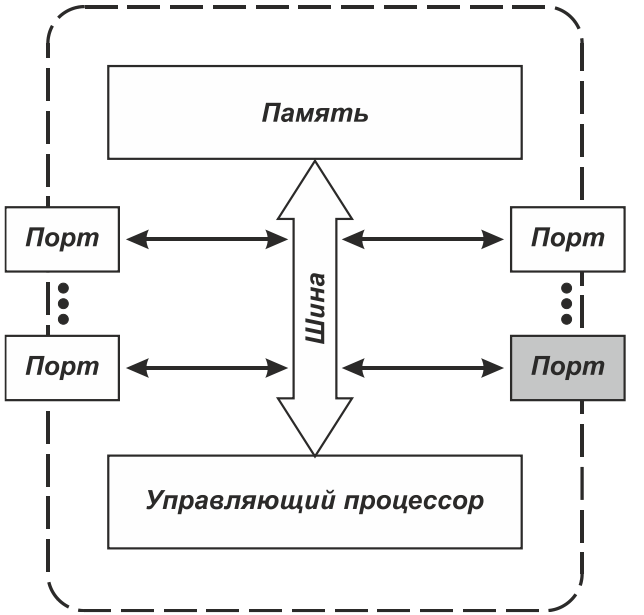
Сервера делят на виртуальные и физические. Последние требуется из-за нехватки вычислительной производительности одного компьютера для поддержки нормального функционирования даже нескольких десятков серверов.

Интересующая нас задержка передачи сигнала возникает при взаимодействии как физических серверов, так и виртуальных. Сосредоточим свое внимание на “межсерверной” задержке, которая значимо превышает “виртуальную”. “Межсерверная” задержка содержит ряд составляющих, которые представлены на круговой диаграмме.

Наращивание вычислительной производительности GPU-кластера требует уменьшения каждой из этих составляющих.

Разработка соответствующих мероприятий по минимизации задержки требует понимания природы возникновения каждой из составляющих, которые рассматриваются далее.

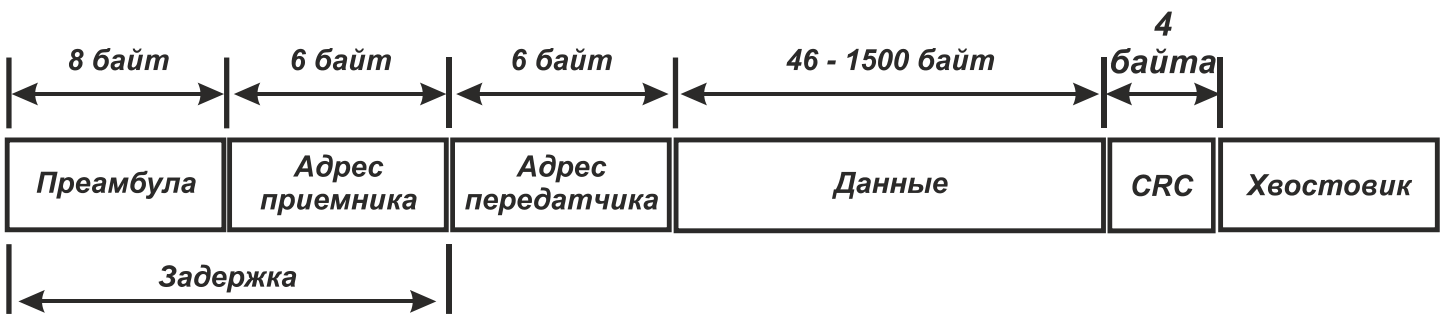
# Задержка сигнала - коммутаторы – 1 (2)

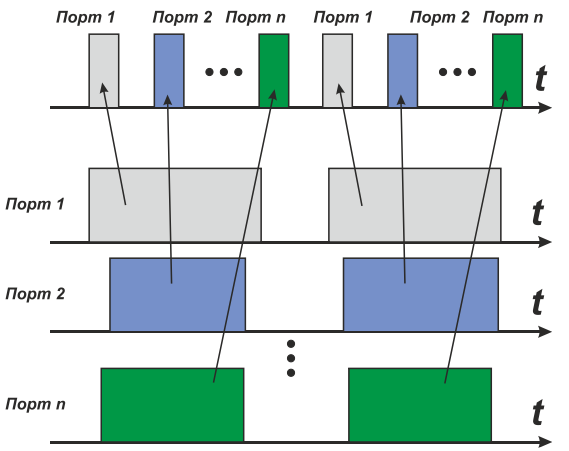



Коммутатор как сетевое устройство представляет собой многопортовый мост, а его упрощенная структурная схема изображена на эскизе в верхней правой части слайда. При работе использует цифровую схему организации информационного обмена и мультиплексирования сигналов отдельных каналов во времени, что

- требует предварительной синхронизации входных потоков;
- вызывает задержку по времени для принятия решения о передаче в данный конкретный момент “0” или “1”;
- вызывает необходимость во временном хранении информации в промежуточной памяти непосредственно перед формированием пакетов перед их передачей на соответствующий выходной порт.

Применение известных методов уменьшения задержки (например, коммутация в режиме “на лету”), а также наращивание тактовой частоты процессора и внутренней шины не решает проблему радикально.



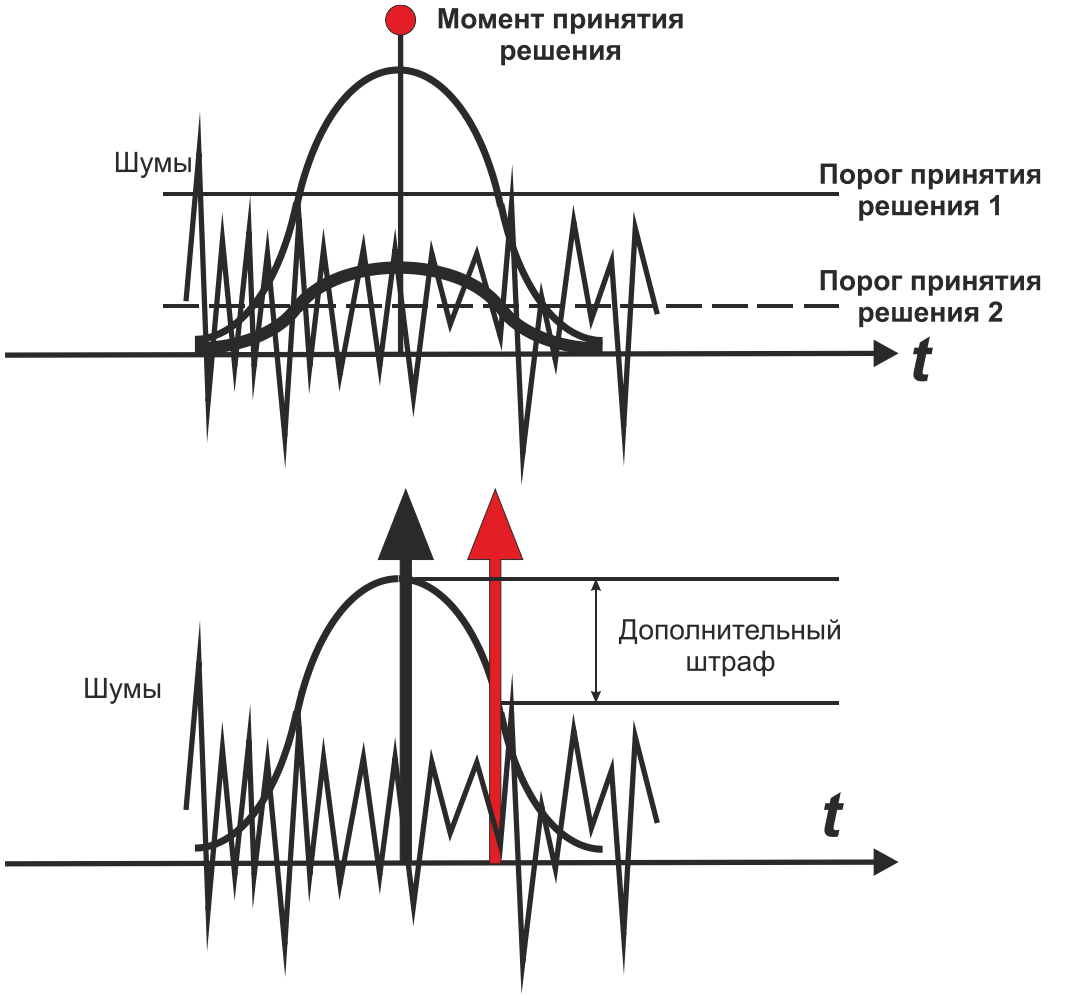



# Задержка сигнала - коммутаторы – 2 (2)

Любой коммутатор даже в синхронных сетях всегда вносит задержку как минимум на половину тактового интервала передаваемого сигнала. Причина такого положения дел заключается в том, что решение о поступлении на вход приемника сигналов логического нуля или единицы принимается в момент максимально возможного мгновенного отношения сигнала к шуму на входе решающего устройства, которое в момент поступления строба просто сравнивает мгновенное значение сигнала к некоторым пороговым значением.

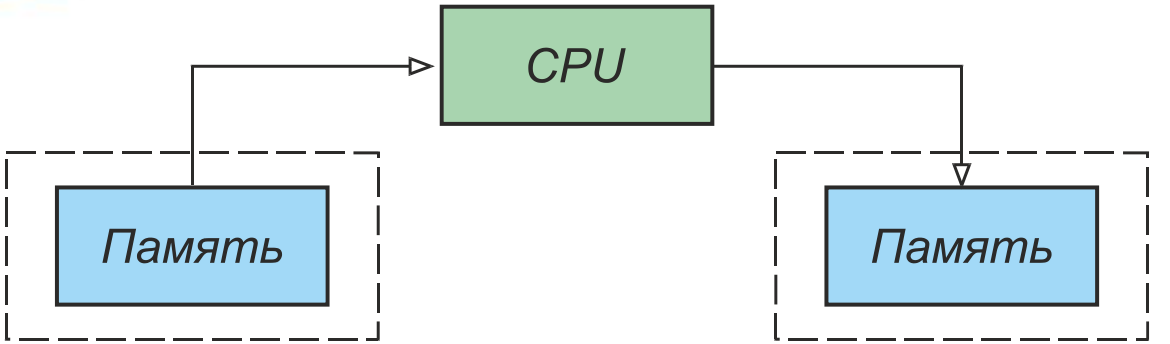
Эскиз в правой части слайда в качественной форме поясняет, что наилучшие условия для приема возникают в центре тактового интервала, где сигнал обычно достигает своего максимума, тогда как шум в среднем сохраняет неизменное значение на всем тактовом интервале.

Соответственно, любое отклонение от центра может трактоваться как появление в тракте передачи дополнительных потерь, а в процесс передачи сигнала неизбежно вносится задержка на половину тактового интервала.



# Прямой доступ к памяти как средство минимизации задержки

Обычно обмен данными происходит под управлением и с участием центрального процессора CPU. В режиме прямого доступа к памяти (Direct Memory Access, DMA) процессор CPU просто исключается из выполнения этой процедуры, а необходимые функции передаются контроллеру прямого доступа к памяти.




*Традиционная схема передачи*

Передача данных происходит с меньшей задержкой за счет того, что

- данные не проходят CPU, который не отличается высоким быстродействием;
- передача данных передается узкоспециализированному устройству (контроллеру DMA).


*Передача методом прямого доступа к памяти*

Иначе говоря, процедура, выполняемая в процессе передачи данных от одного устройства к другому, реализуется не на программном уровне, а заметно более быстром аппаратным.

Платой за быстродействие становится резкое сокращение функциональных возможностей схемы.

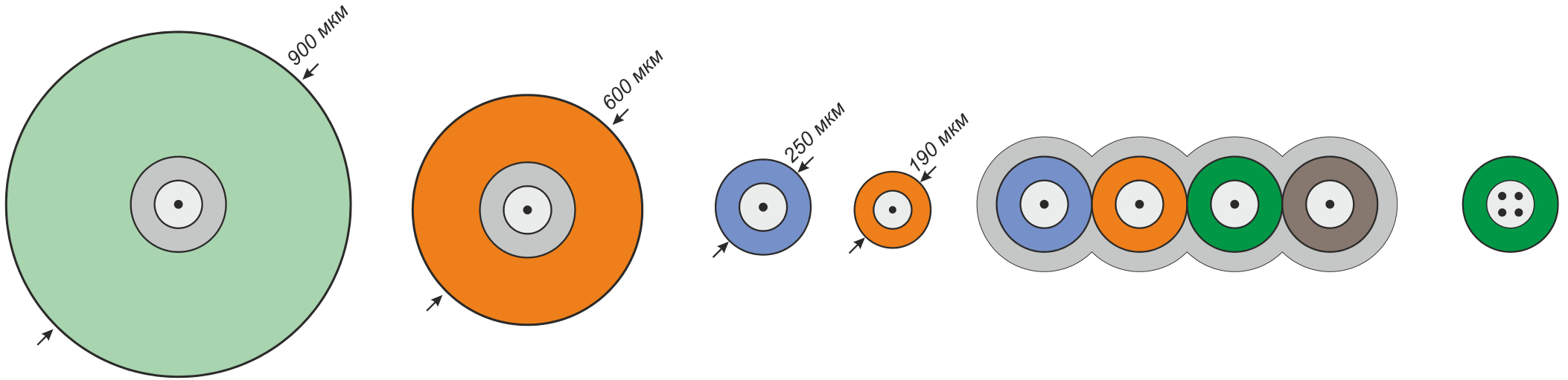
Полноценно реализовать преимущества DMA можно только при наличии высокоскоростных каналов связи.

# Уменьшение диаметра  волокон кабелей для GPU-кластера машинного зала ЦОДа

Из-за большого количества волоконно-оптических кабелей, применяемых при построении GPU-кластеров в машинном зале ЦОД, актуальна задача уменьшение их внешнего диаметра. Для ее решения привлекаются следующие приемы

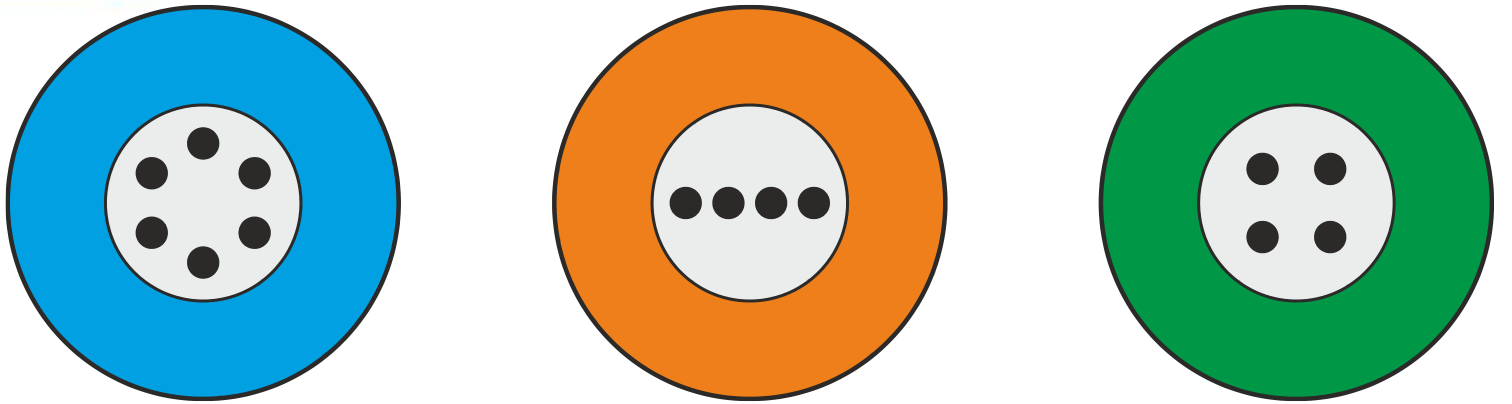
- применение в традиционных конструкциях кабелей волокон преимущественно в первичном в буферном покрытии;
- уменьшение диаметра оболочки и первичного буферного покрытия;
- использование ленточной конструкции;
- переход на многосердцевинное волокно.

Выигрыш от перехода на новые разновидности волокон демонстрирует эскиз в нижней части слайда (при изображении поперечного сечения волокон пропорции соблюдены).



# Особенности конструкции многосердцевинных волокон – 1(2)

Идея применения многосердцевинных волокон MCF (multi core fiber) была высказана еще в 1979 году японской компанией Sumitomo. Первоначально такие волокна предназначались  для доставки излучения накачки к промежуточным оптическим усилителям классических волоконно-оптических сетей.



Техническая возможность создания многосердцевинных волокон (при наличии соответствующей технологии) определяется тем, что диаметр модового поля типового 125-микронного одномодового световода находится в диапазоне 9 – 12 мкм. Это позволяет разместить в предела одной сердцевины несколько световедущих сердцевин. Примеры таких волокон показаны на эскизе.

Известны волокна даже с 19 сердцевинами, хотя с учетом тяготения оптических интерфейсов для эксплуатации в составе аппаратуры машинного зала ЦОДа к применению схем передачи на основе стандарта CWDM на современном этапе достаточно всего четырех сердцевин.

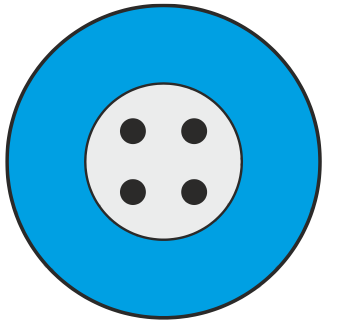
Сердцевины располагаются в структуре оболочки с использованием различных схем, некоторые из которых показаны в нижней части слайда. Волокна для ЦОД оптимизируются для работы в спектральном диапазоне 1310 нм.

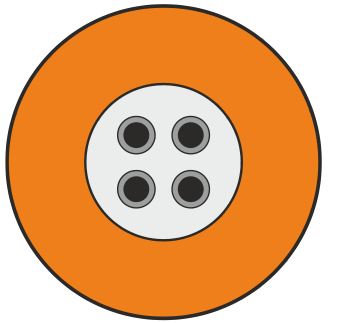
Выполнение отдельных сердцевин по единой технологии и их нахождение в пределах одной оболочки в сочетании с небольшими дальностями связи в машинном зале ЦОДа эффективно решает проблему оптического skew.

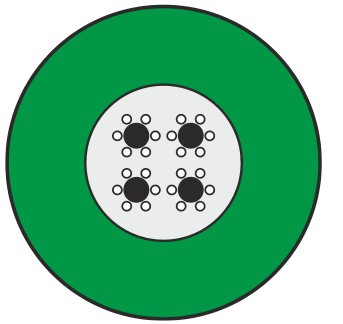
# Особенности конструкции многосердцевинных волокон – 2(2)

Одной из проблем использования многосердцевиных волокон для передачи высокоскоростных информационных потоков, которая усугубляется небольшим энергетическим потенциалом сетевых интерфейсов для машинного зала ЦОДа, является переходная помеха между отдельными сердцевинами. Известно три направления минимизации этого нежелательного влияния

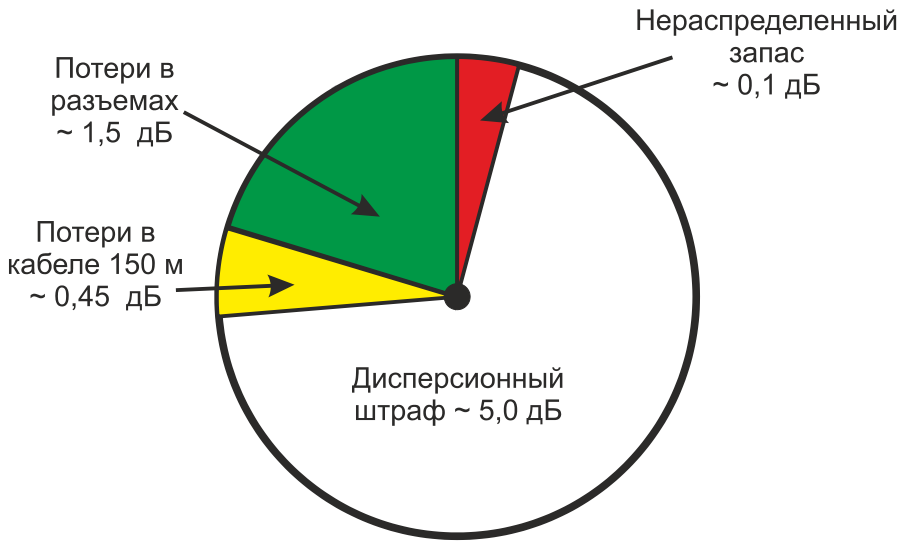
- увеличение расстояния между сердцевинами;
- применение кольцевых запирающих канавок (т.н. TA-MCF);
- формирование круглых воздушных каналов (т.н. HA-MCF);.



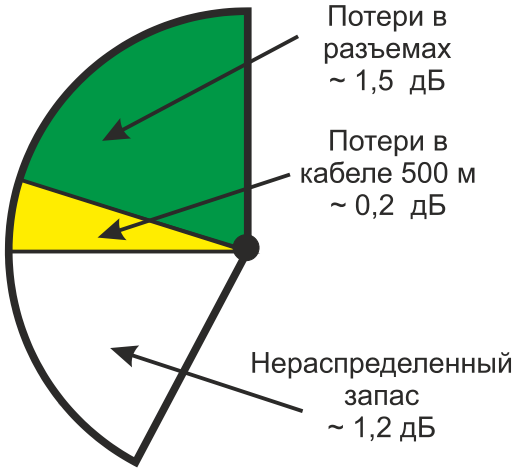




*Многомодовые линии*




*Одномодовые линии*



# Подключение к многосердцевинным оптическим трактам

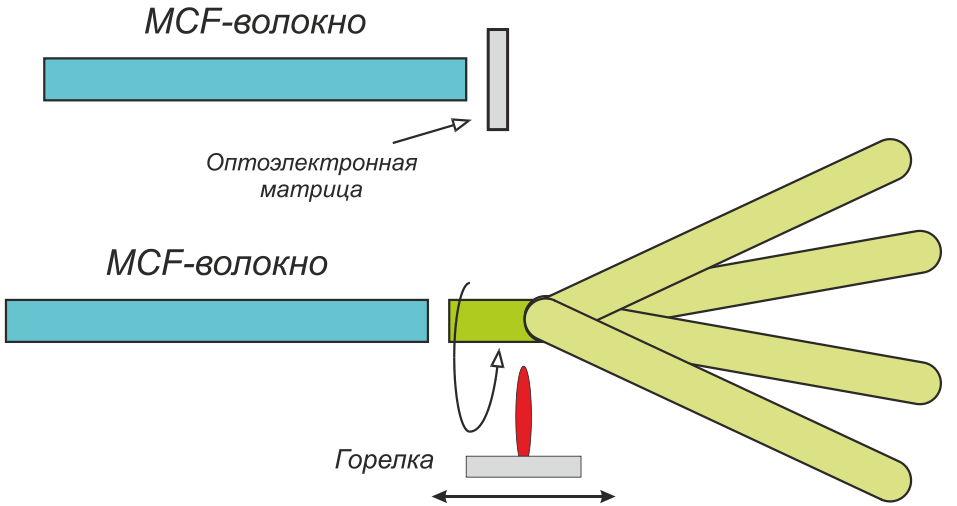
При подключении оптических приемопередатчиков к многосердцевинным волокнам используются два основных решения

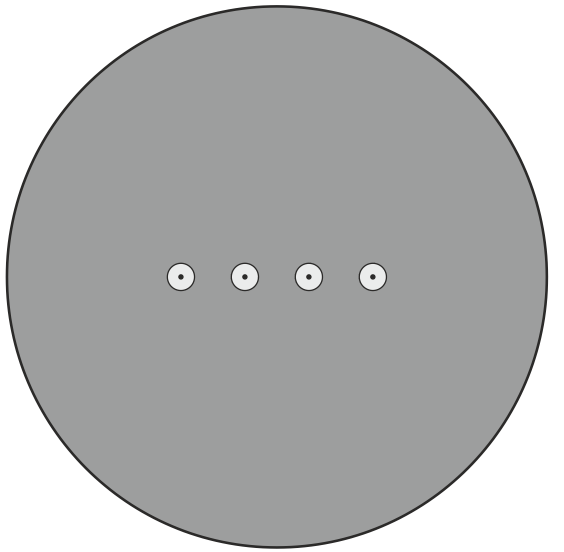
- прямая стыковка волокна с оптоэлектронными компонентами;
- применение разветвительных шнуров.

В основу первого решения положены матричные микроэлектронные сборки излучателей и фотодиодов. По своим геометрическим параметрам они полностью соответствуют MCF-волокнам, что позволяет стыковать их напрямую без использования дополнительных оптических систем. Возможность их формирования практически продемонстрирована еще как минимум в 2011 году.

Второй способ предполагает формирование волоконной “косички”, для чего используется технология, близкая к применяемой при изготовлении сплавных сплиттеров для пассивных оптических сетей.

В качестве разъемов для кабелей с многосердцевинными волокнами могут быть использованы стандартные конструкции с дополнительным контролем углового положения центрирующего группового наконечника со стандартным диаметром 2,5 мм. Известно, например, предложение разъема SC-DC с рядной схемой укладки волокон в структуре такого наконечника.



# Начало работы ассоциации SDM4 MCF MS

О перспективности реализации транковых кабелей на базе многосердцевинных волокон свидетельствует создание ассоциации SDM4 MCF MSA, в состав которой вошли такие авторитетные рыночные игроки как компании Fujikura, Corning, Sumitomo и TeraHop PTE. Официально о создании ассоциации объявлено 11 марта 2026 года. Главной целью этого инициативы SDM4 MCF MSA объявлена разработка спецификации 4-сердцевинного одномодового волокна, которое рассчитано на работу в O-диапазоне (1260–1360 нм).

Выбор данного диапазона определяется в первую очередь его заметной экономичностью по сравнению с диапазонами третьего окна прозрачности (длины волн 1550 нм). Определенное значение имеет также минимальная дисперсия, что может иметь определенное значение при объединении в общую структуру нескольких ЦОДов (интерконнект ЦОДов).


В процессе выполнения работы должно быть

- разработано детальное описание конструкции волокна;
- определены основные геометрические парраметры;
- сформирован перечень и заданы предельные значения параметров передачи.

Выход первого релиза разрабатываемого стандарта намечен на конец года.

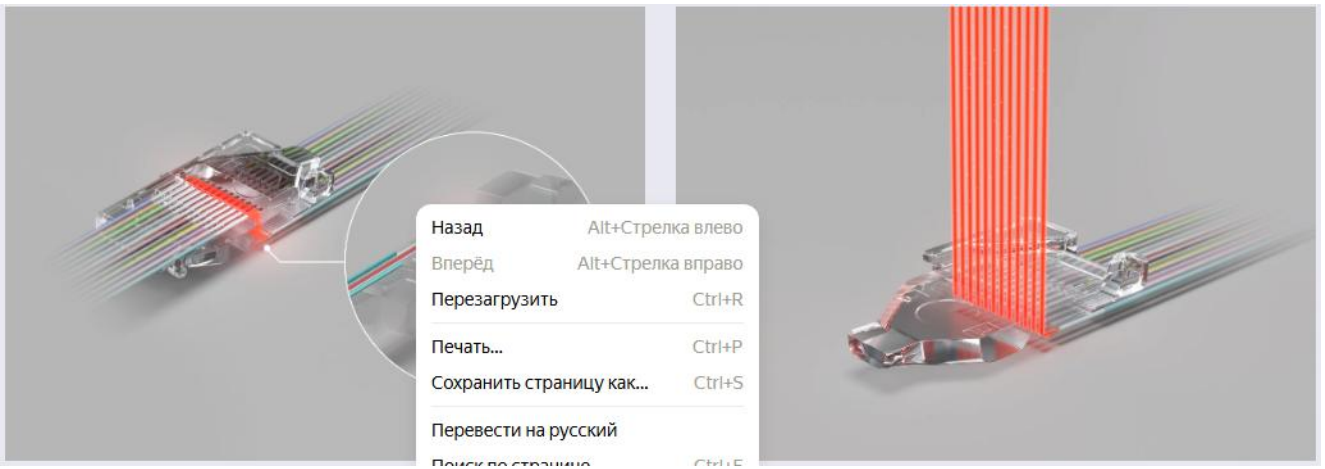
# Совершенствование групповых оптических соединителей

Компанией 3M предложен новый тип оптического наконечника EBO, который автоматически реализует схему снижения влияния загрязнений на оптические потери в разъеме через расширения луча.

Наконечник монтируется в корпусе MPO, в оригинальном корпусе LC-типа с защелкой или в групповом прямоугольном корпусе.

Пока доступны наконечники на 8 и 12 волокон, в перспективе предполагается дополнить линейку 16-волоконным наконечником.

Разъемы на основе подобных наконечников позволяют коммутировать одновременно до 144 волокон. Практические внедрены в кабельную систему такими авторитетными игроками мирового уровня рынка решений физического внутриобъектовых информационных как немецкая компания OSI Rosenberger и американская корпорация Molex (система VersaBeam – анонсирована в марте текущего года). В перечень партнеров 3M входят также  Sumitomo и US Conec.

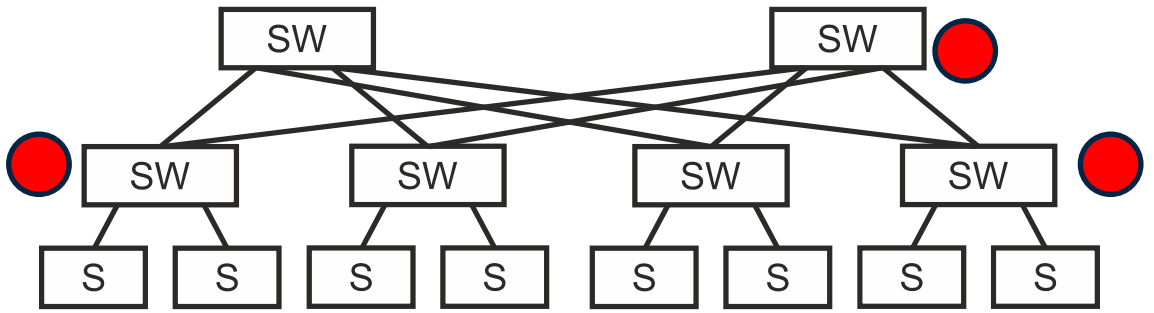



# Переход на полносвязанную топологию как средство уменьшения задержки в линии

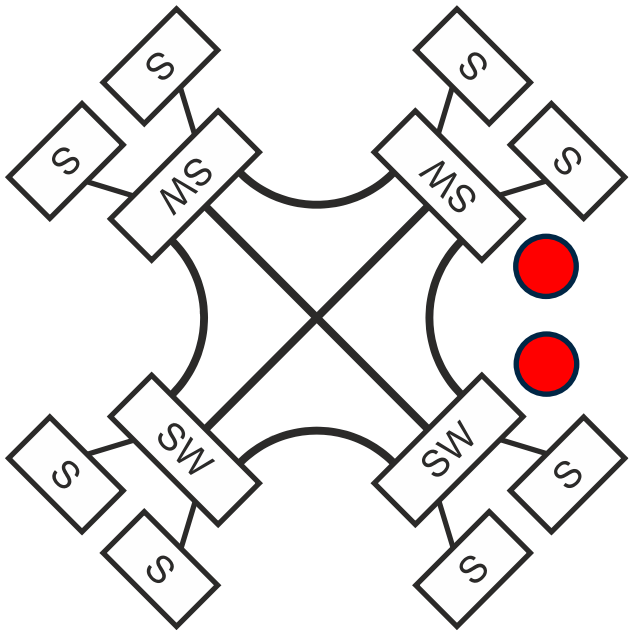
Применение полносвязанной топологии является естественной реакцией отрасли на проблему наращивания быстродействия GPU-структур. Переход на такую схему реализации физического уровня машинного зала ЦОДа (точнее его части) означает устранение еще одного коммутатора из цепи передачи сигналов между двумя серверами  и их сокращение до двух (вместо трех в структурах spine-leaf) в процессе обработки поступающего запроса.

Достигаемый выигрыш по количеству коммутаторов в полном тракте передачи сигнала демонстрирует схема в нижней части слайда. Находящиеся в тракте коммутаторы выделены кружками с красной заливкой.

*Spine-leaf - структура*




*Mesh - структура*




S - *сервер,  SW -  коммутатор*

# Заключение

1. Формирование физического уровня информационной инфраструктуры GPU-кластера может быть при необходимости выполнено на стандартной элементной базе и не требует безусловного внедрения дополнительных вариантов кабельных трактов сверх тех, которые задаются действующими стандартами.
2. Характеристики оптической СКС для GPU-кластеров могут быть улучшены за счет разработки специализированной элементной базы.
3. Переход на кластерную структуру организации ЦОД стимулирует рост объемов применения коммутационного оборудования на базе VSFF-разъемов следующего поколения, а также кабелей прямого соединения.
4. Характеристики информационной кабельной системы, создаваемой в интересах поддержки функционирования GPU-кластера, в первую очередь в части ее быстродействия и вносимой задержки могут быть улучшены в случае применения в оптических кабелях многосердцевинных и микроструктурированных волоконных световодов, которые пока не нормируются действующими стандартами СКС.
5. Применение GPU-кластеров увеличивает объемы применения одномодовой техники.

Спасибо за внимание!