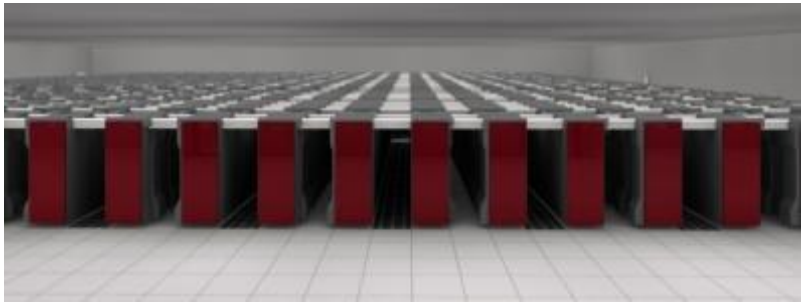
FUJITSU

формируем завтра вместе с вами

FEFS: Масштабируемая кластерная файловая система



K Computer (RIKEN AICS)

”К компьютер” является названием, которое RIKEN использовал для своего суперкомпьютера

PRIMEHPC FX10

PRIMERGY

Схема

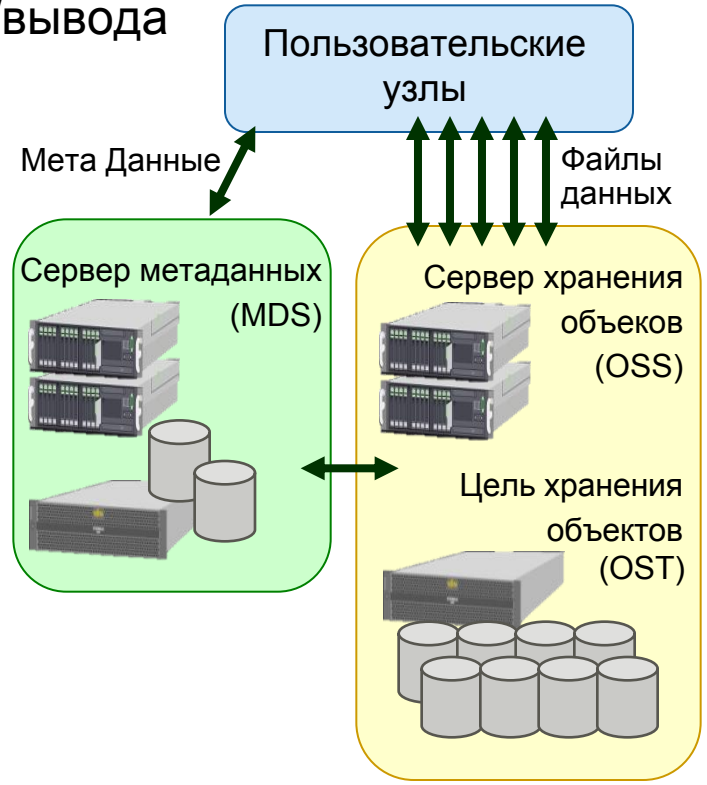
- Обзор FEFS
  - Основные характеристики
  - Целевая система
- Архитектура ввода/вывода
  - Концепция и проект системы
  - Высокая надежность
- Технические вопросы
  - Расширеня Lustre
  - Зонирование ввода/вывода
  - QoS взноса честности, QoS лучшего использования
- Оценка производительности
  - Пропускная способность (IOR)
  - Отклик (mdtest)
- Аннотация
  - Вклад в сообщество Lustre

Характеристики FEFS

FUJITSU

FEFS является масштабируемой кластерной файловой системой на основе Lustre

- Высокая производительность & Высокая масштабируемость
  - Масштабируемая производительность ввода/вывода (~1TB/s) & емкость (~8EB)
- Управление использованием ввода/вывода
  - QoS взноса честности
  - QoS лучшего использования
- Высокая надежность & Высокая доступность
  - Устойчивость к отказам с резервными аппаратными средствами и continuing file system service



Целевая система

- K компьютер
  - RIKEN и Fujitsu совместно сотрудничали в разработке K компьютера
    - Должен быть установлен в RIKEN AICS, Кобе в 2012
- PRIMEHPC FX10
  - Только что опубликованный новый бренд суперкомпьютеров Fujitsu
- PC кластер
  - PRIMERGY и сервера на основе IA/Linux третьих фирм

Super Computer

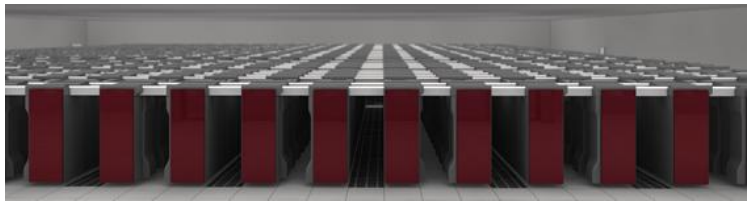
K computer (RIKEN AICS)

PRIMEHPC FX10

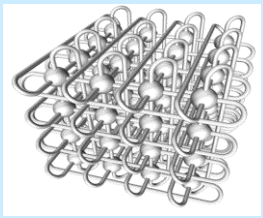
PC Cluster

PRIMERGY

Конфигурация системы



Суперкомпьютер



Tofu: 6D тороидальный интерконнект

PC Cluster

- Передача данных в/от глобальной файловой системы
- Коммуникация данных для управления работой системных заданий

FEFS: локальная файловая система
(Область временного хранения данных заданиями)

I/O network (QDR InfiniBand), management network (GbE)

Пользователь

Узлы авторизации

FEFS: глобальная файловая система
(Область хранения данных)

Узлы управления заданиями | Узлы управления файлами | Управляю-щие узлы | Узел интеграции системы

Управляющие узлы

Администратор

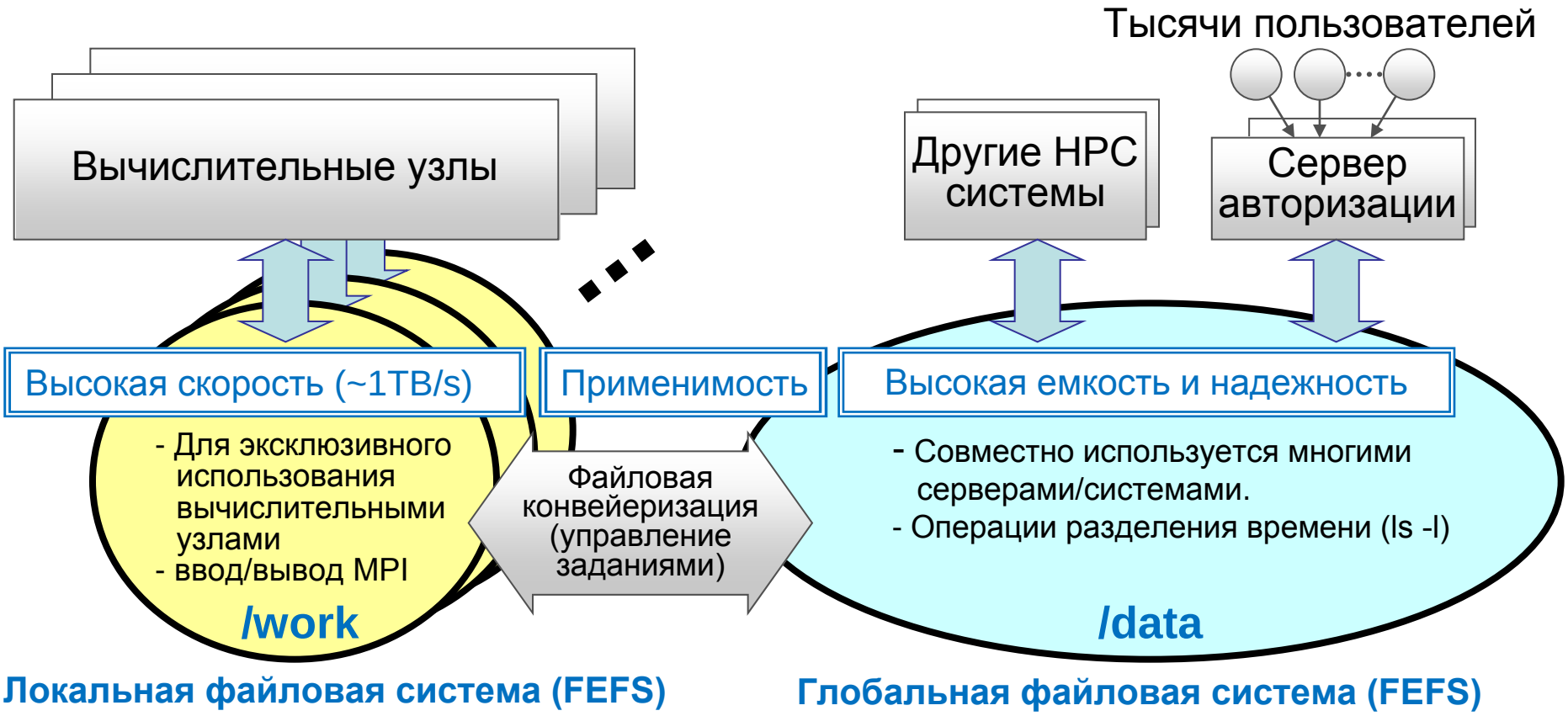
- Авторизация
- Компиляция
- Запуск задания

- Управление работой системы
- Управление работой заданий

Архитектура ввода/вывода: основные концепции

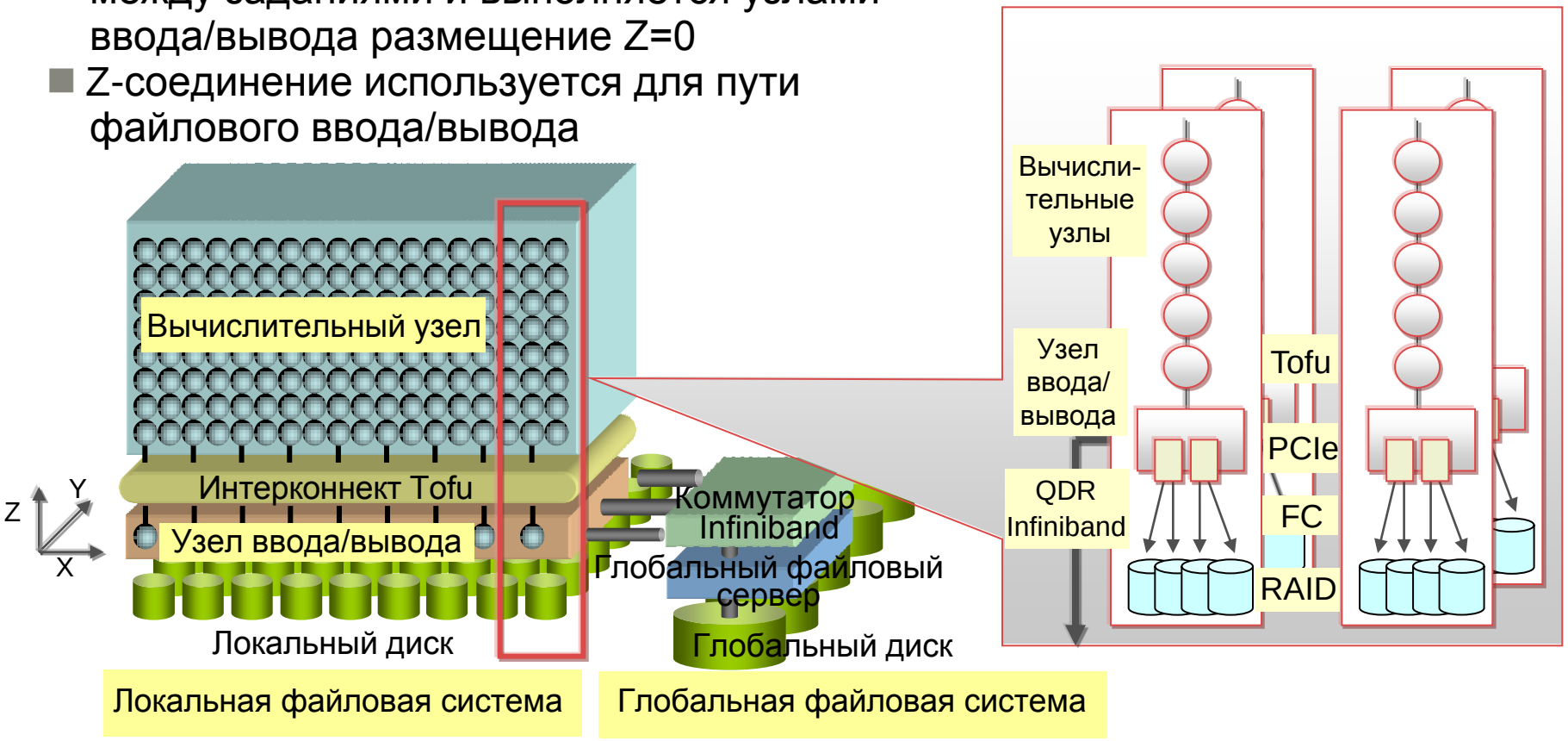
- Несовместимые свойства реализуются путем создания разных уровней

  Файловая система
  - Локальная файловая система (/work): Высокоскоростная ФС для выделенного использования заданиями
  - Глобальная файловая система (/data): ФС большой емкости и надежности для совместного использования



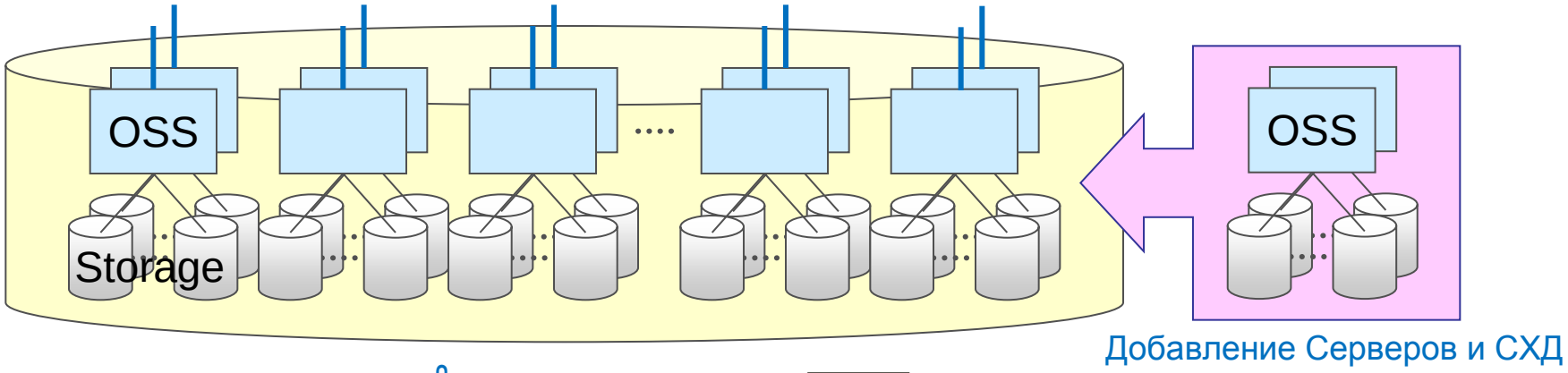
Архитектура ввода/вывода: проект системы

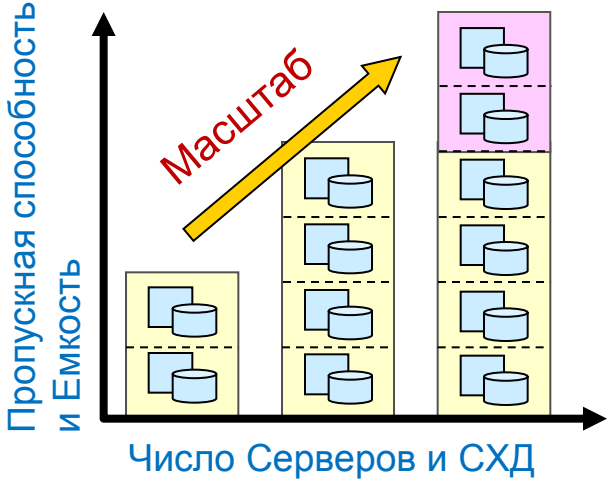
- Оптимизирована для работы масштабируемого файлового ввода/вывода
  - Достижение масштабируемого объемахранения и производительности
  - Исключение конфликтов ввода/вывода для каждой компоненты
- Технология зонирования для локальной файловой системы
  - Файловый ввод/вывод разделяется между заданиями и выполняется узлами ввода/вывода размещение Z=0
  - Z-соединение используется для пути файлового ввода/вывода



Масштабируемые производительность и емкость

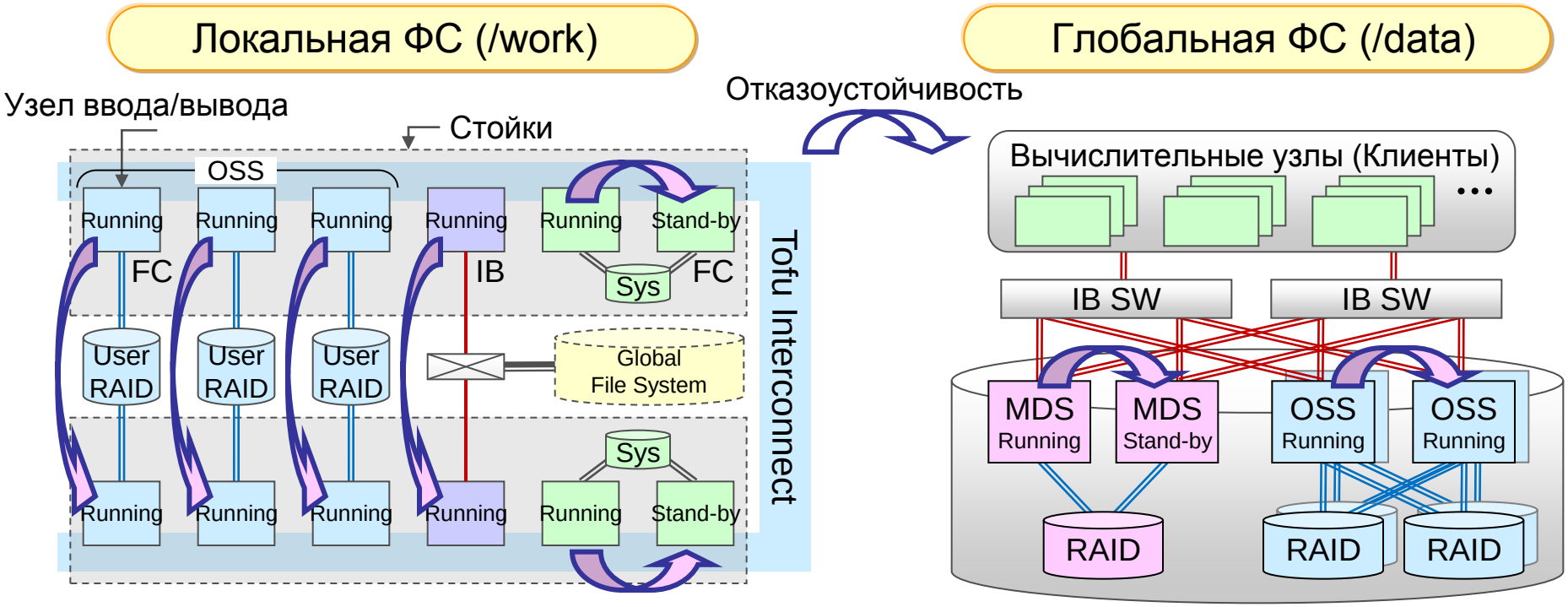
- Высокая пропускная способность и большая емкость достигаются множественностью OSS
  - Увеличение масштаба пропускной способности и емкости добавлением серверов и СХД





Высокая надежность и высокая доступность

- Сохранение устойчивости службы файловой системы к отказам
  - Резервирование аппаратных средств
    - Дублирование путей от InfiniBand, Fibre Channel, серверов ввода/вывода
    - RAID диски (MDS:RAID10, OSS: RAID5/6)
  - Программное обеспечение управления системой
    - Определение отказов и автоматическое переключение на альтернативные пути или сервера



Расширения FEFS для Lustre: свойства

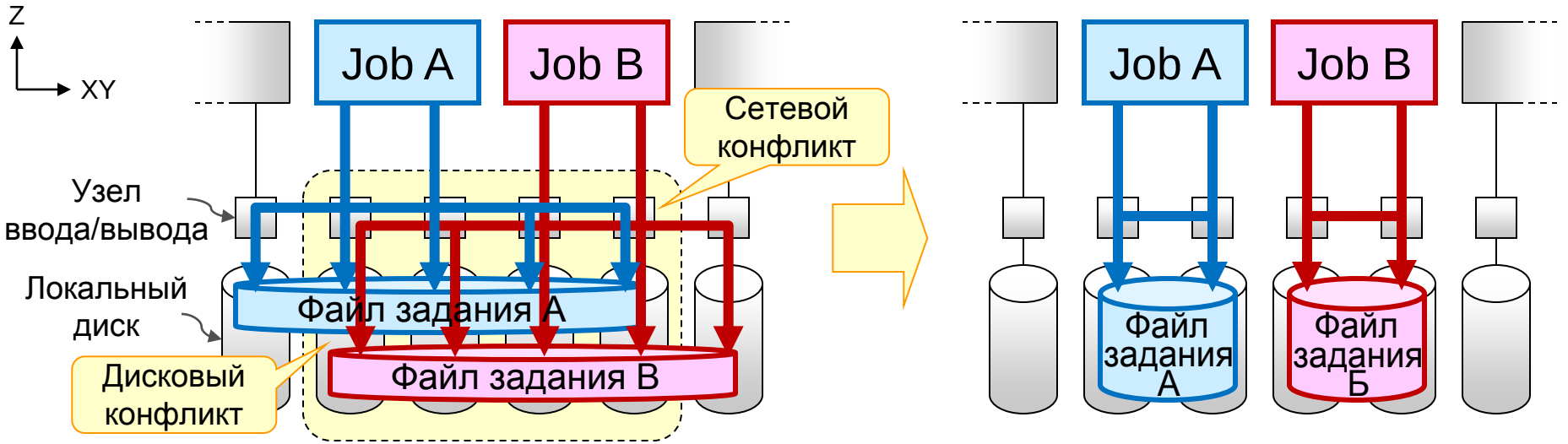
Расширения FEFS для Lustre: Спецификация

FUJITSU

Характеристика		FEFS	Текущая Lustre
Ограничения системы	Макс.размер файловой системы	100PB (8EB)	64PB
	Макс.размер файла	1PB (8EB)	320TB
	Макс. число файлов	32G	4G
	Макс. размер OST	(8E) 100TB	16TB
	Макс.число полос	(1PB) 20k	160
	Макс.число записей ACL	191	32
Масштаби-руемость узлов	Макс.число OST	20k	8150
	Макс.число клиентов	1M	128K
Практичность	QoS	Да	Нет
	Квоты директорий	Да	Нет
Многошинность IB		Да	Нет
Размер блока (внутренняя ФС)		~512KB	4KB

Зонирование ввода/вывода: Разделение ввода/ выода между заданиями

- Вопрос: Аппаратные конфликты ввода/вывода заданий
  - Совместное использование дисковых томов, сетевые связи между заданиями вызывают снижение производительности ввода/вывода из-за их конфликтов
- Наш подход: Разделение аппаратных средств между заданиями
  - Разделение дисковых томов, сетевых связей между заданиями, насколько это возможно



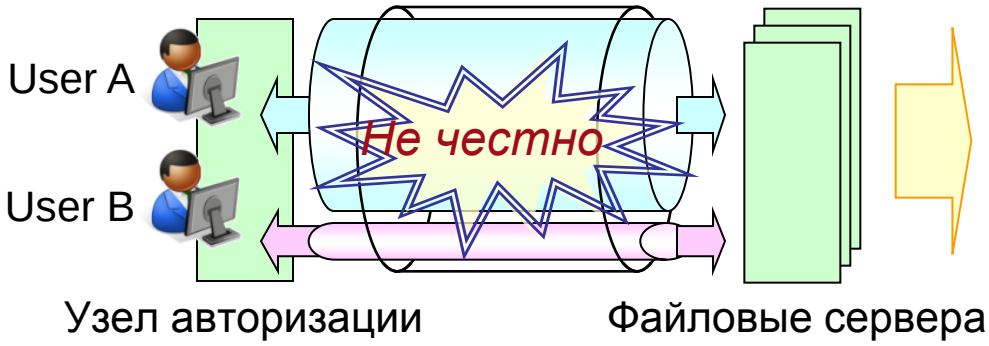
Плохо: с конфликтами ввода/вывода

Хорошо: без конфликта ввода/вывода

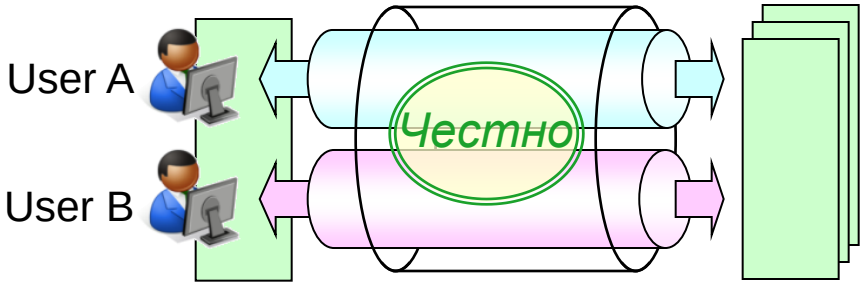
QoS взноса честности

- Вопрос
  - Исключение ситуации, когда кто-то в одиночку занимает ресурсы файлового ввода/вывода
- Наш подход
  - Ограничить число запросов на ввод/вывод, которое каждый пользователь можетвыполнить одновременно на клиентском узле

Без QoS взноса честности



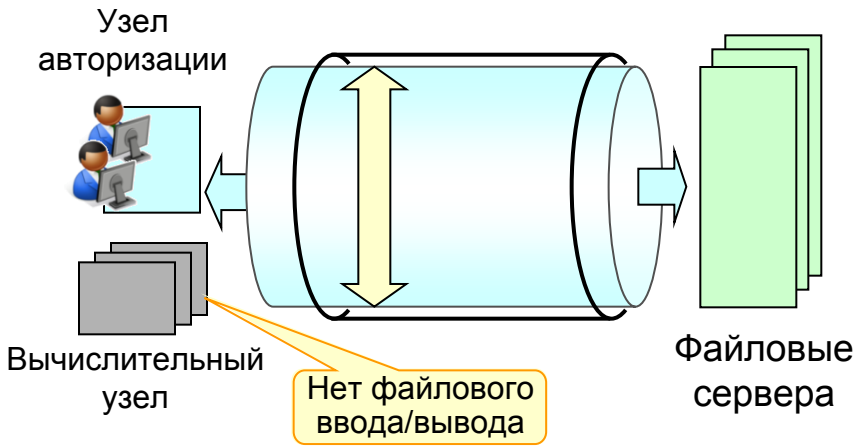
С QoS взноса честности



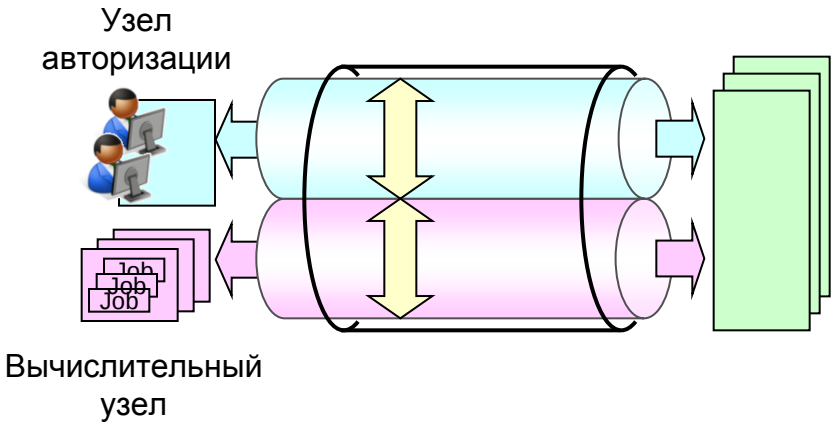
QoS лучшего использования

- Вопрос
  - Эффективное использование всех ресурсов ввода/вывода
- Наш подход
  - Выделение всех серверных ресурсов клиентам, которые выполняют файловый ввод/вывод

Занято одним узлом



Разделяется множеством узлов



Оценка FEFS: QoS

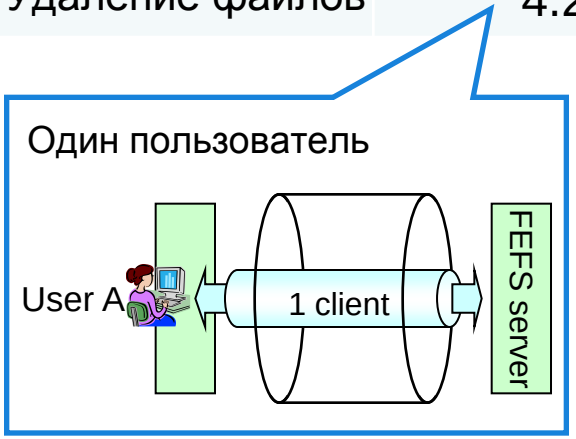
FUJITSU

- QoS эффективность на PC кластере
  - Пользователь A: задание с 1 узлом -> Измерение времени создания/ удаления 10,000 файлов
  - Пользователь B: задание с 19 узлами

Подавляется влияние файловых операций пользователя B

Время работы пользователя A

Пользователь A 10,000 файлов	Без QoS один пользователь	Без QoS много пользователей	С QoS много пользователей
Создание файлов	4.1 sec	10.1 sec	3.9 sec
Удаление файлов	4.2 sec	14.0 sec	5.5 sec



Аннотация и дальнейшая работы

FUJITSU

- Fujitsu разработал кластерную файловую систему FEFS на основе Lustre
  - Высокоскоростной файловый ввод/вывод (~1TB/s), Огромная емкость (~8EB)
  - Высокая надежность и высокая доступность
  - Расширения Lustre: QoS, многошинность IB, квотирование директорий
- Дальнейшие работы
  - Сделать наш вклад в сообщество Lustre
  - Объединить наши расширения в дальнейшие выпуски Lustre

Пресс- релиз

FUJITSU

Whamcloud and Fujitsu to Collaborate on Lustre Development

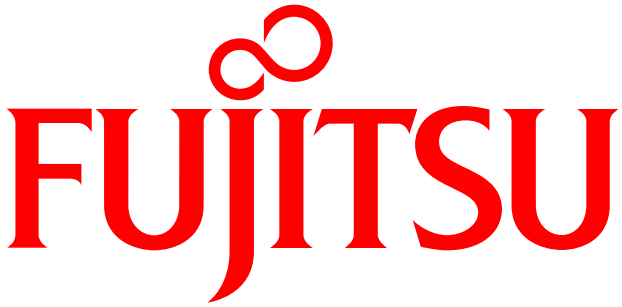
Fujitsu to advance Lustre development for HPC

Danville, CA – November 15, 2011 – Whamcloud, a venture-backed company formed from a worldwide network of high-performance computing (HPC) storage industry veterans, and Fujitsu, the global IT products and services company, and together with RIKEN, the joint developer of the world's fastest supercomputer, the K computer[(1)], announced today that both parties agreed to the principal terms of joint Lustre development. This collaboration will include scalability and file system work for Lustre, and merging Fujitsu's Lustre enhancements into the Lustre 2.x community release.

"Lustre is a central technology in our supercomputing products, and we look forward to working closely with Whamcloud, the leader in file system software technologies, to advance performance, add features and push supercomputing capabilities to new levels," said Yuji Oinaga, Head of Next Generation Technical Computing Unit at Fujitsu. "Fujitsu is committed to being at the forefront of supercomputing technologies."

"Working with Fujitsu is an extreme honor, and we look forward to their Lustre enhancements benefiting the entire community," said Brent Gorda, CEO of Whamcloud. "Lustre is the most widely used file system in HPC and is deployed in the most extreme computing environments. Fujitsu's rigorous quality standards are well-known and this agreement is a great vote of confidence for the future of Lustre.

For more details on Whamcloud and its Lustre support and development services, please see: http://www.whamcloud.com.



формируем завтра вместе с вами