



Доступность и масштабируемость кластерных систем

Безруков Валерий
Технический консультант
Sun Microsystems

bv@sun.com



Доступность

Отношение времени, в течении которого система (сервис) доступен ко времени работы системы

MTBF – время работы системы между сбоями

MTTR – время на ремонт

$$\text{Availability \%} = 100 * \text{MTBF} / (\text{MTBF} + \text{MTTR})$$

В случае кластера ~99.995% (25 минут простоя в год)

Отдельно стоящая система ~99.7 (26 часов простоя в год)

Доступность	Время простоя (в год)
99%	87.7 часов
99.9%	8.8 часов
99.99%	52.6 минут
99.999%	5.25 минут

Масштабируемость

Показатель, отображающий каким образом изменяется производительность системы с добавлением вычислительных ресурсов, отношение совокупной вычислительной мощности к мощности одного процессора/системы

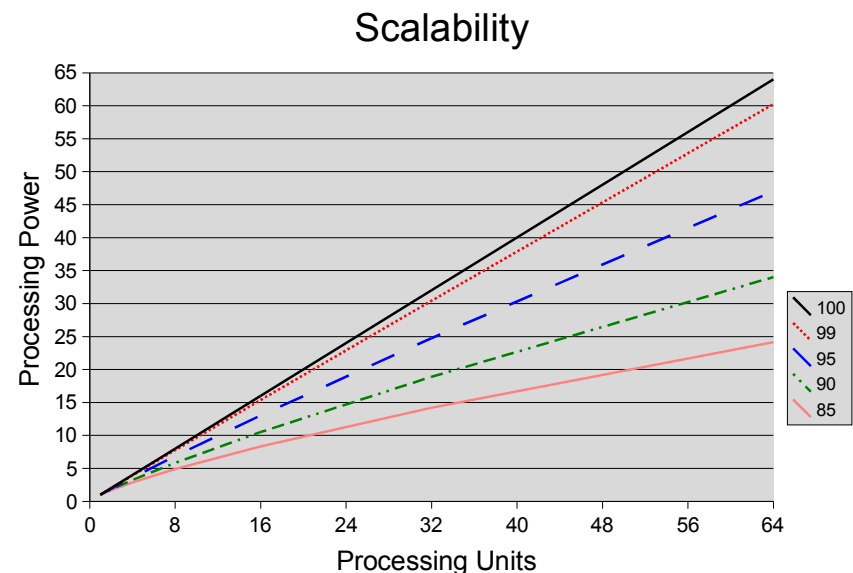
Масштабируемость всегда зависит от набора факторов, например:

- Задачи, которые не подлежат распараллеливанию и должны выполняться на каждой системе/процессоре последовательно
- Координация и синхронизация между процессами и процессорами/системами

Пример: допустим, что масштабируемость (scalability) 90%:

- 1 CPU – 1 CPU
- 2 CPUs – 1.8 CPUs
- 4 CPUs – 3.24 CPUs (1.8*1.8)
- 8 CPUs – 5.83 CPUs (1.8*3.24)

Для систем SMP масштабируемость (scalability) 99% (64 CPUs – 60.25 CPUs)



Виды многопроцессорных систем

- Мультипроцессоры
 - Системы с “сильными” связями
 - Процессоры объединены с памятью быстрой шиной
 - Количество процессоров ограничено (у Sun максимум – Sun Fire 25K: 72 USIV* процессоров)
 - Масштабируемость ~99%
- Мультикомпьютеры
 - Системы со “слабыми” связями
 - Компьютеры объединены через сеть пакетной коммутации (Fast/Gigabit Ethernet+TCP/IP, Myrinet, SCI, ...)
 - Количество компьютеров в вычислительном кластере теоретически неограничено (до 1000...)
 - Масштабируемость... разная

* USIV – CMT процессор (двухядерный)

Мультикомпьютеры

- GRID
- Вычислительные кластеры (HPC)
- Кластеры высокой доступности (HA)
- Кластеры БД (PDB)
- Распределенные кластеры (Scalability)

Мультипроцессоры

- UMA SMP
 - Системы на основе системной шины
 - Sun Enterprise 3x00, 4x00, 6x00, E10K
- NUMA SMP
 - Системы на основе коммутатора
 - Sun Fire, Sun Fire Enterprise

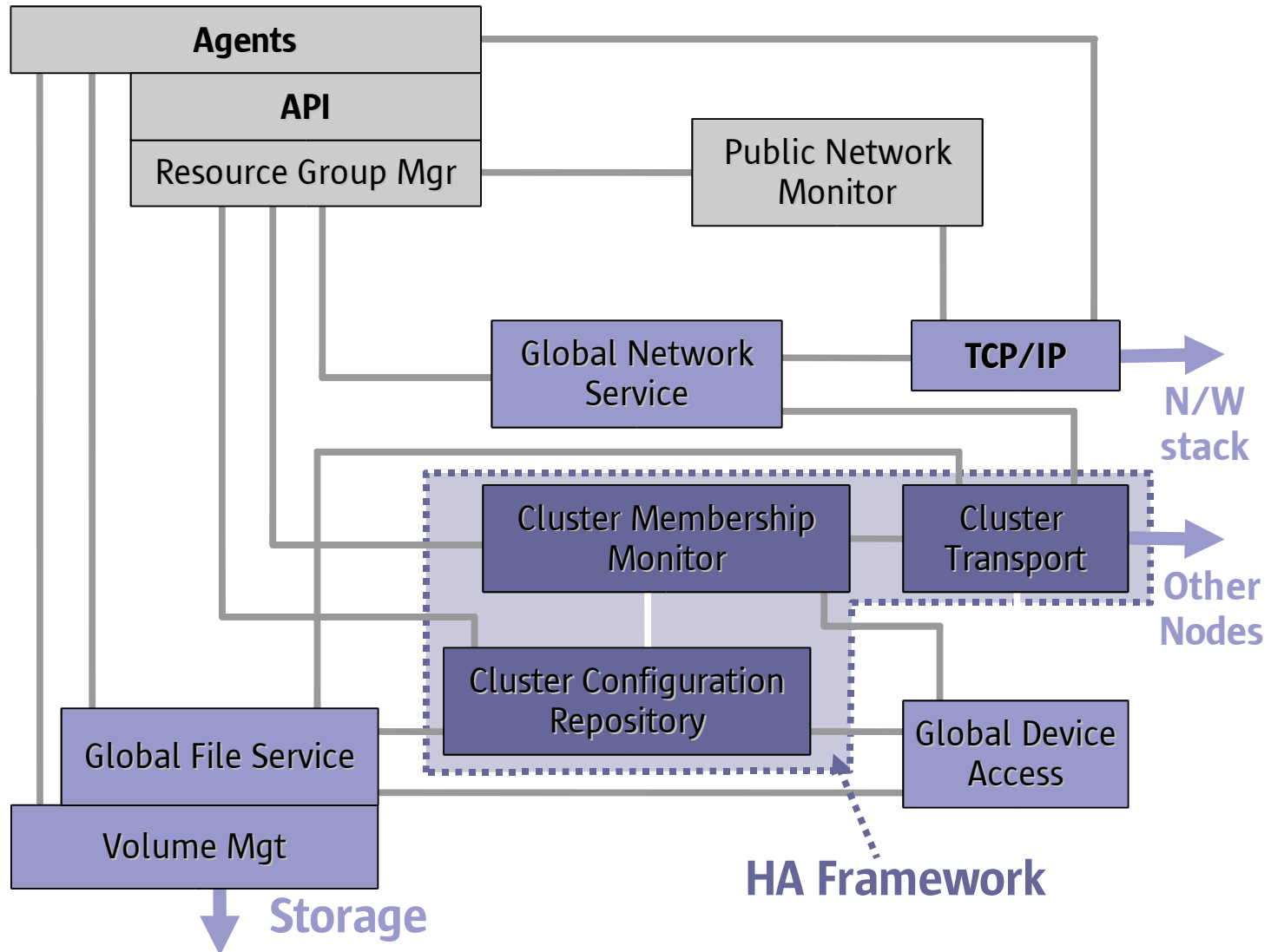
Производительность системной шины серверов Sun и межзловых соединений

	Sunfire 3800	Sunfire 4800/4810	Sunfire 6800	Ethernet DLPI	PCI-SCI RSM	SF Link RSM
Memory Bandwidth	19.2 GB/s	28.8 GB/s	57.6 GB/s			
CPU Bandwidth	19.2 GB/s	28.8 GB/s	57.6 GB/s			
I/O Bandwidth	4.8 GB/s	4.8 GB/s	9.6 GB/s			
Aggregated Bandwidth	24.0 GB/s	33.6 GB/s	67.2 GB/s			
Sustained Bandwidth	9.6 GB/s	9.6 GB/s	9.6 GB/s	60 MB/s 110 MB/s	120 MB/s 200 MB/s	1 GB/s
Latency	180ns-240ns	180ns-240ns	180ns-240ns	60usec	3,6usec	1,7usec

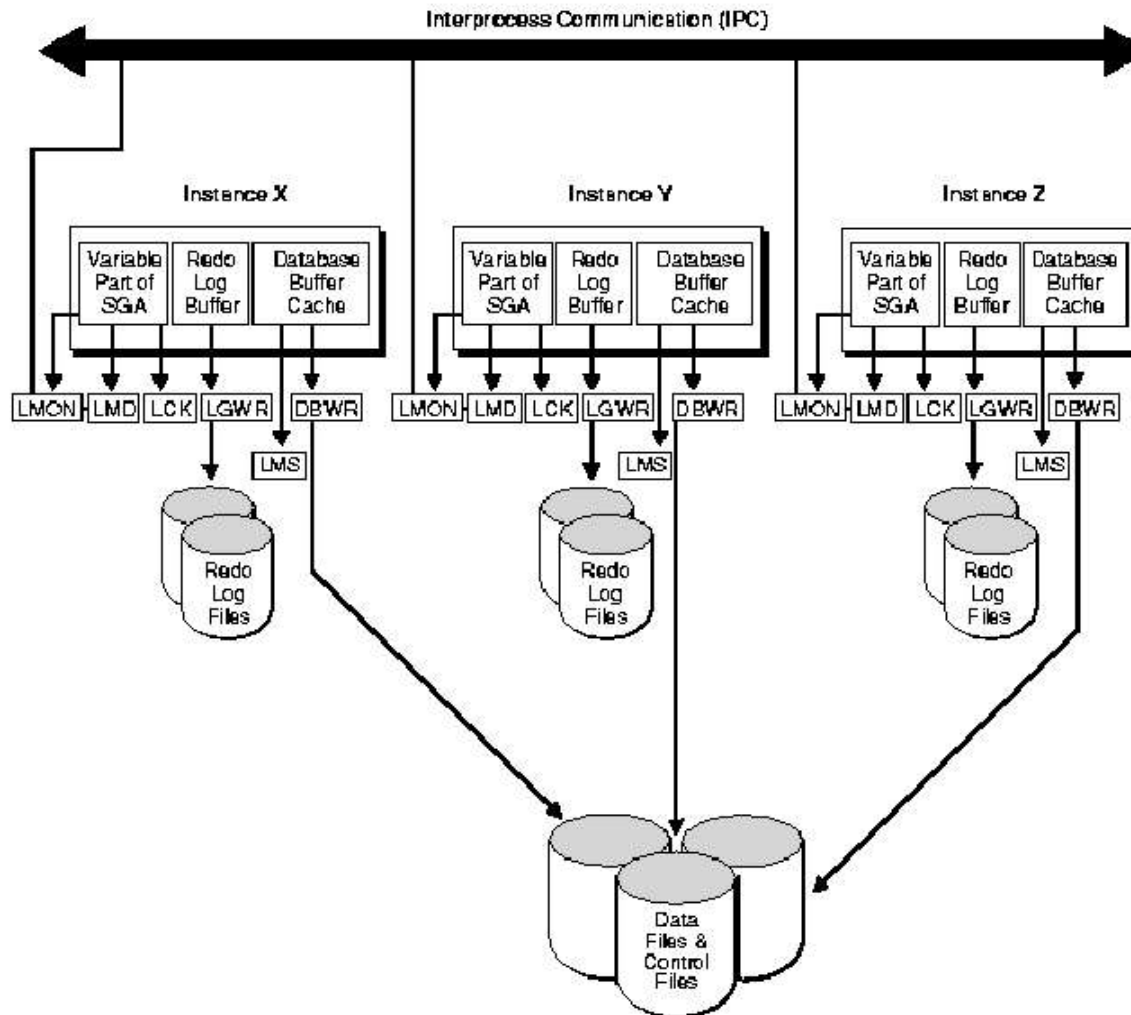
Мультикомпьютеры Sun, или кластерные решения Sun Microsystems

- SunPlex: Sun Cluster 3.1 (Solaris SPARC, Solaris x86 в SC 3.1 u2)
 - 3 режима работы
 - Кластер высокой доступности (HA)
 - Параллельный кластер (Oracle 8i Parallel Server, Oracle 9i RAC, Oracle 10g)
 - Масштабируемый кластер (Scalability)
 - Global Devices
 - Global Files Service (GFS)
 - Global Network Service (IPMP)
 - Возможность пространственного разнесения (до 200 км с применением DWDM)
- Sun HPC ClusterTools (Solaris)
- Sun Grid Engine (Solaris, Linux)

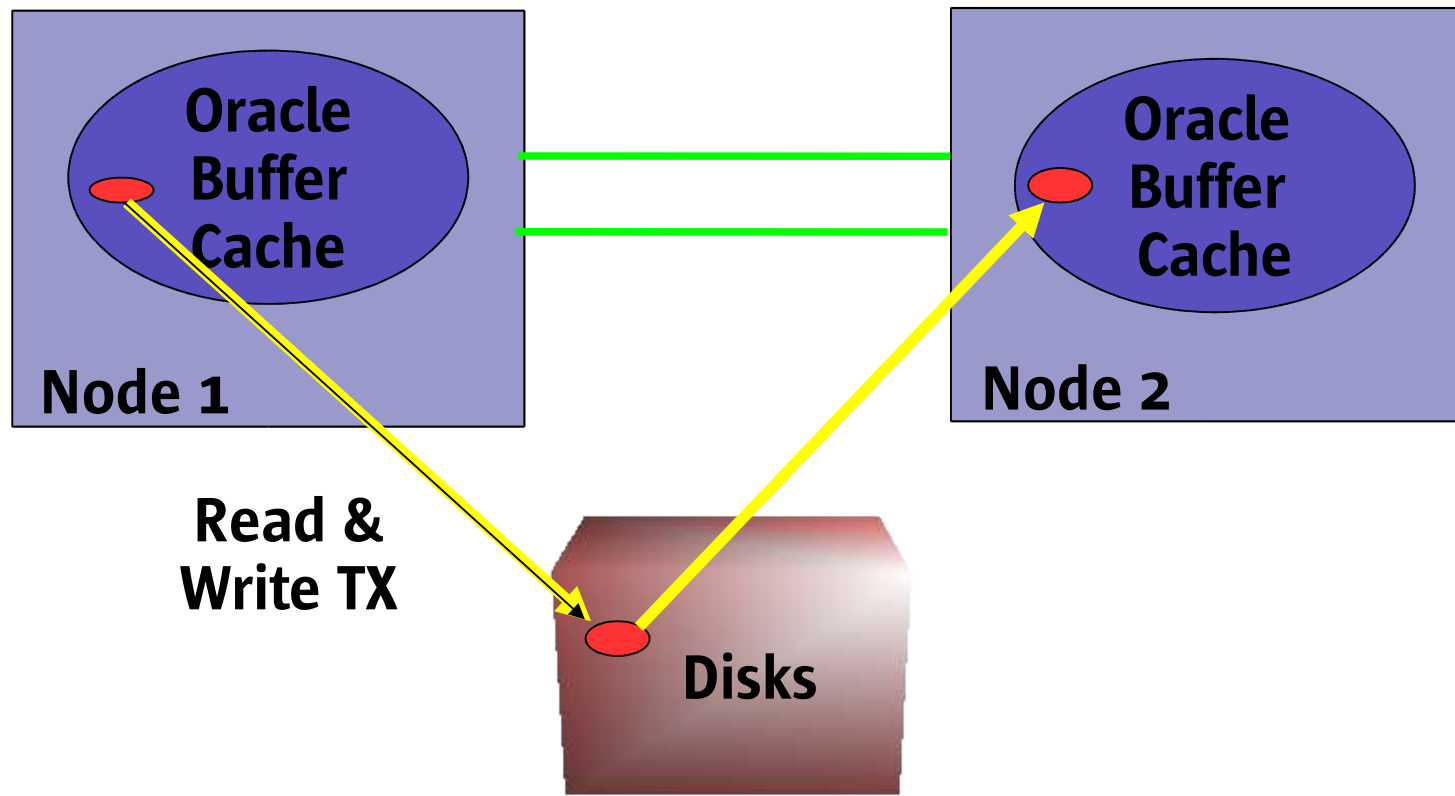
Архитектура SC 3.1



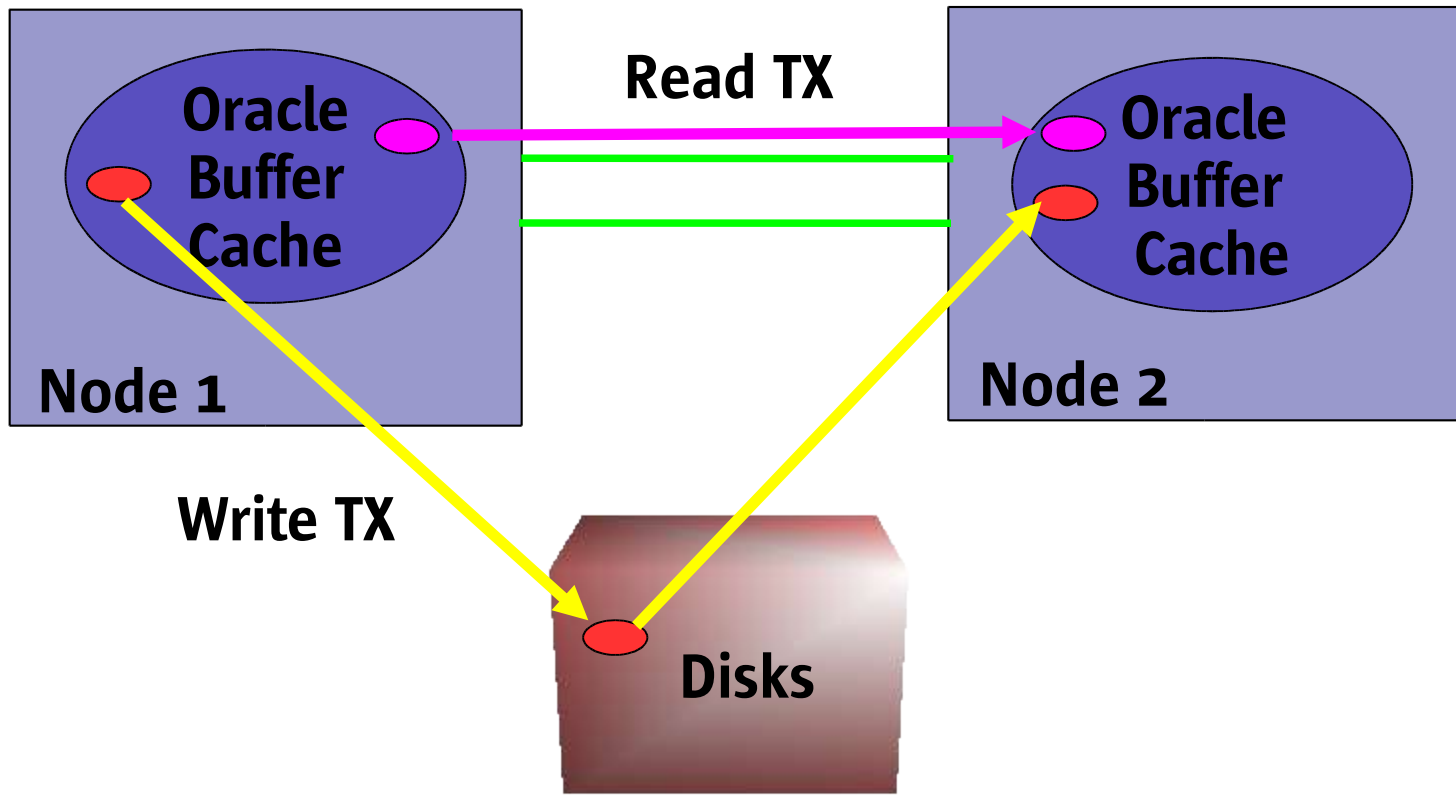
RAC/SC 3.1. Взгляд изнутри.



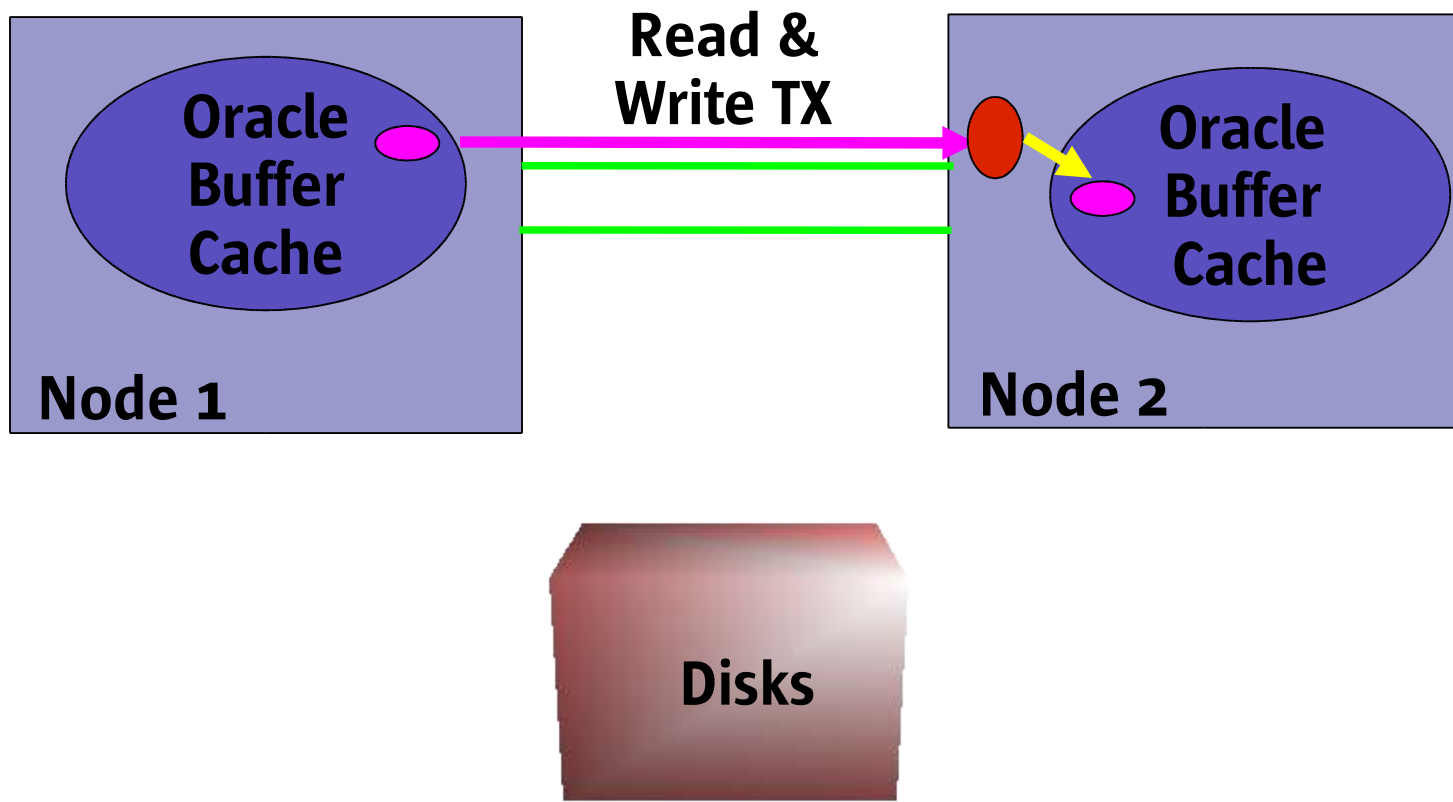
RAC/SC 3.1. Взгляд изнутри.



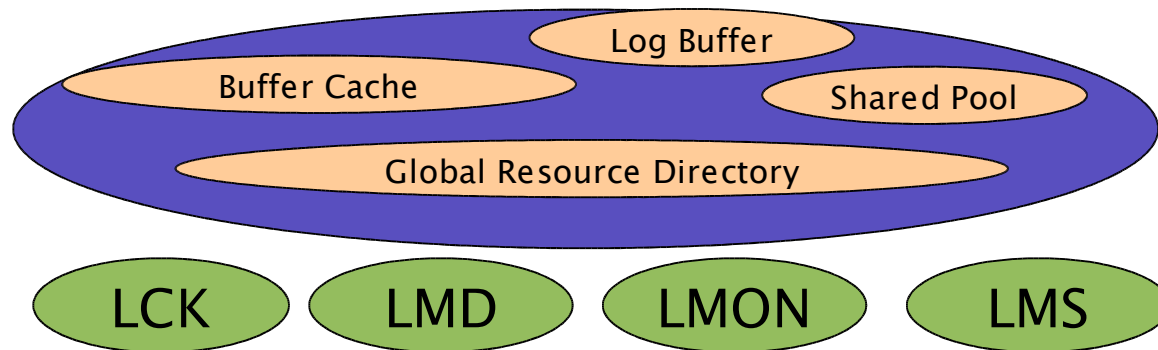
Cache Fusion фаза 1, Oracle 8i OPS



Cache Fusion фаза 2, Oracle 9i RAC



Архитектура Oracle 9i RAC



Oracle Global Cache Services (GCS) and
Oracle Global Enqueue Services (GES)

Oracle – Integrated Unix Distributed Lock Manager

Sun Cluster 3 – Distributed Lock Manager

Sun Cluster 3

Solaris Operating System / Veritas Volume Manager

Oracle 9i RAC: GRD

– Global Resource Directory (GRD)

На каждом инстансе Oracle RAC

Принадлежность ресурса определяется по таблице с использованием алгоритма хэширования (file# и block#)

Синхронизации подлежат данные о блоках, находящихся в режиме “current mode” (update, delete, ...)

Состояние блока (Exclusive / Shared / Null)

Тип блока (Global / Local)

Для каждого блока, считанного в Buffer pool любого из инстансов Oracle 9i RAC, существует запись в GRD

Борьба за ресурсы: Single Oracle Instance

- Блокировки (**enqueues**)
- **в зависимости от состояния блока**
 - ★ Consistent read (R/W conflict)
 - ★ Current block (W/W conflict)
- в случае записи – копирование блока в rollback segment
 - ★ Consistent read
 - ★ Undo
- **блокировки** на уровне транзакций

Борьба за ресурсы: RAC

- Блокировки: Global enqueue service (GES)
 - ★ LMD
- Кэш: Global cache service (GCS)
 - ★ LMS
- Cache Fusion
 - ★ Read / Write Conflict
 - ★ Write / Write Conflict
- Атомарная единица = блок БД

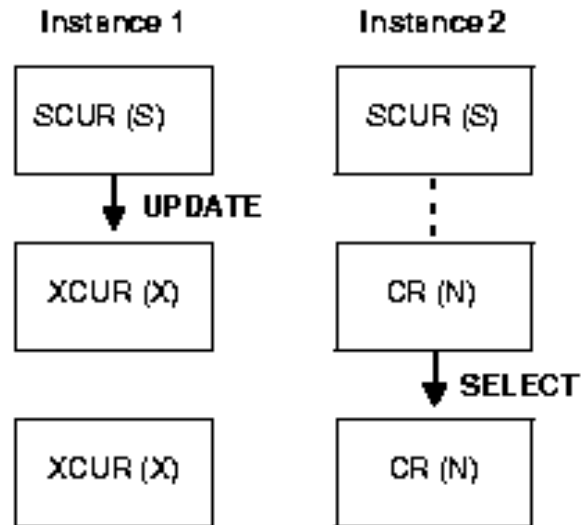
Global Cache Services (LMSn)

Демоны службы Global Cache Services

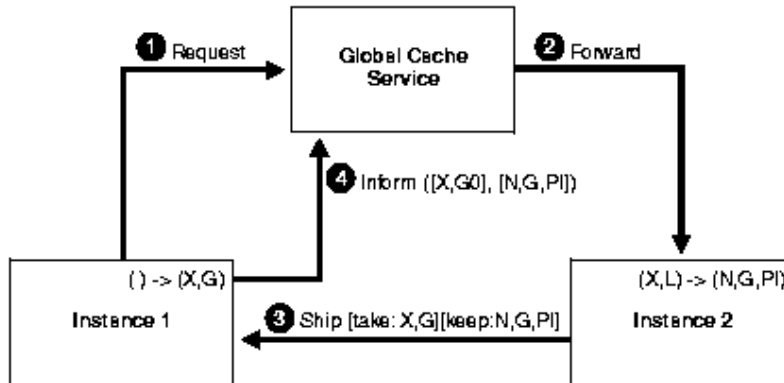
- Управление потоком сообщений между инстансами, управление доступом к блоку и перемещение блока между кэшами инстансов
- Имя процесса LMSn, где n=0..9
- Может быть понадобится увеличить количество процессов с ростом количества процессоров, например
 `_lm_lms = 2` для 4-х процессорной системы
- ... и увеличить приоритет процессов LMS (`priocntl`)

Состояние “current mode” блока

Режим	Статус	Описание	Количество
Null (N)		Информация отсутствует (как правило держателем является другой инстанс)	Много
Shared (S)	Current	Чтение блока	Много
Exclusive (X)	Current	Запись блока	Один на кластер



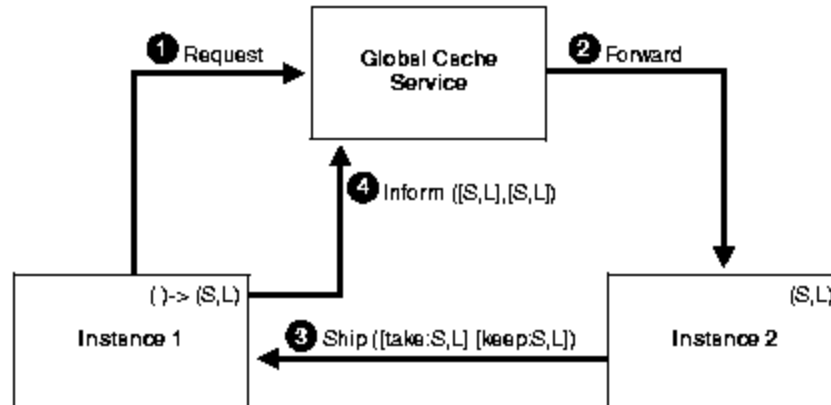
Запрос на изменение блока, находящегося в “current mode”



G = Global
PI = Past Image
X = Exclusive
x = mode
y = role
{ [x,y], [x,y] } = Disposition
 ([requestor disposition],
 [holder disposition])

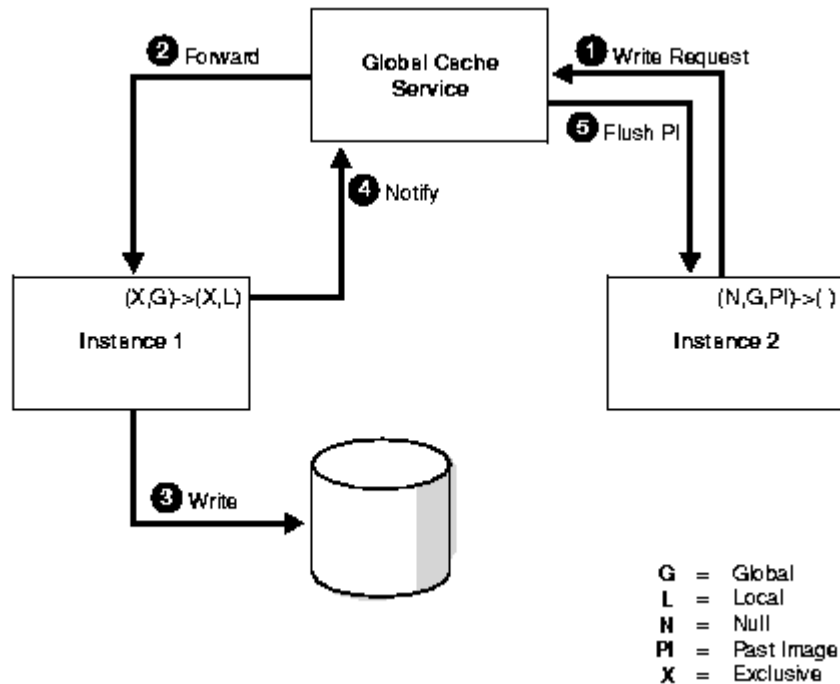
- Блок данных изменен и удерживается Инстансом 2 (X mode) в режиме “local”
- Инстанс 1 хочет его изменить:
 - Посылается запрос GCS
 - Инстанс 2 получает запрос
 - Состояние блока изменяется на нейтральное (N)
 - Копия блока продолжает храниться в кэше Инстанса 2 (Past Image, PI)
- Режим блока меняется на “global”
- LMS пересылает блок Инстансу 1 (сам блок, сообщение о том что Инстанс 2 имеет копию блока, сообщение о том что Инстанс 1 должен изменить состояние блока на заблокированное по чтению/записи (X) в режиме “global”
 - Инстанс 1 информирует GCS о изменении состояния блока (X) в режиме “global”

Запрос на чтение блока



S = Shared
L = Local
{ } = Null
(x,y) = Held
x = mode
y = role
{ [x,y], [x,y] } = Message Content
([requestor disposition], [holder disposition])

Запись блока на диск



- Параллельно может существовать несколько версий блока (на разных инстансах)
- Только последняя версия будет записана на диск, остальные отброшены
- Запись блока на диск:
 - На Инстансе 2 происходит checkpoint
 - Инстанс 2 посылает write request на GCS
 - GCS посылает запрос на Инстанс 1, так как Инстанс 1 является держателем блока
 - Инстанс 1 записывает блок на диск
 - Инстанс 1 отмечает окончание операции записи в GCS и блок переходит в режим "local"
 - GCS сообщает всем держателям копии блока (PI holders) о неактуальности копии блока

Тестирование производительности Oracle 9i RAC

- Тестовый пример (SunPS UK, OLTP application)
 - Задача – понять, каким образом масштабируется Oracle 9i RAC
 - С использованием средства измерения производительности, предоставленного заказчиком
 - Модель БД – OLTP
 - Без изменения кода приложения
 - 90:10 соотношение чтение/запись
 - Распределение записи: 6% insert и 4% update

Тестирование производительности Oracle 9i RAC, продолжение

- Платформа
 - 4 x SFv480R, на каждом
 - 4 x 900+ MHz CPUs
 - 16 GB RAM
 - Solaris 8
 - Oracle 9i (9.2.0.2)
 - SCI / RSM API & RSM RDT
 - Veritas Volume Manager 3.2
 - Дисковая подсистема EMC

Тестирование производительности Oracle 9i RAC, результаты

- Результаты
 - Традиционный тюнинг структуры БД и запросов SQL увеличивает производительность но приводит к худшим результатам с точки зрения масштабирования...
 - Хорошо масштабируются системы с большим объемом чтения (до 1.9 для 2 узлов) либо insert...
 - Приложения, не использующие секционированные таблицы, масштабируются плохо...
 - Производительность инстанса Oracle 9i с подключенной (linked), но выключенной опцией RAC (cluster_database=false), хуже на 10–25% чем одиночного инстанса
- Масштабирование
 - 2 узла: 1.32
 - 3 узла: 1.69
 - 4 узла: 2.09

Тестирование производительности Oracle 9i RAC, выводы

- Использовать межзловые соединения большей пропускной способности и имеющие меньшую задержку по доставке пакета (latency)
 - SCI / SunFire Link
- Использовать Remote Shared Memory (RSM)
 - RSM API для Oracle cache fusion
 - RSMRDT для обмена сообщений между инстансами Oracle
 - RSM позволяет полностью использовать межзловые соединения (round-robin), так как при использовании Ethernet и UDP/IP для работы Oracle используется только одно соединение
- Как разрешить использование RSM в Oracle 9.2.0.2 RAC
 - `_disable_sun_rsm` = FALSE
 - `_reliable_block_sends` = TRUE

Тестирование производительности Oracle 9i RAC, выводы

- Использовать несколько процессов Oracle LMS при работе с RSM
 - Пример: 2 Oracle LMS процесса для 4-х процессорного сервера
 - `_lm_lms = 2`
- Так же может понадобиться увеличить приоритет процессам LMS
 - `prionctl -s -c TS -p 60 -m 60 -i pid {}`

Тестирование производительности Oracle 9i RAC, статистика

- Статистика собрана посредством пакета Oracle **Statspack**:
 - Статистика Global Cache Service (GCS)
 - Статистика Global Enqueue Service (GES)
 - Wait Events
 - Instance Activity
 - Buffer Pool
 - Load Profile

Тестирование производительности Oracle 9i RAC, статистика GCS

Statistic (in ms)	GE	SCI	SFL
CR block build time	0	0.1	0
CR block flush time	0.1	0.15	0.1
CR block send time	0.2	0.3	0.1
CR block processing time	0.3	0.5	0.2
CR block receive time	25.9	19.3	1.4
Global cache get time	21.4	17.25	1.1
Current block pin time	0.4	0.5	0.3
Current block flush time	0	0.05	0
Current block send time	0.2	0.3	0.1
Current block processing time	0.7	0.8	0.4
Current block receive time	60.9	14.9	1.5
Global cache convert time	37.4	26.7	1.25

Тестирование производительности Oracle 9i RAC, статистика GES и Message Statistics

Statistic (in ms)	GE	SCI	SFL
Global lock get time	3.7	3.05	0.35
Global lock convert time	0.1	0.15	0.1
GES message process time	0.1	0.15	0.1

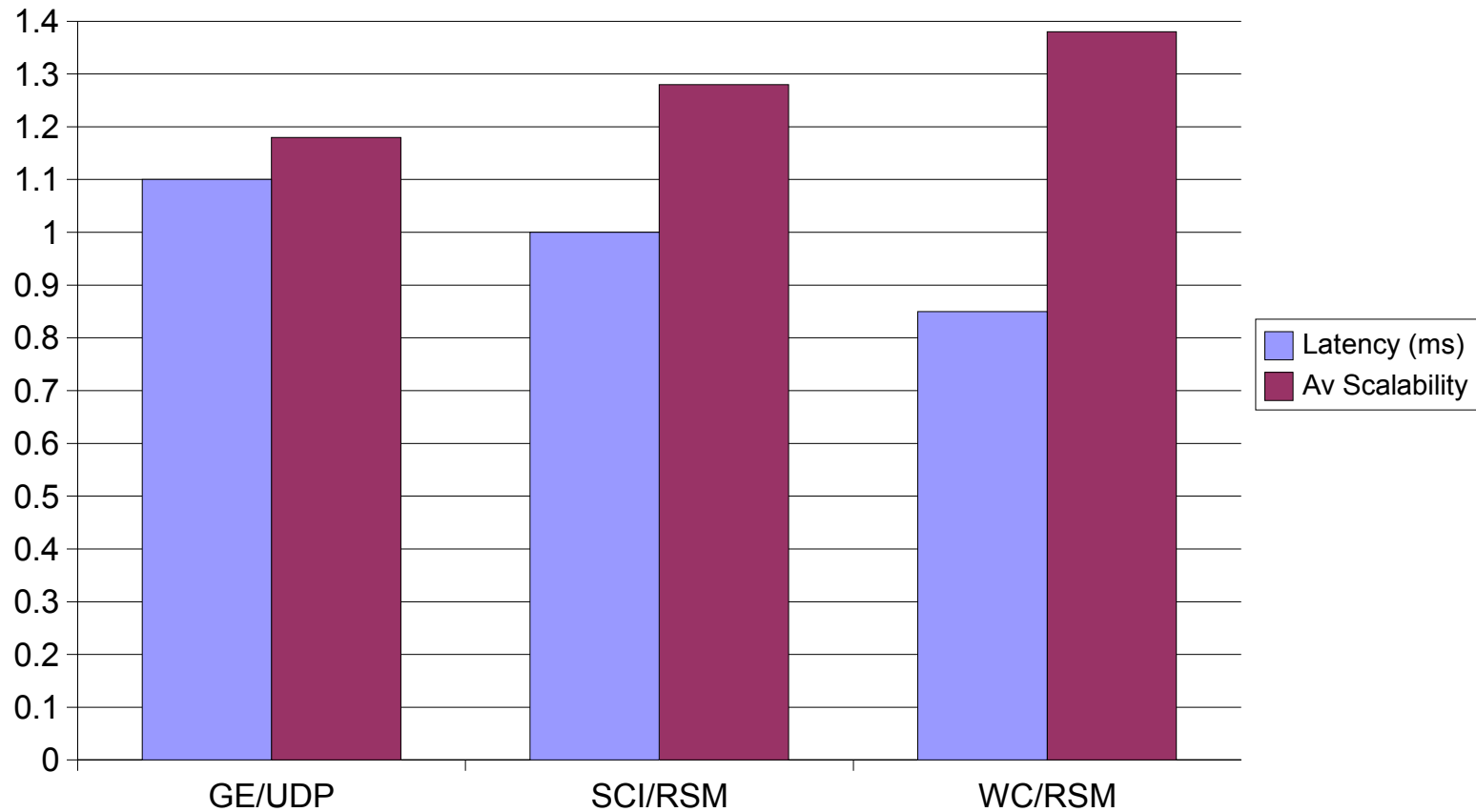
Statistic	GE	SCI	SFL
Message sent queue time	34.7	0.6	0.15
% of flow controlled messages	4.6	0.95	0.25
GCS side channel messages	6.8	0	0

RAC wait events

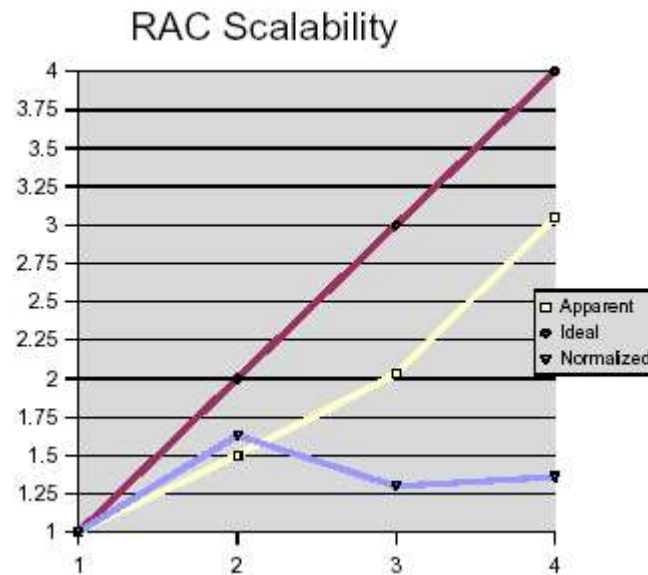
Event	GE	SCI	SFL
Global cache CR request	8750	6418	420
Global cache s to x	3709	5020	277
Buffer busy global cache	284	302	65.5
Enqueue	572	618	31
Global cache open x	134	701	40
Global cache null to x	229	416	136
Global cache busy	31	83	15
Global cache null to s	42	53	11
Buffer busy global CR	257	174	9
Wait for msg sends	3016	8.5	4

Latency vs scalability

Latencies seen by Oracle

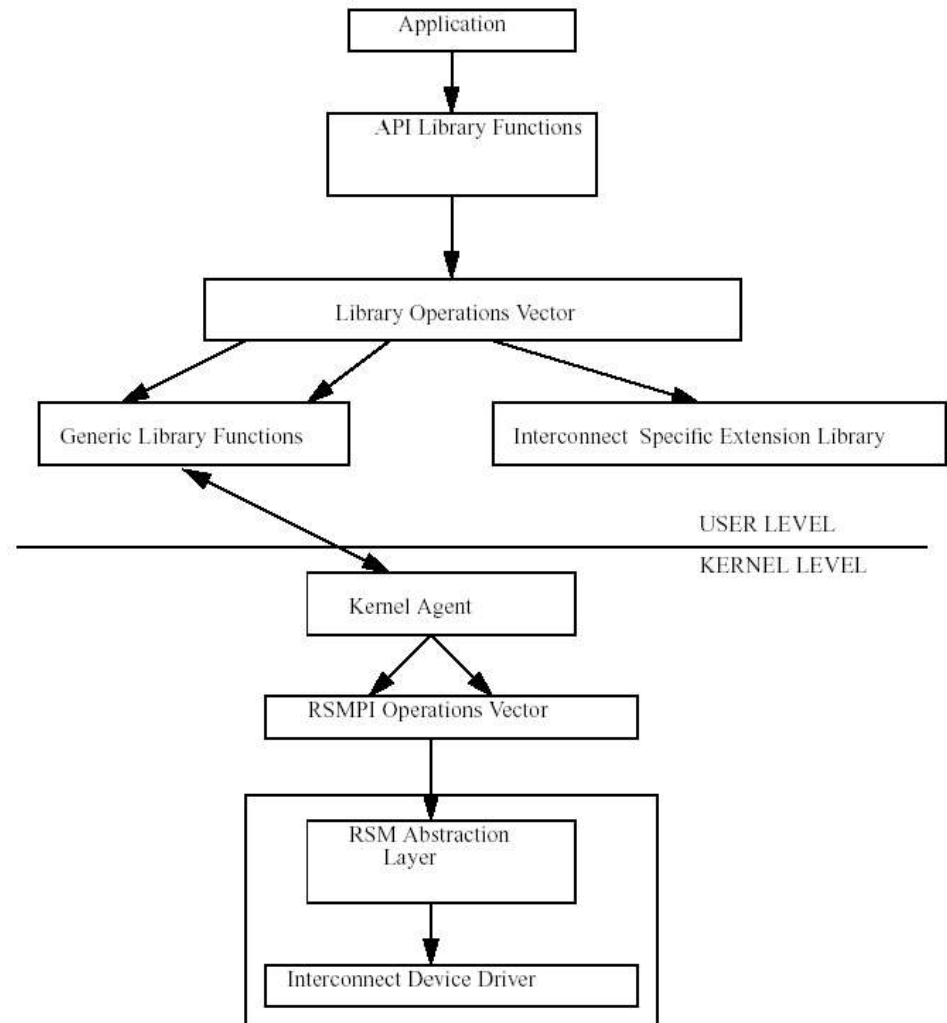


Масштабируемость Oracle 9i RAC: OeBS



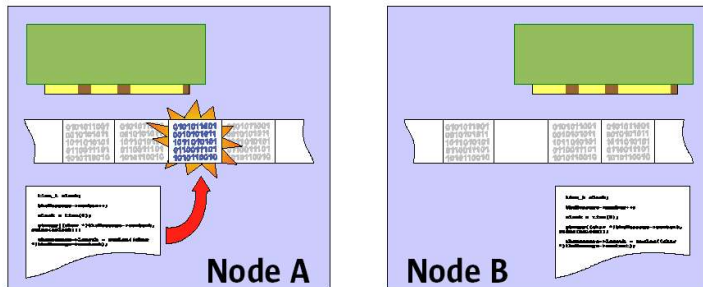
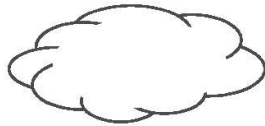
Solaris Remote Shared Memory (RSM)

- Эффективный способ межзвлового взаимодействия в кластере без накладных расходов
- Позволяет получить доступ к сегментам памяти одного узла кластера с другого узла кластера
- Требования к системе:
 - SPARC
 - Solaris 8 10/10 или старше
 - Memory-based межзвловое соединение
 - Sun Fire Link (WildCat)
 - SCI (PCI)
 - Sun Cluster 3.0 12/01 или старше

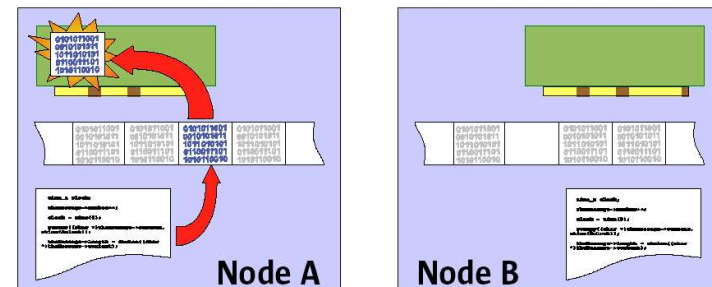


RSM: как это работает

- Allocate some memory on the host node

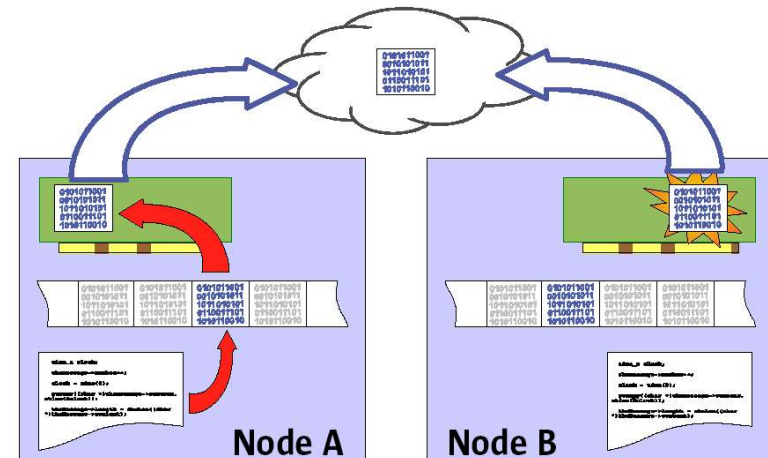
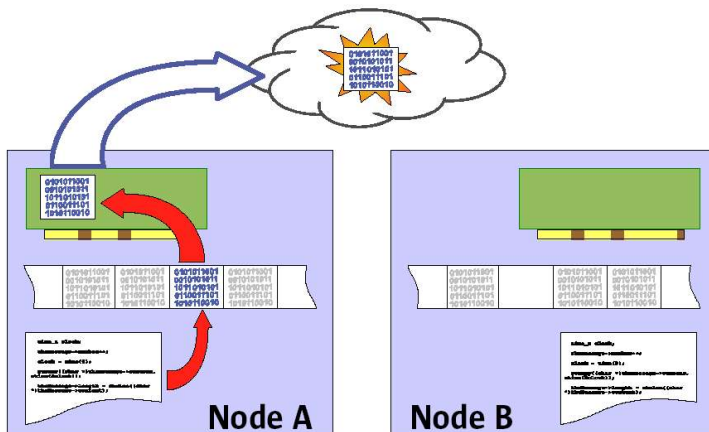


- Assign memory to an exported segment



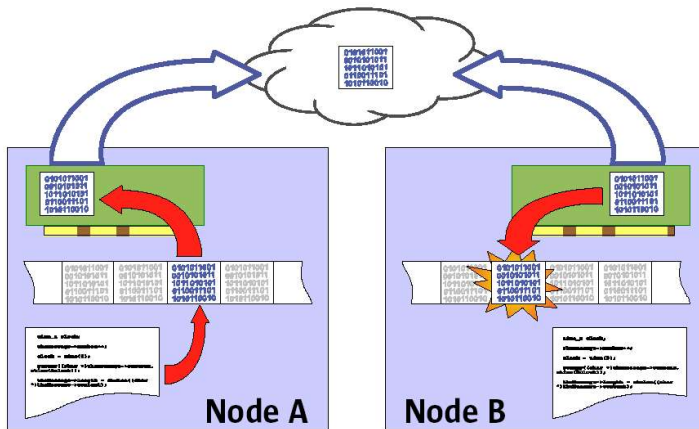
- Publish segment

- Connect to published segment

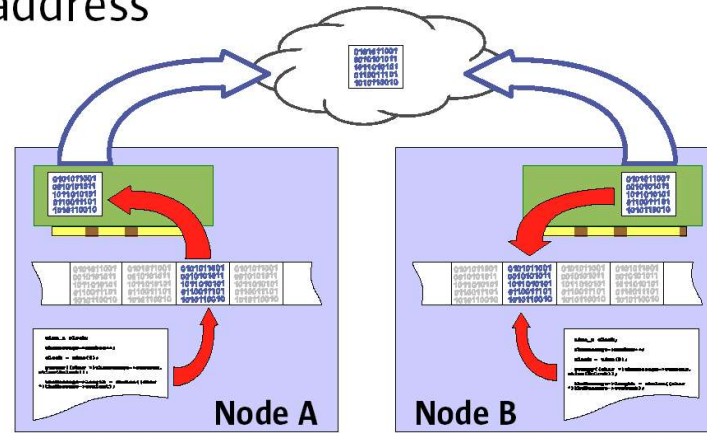


RSM: как это работает...

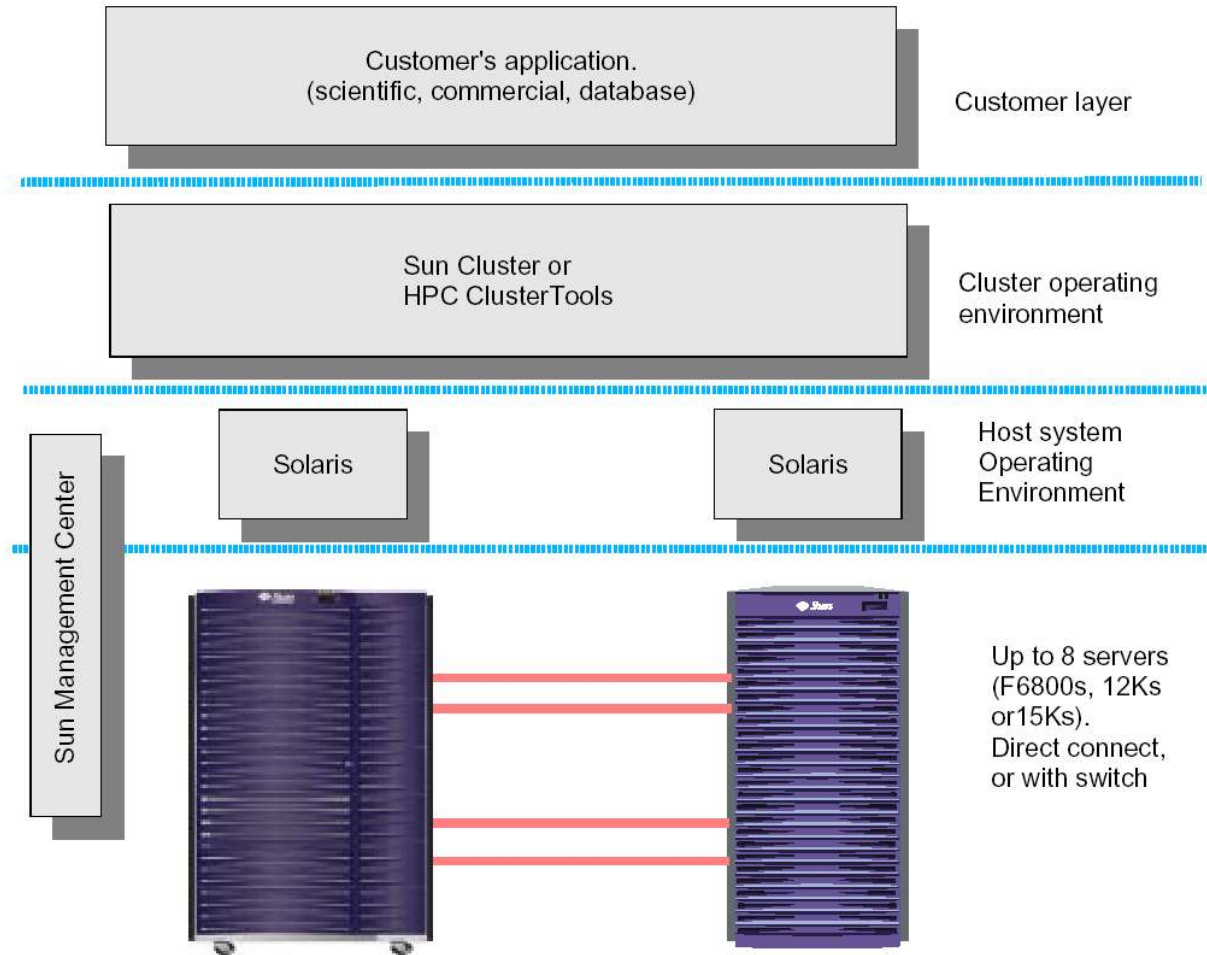
- Map segment into local address



- Access shared memory via local address

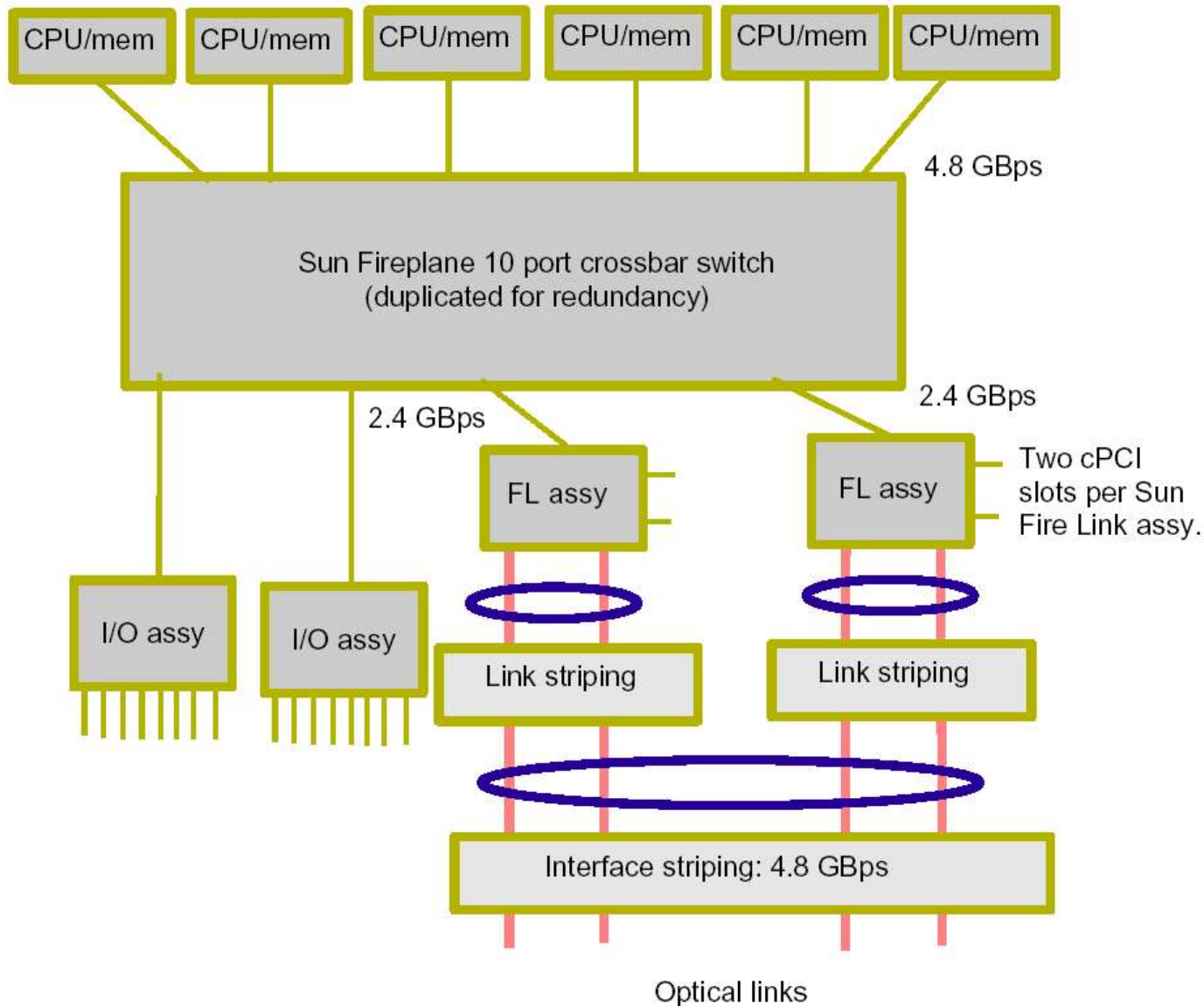


Sun Fire Link...

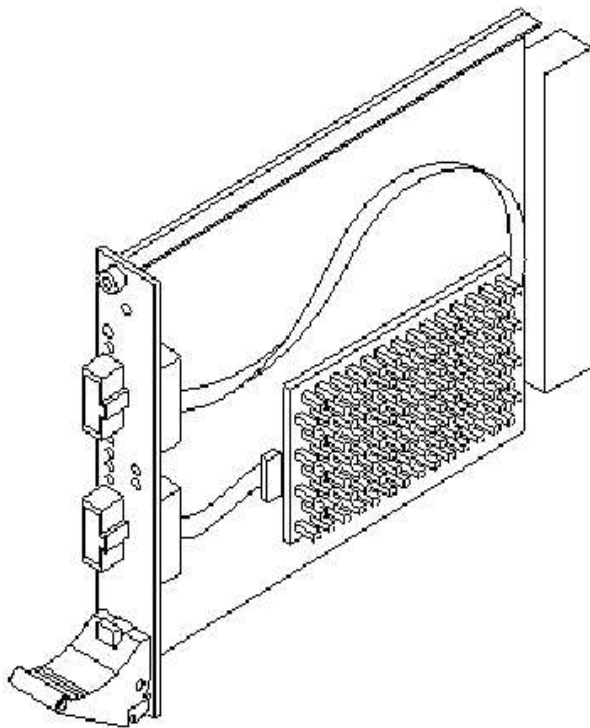


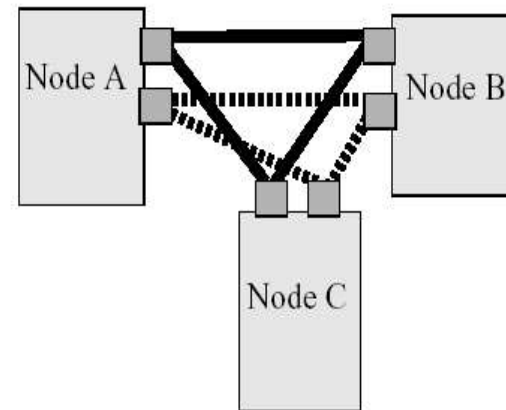
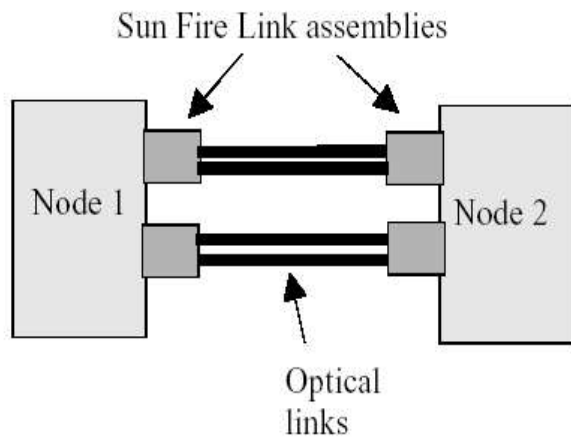
...Sun Fire Link





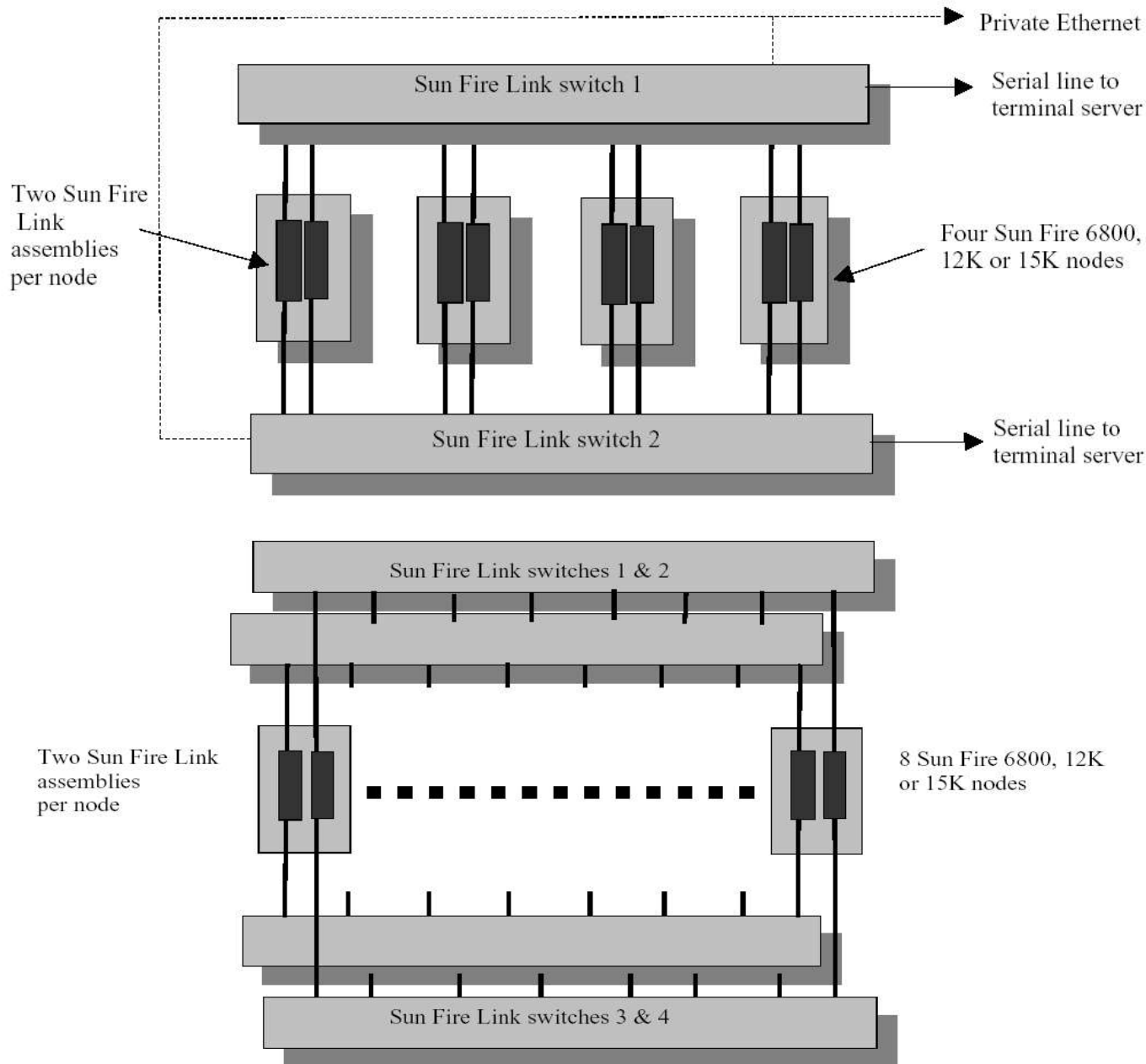
SFL подходит для межзвонкового взаимодействия внутри ВЦ: длина
кабеля до 20м





Коммутатор Sun Fire Link

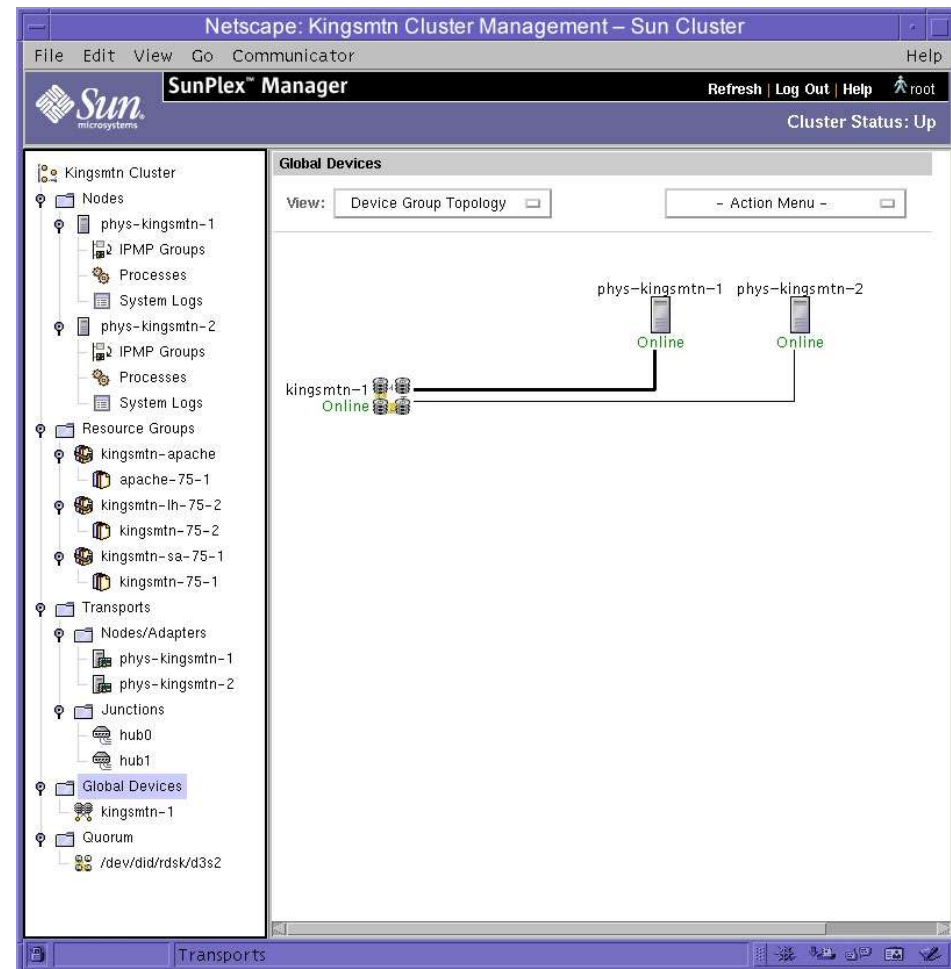




SunPlex manager

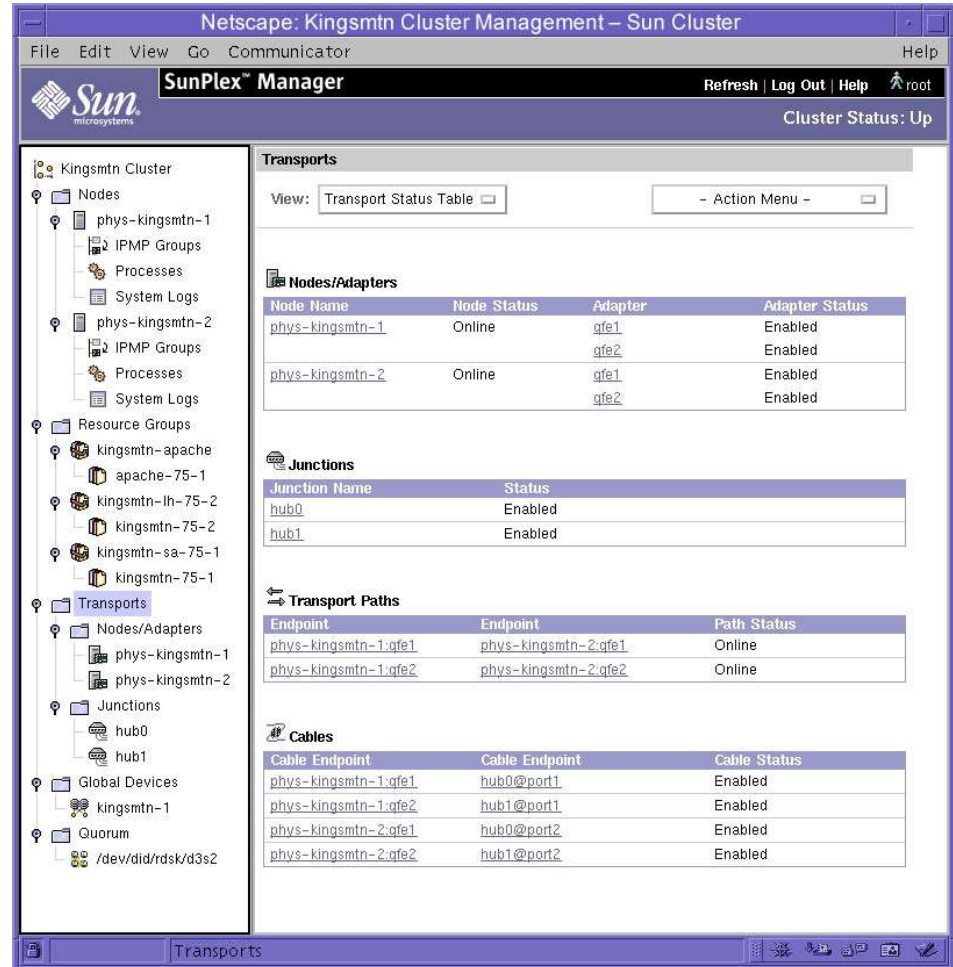
EASY ADMINISTRATION FOR SUNPLEX SYSTEMS

- Быстрая инсталляция и настройка SunCluster
- Управление SunCluster
- Solaris Role Based Access Control New SC 3.1 10/03
- Centralized Install New SC 3.1 10/03



Мониторинг SunPlex

- CLI
- GUI SunMC – мониторинг всех компонентов SunPlex:
 - Аппаратное обеспечение
 - Операционная система
 - Информация о конфигурации кластера: узлы, межузловые соединения, группы устройств, quorum devices, состояние
- Может быть интегрирован с системой управления предприятием (EM) через SNMP



The screenshot displays the SunPlex Manager interface within a Netscape browser window. The main content area is divided into several sections:

- Transports:** View: Transport Status Table. Includes an Action Menu.
- Nodes/Adapters:** A table showing the status of nodes and their adapters.
- Junctions:** A table showing the status of network junctions.
- Transport Paths:** A table showing the status of transport paths between nodes.
- Cables:** A table showing the status of cables connecting nodes.

The left sidebar shows a tree view of the cluster hierarchy, including Nodes, Resource Groups, Transports, Junctions, Global Devices, and Quorum.

Node Name	Node Status	Adapter	Adapter Status
phys-kingsmtn-1	Online	qfe1	Enabled
		qfe2	Enabled
phys-kingsmtn-2	Online	qfe1	Enabled
		qfe2	Enabled

Junction Name	Status
hub0	Enabled
hub1	Enabled

Endpoint	Endpoint	Path Status
phys-kingsmtn-1.qfe1	phys-kingsmtn-2.qfe1	Online
phys-kingsmtn-1.qfe2	phys-kingsmtn-2.qfe2	Online

Cable Endpoint	Cable Endpoint	Cable Status
phys-kingsmtn-1.qfe1	hub0@port1	Enabled
phys-kingsmtn-1.qfe2	hub1@port1	Enabled
phys-kingsmtn-2.qfe1	hub0@port2	Enabled
phys-kingsmtn-2.qfe2	hub1@port2	Enabled



Доступность и масштабируемость кластерных систем

Безруков Валерий
Технический консультант
Sun Microsystems

