Логическая поддержка когерентности в Эльбрус/МЦСТ-XR серверах среднего уровня

Цель и область применения

- ***Область применения – серверы на базе микропроцессоров Эльбрус/МЦСТ-XR***
- ***В серверах среднего уровня т.е. с числом процессоров до 16***

Примечание: в данной презентации термином «процессор» обозначается чип процессоров

- ***Сервер может быть использован как базовый элемент в большой системе***

Основные цели

Для систем с числом процессоров более 4-х надо

1. Рассмотреть варианты протокола когерентности:

- ***с меньшими требованиями к пропускной способности внешних связей;***
- ***позволяющие в полной мере воспользоваться результатами программистов по локализации ресурсов процесса на одном или нескольких близлежащих процессорных чипах.***

2. Предложить аппаратную поддержку для уменьшения числа обращений к удаленной памяти

Построение системы с16-ю процессорами

- Чип_КК использует существующий протокол когерентности для взаимодействия с процессорами кластера
- Для взаимодействия с процессорами удаленных кластеров используется расширенный вариант протокола
- Чип_КК должен выглядеть как еще один процессор для всех локальных процессоров кластера
- Каждый процессорный модуль(кластер) включает чип когерентности и коммутации(Чип_КК)

Основные положения

- Чип-КК расширяет возможности построения систем до 4-х кластеров
- Для быстрого доступа к удаленным данным введен фильтр (Filter)
- Для уменьшения потока Snoop-запросов за пределы кластера введен справочник(Directory)
- Чип-КК использует протокол когерентности, разработанный для однокластерной системы,т.е. точкой синхронизации работы с одной строкой является Home-узел, причем сохраняется полное снупирование всех процессоров Home-кластера; Цель – уменьшить доработки процессора

Протокол MOESI

- Состояния:
- **I** – **Invalid** – нет данных;
- **E** – **Exclusive** – данные есть только у одного владельца; копия не изменена относительно данных в памяти;
- **S- Shared** – данные есть у нескольких совладельцев;
- **M**– **Modified** - данные есть только у одного владельца; копия изменена относительно данных в памяти;
- **O** – **Owned** - данные есть у нескольких совладельцев;копия изменена относительно данных в памяти;

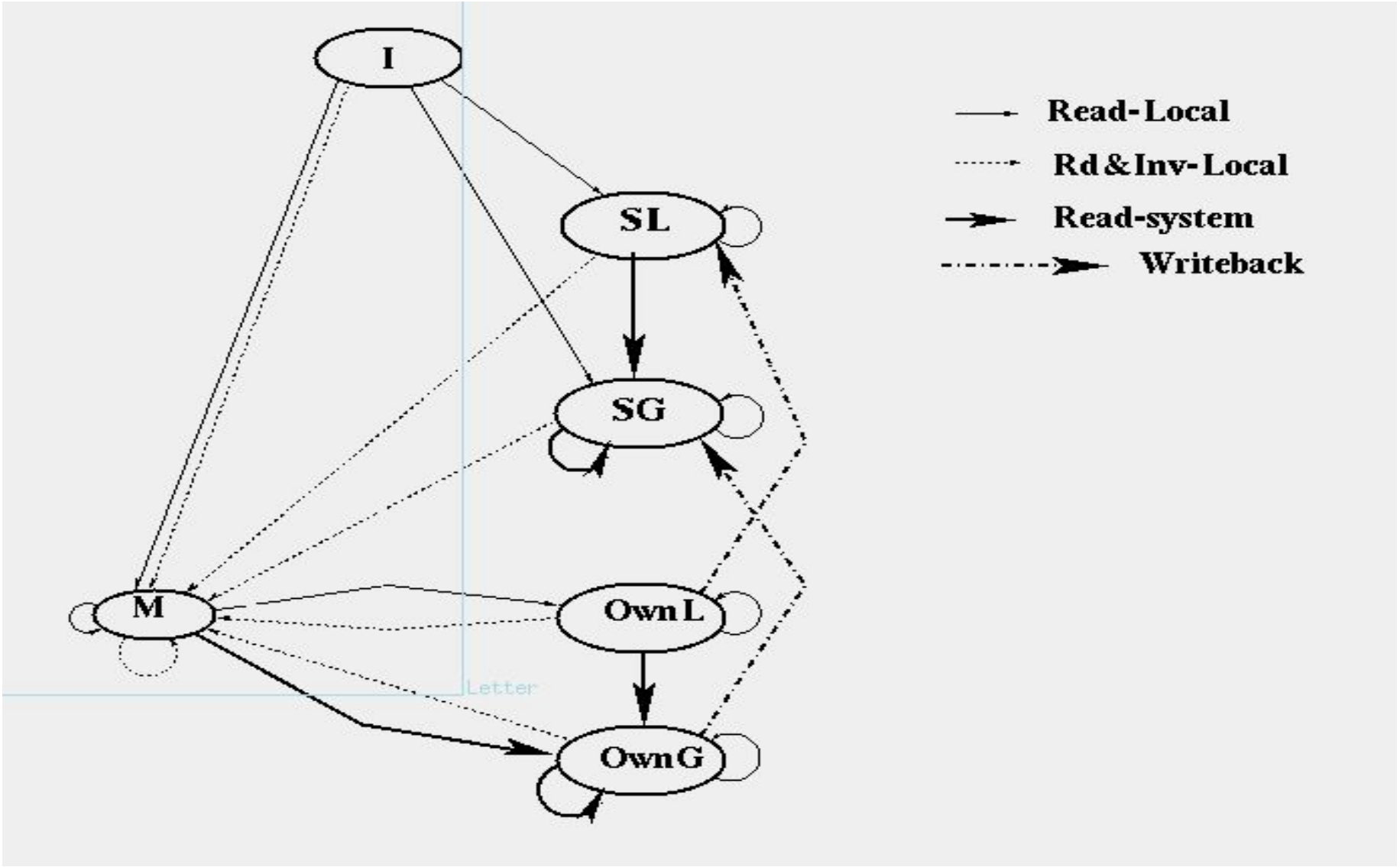
Способы решения проблемы пропускной способности в линках

- 1. Фильтр исключает обращения за пределы кластера для данных из удаленных кластеров, если строка в состояниях M. Данные пересылаются внутри кластера.
- 2. Промахи в справочнике исключают обращения за пределы кластера для локальных данных. Данные пересылаются внутри кластера.
- 3. При попадании в справочник запросы высылаются НЕ ВСЕМ кластерам, но лишь только тем кластерам, где действительно находятся копии данных
- 4. Совместное использование фильтра и справочника должно сократить требования к пропускной способности внутрикластерных и межкластерных линков.

Состояния строки в фильтре

- Состояния:
- **I** – **Invalid** – нет данных;
- **S_G** - **Shared_global** – данные есть у нескольких совладельцев, причем НЕ только данного кластера;
- **S_L** - **Shared_local** – данные есть у нескольких совладельцев, но только данного кластера;
- **M** – **Modified** - данные есть только у одного владельца; копия изменена относительно данных в памяти;
- **O_L** – **Owned_local** - данные есть у нескольких совладельцев, но только данного кластера;копия изменена относительно данных в памяти;
- **O_G** – **Owned_global** - данные есть у нескольких совладельцев, причем НЕ только данного кластера;копия изменена относительно данных в памяти;

Автомат состояний фильтра



Время доступа к локальной памяти с использованием справочника

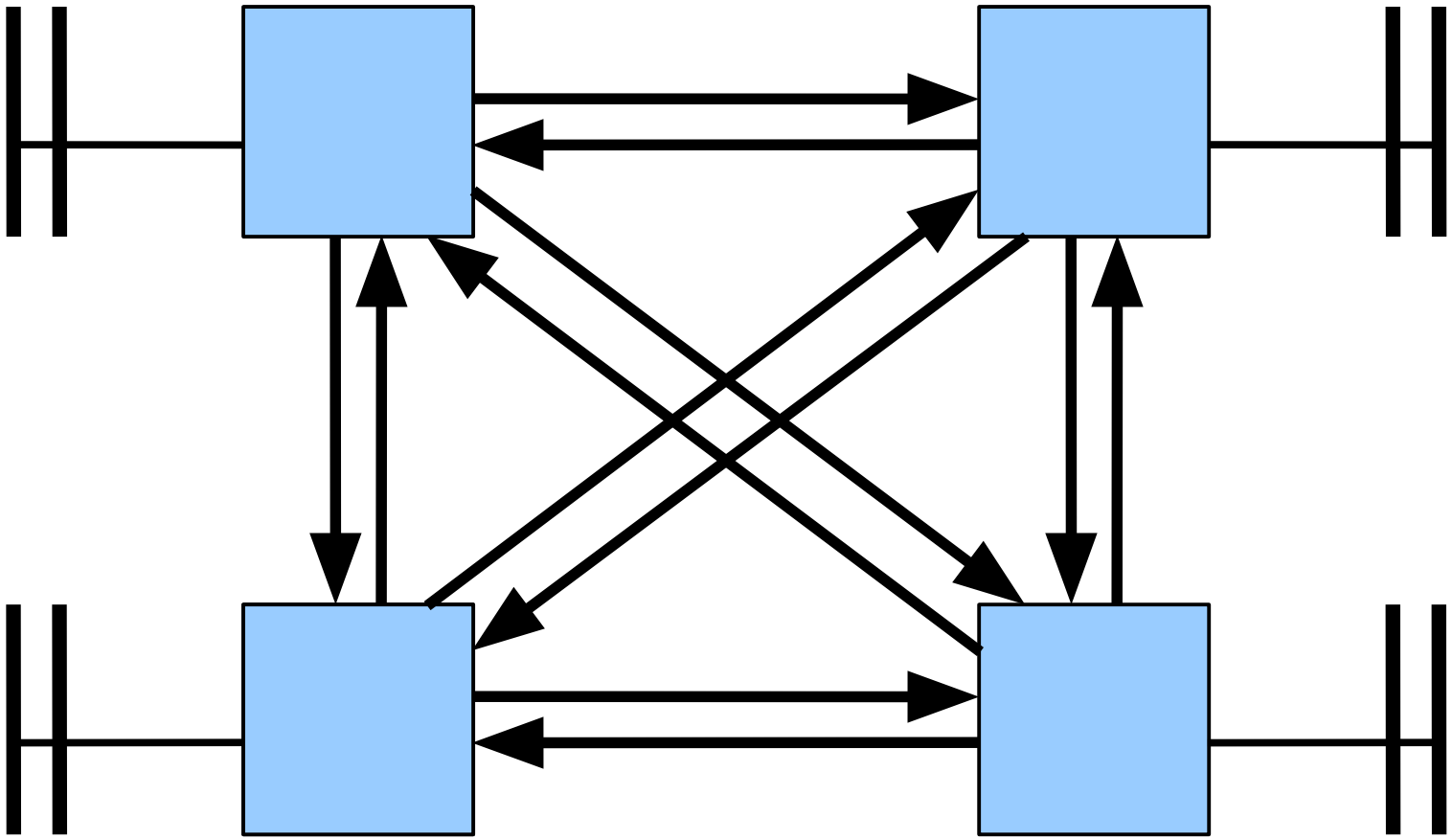
- Справочник отслеживает состояния строк локальной памяти кластера
- Для каждой строки, которая кэширована в удаленном кластере в справочнике поддерживается ее состояние (Shared, Modified, Owned) и бит-вектор кластеров совладельцев данных строки
- Справочник, в случае конфликта по ресурсам, требует вытеснения строки -”жертвы” из кэш-памятей всех совладельцев
- Для запросов от локальных процессоров за локальными данными в случае промаха в справочнике уменьшается время доступа к данным за счет исключения выхода за пределы кластера

Состояния строки в справочнике

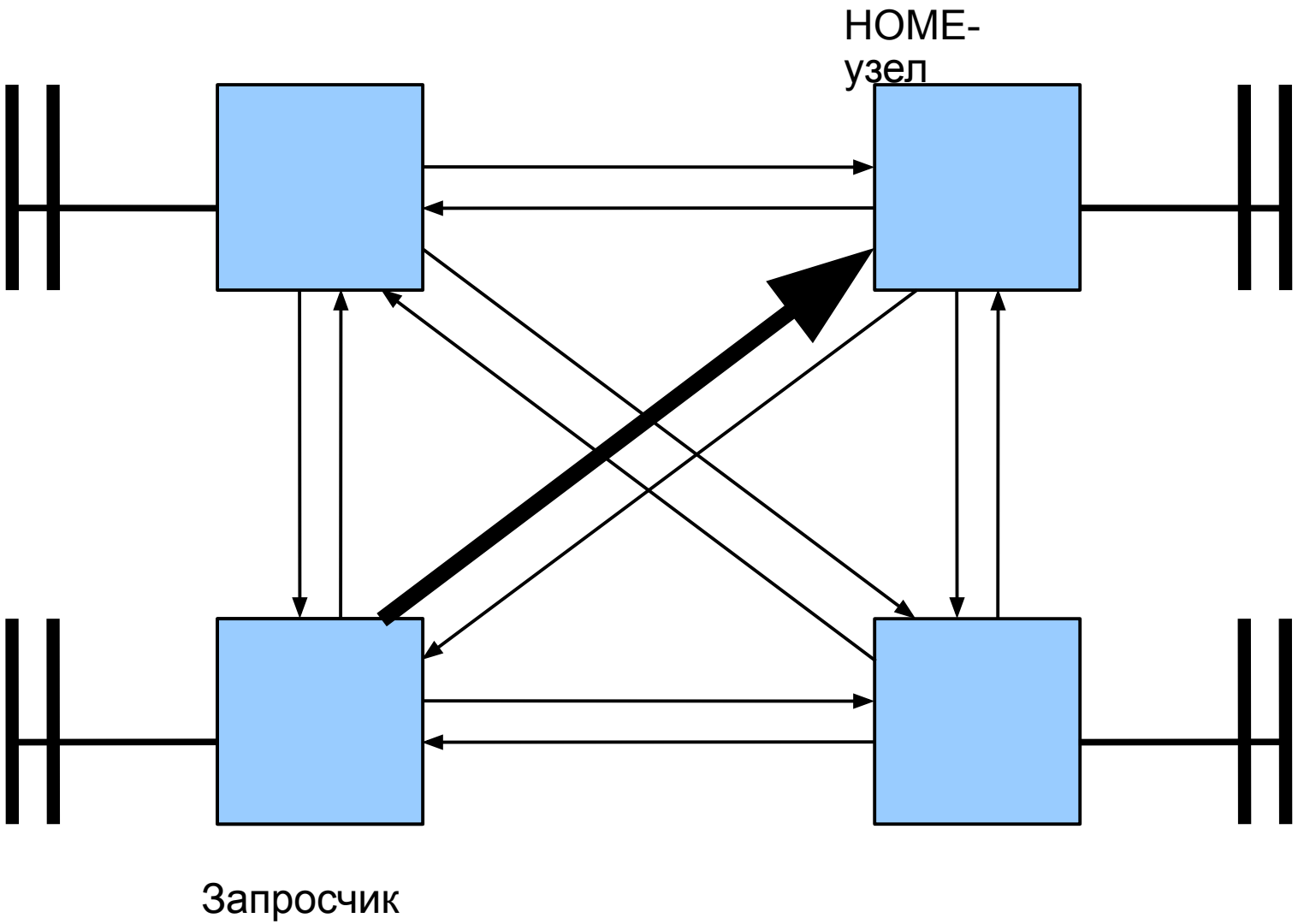
- Состояния:
- **I** – **Invalid** – нет данных;
- **S** - **Shared** – данные есть у нескольких кластеров-совладельцев,
- **M** – **Modified** - данные есть только у одного кластера-владельца;
  копия изменена относительно данных в памяти;

Время доступа к удаленной памяти с использованием фильтра

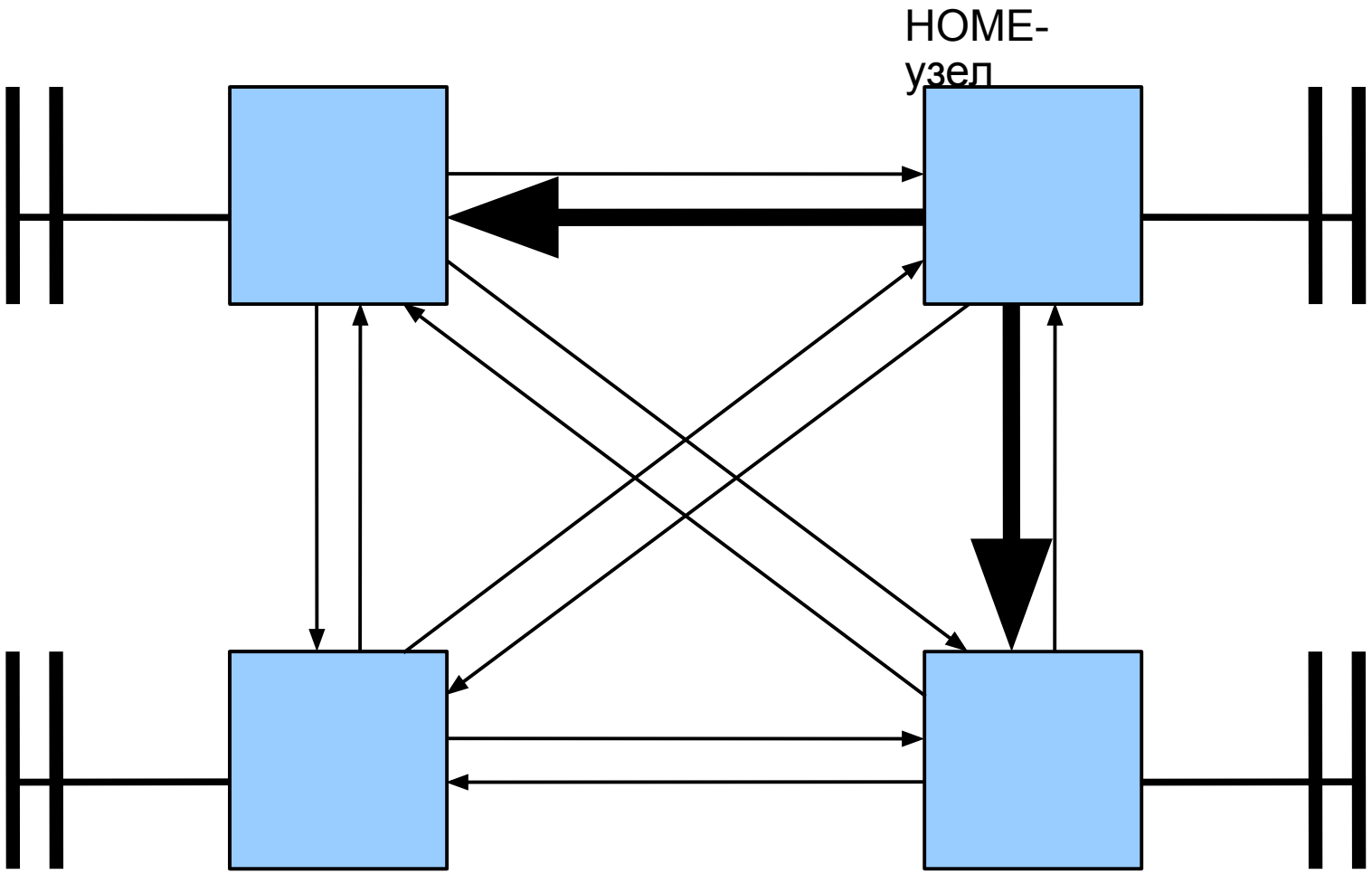
- Фильтр отслеживает состояния строк удаленной памяти внутри данного кластера
- Для каждой строки, которая кэширована в кластере в фильтре поддерживается ее состояние (Shared, Modified, Owned, причем имеется подсказка о наличии копий этой строки в других кластерах т.е. является ли копия только локально или глобально кэшируемой) и бит-вектор процессоров совладельцев данных строки
- Фильтр, в случае конфликта по ресурсам, требует вытеснения строки -”жертвы” из кэш-памятей всех процессоров-совладельцев данного кластера
- Для запросов от локальных процессоров за удаленными данными в случае попадания в фильтр уменьшается время доступа к данным за счет исключения выхода за пределы кластера в состояниях строки Shared_лок, Modified, Owned_лок



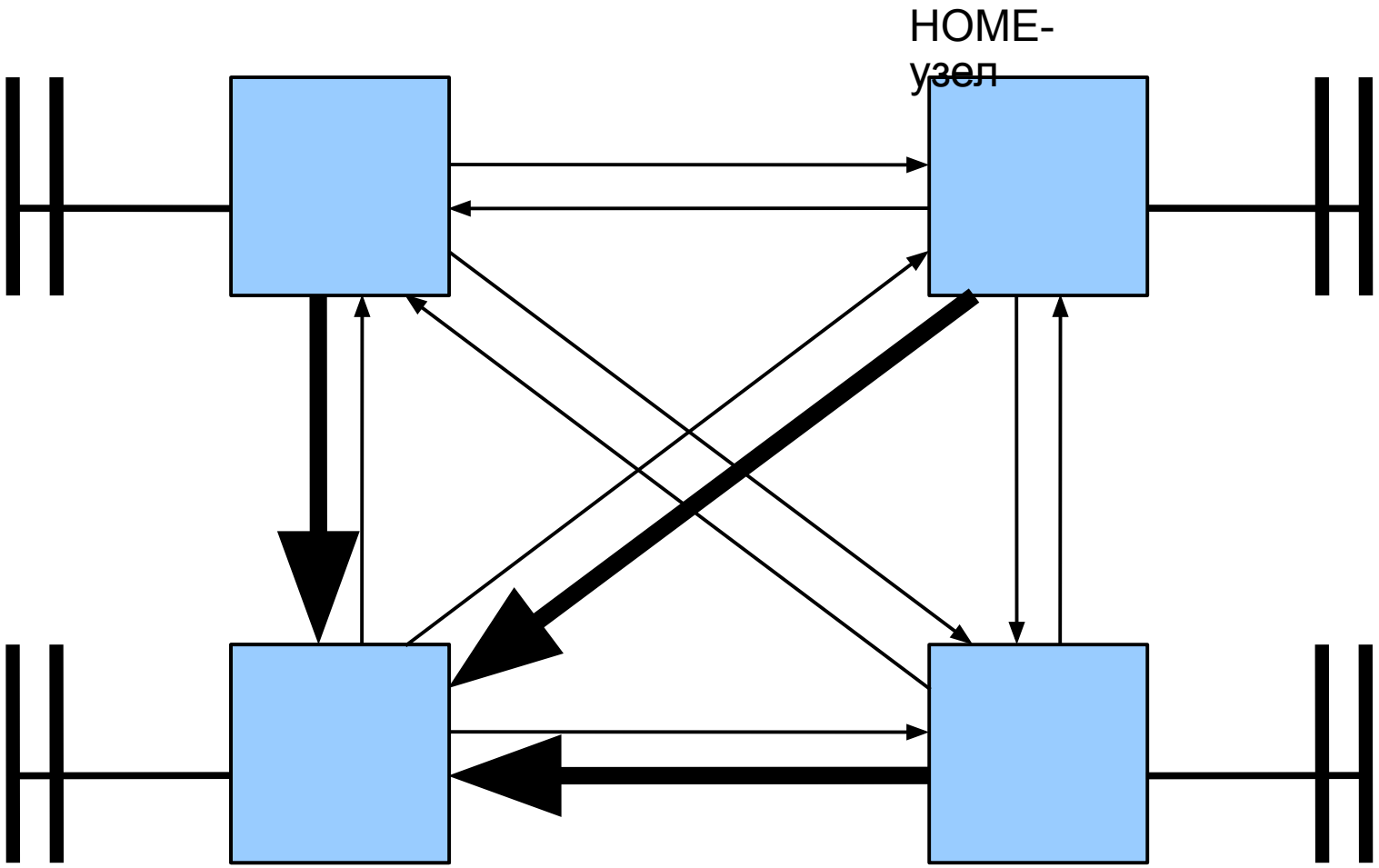
кластер с 4-мя процессорами



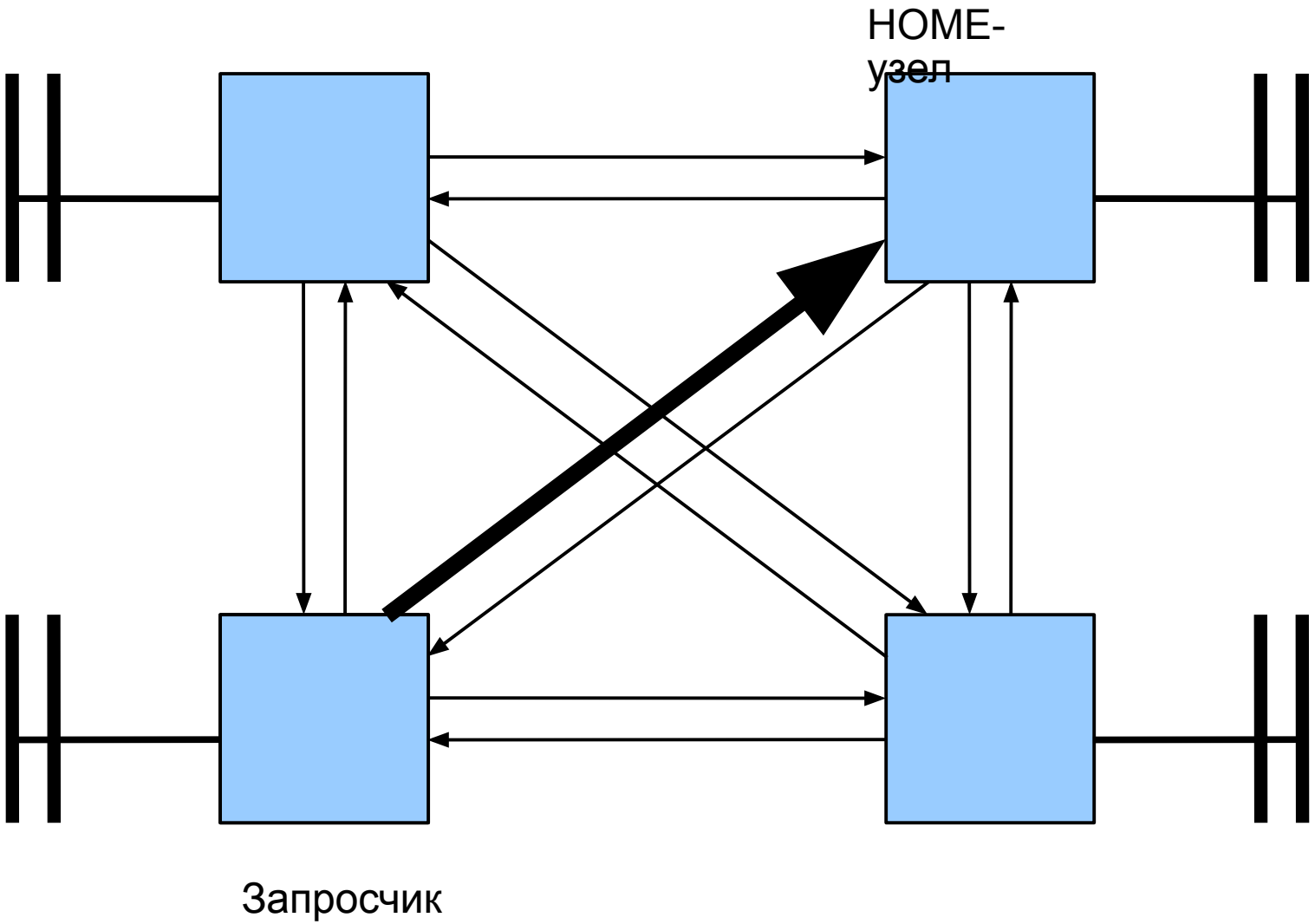
1-ая стадия: Запрос Read_64 в HOME-узел



2-ая стадия: Выдача Snoop-запросов



3-я стадия: сбор Snoop-ответов



4-ая стадия: Сообщение о завершении операции

Особенности существующего протокола когерентности

- ***Snoop-based – т.е. используется опрос кеш -памяти всех процессоров системы***
- ***Требует минимального дополнительного оборудования в процессоре***
- ***Логическая простота протокола позволяет в кратчайшие сроки произвести разработку и верификацию RTL-описания***
- ***Система внешних связей (линков) позволят создавать конфигурации от 2-х до 4-х процессоров с минимальным числом проходов (hop'ов) между процессорами***

Локальность ресурсов процесса

Ресурсы процесса включают:

- ***процессор(ы)***
- ***память***

Варианты локальности:

- процессор обращается только к памяти “своего” чипа;
- процессор обращается только к памяти “своего” чипа и памяти близлежащих (по числу hop'ов доступа) чипов – это позволяет увеличить используемую данным процессором пропускную способность памяти;
- группа процессоров обращается только к памяти “своих” чипов и памяти близлежащих (по числу hop'ов доступа) чипов.

Что предлагается

Для
а) уменьшения времени доступа к данным памяти и

б) исключения непроизводительных потерь пропускной способности линков внешних связей процессора

Ввести в кластер (модуль процессора) два устройства:

- справочник (Directory) и
- фильтр (Filter)

1. Использовать справочник (Directory) для отслеживания состояния и местонахождения локальных данных кластера, взятых в удаленные кластеры

2. Использовать фильтр (Filter) для отслеживания состояния и местонахождения внутри кластера данных, взятых из удаленных кластеров

Варианты справочников

- *Полный справочник* – имеет информацию о каждой строке памяти
- *Усеченный справочник* - имеет информацию не о каждой строке памяти, а только о тех строках данных, копии которых взяты в кэш-памяти процессоров; организуется в виде кэш-памяти; возможные проблемы связаны
  - с размерами необходимой кэш-памяти для большой системы и
  - как следствие с возможностью принудительного вытеснения данных из кэш какого-либо процессора для освобождения памяти для размещения информации о строке данных размещаемой в другом процессоре

Организация справочника (Directory)

- Элемент справочника имеется для каждой строки локальной памяти, взятой в удаленные кластеры
- Справочник организован в виде множественно-ассоциативной кэш-памяти с числом колонок не менее их суммарного числа во всех процессорах удаленных кластеров (~12X4=48)

Организация справочника (Directory) (продолжение)

Структура элемента справочника(MOESI протокол):

- состояние строки данных (2бита);

- указатель на владельца модифицированной

  копии данных (logN бит);

- бит-вектор указателей на совладельцев копий данных (N бит),

  где N – число кластеров в системе

- адресный тег (40бит адреса, исключая 6 разрядов адреса внутри строки и число разрядов индекса при обращении в кэш-память)

Примечание: Кластер (процессорный модуль) должен быть одним абонентом справочника независимо от числа процессоров внутри него, состояние и местоположение строки внутри кластера уточняется по информации фильтра

Организация фильтра (Filter)

- Элемент фильтра имеется для каждой строки памяти, взятой из удаленных кластеров
- Состояние строки отражает наличие копий строки в других кластерах (признак локальности или глобальности копии)
- Фильтр организован в виде множественно-ассоциативной кэш-памяти с числом колонок не менее их суммарного числа во всех процессорах кластера (~4X4=16)

Организация фильтра (Filter) (продолжение)

Структура элемента фильтра (модифицированный MOESI протокол):

- состояние строки данных (2/3 бита);
- указатель на владельца модифицированной копии данных (logN бит);
- бит-вектор указателей на совладельцев копий данных (N бит), где N – число процессоров в кластере
- адресный тег (40бит адреса, исключая 6 разрядов адреса внутри строки и число разрядов индекса при обращении в кэш-память)

Аппаратные затраты на справочник и фильтр

- СПРАВОЧНИК
- Число строк в процессорах удаленных кластеров: [2M/64(байта в строке)] x 12 = 24/64 M(строк) = 0.375M(строк)
- Размер элемента справочника: ~4байта
- Размер справочника: 0.375 M(строк) x 4(байта на строку) = 1.5 Мбайт

- ФИЛЬТР
- Число строк в процессорах локального кластера: [2M/64(байта в строке)] x 4 = 8/64 M(строк) = 0.125M(строк)
- Размер элемента фильтра: ~4байта
- Размер фильтра: 0.125M (строк) x 4(байта на строку) = 0.5 Мбайт

Возможности встроенной памяти в ALTERA Stratix FPGA Family

- StratixIII FPGA Family – 65nm process
- M9K Memory Blocks – 1040
- M144 Memory Blocks – 48
- Embedded Memory(Kbits) – 16,272 ~ 2MBytes
- Package - F1760

- StratixIV FPGA Family – 40nm process
- M9K Memory Blocks – 1280
- M144 Memory Blocks – 64
- Embedded Memory(Kbits) – 20,736 ~ 2.5MBytes
- Package - F1932

Возможности IO в ALTERA StratixIV FPGA Family

- User I/O - 904
- Full-Duplex LVDS(Receive/Transmit) – 98
- Medium performance LVDS Channels – 256
- Transceivers - 48
  (full-duplex CDR-based transceivers at up to 8.5 Gbps)
- Package - F1932

Особенность *Эльбрус/МЦСТ-XR NUMA архитектуры*

с учетом добавляемых аппаратных средств: справочника (directory) и фильтра для строк данных, взятых из удаленных узлов

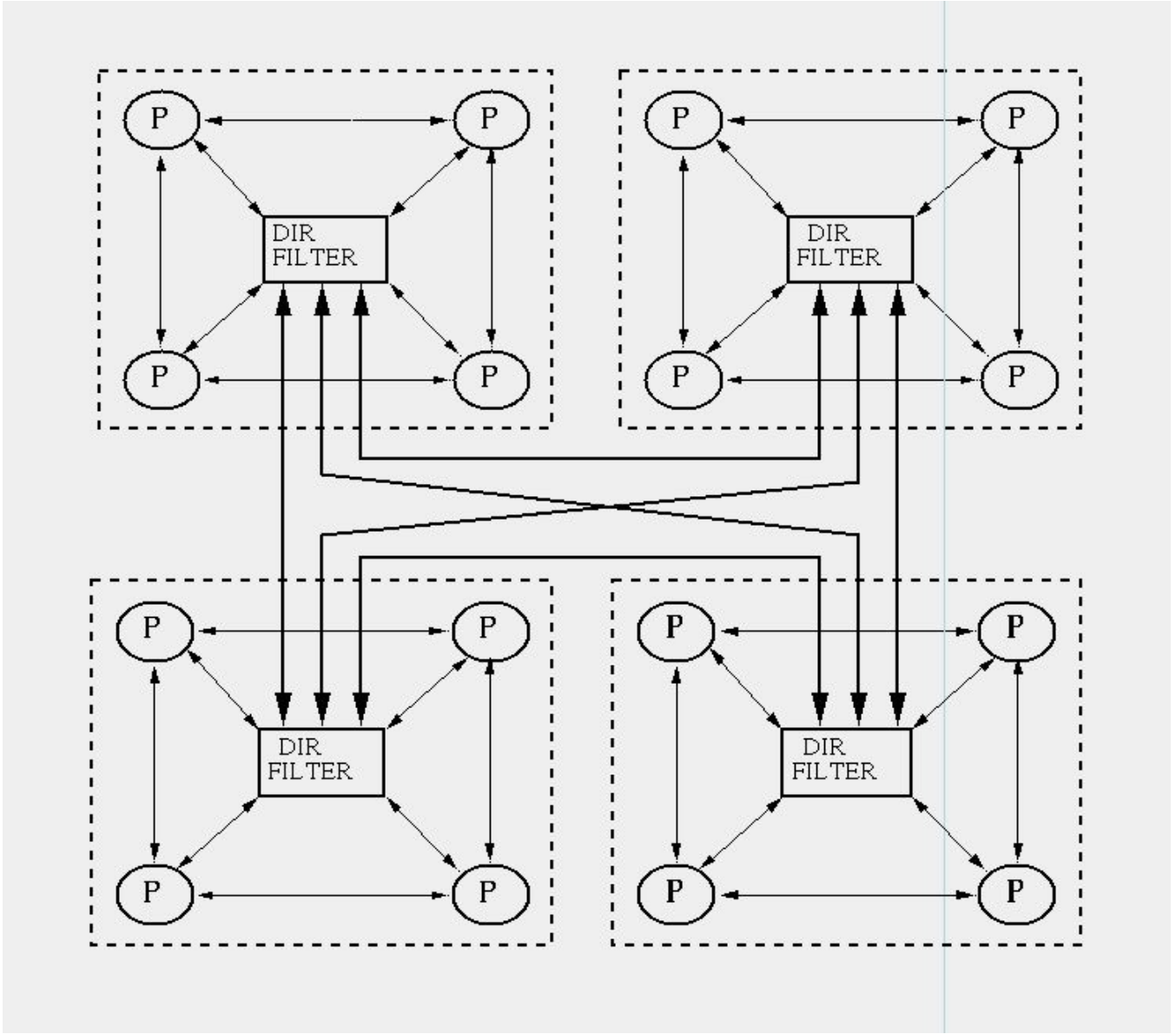
Протокол когерентности с использованием добавляемых аппаратных средств:
справочника (directory) и
фильтра для строк данных, взятых из удаленных узлов

позволяет использовать локальность ресурсов процесса т.е. процессора и памяти, ограничивая обращения к памяти пределами одного кластера

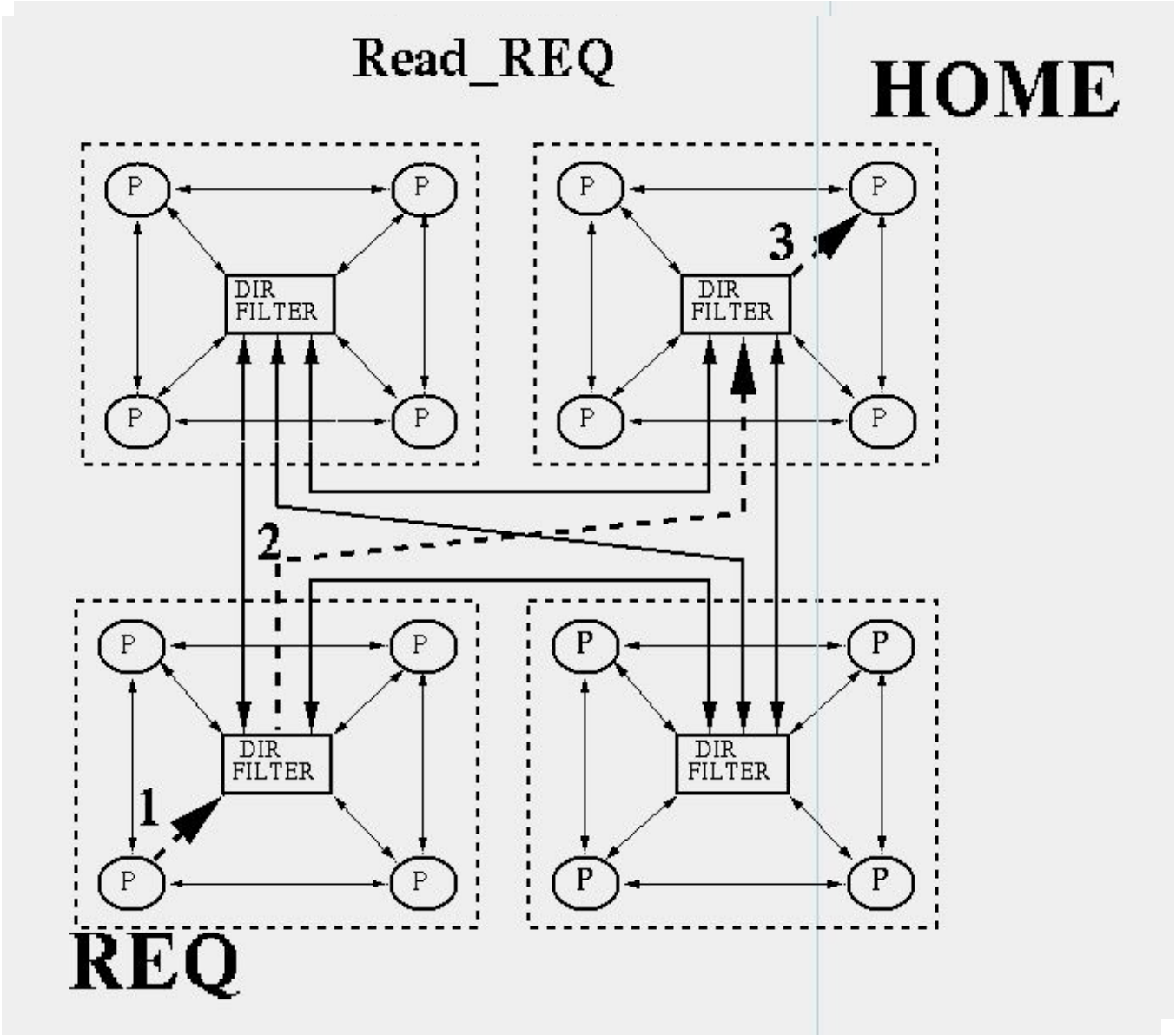
Read_64 операция в 16-ти процессорной системе

- Передача Запроса Read_64 от Запросчика(REQ) в Home – 3 пакета
- Snoop_Req в Home-кластере – 4 пакета
- Snoop_Resp в Home-кластере – 3 коротких пакета
- Data_Resp в Home-кластере – 2 пакета данных
- Data_Resp от Home-кластера – Запросчику – 2 пакета данных по 2-м линкам
- Ответ от Запросчика(REQ) в Home - 3 коротких пакета

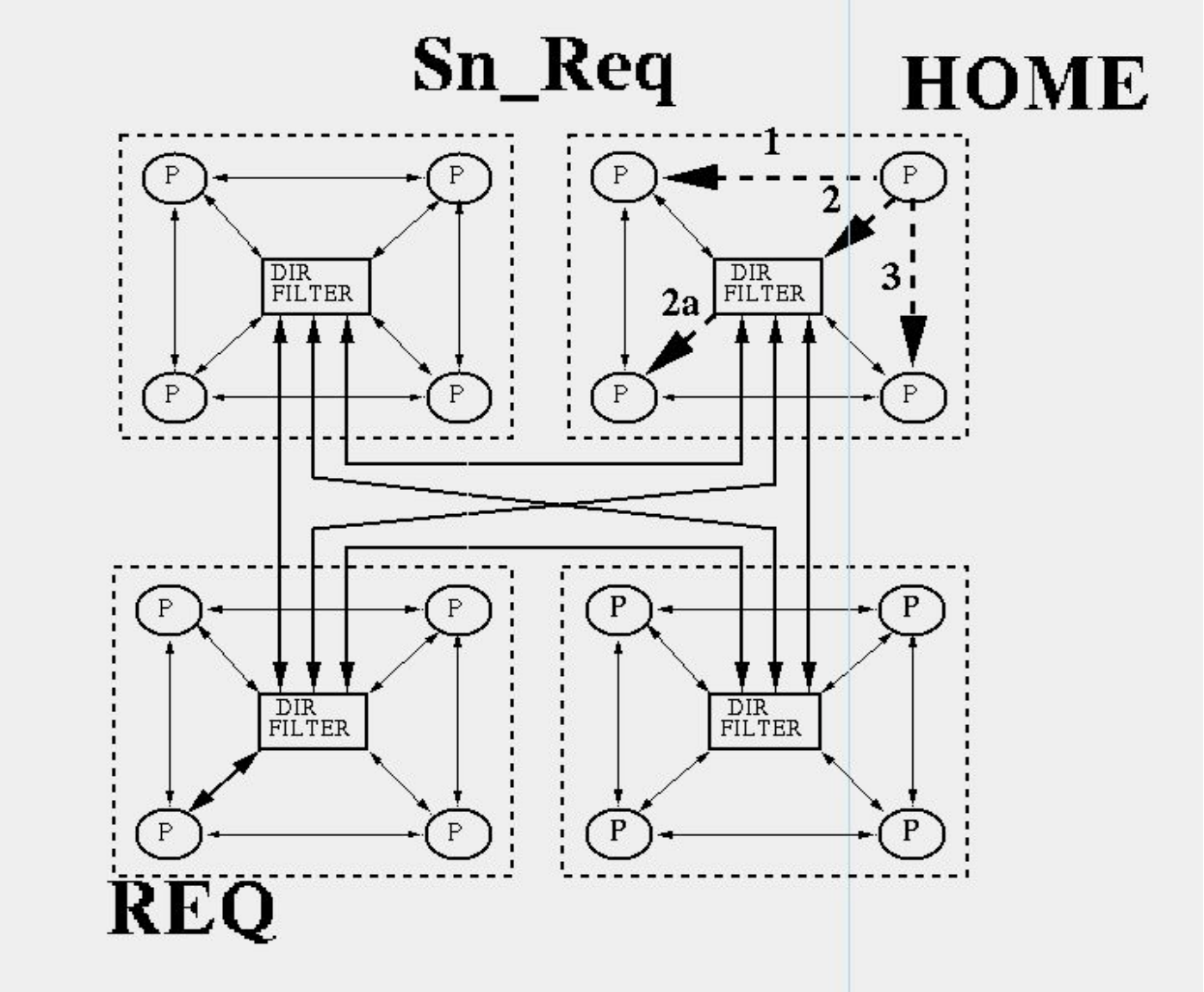
16-процессорная система



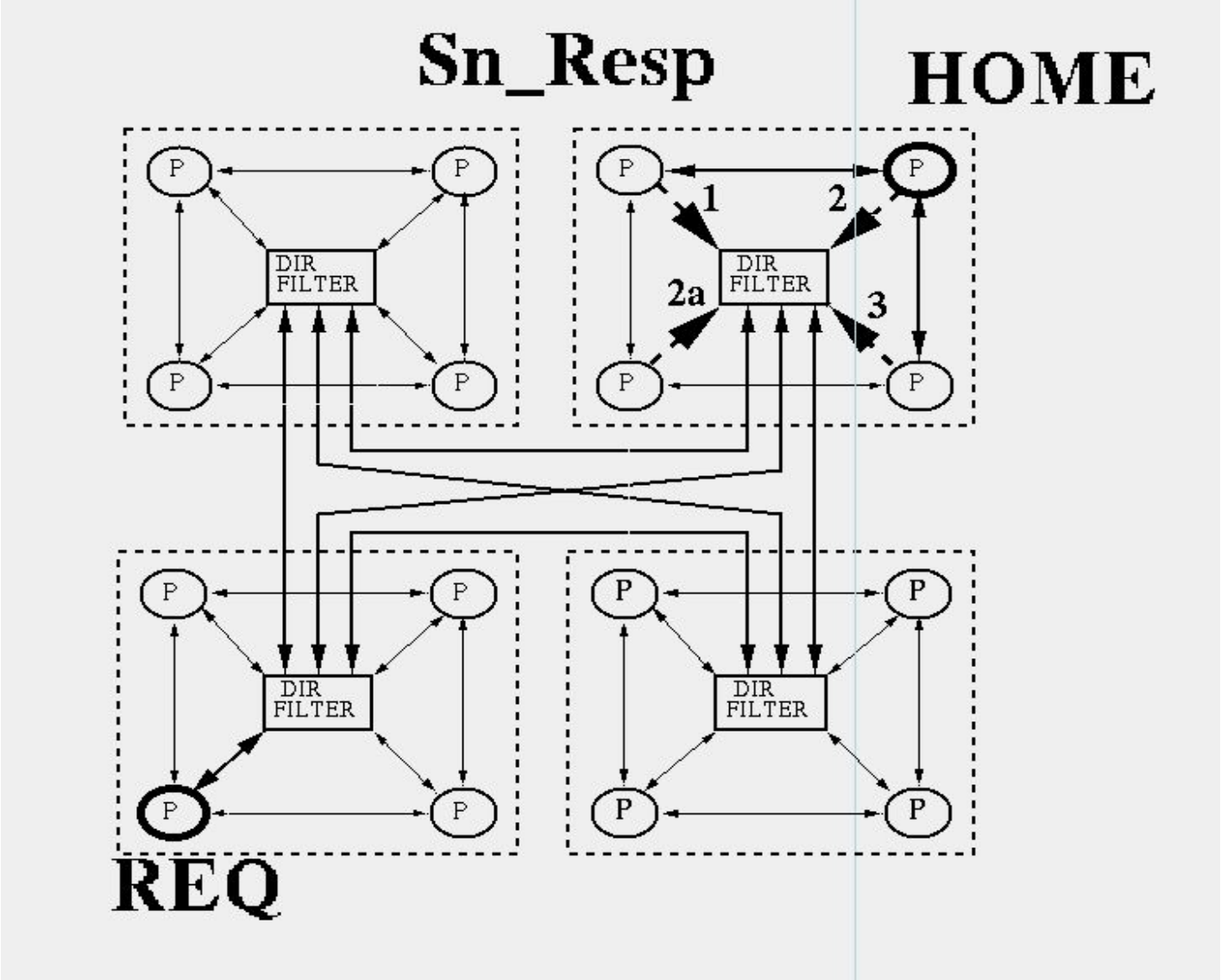
Read_64 операция в 16-ти процессорной системе(1 из 5)



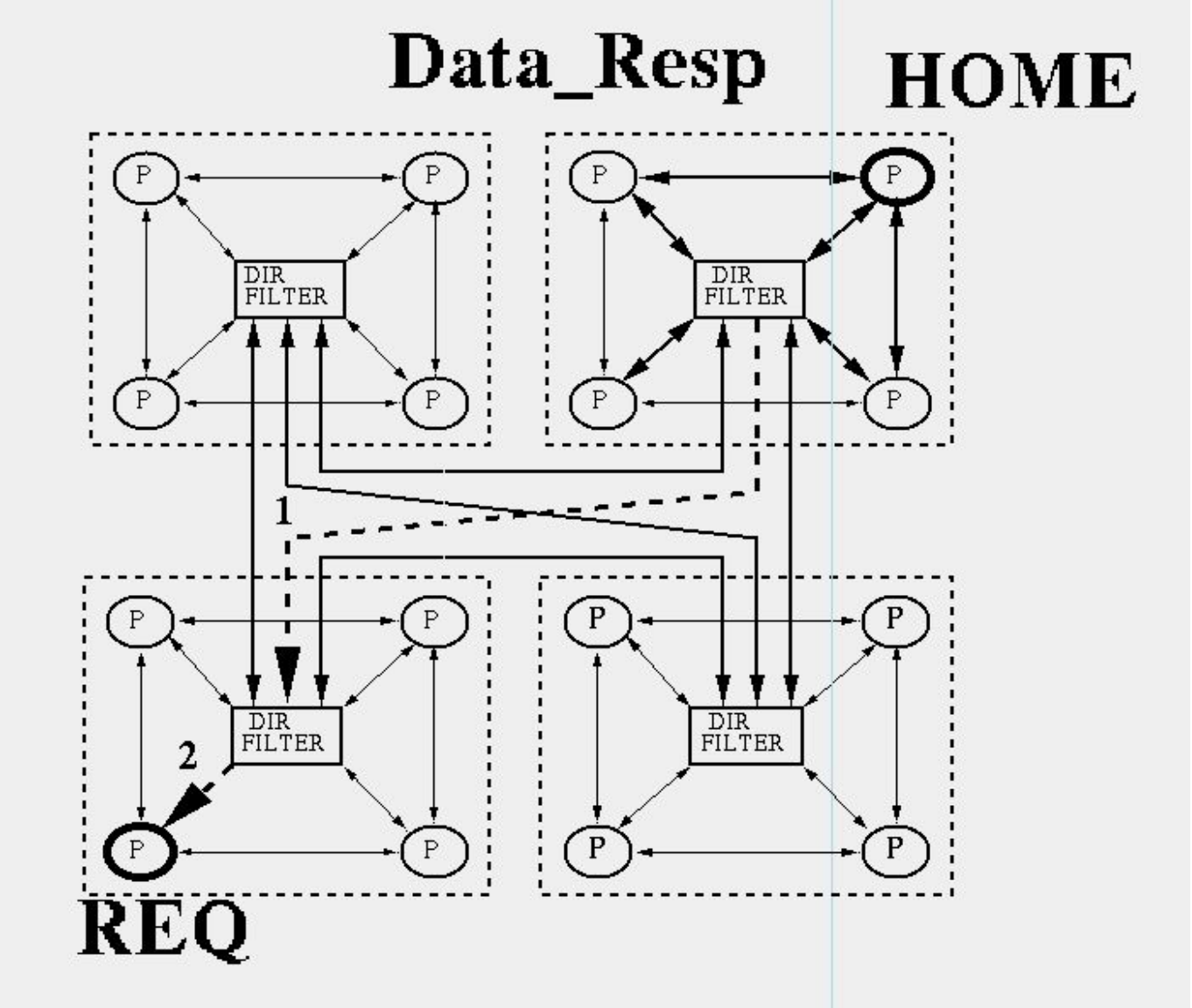
Read_64 операция в 16-ти процессорной системе(2 из 5)



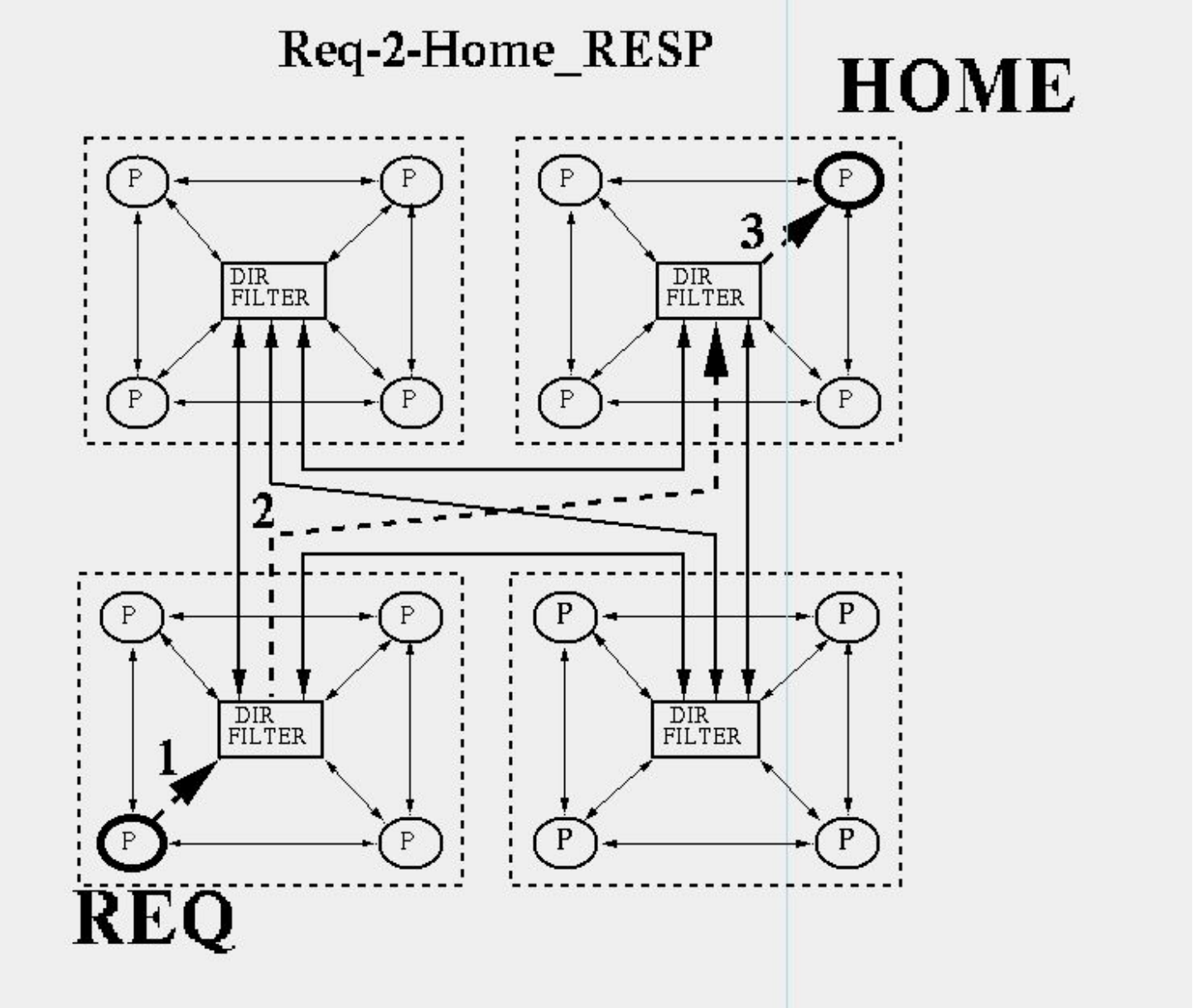
Read_64 операция в 16-ти процессорной системе(3 из 5)



Read_64 операция в 16-ти процессорной системе(4 из 5)



Read_64 операция в 16-ти процессорной системе(5 из 5)



Расширение аппаратной поддержки когерентности на 64-х процессорный вариант

- СПРАВОЧНИК
- Число строк в процессорах удаленных кластеров: [2M/64(байта в строке)] x 60 = 120/64 M(строк) = 1.875M(строк)
- Размер элемента справочника(с учетом увеличения бит-вектора совладельцев копии строки): ~6байт
- Размер справочника: 1.875M(строк) x 6(байта на строку) =11.25 Мбайт
- Разреженный в два раза справочник:1.875M/2(строк) x 8(байт на строку) = 7.5Мбайт(с учетом добавления еще одного бит-вектора совладельцев копии строки)
- Разреженный в 4 раза справочник:1.875M/4(строк) x 12(байт на строку) = 5.625Мбайт(с учетом добавления 4-х бит-векторов совладельцев копии строки)

Расширение аппаратной поддержки когерентности на 64-х процессорный вариант

Вариант справочника	Размер справочника (Мбайт)	Процент покрытия при размере справочника в 1.5Мбайта	Общий размер справочников в системе (х 16 чипов) в Мбайтах	Суммарный процент перекрытия
Полный справочник	11.25	13.3%	24	213%
Разреженный в 2 раза справочник	7.5	20%	24	320%
Разреженный в 4 раза справочник	5.625	26.7%	24	427%

Расширение аппаратной поддержки когерентности на 64-х процессорный вариант(прогноз для 32nm)

Вариант справочника	Размер справочника (Мбайт)	Процент покрытия при размере справочника в 2.0Мбайта (мой прогноз для 32nm процесса)	Общий размер справочников в системе (х 16 чипов) в Мбайтах	Суммарный процент перекрытия
Полный справочник	11.25	17.8%	32	284%
Разреженный в 2 раза справочник	7.5	26.7%	32	427%
Разреженный в 4 раза справочник	5.625	35.5%	32	569%

Расширение аппаратной поддержки когерентности на 64-х процессорный вариант

- Вариант с поддержкой справочника на внешней памяти, подключаемой к Чип-КК вопросы:
- - по числу дополнительных контактов FPGA-чипа;
- - по временным параметрам т.к. обращение к справочнику на внешней памяти увеличивает задержку в доступе к общей памяти;
- По коммутации 16-ти кластеров в единую систему т.к. требуются коммутационные элементы

Расширение аппаратной поддержки когерентности на 64-х процессорный вариант

- ВЫВОД: прогресс в области FPGA не позволяет рассчитывать на возможность реализации полного справочника такого объема на одном чипе FPGA
- Варианты разреженных справочников также не дают гарантированного решения, требуются дальнейшие проработки
- Варианты справочников на внешних элементах памяти требуют увеличения числа используемых контактов и приводят к увеличению времени доступа к памяти, что также требует дальнейшей проработки

Figure 1. HP sx2000 cell board architecture

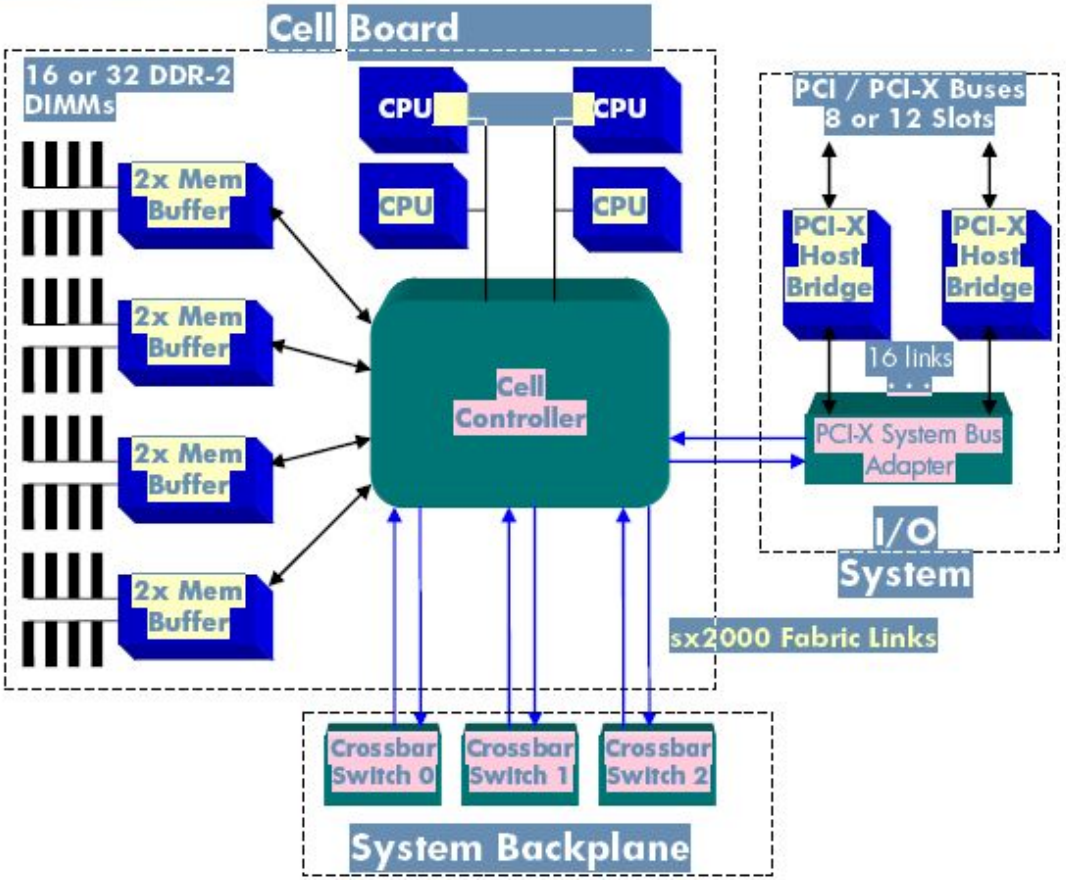


Figure 2. System topology

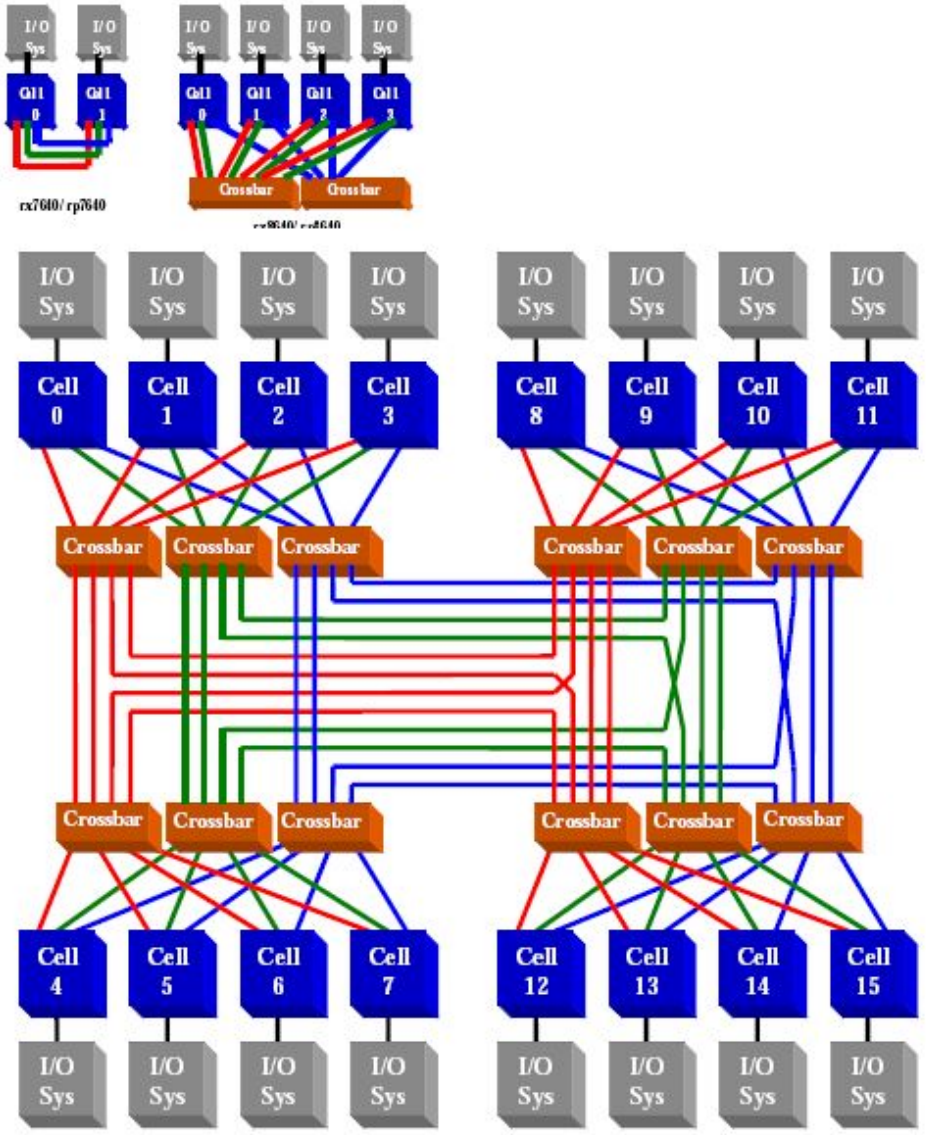
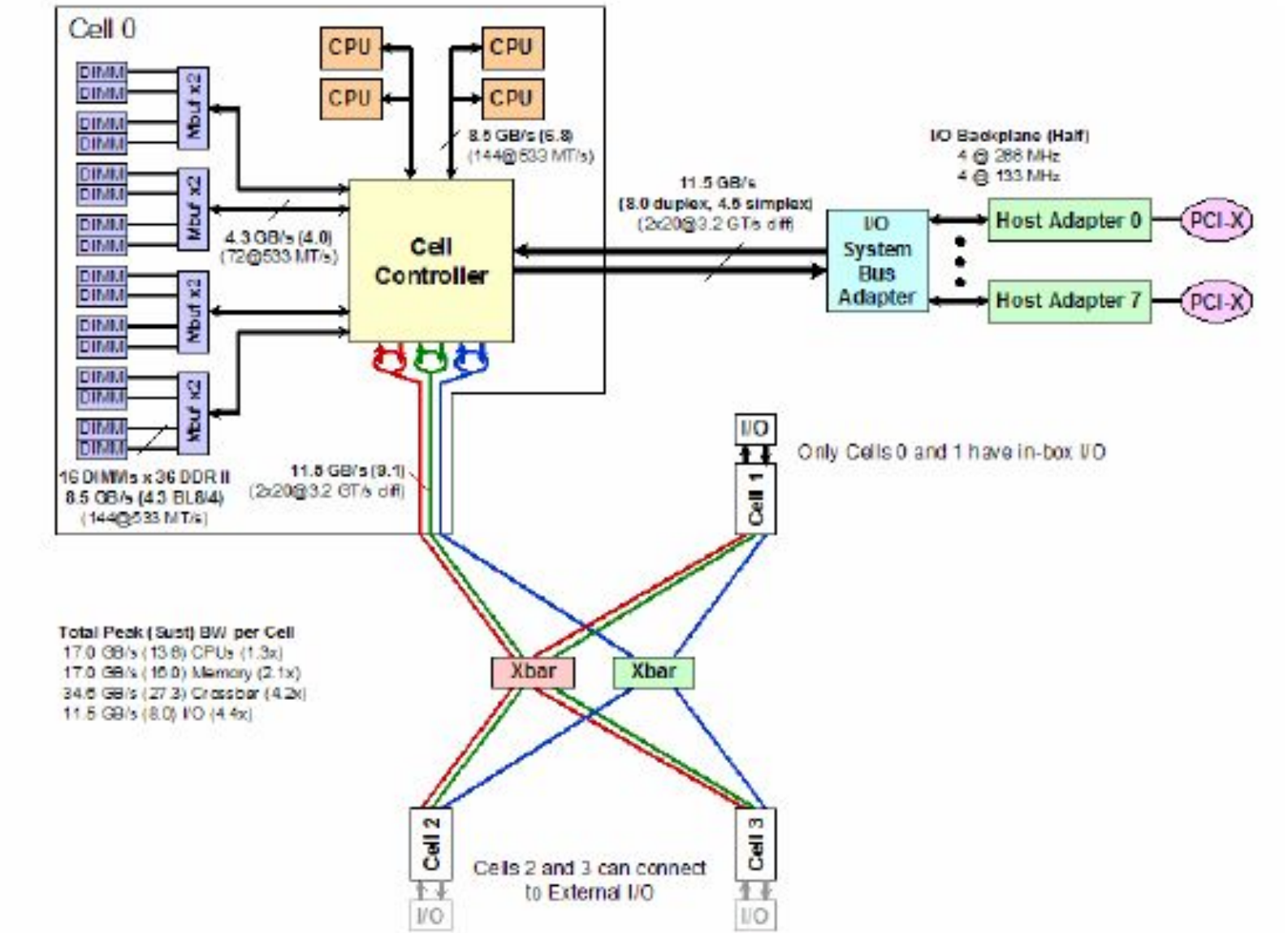


Figure 5. Architecture block diagram of the HP Integrity rx8640 Server, showing the modularity of the system



There are two types of memory latency within the HP Integrity rx8640 Server:

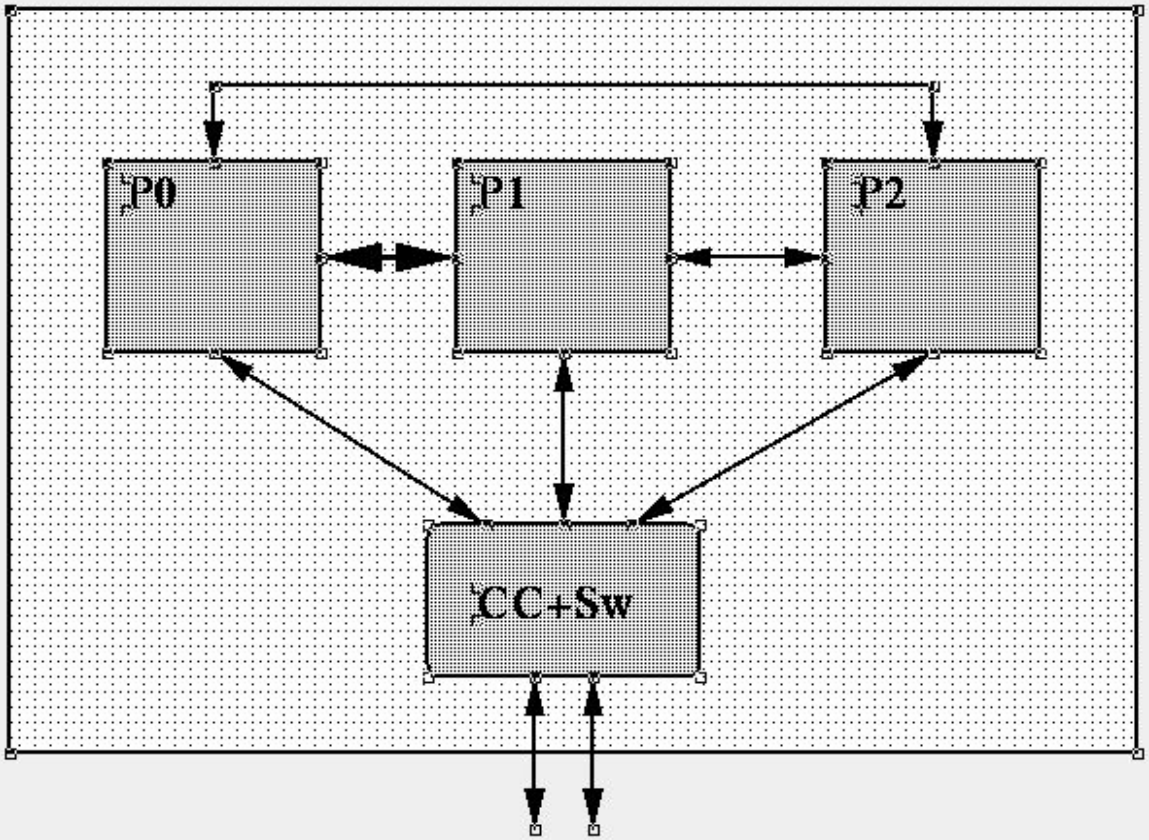
- Memory latency within the cell refers to the case in which an application runs on a partition that consists of a

Number of processors per partition	Average memory latency
Four processors (one cell)	~185 ns
Eight processors (two cells)	~249 ns
Sixteen processors (four cells)	~334 ns

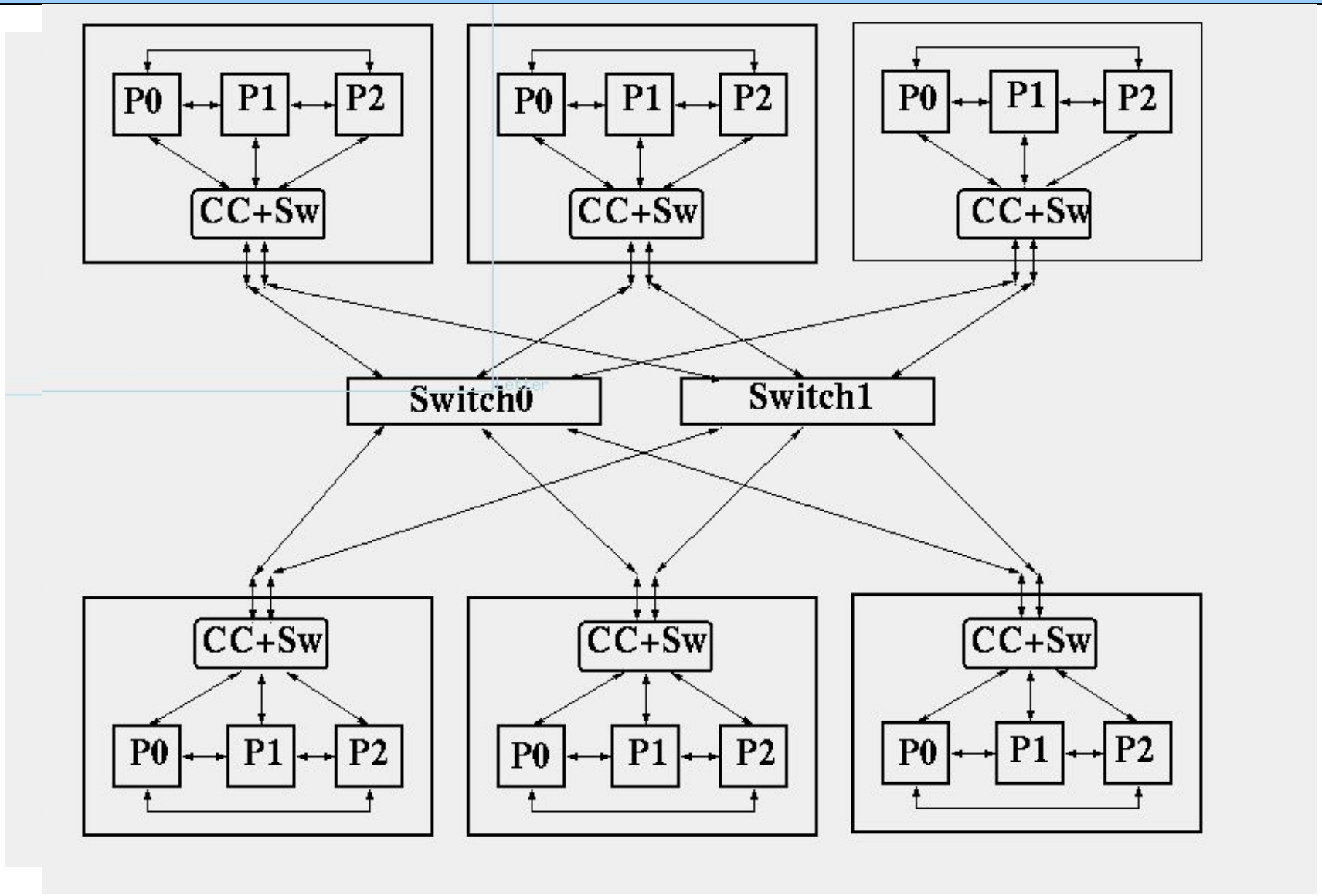
Открытые вопросы

- Работа с E состоянием
  (м.б. изменения в процессоре!?)
- Обеспечение надежности 16-ти процессорной системы – Чип-КК является чипом определяющим отказ сразу всего кластера!?
- Ввод-вывод, как основа реального времени, должен иметь возможности по сокращенному (некогерентному) доступу в память

3-х процессорный модуль



16-процессорная система с избыточностью (18 процессоров)



16-процессорная система с избыточностью (18 процессоров) с межсоединениями типа “Кольцо”