PCI-SCI мост с оптимиран потребителски интерфейс

Проф. Д-р. Волфганг Рем
ТУ Кемнитц, ФРГермания
e-Mail rehm@informatik.tu-chemnitz.de
Ас. Д-р. Станислав Симеонов
Бургаски Свободен университет
e-Mail stan@bfu.bg

Woflgang Rehm, Stanislav Simeonov. PCI-SCI Bridge with an Optimized User Interface

The possibilities of the physical layers of the new built networks increase all the time. This speed is not completely visible for the application level. With the availability of Distributed Shared Memory realized bye the SCI technology it is necessary latency time to be decreased to several microseconds. The goal of this article is to precise a concept for the PCI-SCI-bridge with the important features. Apart from all nice and new concepts there is one major guideline. It must be realizable using common FPGA/CPLD technique.

Достижимата пропускателна способност на физическите слоеве на новосъздадените мрежи все повече нараства. В тази връзка се поставя въпроса за по ефективното използване на компютърната техника и по специално на персоналните компютри. Основни причини за това са:

- Все по високата производителност на персоналните компютри;
- Все по ниската цена на хардуера;
- Голямото разнообрази от програмни системи, развойни среди и приложения във всички области.

Мотивация и въведение

Много от задачите решавани в електрониката, електротехниката и микроелектрониката са свързани с изграждането на сложни математически модели, съдържащи голям брой изчисления. Не рядко изчислителната мощност на съвременните работни станции е недостатъчна за ефективното решаване на проблемите. Приложението на многопроцесорни системи не винаги е достатъчно ефективно. От тази гледна точка скалирането на персонални компютри се явява едно възможно решение на възникналите проблеми.

Понятието скалируемост се дефинира по различен начин. Би могло да се каже, че скалируемост на дадена паралелна система това е способността за увеличаване на производителността на клъстърната компютърна система като функция на увеличаването броя на съставните и елементи, в частност персонални компютри за клъстър от PC. В последните години скалируемостта на паралелните системи придоби особено значение при разработката на съвременните архитектури. Увеличаването на броя на ресурсите, води до увеличаване на оверхеада в системата, породен от необходимостта за междупроцесорна комуникация и синхронизация.

Главната цел на разработката е да дефинира една концепция за моста PCI-SCI. Най – важните особенности които трябва да бъдат реализирани с този мост са:

- Поддръжка за Разпределената Съвместна Памет(Транспарентен способ);
- Механизми за предотвратяване на производствена авария когато множествени процеси на една SMP машина достигат (имат достъп) до отдалечена памет посредством транспарентния способ (поточни буфери,подобни на известните от моста PCI-SCI на Dolphins).
- И последно, но не най-маловажно, наличността на защитен на потребителско ниво директен достъп до паметта за предоставяне на максимална пропускателна способност.

Анализ на натоварването на централния процесор

При този анализ се определя до каква степен е влиянието на DMA-трансфера, работещ във фонов режим върху натоварването на процесора. Тестовете са изпълнени с използването на различни нива на включване на памет от тип кеш. Те са изпълнени с помощта на два двупроцесорни компютъра с процесори PII на 350 MHz, чипсет 440BX, с PCI-SCI мост от тип Dolphins D310, операционна система Linux. Същите са проведени в експерименталната лаборатория на катедра "Компютърни архитектури и микропрограмиране", към факултет по Информатика на ТУ Кемниц, ФРГермания.

Резултатите са обобщени в таблица 1. Тези резултати показват незначителното влияние на DMA върху работата на процесора. Малко по долу те ще бъдат интерпретирани.

Използване на кеш-памет	Без DMA	Със DMA	Процентно съотношение
Изключен	51,2 MB/sec	43,9 MB/sec	85,7%
Средно	136,4MB/sec	122,0 MB/sec	89,4%
Пълно	408,8MB/sec	402,3 MB/sec	98,4%

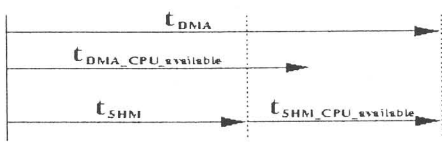
Таблица 1.

Горепоказаните резултати от тестовете дават възможност за следните разсъждения:

В най-лошият случай при DMA-трансфер, пропусквателната способност е с 15% по бавен от колкото при разпределената, съвместно използвана памет. По време на DMA-трансфера, централният процесор не е зает и може да се използва за изчислителна дейност в рамките на приложението. Следователно:

$$t_{DMA\_CPU\_available}=0,85.t_{DMA}$$

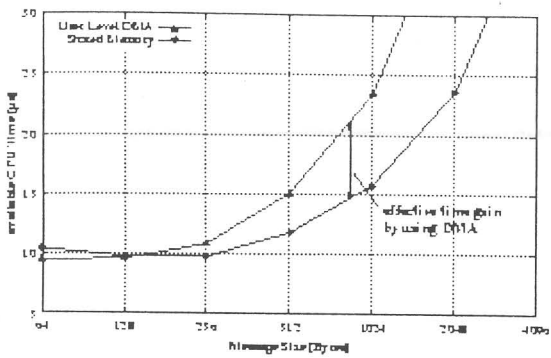
В този случай става дума за трансферирани данни с дължина повече от 64Byte. При трансфер на по малки по размери масиви, няма разлика в пропусквателната способност[1].



Фиг. 1. Сравнителна схема за заетост на централният процесор

На фиг. 1 е показано принципното съотношение на времената за трансфер. За да се осъществи трансфер на данни с помощта на разпределена съвместно използвана памет е необходимо по малко време, но за сметка на това централният процесор е зает. Заетостта на процесора се предизвиква от начина по който централният процесор (CPU) изпълнява копирен цикъл за запис на блока данни в паметта на един отдалечен възел. Всяка дума трябва да бъде прочетена от CPU преди тя да бъде записана в отдалечената памет. Това е неефикасно и възпрепятствува CPU в изпълнението на някоя по належаща работа. Следователно:

$$t_{SHM\_available}=t_{DMA}-t_{SHM}$$



Фиг. 2 Налично време за централният процесор при трансфер

На фиг. 2, тези прости формули са представени графично. При по големи дължини на трансферираните данни се забелязва все по голяма загуба на процесорно време.

Нека все пак се има в предвид, че тези резултати са повече ориентировъчни. За пълен анализ на работата на DMA е необходима по прецизно избрана стъпка на измерванията, което не винаги е тривиална задача.

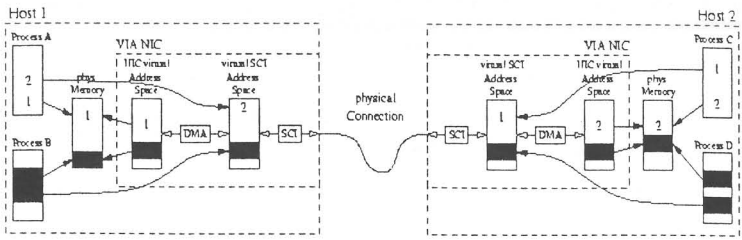
Нова архитектурна концепция за PCI-SCI мост

В съществуващите в момента PCI-SCI мостове се реализира статично управление на паметта. Те предлагат 1:1 картографиране на прозорец от паметта за цялата експортирана памет. Такова решение е свързано доста неудобства.

Спецификацията на Виртуалната интерфейсна архитектура премахва традиционните протоколни стекове протоколни стекове между приложенията и хардуера, които по принцип допринасят за забавяния на комуникациите.

VIA реализира механизъм за изпълнение на DMA – дескриптори от потребителско ниво, като същевременно осигурява защита между множествени процеси, използващи един и същ VIA – хардуер. Когато множествени процеси искат достъп до хардуера директно, тази заявка се обработва от мрежовия интерфейсен контролер (NIC). На всеки потребителски процес се предоставя съответен виртуален интерфейс (VI). Всеки VI притежава собствен контекст. NIC пази контекста за всеки обезпечен виртуален интерфейс. NIC обезпечава даден брой VI

контексти. Всеки един контекст принадлежи към една крайна точка на VI.



Фиг. 3. Функционална схема на модифициран PCI-SCI мост.

Въз основа на направените по горе разглеждания се налага идеята да се комбинират двете единици (SCI със Съвместно Използваемата Памет от една страна и Виртуалната Интерфейсна Архитектура със своята защитена на потребителско ниво DMA от друга страна) в един комуникационен модул.

Фиг. 3 показва принципната схема на реализация на моста. За да се запази яснотата на фигурата, не са показани всички картографиращи стрелки. В допълнение, процеси са способни да картографират отдалечена памет в собственото си пространство и могат да достигнат тази памет използувайки прости трансакции. Това не е възможно с един конвенционален VIA-NIC. Но за трансфери, посредством отдалечен директен достъп до паметта (RDMA) картографирането на внесена SCI памет в процесното виртуално пространство не се изисква.

Главната разлика за модифицирани RDMA трансфери е, че данните се трансферират между NIC виртуалното адресно пространство (или изнесената локална памет) и виртуалното SCI адресно пространство (или внесената SCI памет). Проверка за право на достъп за дистанционни адреси не се прави в отдалеченият възел, а на локалния възел по време на адресната транслация по посока на трансфера(по посока на импортираната памет).

Отнасяйки се към фиг. 3, процес A е регистрирал части от своята памет във VIA NIC и защитния tag е определен за тази зона. Внесената SCI област памет (от процес C) приема същия защитен tag X. Когато процес A инициира един RDMA трансфер, този трансфер може да се извърши само между (1) и (2) адресни области. Всички други области са

с други защитни tag и опити на процес А да ги достигне ще пропаднат с една защитна грешка. Но какво се случва в дистанционния край ? Дистанционния възел просто получава записни или прочитни пакети достигащи неговата локална памет. На отдалеченият VIA NIC не е необходимо да определя VI контекст за идващите трансакции. Обаче, едно одобрение на право за достъп може да бъде една проста проверка на прочит/запис разрешаващ бит. Отсъствието на проверка на защитния tag в дистанционния край изглежда че повдига един потенциален адресно защитен проблем. Когато даден възел получен пакет, който е замислен за прочит или запис на данни извън локалната памет, как може NIC да бъде осигурен че този достъп е правилен? Отговорът изглежда фатален: Той не може да бъде сигурен. NIC трябва да разчита на проверката на правото на достъп при източника на тази операция. И ако в източника нещо е вървяло не както трябва, тогава приемника не може да направи нищо срещу това. Типично, такива неизправности ще се случват поради дефекти в ядрения софтуер контролиращ VIA хардуера. Но това не е проблем на този модифициран RDMA модел. Също в случай на конвенционалния VIA RDMA модел такива проблеми могат да се случат поради неправилни настройки – например на идентификатора на VI назначението. Въпреки, че вероятността да се открият неправилни адреси изглежда по-голяма в този случай, тъй като даденият дистанционен NIC виртуален адрес трябва случайно да съответствува на защитния край на неправилния VI контекст така, че неправилният достъп не се регистрира. Накрая, това "базирано на разчитане" приемане на пакети е един общ проблем на разпределените системи. Но ако е наличен един добре работещ софтуер за ядрото, модифицираният RDMA модел предлага достатъчно добро защитно ниво.

Заключение

За да се избере добра отправна точка за проектиране и създаване на PCI-SCI-мост е необходим задълбочен функционален анализ. За отбелязване е, че в момента съществуващите мостове показват по ниска пропусквателна способност посредством отдалечен директен достъп до паметта в сравнение с разпределената съвместно използвана памет. Тук е мястото да се отбележи, че независимо от това, производителността на системата като цяло се повишава, поради осво-

бождаване на централния процесор от комуникационни задачи. Това е и основният довод за проектиране на мост и осъществяване на машина за отдалечен директен достъп от потребителско ниво.

Използвана литература

1. **С. СИМЕОНОВ. Характеристика на паралелни системи, на базата на персонални компютри и адаптери за тях, под печат**
2. IEEE Standard for Scalable Coherent Interface (SCI). IEEE Std 1596-1992, IEEE Computer Sociate, 1993
3. M. Blumrich, C. Dubnichki, E.W. Felten, K. Li, and M.R. Mesarina. Two virtual memory mapped network interface design. In Proceeding of Hot Interconnects II Symposium, page 134-142, August 1994.
4. M. Blumrich, C. Dubnichki, E.W. Felten, K. Li. Protected, User-Level DMA for the SHRIMP Network Interface. Dept. of Computer Science. Princenton University, 1996.
5. M. Welsh, A. Basu, Th. Von Eicken. Incorporating Memory Management into User-Level Network Interface. Dept. of Computer Science. Cornell University, 1997.
6. Intel, Compac and Microsoft. Virtual Interface Architecture Specification V1.0
7. W. Rehm, Parallelrechner und Parallelprogrammierung, Komplexband '94. TU-Chemnitz.
8. M. Trams. Design einer systemfreundlichen PCI-SCI Bridge mit optimierten User-Interface, Diplomarbeit, TU-Chemnitz, 1998.