



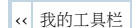
搜索

CUDA的合并存储器访问(coalesced Memory Access)与half- ...

下一页

1# 电梯直达 ☐ 





管理员

注册时间 2007-7-11

积分 32858

串个门 打招呼

加好友 发消息

MPI



骑都尉 (从五品)

注册时间 2010-3-29

积分 205

串个门 打招呼

加好友 发消息

```
struct float4 {
    float a, b, c, d;
    ...
}
```

cuda2010 发表于 2010-5-2 13:10

对的，在术语中我们称之这为所谓的AOS (Array of Sturcture) 改成SOA (Structure of Array) 。AOS是适合标量处理器进行处理的一个数据结构，在GPU这种向量处理器中情况会有所变化，其基本的最底层的数据结构一定是数组。每个Warp处理数组的一个切片 (Slice) 。

这不是NV在CUDA中首次使用的，在传统的基于OGL的GPGPU时代开始，就是这样了。我们通过给PS绑定不同的纹理（即数据）来构成输入数据的SOA。

我相信只要Warp机制还在，那么以后在GPU上无论搭建什么更NB的数据结构，最底层的数据结构都是数组。

使用道具 举报

发表于 2010-5-2 15:03:23 | 只看该作者 4#

本帖最后由 MPI 于 2010-5-2 15:04 编辑

这是一个常见问题，这种情况一般用分量数组效果比较好，也就是把

```
struct float4 {
    float a, b, c, d;
    ...
}
```

cuda2010 发表于 2010-5-2 13:10

多谢大大，我感觉CUDA计算我以前做的OpenMP有点类似。顺序运算都在CPU上，然后要并行了就fork到gpu上，不过比OpenMP复杂，这里还要分配GPU存储器空间。

另外，还有一些不明白的地方。如下：

```
For best performance, global memory accesses should be coalesced:

A memory access coordinated within a warp

A contiguous, aligned, region of global memory
    • 128 bytes -- each thread reads a float or int
    • 256 bytes -- each thread reads a float2 or int2
    • 512 bytes -- each thread reads a float4 or int4
    • float3s are not aligned!

Warp base address (WBA) must be a multiple of 16*sizeof(type)

The kth thread should access the element at WBA + k

Not all threads need to participate

These restrictions apply to both reading and writing
```

<< 我的工具栏

- Use shared memory to achieve this

上面说了float2、float4也可以coalescing，该怎么理解呢？

使用道具

举报

5#

cuda2010

发表于 2010-5-2 16:47:02 | 只看该作者

本帖最后由 cuda2010 于 2010-5-2 17:10 编辑

按编程手册上的描述，float2和float4也可以coalescing，不过也要half-warp全部按顺序访问，“一个访存请求只包含4个Thread”似乎是不行的。此外，这些非4字节操作时的coalescing是否是真的我觉得还得打个问号，特别是16字节word的情况(如float4)，因为手册上提到了此时即使对齐带宽也将显著降低，和不对齐差不了多少。所以我觉得如果能做到4字节word对齐访问那是最好的。

Coalesced 8-byte accesses deliver a little lower bandwidth than coalesced 4-byte accesses and coalesced 16-byte accesses deliver a noticeably lower bandwidth than coalesced 4-byte accesses. But, while bandwidth for non-coalesced accesses is around an order of magnitude lower than for coalesced accesses when these accesses are 4-byte, it is only around four times lower when they are 8-byte and around two times when they are 16-byte. (cuda编程手册v2.3.1, p90)

1



查看全部评分 daming

使用道具

举报

6#

风辰

发表于 2010-5-2 20:50:01 | 只看该作者

本帖最后由 风辰 于 2010-5-2 20:56 编辑

在c u d a中，上面的两位是正确的，但是在s t r e a m中，结果反了过来，所以如果用的是o p e n C L，就要分别对n v的和a m d的分别优化了。

另，如果是完美对齐的话，就是满足1 . 0和1 . 1要求的对齐，性能好像刚好相反，因为此时8字节对齐可以一次读1 2 8个字节，相比每次4字节读的次数减少了一半

另，在c u d a中线程的最小单位是3 2个，所以你要控制4个是办不到的，呵呵！

1



查看全部评分 daming

使用道具

举报

7#

MPI

发表于 2010-5-3 00:47:09 | 只看该作者

本帖最后由 MPI 于 2010-5-3 00:49 编辑

多谢楼上个人大哥。我听人说CUDA程序的加速的第一步就是Coalescing Global Memory 访问以及是合理利用

<< 我的工具栏



上骑都尉 (正五品)  
注册时间 2010-4-10  
积分 834  
串个门 加好友  
打招呼 发消息



版主 | 兼技术专家  
注册时间 2010-3-8  
积分 1797  
串个门 加好友  
打招呼 发消息



骑都尉（从五品）

注册时间2010-3-29

积分205

串个门

加好友

打招呼

发消息

Shared Memory。不知道这个理解是否正确？

我这里有一份关于coalescing访问的资料，贡献出来。

1



查看全部评分


daming



使用道具

举报

MPI

 发表于 2010-5-3 01:27:17 | 只看该作者

8#



骑都尉（从五品）

注册时间2010-3-29

积分205

串个门

加好友

打招呼

发消息

本帖最后由 MPI 于 2010-5-3 01:43 编辑

在c u d a中，上面的两位是正确的，但是在s t r e a m中，结果反了过来，所以如果用的是o p e n C L，就 ...

风辰 发表于 2010-5-2 20:50


知道这里高人多，大牛说的我不是特别能理解，大牛能详细说说么？

这个图说明了什么呢？ 在1.2设备上，GPU底层硬件启动一次Global Memory的最小粒度是128B ？ 所以，我们最好使用float2来发起GMEM访问？

使用道具

举报

17331014

 发表于 2010-5-3 08:27:39 | 只看该作者

9#



超级版主

注册时间2010-3-7

积分2903

串个门

加好友

打招呼

发消息

float类型.32\*4=128B.

老程序员

使用道具

举报

风辰

 发表于 2010-5-3 09:09:56 | 只看该作者

10#

回复 [8#](#) MPI

因为A M D的相比N V的在处理矢量方面要快

<<

我的工具栏



版主 | 兼技术专家

注册时间 2010-3-8

积分 1797

串个门 加好友

打招呼 发消息

cuda2010



上骑都尉 (正五品)

注册时间 2010-4-10

积分 834

串个门 加好友

打招呼 发消息

ic.expert



管理员

而每次读128B要完美对齐才能读到float2

我翻译的CUDA3.0中文版正式版<http://focus.it168.com/201003/cuda3/index.html>

发表于 2010-5-3 10:01:22 | 只看该作者

4#资料中提到的"Warp base address (WBA) must be a multiple of 16\*sizeof(type)"我觉得不大准确。例如对int/float型来说16\*4=64字节是不够的，我认为这个起始地址对齐与类型无关，不论哪种类型，都应该是128(或256)字节对齐。

For best performance, global memory accesses should be coalesced:

A memory access coordinated within a warp

A contiguous, aligned, region of global memory

- 128 bytes -- each thread reads a float or int
- 256 bytes -- each thread reads a float2 or int2
- 512 bytes -- each thread reads a float4 or int4
- float3s are not aligned!

Warp base address (WBA) must be a multiple of 16\*sizeof(type)

The kth thread should access the element at WBA + k

Not all threads need to participate

These restrictions apply to both reading and writing

- Use shared memory to achieve this

MPI 发表于 2010-5-2 15:03

发表于 2010-5-4 03:01:20 | 只看该作者

4#资料中提到的"Warp base address (WBA) must be a multiple of 16\*sizeof(type)"

我觉得不大准确。例如对i ...

cuda2010 发表于 2010-5-3 10:01

下面是手册上的图，对齐的唯一意义就是不让一个memory Traffic变成两个。主要是提高memory Access

我的工具栏

<div>注册时间 2007-7-11</div> <div>积分 32858</div> <div>串个门</div> <div>加好友</div> <div>打招呼</div> <div>发消息</div>	<div>一种办法。在CPU上我们一般不大重视又一个原因是因为X86指令集本身CISC特性引起的。但是在GPU上，对齐是提高带宽使用效率的重要手段。</div> <div>这样说吧，在Fermi GPU硬件上，Memory Controller是按照1KB进行低位交叉编码来组建存储空间的，也就是说每增长1KB，存储器访问所使用的Memory Controller就会跳到下一个。这样，即使是访问尺寸很少的存储区域，也能完全开发6个memory Controller对应的存储带宽，而不是64bit对应的存储带宽。</div> <div>使用道具 举报</div> <div>13#</div>
<div>MPI</div> <div></div> <div>骑都尉（从五品）</div> <div>注册时间 2010-3-29</div> <div>积分 205</div> <div>串个门</div> <div>加好友</div> <div>打招呼</div> <div>发消息</div>	<div>发表于 2010-5-4 06:37:44   只看该作者</div> <div>float类型,32*4=128B.</div> <div>17331014 发表于 2010-5-3 08:27</div> <div>这里引出了我的另一个疑问，为什么很多资料都是使用half-warp，即，16个Thread来解释粗粒度访问呢？half-warp这个概念有什么用处？？？</div> <div>所以我理解应该是float2类型：16*8 BYTE =128B</div> <div>不知道这样理解是否正确？</div> <div>使用道具 举报</div> <div>14#</div>
<div>cuda2010</div> <div></div> <div>上骑都尉（正五品）</div> <div>注册时间 2010-4-10</div> <div>积分 834</div> <div>串个门</div> <div>加好友</div> <div>打招呼</div> <div>发消息</div>	<div>发表于 2010-5-4 10:30:35   只看该作者</div> <div>我觉得12#手册Fig.g-1这个图会给人误导，这个图给人的印象是对于2.0以下硬件64字节对齐就可以了，实际上至少需要128字节，最好256字节。更高级别的对齐会带来额外的性能提升。</div> <div>另外其中第3个图里面1.2 and 1.3下面的1x128B at 128不知道是怎么回事，我觉得是写错了，应该是1x64B at 128。</div> <div>下面是手册上的图，对齐的唯一意义就是不让一个memory Traffic变成两个。主要是提高memory Access效率的一种办法。在CPU上我们一般不大重视又一个原因是因为X86指令集本身CISC特性引起的。但是在GPU上，对齐是提高带宽使用效率的重要手段。</div> <div>这样说吧，在Fermi GPU硬件上，Memory Controller是按照1KB进行低位交叉编码来组建存储空间的，也就是说每增长1KB，存储器访问所使用的Memory Controller就会跳到下一个。这样，即使是访问尺寸很少的存储区域，也能完全开发6个memory Controller对应的存储带宽，而不是64bit对应的存储带宽。</div> <div>ic.expert 发表于 2010-5-4 03:01</div> <div>使用道具 举报</div> <div>15#</div>
<div>cuda2010</div> <div></div> <div>上骑都尉（正五品）</div> <div>注册时间 2010-4-10</div>	<div>发表于 2010-5-4 11:07:00   只看该作者</div> <div>关于float2和float4类型，即使满足了对齐条件，还会遇到smem的bank conflict问题。这个也会严重影响性能，而且和内存对齐不同，这个问题就float2和float4来说似乎是无解的。</div> <div>此外，现在觉得我在5#引用手册中提到的"it is only around four times lower when they are 8-byte and around two times when they are 16-byte"也很可能有问题。测试了一下SDK中的reduction，当type=double时，带宽的确下降为只有type=float时的一半左右。但这个不一定全部是对齐问题引起的，在其他一些测试中double带宽下降没有这么明显。</div> <div>我的工具栏</div>

<div>积分834</div> <div><div>串个门</div><div>加好友</div><div>打招呼</div><div>发消息</div></div>	
17331014	<div><div>发表于 2010-5-4 16:18:35   只看该作者</div></div>
<div></div> <div>超级版主</div> <div><div>注册时间2010-3-7</div><div>积分2903</div></div> <div><div>串个门</div><div>加好友</div><div>打招呼</div><div>发消息</div></div>	<div><div>使用道具</div><div>举报</div></div> <div>16#</div> <div><div>这里引出了我的另一个疑问，为什么很多资料都是使用half-warp，即，16个Thread来解释粗粒度访问呢？ ha ...</div><div>MPI 发表于 2010-5-4 06:37</div></div> <div><div>个人认为half-warp的意义只在smem访问中. 对gmem访问,应该还是以warp单位对齐的....</div><div>老程序员</div></div>
ic.expert	<div><div>发表于 2010-5-5 12:10:37   只看该作者</div></div>
<div></div> <div>管理员</div> <div><div>注册时间2007-7-11</div><div>积分32858</div></div> <div><div>串个门</div><div>加好友</div><div>打招呼</div><div>发消息</div></div>	<div><div>使用道具</div><div>举报</div></div> <div>17#</div> <div><div>大牛可以测试一下，看看单纯的halfwarp对齐的访问，当一个warp的两个halfwarp都访问都不同的地址的时候，会不会导致性能下降：&gt;</div></div>
cuda2010	<div><div>发表于 2010-5-5 17:42:40   只看该作者</div></div>
<div></div> <div>上骑都尉（正五品）</div> <div><div>注册时间2010-4-10</div><div>积分834</div></div> <div><div>串个门</div><div>加好友</div><div>打招呼</div><div>发消息</div></div>	<div><div>使用道具</div><div>举报</div></div> <div>18#</div> <div><div>呵呵，17331014网友现在也不相信手册了？目前的手册上仍然明确说明gmem读写的基本对齐单位是half-warp(如 cuda 3.0 guide, g.3.2.1, g.3.2.2)。不过我也觉得还是有必要实测一下。</div></div>
cuda2010	<div><div>发表于 2010-5-5 18:15:48   只看该作者</div></div>
	<div><div>19#</div><div>&lt;&lt; 我的工具栏</div></div>



上骑都尉（正五品）

注册时间2010-4-10

积分834

串个门加好友

打招呼发消息

17331014



超级版主

注册时间2010-3-7

积分2903

串个门加好友

打招呼发消息

17331014



超级版主

注册时间2010-3-7

积分2903

串个门加好友

打招呼发消息

ic.expert

最近尝试对<http://www.opengpu.org/viewthread.php?tid=2580&extra=page%3D2>（随机gmem访问）中的gpu程序做一些改进，但都没什么效果，实测带宽2.3GB/s只有理论带宽峰值的大约1/48。而如果half-warp对齐机制正确那么此时最大带宽应该能达到理论带宽峰值的大约1/16。大胆猜测一下，难道实际的内存对齐机制是以1.5个warp(3个half-warp)为单位对齐的？呵呵。

发表于 2010-5-5 19:14:42 | 只看该作者

使用道具 举报

20#

呵呵，17331014网友现在也不相信手册了？目前的手册上仍然明确说明gmem读写的基本对齐单位是half-warp(如cu ...  
cuda2010 发表于 2010-5-5 17:42

半信半疑吧.主意是N太会夸张了点.呵呵.

老程序员

使用道具 举报

21#

发表于 2010-5-5 19:20:04 | 只看该作者

最近尝试对（随机gmem访问）中的gpu程序做一些改进，但都没什么效果，实测带宽2.3GB/s只有理论带宽峰值的大 ...  
cuda2010 发表于 2010-5-5 18:15

对齐其实应该和mc有关....但分析太复杂了.  
换个思路:  
如果half-warp对齐是满性能的话,那么warp对齐更没问题了.  
而有时(或者说经常)可以看到warp对齐的性能比half-warp高点(换句话说光half-warp对齐不够)...  
因此,个人感觉还是认为gmem访问按warp对齐更好些.呵呵.

老程序员

使用道具 举报

22#

发表于 2010-5-5 19:48:07 | 只看该作者

对齐其实应该和mc有关....但分析太复杂了.  
换个思路:  
如果half-warp对齐是满性能的话,那么warp对齐更没 ...

我的工具栏





管理员

注册时间 2007-7-11

积分 32858

[串个门](#) [加好友](#)

[打招呼](#) [发消息](#)

ic.expert



管理员

注册时间 2007-7-11

积分 32858

[串个门](#) [加好友](#)

[打招呼](#) [发消息](#)

17331014



超级版主

注册时间 2010-3-7

积分 2903

[串个门](#) [加好友](#)

[打招呼](#) [发消息](#)

17331014

17331014 发表于 2010-5-5 19:20

恩，我再帮大牛计算一下。

GDDR3允许的最大Burst长度是4此连续的访问。

GT200的GDDR3接口是512bit宽度的，总计8个MC，那么每个MC的宽度是64bit。

由于GDDR3也是上升、下降边界同时传送数据了，也就是一个clock（一次的访问导致）传输两笔数据。

所以每个单独的MC启动一次，最佳的访存粒度就是4 burst \* 2 data/clock \* 64bit/mc = 16\*32bit =halfwarp宽度

:>

但是Fermi 我不肯定，因为Fermi的L2 Cache Line大小等于1KB =32\*32bit=warp宽度。所以不确定halfwarp会不会导致Fermi L2 Cache的性能下降。

使用道具 举报

发表于 2010-5-5 19:51:29 | 只看该作者

23<sup>#</sup>

呵呵，17331014网友现在也不相信手册了？目前的手册上仍然明确说明gmem读写的基本对齐单位是half-warp(如cu ...

cuda2010 发表于 2010-5-5 17:42

恩，以后NV手册上的数据都得实际测试一下，呵呵。NV这家公司太不像话了，到处学术/商业造假，哈哈{:4\_176:}

使用道具 举报

发表于 2010-5-5 19:52:48 | 只看该作者

24<sup>#</sup>

本帖最后由 17331014 于 2010-5-5 19:55 编辑

GT200的GDDR3接口是512bit宽度的，总计8个MC，那么每个MC的宽度是64bit。

ic.expert 发表于 2010-5-5 19:48

这句前提有问题，被阉割的产品太多了。不能说是512bit宽度。。。只能说最大是512bit宽度。当然，也可能阉割的是mc,每个mc的宽度仍是64bit.但谁知道哪？呵呵。

老程序员

使用道具 举报

发表于 2010-5-5 19:59:14 | 只看该作者

25<sup>#</sup>

恩，以后NV手册上的数据都得实际测试一下，呵呵。NV这家公司太不像话了，到处学术/商业造假，哈哈 ...

我的工具栏



超级版主

注册时间2010-3-7

积分2903

串个门

加好友

打招呼

发消息

ic.expert



管理员

注册时间2007-7-11

积分32858

串个门

加好友

打招呼

发消息

17331014



超级版主

注册时间2010-3-7

积分2903

串个门

加好友

打招呼

发消息

cuda2010

ic.expert 发表于 2010-5-5 19:51

因此，我等平民百姓只能姑且听之，凑合用之。多留点余地，以免自己挖坑自己埋。哈哈。{:4\_203:}

老程序员

使用道具 举报

发表于 2010-5-5 20:06:57 | 只看该作者

26#

这句前提有问题，被阉割的产品太多了。不能说是512bit宽度。。。只能说最大是512bit宽度。  
当然，也可...

17331014 发表于 2010-5-5 19:52

的确，每个MC仍然是64 bit 。并且从NV50时代开始，MC的宽度就没有变过~

别的不能肯定，这个我能肯定：> 谁有精力的话可以找来每块卡的参数来看看 ~~

使用道具 举报

发表于 2010-5-5 20:20:13 | 只看该作者

27#

的确，每个MC仍然是64 bit 。并且从NV50时代开始，MC的宽度就没有变过~

别的不能肯定，这个我能肯定：...

ic.expert 发表于 2010-5-5 20:06

老程序员

使用道具 举报

发表于 2010-5-5 20:27:18 | 只看该作者

28#

刚才我做了一个这样的测试，以half-warp为单位对齐，但是结果是对带宽没有影响。

具体做法是启动足够多的block，每个block 64个threads(4个half-warp)。然后让每个block第1个和第4个half-warp访问的内容交换，第2个和第3个half-warp访问的内容交换，这样访存操作对half-warp来说是对齐的但对warp来说就没有全部对齐。测试结果是和没有交换的带宽完全一样。但我相信也许其他一些测试能得到不同结论。就比如我上面提到的那个随机测试。

<< 我的工具栏

CUDA  
2010

上骑都尉 (正五品)  
注册时间 2010-4-10  
积分 834  
串个门 加好友  
打招呼 发消息

17331014

发表于 2010-5-5 21:26:07 | 只看该作者

大牛可以测试一下，看看单纯的halfwarp对齐的访问，当一个warp的两个halfwarp都访问都不不同的地址的时候，会 ...  
ic.expert 发表于 2010-5-5 12:10

使用道具 举报



超级版主  
注册时间 2010-3-7  
积分 2903  
串个门 加好友  
打招呼 发消息

cuda2010

发表于 2010-5-5 21:57:49 | 只看该作者

刚才我做了一个这样的测试，以half-warp为单位对齐，但是结果是对带宽没有影响。  
  
具体做法是启动足够多的b ...  
cuda2010 发表于 2010-5-5 20:27

使用道具 举报

CUDA  
2010

上骑都尉 (正五品)  
注册时间 2010-4-10  
积分 834  
串个门 加好友  
打招呼 发消息

本帖最后由 cuda2010 于 2010-5-5 22:00 编辑

有这个可能。  
我在28#测试的交换范围限于256字节以内(但超过了128字节)，官方资料中似乎只提到了half-warp内部任意交换可合并(64字节)，256字节已经超出了GT200的最大memory Traffic尺寸，没听说过存在这种范围的合并。(不知ic.expert网友有没有这方面的内部资料?)  
明天我会对更大范围内的交换做测试，进一步澄清这个问题。

使用道具 举报

发帖

返回列表 1 2 下一页

最近看过此主题的会员

zBear

sunmoon9  
898

Velaceec  
s

rainway

llyanwen  
yuan

tj12

jm1007

zyt6575

juyhnm

yuquan08

jerryww1

luxu

xiaozhan  
gyu

nancyf11  
4

oucws201  
1

我的工具栏

高级模式

您需要登录后才可以回帖 登录 | 注册

发表回复

☐ 回帖后跳转到最后一页

↑

↶

⊖

↷

↓