(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106489131 A
(43)申请公布日 2017.03.08

(21)申请号 201580036964.4

(22)申请日 2015.06.22

(30)优先权数据
14/331,991 2014.07.15 US

(85)PCT国际申请进入国家阶段日
2017.01.06

(86)PCT国际申请的申请数据
PCT/US2015/036960 2015.06.22

(87)PCT国际申请的公布数据
WO2016/010681 EN 2016.01.21

(71)申请人 高通股份有限公司
地址 美国加利福尼亚州

(72)发明人 陈林 安德鲁·埃文·格鲁贝尔
焦国方 钱特·霍
普拉莫德·瓦桑特·阿拉加迪

(74)专利代理机构 北京律盟知识产权代理有限责任公司 11287
代理人 宋献涛

(51)Int.Cl.
G06F 7/499(2006.01)
G06F 7/552(2006.01)
G09C 1/00(2006.01)

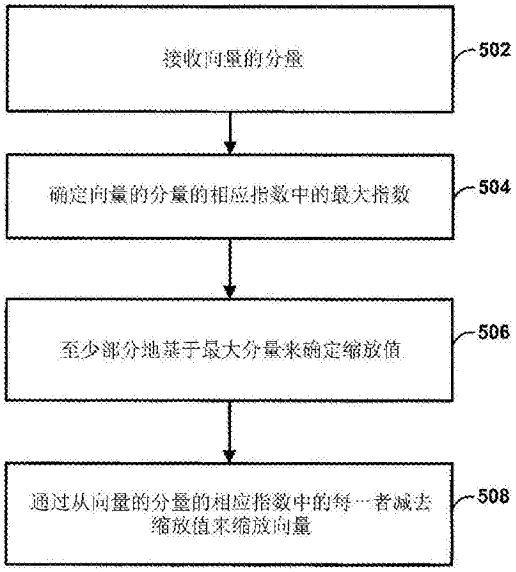
权利要求书3页 说明书10页 附图4页

(54)发明名称

用在算术逻辑单元中的向量缩放指令

(57)摘要

至少一个处理器可接收向量的分量，其中所述向量的所述分量中的每一者包括至少一个指数。所述至少一个处理器可进一步确定所述向量的所述分量的相应指数中的最大指数，且可至少部分地基于所述最大指数来确定缩放值。所述至少一个处理器的算术逻辑单元可通过从所述向量的所述分量的所述相应指数中的每一者减去所述缩放值来缩放所述向量。



CN 106489131 A

1.一种用于缩放向量的方法,所述方法包括:

通过至少一个处理器接收向量的分量,其中所述向量的所述分量中的每一者包括至少一个指数;

通过所述至少一个处理器确定所述向量的所述分量的相应指数中的最大指数;

至少部分地基于所述最大指数来确定缩放值;及

利用所述至少一个处理器的算术逻辑单元ALU通过从所述向量的所述分量的所述相应指数中的每一者减去所述缩放值来缩放所述向量。

2.根据权利要求1所述的方法,其中所述向量的所述分量中的每一者包括浮点数,且其中所述浮点数表示为正负号位、有效数及所述指数。

3.根据权利要求1所述的方法,其中:

所述向量包括三维向量;且

所述向量的所述分量包括x分量、y分量及z分量。

4.根据权利要求3所述的方法,其中缩放所述向量进一步包括:

在第一时钟周期中利用所述ALU通过从所述向量的所述x分量的第一指数减去所述缩放值来对所述向量的所述x分量进行缩放;

在第二时钟周期中利用所述ALU通过从所述向量的所述y分量的第二指数减去所述缩放值来对所述向量的所述y分量进行缩放;及

在第三时钟周期中利用所述ALU通过从所述向量的所述z分量的第三指数减去所述缩放值来对所述向量的所述z分量进行缩放。

5.根据权利要求4所述的方法,其进一步包括:

将所述经缩放的x分量、所述经缩放的y分量及所述经缩放的z分量输出至存储器中的连续存储位置中。

6.根据权利要求1所述的方法,其中所述ALU包括硬件数字电路。

7.根据权利要求1所述的方法,其中至少部分地基于所述最大指数来确定缩放值包括:

确定所述缩放值为所述最大指数。

8.根据权利要求1所述的方法,其中至少部分地基于所述最大指数来确定缩放值包括:

至少部分地基于所述最大指数及最大代表性指数来确定所述缩放值。

9.一种用于缩放向量的设备,所述设备包括:

经配置以存储向量的分量的存储器,其中所述向量的所述分量中的每一者包括至少一个指数;

至少一个处理器,其经配置以:

确定所述向量的所述分量的相应指数中的最大指数,且

至少部分地基于所述最大指数来确定缩放值;及

算术逻辑单元ALU,其经配置以通过从所述向量的所述分量的相应指数中的每一者减去所述缩放值来缩放所述向量。

10.根据权利要求9所述的设备,其中所述向量的所述分量中的每一者包括浮点数,且其中所述浮点数表示为正负号位、有效数及所述指数。

11.根据权利要求9所述的设备,其中:

所述向量包括三维向量;且

所述向量的所述分量包括x分量、y分量及z分量。

12.根据权利要求11所述的设备，其中所述ALU经配置以：

在第一时钟周期中通过从所述向量的所述x分量的第一指数减去所述缩放值来对所述向量的所述x分量进行缩放；

在第二时钟周期中通过从所述向量的所述y分量的第二指数减去所述缩放值来对所述向量的所述y分量进行缩放；且

在第三时钟周期中通过从所述向量的所述z分量的第三指数减去所述缩放值来对所述向量的所述z分量进行缩放。

13.根据权利要求12所述的设备，其中所述ALU经配置以：

将所述经缩放的x分量、所述经缩放的y分量及所述经缩放的z分量输出至所述存储器中的连续存储位置中。

14.根据权利要求9所述的设备，其中所述ALU包括硬件数字电路。

15.根据权利要求9所述的设备，其中所述至少一个处理器经配置以：

确定所述缩放值为所述最大指数。

16.根据权利要求9所述的设备，其中所述至少一个处理器经配置以：

至少部分地基于所述最大指数及最大代表性指数来确定所述缩放值。

17.一种用于缩放向量的设备，所述设备包括：

用于接收向量的分量的装置，其中所述向量的所述分量中的每一者包括至少一个指数；

用于确定所述向量的所述分量的相应指数中的最大指数的装置；

用于至少部分基于所述最大指数来确定缩放值的装置；及

用于通过从所述向量的所述分量的所述相应指数中的每一者减去所述缩放值来缩放所述向量的装置。

18.根据权利要求17所述的设备，其中所述向量的所述分量中的每一者包括浮点数，且其中所述浮点数表示为正负号位、有效数及所述指数。

19.根据权利要求18所述的设备，其中：

所述向量包括三维向量；且

所述向量的所述分量包括x分量、y分量及z分量。

20.根据权利要求19所述的设备，其中所述用于缩放所述向量的装置进一步包括：

用于在第一时钟周期中通过从所述向量的所述x分量的第一指数减去所述缩放值来对所述向量的所述x分量进行缩放的装置；

用于在第二时钟周期中通过从所述向量的所述y分量的第二指数减去所述缩放值来对所述向量的所述y分量进行缩放的装置；及

用于在第三时钟周期中通过从所述向量的所述z分量的第三指数减去所述缩放值来对所述向量的所述z分量进行缩放的装置。

21.根据权利要求20所述的设备，其中所述用于缩放所述向量的装置进一步包括：

用于将所述经缩放的x分量、所述经缩放的y分量及所述经缩放的z分量输出至存储器中的连续存储位置的装置。

22.根据权利要求17所述的设备，其中所述用于至少部分地基于所述最大指数来确定

缩放值的装置包括：

用于确定所述缩放值为所述最大指数的装置。

23.根据权利要求17所述的设备，其中所述用于至少部分地基于所述最大指数来确定缩放值的装置包括：

用于至少部分地基于所述最大指数及最大代表性指数来确定所述缩放值的装置。

24.一种存储指令的计算机可读存储媒体，所述指令在执行时致使一或多个可编程处理器：

接收向量的分量，其中所述向量的所述分量中的每一者包括至少一个指数；

确定所述向量的所述分量的相应指数中的最大指数；

至少部分地基于所述最大指数来确定缩放值；且

通过从所述向量的所述分量的所述相应指数中的每一者减去所述缩放值来缩放所述向量。

25.根据权利要求24所述的计算机可读存储媒体，其中所述向量的所述分量中的每一者包括浮点数，且其中所述浮点数表示为正负号位、有效数及所述指数。

26.根据权利要求24所述的计算机可读存储媒体，其中：

所述向量包括三维向量；且

所述向量的所述分量包括x分量、y分量及z分量。

27.根据权利要求26所述的计算机可读存储媒体，其中所述指令进一步致使所述一或多个可编程处理器：

在第一时钟周期中通过从所述向量的所述x分量的第一指数减去所述缩放值来对所述向量的所述x分量进行缩放；

在第二时钟周期中通过从所述向量的所述y分量的第二指数减去所述缩放值来对所述向量的所述y分量进行缩放；且

在第三时钟周期中通过从所述向量的所述z分量的第三指数减去所述缩放值来对所述向量的所述z分量进行缩放。

28.根据权利要求27所述的计算机可读存储媒体，其中所述指令进一步致使所述一或多个可编程处理器：

将所述经缩放的x分量、所述经缩放的y分量及所述经缩放的z分量输出至存储器中的连续存储位置中。

29.根据权利要求24所述的计算机可读存储媒体，其中所述指令进一步致使所述一或多个可编程处理器：

确定所述缩放值为所述最大指数。

30.根据权利要求24所述的计算机可读存储媒体，其中所述指令进一步致使所述一或多个可编程处理器：

至少部分地基于所述最大指数及最大代表性指数来确定所述缩放值。

用在算术逻辑单元中的向量缩放指令

技术领域

[0001] 本发明涉及计算机处理中的向量缩放。

背景技术

[0002] 向量标准化是对向量进行的运算，其需要计算向量的长度且接着将向量的每一分量除以所计算的向量长度。如果三维向量(x,y,z)的长度计算为$(x^2+y^2+z^2)$的平方根，那么要是所述向量的(x,y,z)值较大，则此类计算可能溢出存储所述计算的中间结果的寄存器。

发明内容

[0003] 本发明呈现用于计算机处理中的向量缩放的技术。根据本发明的技术，在将向量标准化之前，可缩放向量，以使得在标准化期间计算向量的长度不会溢出存储计算向量的长度的中间结果的寄存器。一种图形处理单元(GPU)的算术及逻辑单元(ALU)可经配置以执行三周期缩放指令以供执行向量缩小。与基于软件的向量缩小相比，由ALU所提供的用于执行向量缩小的指令可潜在地相对更高效地执行向量缩小。

[0004] 在本发明的一个实例中，一种用于缩放向量的方法可包含通过至少一个处理器接收向量的分量，其中所述向量的分量中的每一者包括至少一个指数。所述方法可进一步包含通过所述至少一个处理器确定所述向量的分量的相应指数中的最大指数。所述方法可进一步包含至少部分地基于所述最大指数而通过所述至少一个处理器来确定缩放值。所述方法可进一步包含利用所述至少一个处理器的算术逻辑单元(ALU)通过从所述向量的分量的相应指数中的每一者减去所述缩放值来缩放所述向量。

[0005] 在本发明的另一实例中，一种用于缩放向量的设备可包含经配置以存储向量的分量的存储器，其中所述向量的分量中的每一者包括至少一个指数。所述设备可进一步包含至少一个处理器，所述至少一个处理器经配置以确定所述向量的分量的相应指数中的最大指数且至少部分地基于所述最大指数来确定缩放值。所述设备可进一步包含算术逻辑单元(ALU)，所述算术逻辑单元经配置以通过从所述向量的分量的相应指数中的每一者减去所述缩放值来缩放所述向量。

[0006] 在本发明的另一实例中，一种用于缩放向量的设备可包含用于接收向量的分量的装置，其中所述向量的分量中的每一者包括至少一个指数。所述设备可进一步包含用于确定所述向量的分量的相应指数中的最大指数的装置。所述设备可进一步包含用于至少部分地基于所述最大指数来确定缩放值的装置。所述设备可进一步包含用于通过从所述向量的分量的相应指数中的每一者减去所述缩放值来缩放所述向量的装置。

[0007] 在本发明的另一实例中，一种计算机可读存储媒体可存储指令，所述指令在执行时致使一或多个可编程处理器：接收向量的分量，其中所述向量的分量中的每一者包括至少一个指数；确定所述向量的分量的相应指数中的最大指数；至少部分地基于所述最大指数来确定缩放值；且通过从所述向量的分量的相应指数中的每一者减去所述最大指数来缩放所述向量。

[0008] 在随附图式及以下描述中阐述一或多个实例的细节。其它特征、目标及优点将从所述描述及图式以及从权利要求书显而易见。

附图说明

[0009] 图1为说明可经配置以实施本发明的一或多个方面的实例计算装置的框图。

[0010] 图2为进一步详细地说明图1的CPU、GPU及系统存储器的实例实施方案的框图。

[0011] 图3为说明可根据本发明中所揭示的技术缩放的实例三维向量的概念图。

[0012] 图4为说明用于表示向量的每一分量的实例浮点格式的概念图。

[0013] 图5为说明用于缩放向量的实例过程的流程图。

具体实施方式

[0014] 总的来说，本发明描述用于通过硬件缩放向量以使得向量标准化运算不会溢出存储所述运算的中间结果的寄存器的技术。在一个实例中，处理器可执行用于缩放向量的软件代码。对于三维向量，所述软件代码可包含用于确定三维向量中的最大分量且将三维向量的每一分量除以所述最大分量的代码。然而，用于缩放向量的软件代码可能比基于硬件的用于缩放向量的技术更慢。照此，基于硬件的用于缩放向量的方法可提高性能。

[0015] 例如中央处理单元(CPU)或图形处理单元(GPU)等处理器可包含硬件算术逻辑单元(ALU)。ALU可为能够快速执行整数算术及逻辑运算的数字电路。照此，ALU可为一件更高效地缩放向量的理想硬件。然而，因为ALU可能通常设计成仅执行例如加法、减法、“与”运算及“或”运算等简单运算，所以ALU可能不支持实施上文所描述的缩放向量的技术所必要的乘法或除法运算，所述必要运算通常包含将向量的分量除以最大分量。

[0016] 根据本发明的各方面，例如GPU或CPU等至少一个处理器可接收向量的分量(其中所述向量的分量中的每一者包括至少一个指数)且可确定所述向量的分量的相应指数中的最大指数。所述至少一个处理器可至少部分地基于所述最大指数来进一步确定缩放值。CPU或GPU的ALU可通过从所述向量的分量的相应指数中的每一者减去所述缩放因数来缩放所述向量。

[0017] 图1为说明可经配置以实施本发明的一或多个方面的实例计算装置的框图。如图1中所展示，计算装置2可为计算装置，包含(但不限于)视频装置、媒体播放器、机顶盒、无线手持机(例如移动电话及所谓的智能电话)、个人数字助理(PDA)、台式计算机、膝上型计算机、游戏控制台、视频会议单元、平板计算装置，及其类似者。在图1的实例中，计算装置2可包含中央处理单元(CPU)6、系统存储器10及GPU 12。计算装置2还可包含显示处理器14、收发器模块3、用户接口4及显示器8。收发器模块3及显示处理器14均可为与CPU 6及/或GPU 12相同的集成电路(IC)的部分，均可在包含CPU 6及/或GPU 12的一或多个IC的外部，或可形成于在包含CPU 6及/或GPU 12的IC外部的IC中。

[0018] 计算装置2可包含图1中出于清楚起见而未展示的额外模块或单元。举例来说，计算装置2可包含扬声器及麦克风(图1中均未展示)以在计算装置2为移动无线电话的实例中实行电话通信，或在计算装置2为媒体播放器的情况下包含扬声器。计算装置2还可包含摄像机。此外，计算装置2中所展示的各种模块及单元可能不是计算装置2的每个实例中所必需的。举例来说，在计算装置2为台式计算机或经装备以与外部用户接口或显示器介接的其

它装置的实例中，用户接口4及显示器8可在计算装置2外部。

[0019] 用户接口4的实例包含(但不限于)轨迹球、鼠标、键盘及其它类型的输入装置。用户接口4还可为触摸屏，且可作为显示器8的部分并入。收发器模块3可包含允许计算装置2与另一装置或网络之间的无线或有线通信的电路。收发器模块3可包含调制器、解调器、放大器及其它用于有线或无线通信的此类电路。

[0020] 处理器6可为微处理器，例如中央处理单元(CPU)，其经配置以处理供执行的计算机程序指令。处理器6可包括控制计算装置2的操作的通用或专用处理器。用户可将输入提供至计算装置2，以使处理器6执行一或多个软件应用程序。在处理器6上执行的软件应用程序可包含(例如)操作系统、字处理器应用程序、电子邮件应用程序、电子表格应用程序、媒体播放器应用程序、视频游戏应用程序、图形用户接口应用程序或另一程序。另外，处理器6可执行用于控制GPU 12的操作的GPU驱动程序22。用户可经由一或多个输入装置(未展示)(例如，键盘、鼠标、麦克风、触摸垫或经由用户输入接口4耦合至计算装置2的另一输入装置)将输入提供至计算装置2。

[0021] 在处理器6上执行的软件应用程序可包含一或多个图形渲染指令，其指令处理器6致使图形数据渲染至显示器8。在一些实例中，软件指令可符合图形应用程序编程接口(API)，例如开放图形库(OpenGL®)API、开放图形库嵌入系统(OpenGL ES)API、Direct3D API、X3D API、RenderMan API、WebGL API或任何其它公共或专有标准图形API。为了处理图形渲染指令，处理器6可将一或多个图形渲染命令(例如，通过GPU驱动程序22)发布至GPU 12，以使GPU 12执行图形数据的渲染中的一些或全部。在一些实例中，待渲染的图形数据可包含例如点、线、三角形、四边形、三角带等的图形图元的列表。

[0022] GPU 12可经配置以执行图形运算，从而将一或多个图形图元渲染至显示器8。因此，当在处理器6上执行的软件应用程序中的一者需要图形处理时，处理器6可将图形命令及图形数据提供至GPU 12以供渲染至显示器8。图形数据可包含(例如)绘制命令、状态信息、图元信息、纹理信息等。在一些情况下，GPU 12可内置有高度并行结构，所述高度并行结构提供比处理器6更有效的复杂图形相关操作的处理。举例来说，GPU 12可包含经配置而以并行方式对多个顶点或像素进行操作的多个处理元件，例如着色器单元。在一些情况下，GPU 12的高度并行性质允许GPU 12比使用处理器6直接将场景绘制至显示器8更快速地将图形图像(例如，GUI及二维(2D)及/或三维(3D)图形场景)绘制至显示器8上。

[0023] 在一些情况下，可将GPU 12整合至计算装置2的主板中。在其它情况下，GPU 12可存在于安装在计算装置2的主板中的端口中的图形卡上，或可以其它方式并入被经配置以与计算装置2交互操作的外围装置内。GPU 12可包含一或多个处理器，例如，一或多个微处理器、专用集成电路(ASIC)、现场可编程门阵列(FPGA)、数字信号处理器(DSP)或其它等效的集成或离散逻辑电路。GPU 12还可包含一或多个处理器核心，使得GPU 12可称为多核处理器。

[0024] GPU 12可直接耦合至图形存储器40。因此，GPU 12可在不使用总线的情况下从图形存储器40读取数据且将数据写入至图形存储器40。换句话说，GPU 12可使用本地存储装置而非芯片外存储器来在本地处理数据。此类图形存储器40可称为片上存储器。这允许GPU 12通过消除对GPU 12经由总线读取及写入数据(这可能经历繁重的总线通信量)的需要而以更有效的方式操作。然而，在一些情况下，GPU 12可能不包含单独的存储器，而是经由总

线利用系统存储器10。图形存储器40可包含一或多个易失性或非易失性存储器或存储装置，例如，随机存取存储器(RAM)、静态RAM(SRAM)、动态RAM(DRAM)、可擦除可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)、闪存、磁性数据媒体或光学存储媒体。

[0025] 在一些实例中，GPU 12可将完全形成的图像存储在系统存储器10中。显示处理器14可从系统存储器10检索图像，且输出使显示器8的像素发光的值以显示图像。显示器8可为计算装置2的显示器，其显示由GPU 12产生的图像内容。显示器8可为液晶显示器(LCD)、有机发光二极管显示器(OLED)、阴极射线管(CRT)显示器、等离子显示器或另一类型的显示装置。

[0026] 如上文所论述，GPU 12可包含ALU 24，所述ALU 24可为执行整数算术、浮点及逻辑运算的数字电路。可由ALU 24执行的运算可包含加法、减法及逐位运算。在一些实例中，ALU 24可能不能够执行例如乘法及除法等运算。在一些例子中，处理器6还可包含可如ALU 24类似地操作的ALU，其中其可为执行算术及逻辑运算的数字电路。

[0027] 图2为进一步详细说明图1的处理器6、GPU 12及系统存储器10的实例实施方案的框图。如图2中所展示，处理器6可包含至少一个软件应用程序18、图形API 20及GPU驱动程序22，其中的每一者可为在处理器6上执行的一或多个软件应用程序或服务。

[0028] 可供处理器6及GPU 12使用的存储器可包含系统存储器10及帧缓冲器16。帧缓冲器16可为系统存储器10的一部分或可与系统存储器10分离。帧缓冲器16可存储经渲染图像数据。

[0029] 软件应用程序18可为利用GPU 12的功能性的任何应用程序。举例来说，软件应用程序18可为GUI应用程序、操作系统、便携式制图应用程序、用于工程或艺术应用的计算机辅助设计程序、视频游戏应用程序，或使用2D或3D图形的另一类型的软件应用程序。

[0030] 软件应用程序18可包含指令GPU 12渲染图形用户界面(GUI)及/或图形场景的一或多个绘制指令。举例来说，绘制指令可包含界定待由GPU 12渲染的一或多个图形图元的集合的指令。在一些实例中，绘制指令可共同地界定用于GUI中的多个开窗表面的全部或部分。在额外实例中，绘制指令可共同地界定图形场景的全部或部分，所述图形场景包含在由应用程序界定的模型空间或世界空间内的一或多个图形对象。

[0031] 软件应用程序18可经由图形API 20调用GPU驱动程序22，以将一或多个命令发布至GPU 12以供将一或多个图形图元渲染为可显示的图形图像。举例来说，软件应用程序18可经由图形API 20调用GPU驱动程序22以将图元定义提供至GPU 12。在一些情况下，可采用绘制图元(例如，三角形、矩形、三角扇、三角带等)的列表的形式来将图元定义提供至GPU 12。所述图元定义可包含指定与待渲染的图元相关联的一或多个顶点的顶点规格。所述顶点规格可包含每一顶点的位置坐标，且在一些情况下包含与顶点相关联的其它属性，例如色坐标、法向量及纹理坐标。所述图元定义还可包含图元类型信息(例如，三角形、矩形、三角扇、三角形带等)、缩放信息、旋转信息，及其类似者。

[0032] 基于由软件应用程序18发布至GPU驱动程序22的指令，GPU驱动程序22可调配一或多个命令，所述一或多个命令指定供GPU 12执行以便渲染图元的一或多个操作。当GPU 12从CPU 6接收命令时，图形处理管线对命令进行解码且对图形处理管线进行配置以执行命令中所指定的操作。举例来说，图形处理管线中的输入汇编程序可读取图元数据，且将数据汇编成供图形处理管线中的其它图形管线阶段使用的图元。在执行指定操作之后，图形处

理管线将经渲染的数据输出至与显示装置相关联的帧缓冲器16。

[0033] 帧缓冲器16存储用于GPU 12的目的地像素。每一目的地像素可与唯一的屏幕像素位置相关联。在一些实例中，帧缓冲器16可存储每一目的地像素的色彩分量及目的地阿尔法值。举例来说，帧缓冲器16可存储每一像素的红、绿、蓝、阿尔法(RGBA)分量，其中“RGB”分量对应于色彩值且“A”分量对应于目的地阿尔法值。虽然将帧缓冲器16及系统存储器10说明为单独的存储器单元，但在其它实例中，帧缓冲器16可为系统存储器10的部分。

[0034] 在一些实例中，图形处理管线可包含顶点着色器阶段、外壳着色器阶段、域着色器阶段、几何着色器阶段及像素着色器阶段中的一或多者。所述图形处理管线的这些阶段可视为着色器阶段。这些着色器阶段可实施为在GPU 12中的着色器单元46上执行的一或多个着色器程序。着色器单元46可包括一或多个配置为处理分量的可编程管线的着色器单元。在一些实例中，着色器单元46可称为“着色器处理器”或“统一着色器”，且可执行几何、顶点、像素或其它着色操作以渲染图形。

[0035] 通过将命令发送至着色器单元46以执行图形处理管线中的顶点着色器阶段、外壳着色器阶段、域着色器阶段、几何着色器阶段及像素着色器级中的一或多者，GPU 12可指定着色器单元46执行多种着色操作，例如顶点着色、外壳着色、域着色、几何着色、像素着色及其类似者。在一些实例中，GPU驱动程序22可经配置以编译一或多个着色器程序且将所编译的着色器程序下载至含于GPU 12内的一或多个可编程着色器单元上。可以高级着色语言编写着色器程序，例如OpenGL着色语言(GLSL)、高级着色语言(HLSL)、图形C语言(Cg)着色语言等。经编译的着色器程序可包含一或多个控制GPU 12内的着色器单元46的操作的指令。举例来说，着色器程序可包含：顶点着色器程序，其可由着色器单元46执行以实施顶点着色器阶段的功能；外壳着色器程序，其可由着色器单元46执行以实施外壳着色器阶段的功能；域着色器程序，其可由着色器单元46执行以实施域着色器阶段的功能；几何着色器程序，其可由着色器单元46执行以实施几何着色器阶段的功能；及/或像素着色器程序，其可由着色器单元46执行以实施像素着色器的功能。顶点着色器程序可对可编程顶点着色器单元或统一着色器单元的执行进行控制，且包含指定一或多个逐顶点操作的指令。

[0036] 着色器单元46可包含处理器核心48，所述处理器核心中的每一者可包含用于获取及解码操作的一或多个组件、用于执行算术计算的一或多个算术逻辑单元、一或多个存储器、缓存及寄存器。在一些实例中，处理器核心48也可称作标量处理元件。处理器核心48中的每一者可包含通用寄存器25。通用寄存器25可存储直接由处理器核心48中的ALU 24存取的数据。举例来说，通用寄存器25可存储待由ALU 24缩放的向量分量，且也可存储由ALU 24输出的经缩放向量分量。

[0037] 处理器核心48中的每一者可包含标量ALU，例如ALU 24。如上文所论述，ALU 24可为执行整数算术、浮点及逻辑运算的数字电路。可由ALU 24执行的运算可包含加法、减法及逐位运算。在一些实例中，ALU 24可能不能够执行例如乘法及除法等运算。根据本发明的各方面，ALU 24可通过缩放向量的分量来缩放向量。ALU 24也可将向量的经缩放分量输出至图形存储器40或通用寄存器25。

[0038] 图形存储器40为物理地整合至GPU 12的集成电路中的片上存储装置或存储器。因为图形存储器40为片上存储器，所以与经由系统总线从系统存储器10读取值或将值写入至系统存储器10相比，GPU 12可能能够更快速地从图形存储器40读取值或将值写入至图形存

储器40。图形存储器40可存储向量的分量，且也可存储由ALU 24缩放之后的向量的经缩放分量。图形存储器40可以浮点格式存储向量的分量，以使得每一分量可作为有正负号的位、有效及指数存储在图形存储器40中。

[0039] 图3为说明可由CPU 6或GPU 12缩放的实例三维向量的框图。如图3中所展示，三维笛卡尔(Cartesian)坐标系统52中的向量50可经由指示向量50的相应分量54A至54C(“分量54”)的值的元组(x,y,z)来表示。向量50的分量54可包含x分量54A、y分量54B及z分量54C。

[0040] 如上文所论述，向量50的分量54中的每一分量可为作为有正负号的位、有效数及指数存储在图形存储器40或通用寄存器25中的浮点值。举例来说，实例浮点值1.2345可等于$12345*10^{-4}$，因此12345可为有效数或尾数，且-4可为底数10的指数。在其它实例中，所述指数可为底数2的指数。为了表示负指数值，可对指数进行偏置或偏移，以使得指数值转换成正值。举例来说，可将值15加至指数，以使得指数值-4可在存储器中存储为11。

[0041] 图4为说明用于表示向量50的分量54中的每一分量的实例浮点格式的概念图。如上文所论述，向量50的分量54中的每一分量可为浮点值。如图4中所展示，分量54中的每一分量可以浮点格式60表示。浮点格式60可包含正负号位62，其指示由浮点格式60表示的浮点值的正负号。正负号位62在浮点值的正负号为负的情况下可为1，且在浮点值的正负号为正的情况下可为0。浮点格式60可进一步包含指数64及有效数66。在一个实例中，对于32位IEEE浮点格式60，正负号位62可为一个位，指数64可为8位(伴以127的偏差)，且有效数66可为23位(伴以整数隐藏)。举例来说，浮点值-82.3125可等于-1.0100100101_2*2^6。在本实例中，正负号位62可设置成1。指数64可为10000101，归因于127的偏差，其为133_{10}，且因为可隐藏整数位，所以有效数66可为01001001010000000000000。

[0042] 根据本发明的各方面，处理器6或GPU 12可使用ALU 24缩放向量50，以使得处理器6或GPU 12可执行向量50的向量标准化而不溢出存储所述向量标准化运算的中间结果的寄存器。因为ALU 24可为硬件电路，所以此类向量50的缩放可在硬件中执行，而非在由(例如)着色器单元46执行的软件中执行。此外，因为ALU 24可包含用于执行加法及减法运算的功能性但可能不包括用于执行乘法及/或除法运算的功能性，所以ALU 24可能能够在不执行乘法或者除法运算的情况下缩放向量50。为了缩放向量50，GPU 12可接收向量50的分量54。举例来说，如果向量50为三维向量，那么GPU 12可接收向量50的x分量54A、y分量54B及z分量54C。向量50的分量54可存储在存储器中，例如系统存储器10、图形存储器40、着色器单元46的存储器、通用寄存器25，及其类似者。

[0043] 如上文所论述，向量50的分量54可各自为包含至少一有效数及一指数的浮点值。GPU 12可确定分量54的指数中的最大指数。举例来说，如果分量54的指数为-1、2及5，那么GPU 12可确定分量54的指数中的最大指数为5。在一些实例中，确定分量54的指数中的最大指数可包含确定分量54的指数中的最大值指数。因此，(例如)如果分量54的指数为-1、2及5，GPU 12可确定分量54中的最大值指数为5，因为5大于2或-1。

[0044] 响应于GPU 12确定分量54的指数中的最大指数，GPU 12可确定用于缩放分量54的指数中的每一者的缩放值。在一个实例中，缩放值可等于最大指数，因此对于分量54的指数，缩放值可为5。在另一个实例中，GPU 12可确定缩放值以防止在缩放之后经缩放指数的下溢及/或溢出在这种情况下，GPU 12可至少部分地基于所述最大指数来确定所述缩放值。举例来说，缩放值可为最大指数+常量。举例来说，缩放值可为最大指数-(maximum_

representable_exponent-1)/2+1。maximum_representable_exponent可为从分量54的浮点格式导出的常量的最大可表示的指数。举例来说，对于32位IEEE浮点数，最大可表示的指数为128。指数范围为[-127,128](即，从-127至128且包含端值)，因为32位浮点数中的指数利用8位表示。在另一个实例中，GPU 12可确定缩放值为(最大指数-1)/2-2。因此，给定最大指数15，缩放值可为(15-1)/2-2，结果可为5。

[0045] 响应于GPU 12确定缩放值，ALU 24可经配置以通过从分量54的每一指数减去所述缩放值来对分量54中的每一分量进行缩放。举例来说，针对向量50的x分量54A、y分量54B及z分量54C给定指数值-1、2及5，鉴于GPU 12确定5为分量54的指数中的最大指数且鉴于GPU 12确定缩放值为所述最大指数的值(即，将缩放值设置为5)，则ALU 24可从x分量54A的指数值-1减去5，ALU 24可从y分量54B的指数值2减去5，且ALU 24可从z分量54C的指数值5减去5，从而得到具有分别针对x分量54A、y分量54B及z分量54C的指数值-6、-3及0的经缩放分量54。所得包含由ALU 24输出的指数的经缩放分量可存储在存储器中，例如图形存储器40、系统存储器10、通用寄存器25，及其类似者。

[0046] 如上文所论述，在一些实例中，分量54的指数可为偏置指数。GPU 12及ALU 24可按未偏置指数的类似方式处理偏置分量。举例来说，如果分量54的指数的值-1、2及5通过15加以偏置，使得将15加至分量54的每一指数，那么分量54的偏置指数的值可为14、17及20。因此，GPU 12可确定分量54的偏置指数中的最大指数为20。响应于GPU 12确定分量54的偏置指数中的最大指数，GPU 12可至少部分地基于所述最大指数确定缩放值。在本实例中，，GPU 12可将缩放值设置为所述最大指数的值。响应于GPU 12确定缩放值，ALU 24可通过从分量54的每一指数减去所述缩放值来对分量54中的每一分量进行缩放。举例来说，针对向量50的x分量54A、y分量54B及z分量54C给定偏置指数值14、17及20，且鉴于GPU 12确定20为分量54的指数中的最大指数及缩放值被设置成所述最大指数的值，则ALU 24可从x分量54A的指数值14减去20，ALU 24可从y分量54B的值17减去20，且ALU 24可从z分量54C的指数值20减去20。ALU 24可将偏差15加至分量54的每一指数，从而得到具有分别针对x分量54A、y分量54B及z分量54C的偏置指数值9、12及15的经缩放分量54。

[0047] ALU 24可经配置以每一时钟周期输出一个经缩放分量，因此ALU 24可在第一时钟周期输出经缩放的x分量，在第二时钟周期输出经缩放的y分量且在第三时钟周期输出经缩放的z分量。用于执行向量50的缩放的实例伪码可如下表达：

```
if (src0, src1, or src2 is INF or NaN){ // output = input
        dst = src0;
        dst+1 = src1;
        dst+2 = src2;
}
else maxexp = max(src0.exp, src1.exp, src2.exp);

if(src0 is 0 or denorm)
        dst = 0 preserving src0 sign;
else dst.exp = src0.exp - maxexp;

if(src1 is 0 or denorm)
        dst = 0 preserving src1 sign;
else (dst+1).exp = src1.exp - maxexp;

if(src2 is 0 or denorm)
        dst = 0 preserving src2 sign;
else (dst+2).exp = src2.exp - maxexp;
```

[0048]

[0049] 如在上文伪码中所展示，src0、src1及src2可为三维向量的分量的源存储位置。dst可为存储器中用于第一经缩放分量的目标位置，dst+1可为存储器中用于第二经缩放分量的下一个连续目标位置，且dst+2可为存储器中用于第三经缩放分量的下一个连续目标位置。如可见，经缩放分量可存储在存储器中连续的存储位置中。

[0050] 如上所述，GPU 12可确定分量中的任何者是否为无穷大或不是数值。如果分量为未被定义或不可表示的值，那么分量可能不是数值。不是数值的分量可用1s填充指数64，且可使有效数66为非零值。如果如此，那么不对向量的分量进行缩放。否则的话，GPU 12可至少部分地基于分量的指数中的最大指数来确定缩放值。对于src0、src1及src2中的分量中的每一者，如果指数为零或反常数值，那么不对分量进行缩放。反常数值可为具有小于最小正常数值的量值的任何非零数值。否则的话，ALU 24可通过从所述缩放值减去分量的指数来对分量进行缩放。在一些实例中，目标存储位置dst、dst+1或dst+2中的任一者可与源存储位置src0、src1或src2重叠。

[0051] GPU 12可提供缩放功能以供将向量的经缩放分量输出为(rpt2)scale.x,(r)x,(x,y,z);。(rpt2)可指示缩放指令将在初始执行之后重复两次，因此其可执行总计三次以输出三维向量的经缩放的x、y及z分量。scale.x可为缩放功能的功能名称，其中scale.x中的x可为存储向量的未缩放分量的源存储位置的开始位置。(r)x可指示缩放指令将被重复，且(r)x中的x可为用于存储向量的经缩放分量的目标存储位置的开始位置。(x,y,z)可为所述向量待由缩放功能予以缩放的分量。ALU 24的吞吐量(其可为在一定量的时间中能够发

布至ALU 24的管线的缩放指令的数目)可为3周期/缩放指令。ALU 24在执行缩放指令中的时延(其可为从发布至完成用以执行缩放指令的时间总量)可取决于缩放指令的实施方案。

[0052] 图5为说明用于缩放向量的实例过程的流程图。如图5中所展示,所述过程可包含通过处理器6或GPU 12接收向量的分量,其中所述向量的分量中的每一者包括至少一个指数(502)。所述过程可进一步包含通过处理器6或GPU 12确定所述向量的分量的相应指数中的最大指数(504)。所述过程可进一步包含通过处理器6或GPU 12至少部分地基于所述最大分量来确定缩放值(506)。所述过程可进一步包含利用处理器6或GPU 12的ALU 24通过从所述向量的分量的相应指数中的每一者减去所述缩放值来缩放所述向量(508)。

[0053] 在一些实例中,缩放向量可进一步包含通过ALU 24在不执行乘法运算或除法运算的情况下对向量进行缩放。在一些实例中,所述向量的分量中的每一者可为浮点数,且其中所述浮点数可表示为正负号位、有效数及指数。在一些实例中,所述向量可包括三维向量,且所述向量的分量可包括x分量、y分量及z分量。在一些实例中,缩放所述向量可包含在第一时钟周期中利用ALU 24通过从所述向量的x分量的第一指数减去缩放值来对所述向量的x分量进行缩放,在第二时钟周期中利用ALU 24通过从所述向量的y分量的第二指数减去缩放值来对所述向量的y分量进行缩放,且在第三时钟周期中利用ALU 24通过从所述向量的z分量的第三指数减去缩放值来对所述向量的z分量进行缩放。在一些实例中,所述过程可进一步包含将经缩放的x分量、经缩放的y分量及经缩放的z分量输出至存储器中的连续存储位置中。在一些实例中,ALU 24可为硬件数字电路。

[0054] 在一些实例中,至少部分地基于最大指数来确定缩放值可包含确定缩放值为所述最大指数。在一些其它实例中,至少部分地基于最大指数来确定缩放值可包含至少部分地基于所述最大指数及最大代表性指数来确定缩放值。

[0055] 在一或多个实例中,所描述的功能可实施于硬件、软件、固件或其任何组合中。如果实施于软件中,那么可将所述功能作为一或多个指令或代码存储在计算机可读媒体上或经由计算机可读媒体传输。计算机可读媒体可包含计算机数据存储媒体或通信媒体,所述通信媒体包含促进计算机程序从一处传递至另一处的任何媒体。数据存储媒体可为任何可由一或多个计算机或一个或多个处理器存取以检索用于实施本发明中所描述的技术的指令、代码及/或数据结构的可用媒体。借助于实例而非限制,此类计算机可读媒体可包括RAM、ROM、EEPROM、CD-ROM或其它光盘存储装置、磁盘存储装置或其它磁性存储装置,或任何可用于携载或存储呈指令或数据结构形式的所要程序代码且可由计算机存取的其它媒体。并且,适当地将任何连接称作计算机可读媒体。举例来说,如果使用同轴电缆、光纤电缆、双绞线、数字订户线(DSL)或例如红外线、无线电及微波等无线技术从网站、服务器或其它远程源传输软件,那么所述同轴电缆、光纤电缆、双绞线、DSL或例如红外线、无线电及微波等无线技术包含于媒体的定义中。如本文所使用的磁盘及光盘包含压缩光盘(CD)、激光光盘、光学光盘、数字多功能光盘(DVD)、软盘及蓝光光盘,其中磁盘通常是以磁性方式再现数据,而光盘是用激光以光学方式再现数据。以上各项的组合也应包含在计算机可读媒体的范围内。

[0056] 代码可由一或多个处理器(例如一或多个数字信号处理器(DSP)、通用微处理器、专用集成电路(ASIC)、现场可编程逻辑阵列(FPGA)或其它等效的集成或离散逻辑电路)执行。因此,如本文中所使用的术语“处理器”及“处理单元”可指代前述结构或适于实施本文

中所描述的技术的任何其它结构中的任一者。此外，在一些方面，本文中所描述的功能性可提供于经配置以用于编码及解码的专用硬件及/或软件模块内，或者并入于组合编解码器中。并且，可以将所述技术完全实施于一或多个电路或逻辑元件中。

[0057] 本发明的技术可以在广泛多种装置或设备中实施，包含无线手持机、集成电路(IC)或一组IC(即，芯片组)。本发明中描述各种组件、模块或单元以强调经配置以执行所揭示的技术的装置的功能方面，但未必需要由不同硬件单元实现。实际上，如上文所描述，各种单元可结合合适的软件及/或固件组合在编解码器硬件单元中或通过交互操作硬件单元的集合来提供，所述硬件单元包含如上文所描述的一或多个处理器。

[0058] 已描述各种实例。这些实例及其它实例属于所附权利要求书的范围内。

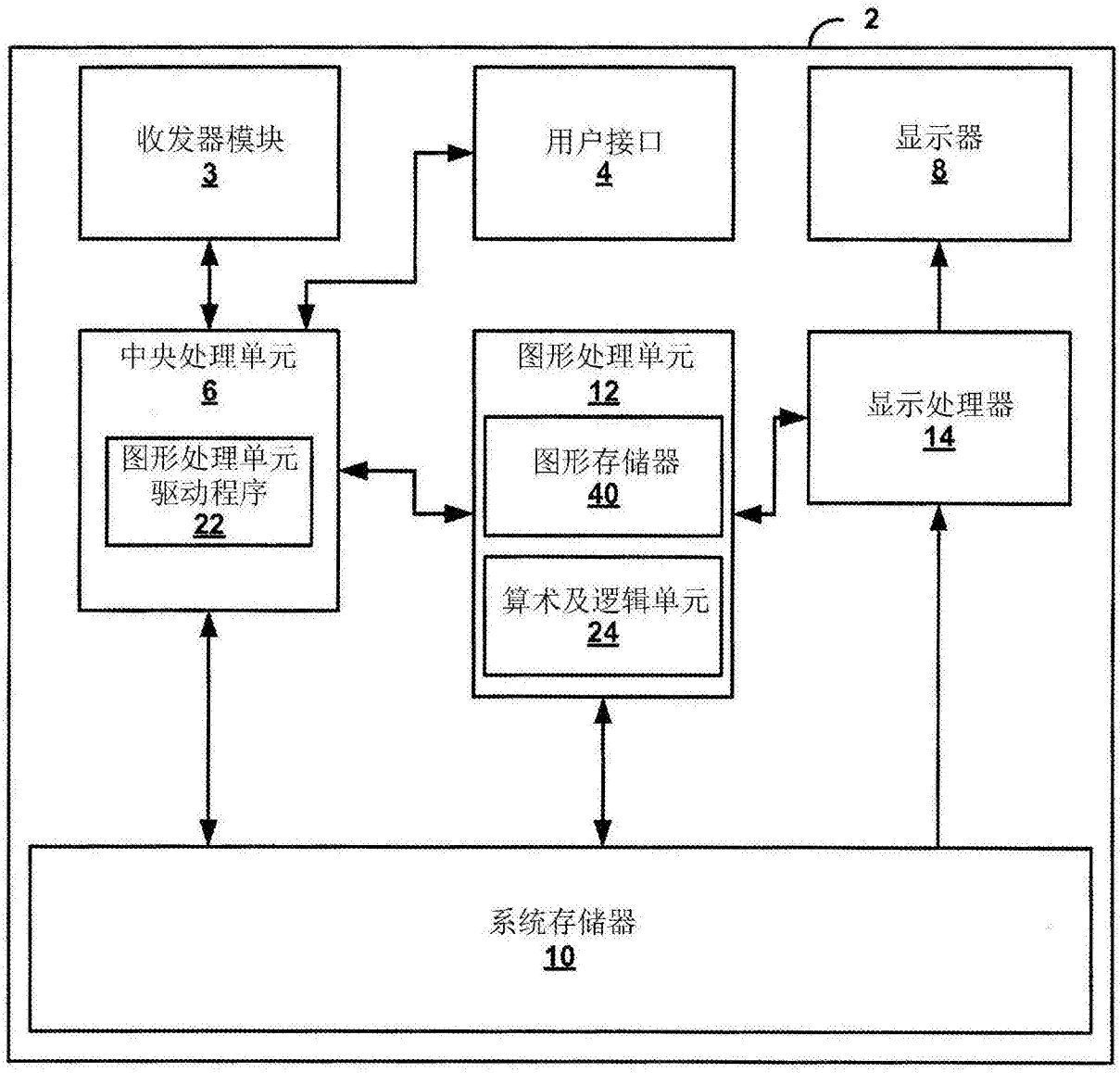

2
收发器模块
3
用户接口
4
显示器
8
中央处理单元
6
图形处理单元
驱动程序
22
图形处理单元
12
图形存储器
40
算术及逻辑单元
24
显示处理器
14
系统存储器
10

图1

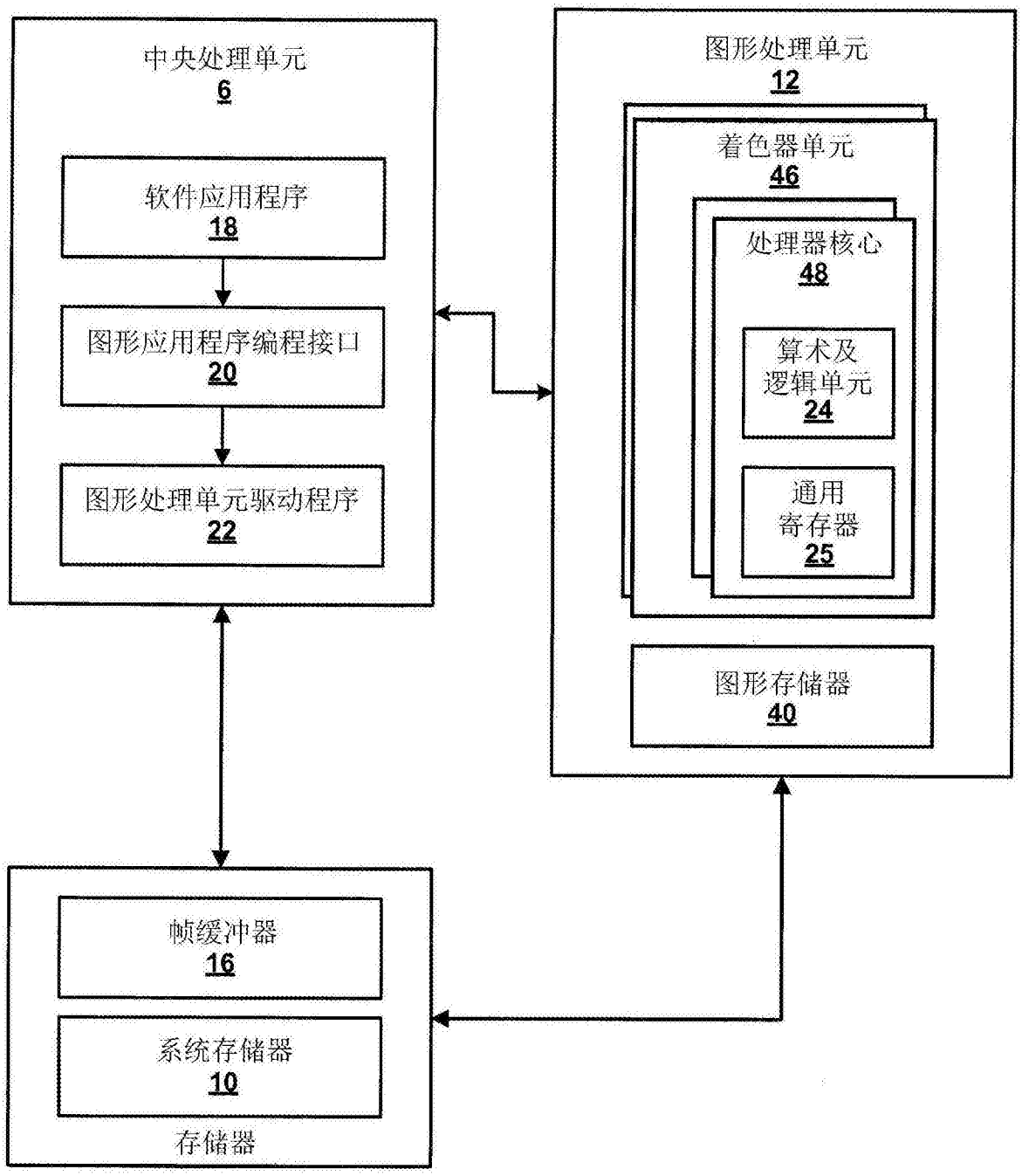

中央处理单元
6
软件应用程序
18
图形应用程序编程接口
20
图形处理单元驱动程序
22
图形处理单元
12
着色器单元
46
处理器核心
48
算术及
逻辑单元
24
通用
寄存器
25
图形存储器
40
帧缓冲器
16
系统存储器
10
存储器

图2

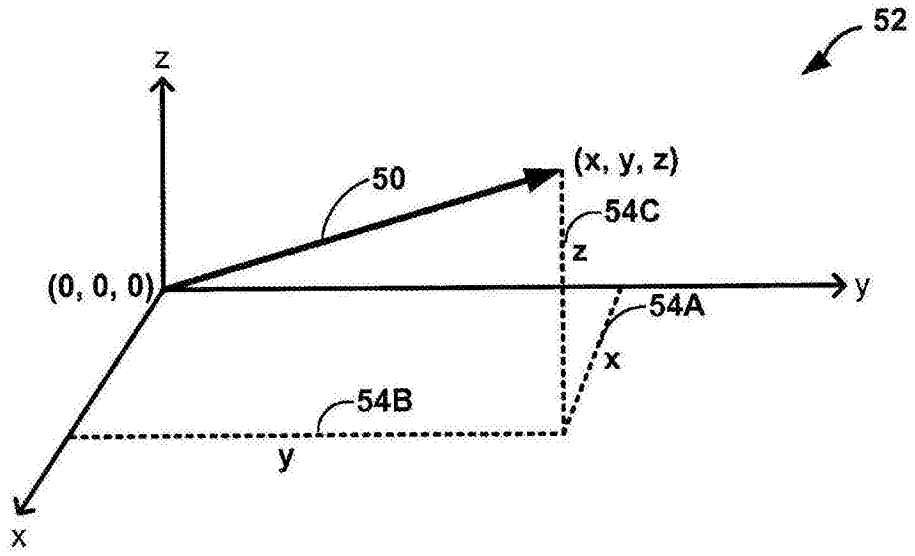


图3

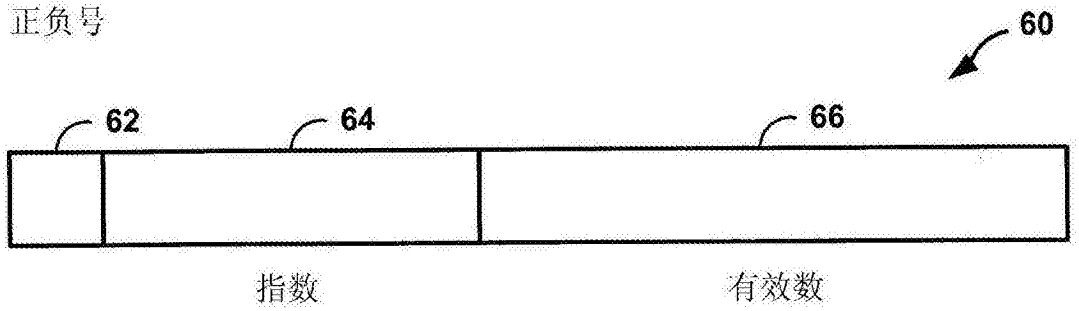


图4

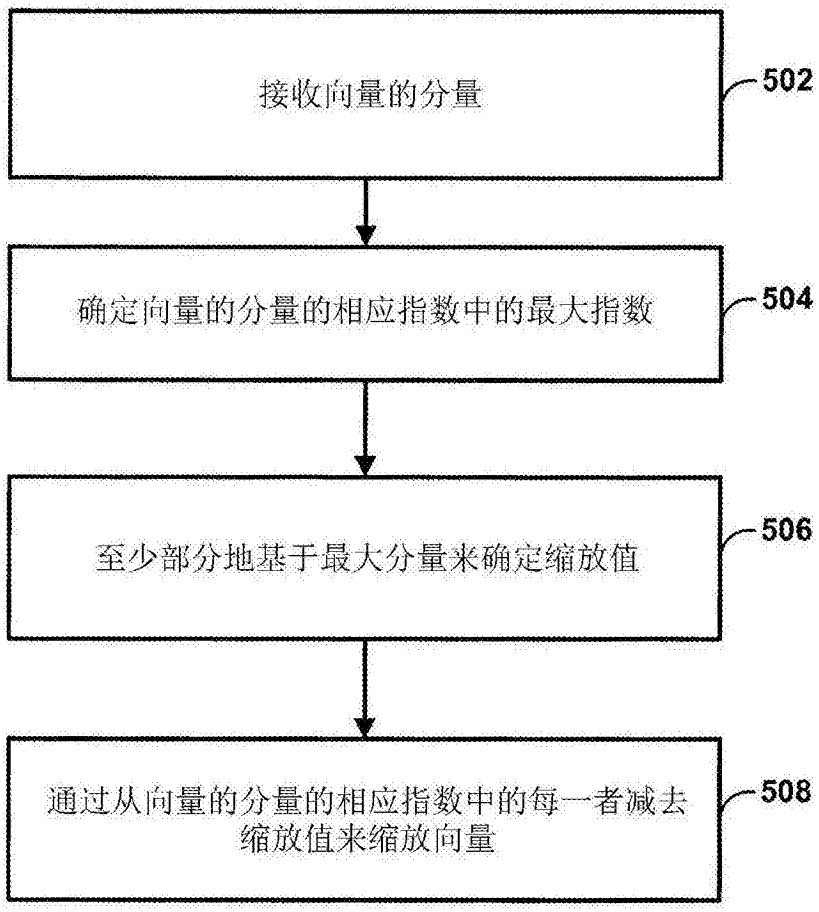

接收向量的分量
502
确定向量的分量的相应指数中的最大指数
504
至少部分地基于最大分量来确定缩放值
506
通过从向量的分量的相应指数中的每一者减去缩放值来缩放向量
508

图5