# 进程

在这一章中，我们将研究各种不同类型的进程是如何影响分布式系统工作的。进程的概念起源于操作系统中，它是用来定义一个程序的运行的。在操作系统层面，进程的管理和调度可能是最重要的课题。然而，当在分布式系统中讨论时，进程的一些其他课题变得更为重要。

我们首先重点讨论线程和它在分布式系统的作用。事实证明，线程不但在多核多处理器环境中对性能有重要的作用，更有助于构造客户端和服务端。在很多情况下，我们看到线程被进程所替代，并且使用底层操作系统来保证保护和促进通信。然而，当性能受到威胁时，线程将扮演着重要的角色。

近些年，虚拟化的概念开始流行起来。虚拟化允许应用程序及其完整的环境(包括操作系统)与其他应用程序并发运行，但高度独立于底层硬件和平台，从而实现高度的可移植性。另外，虚拟化可以帮助隔离由故障和安全引起的错误。这对于分布式系统是很重要的概念，我们将单独讨论。

客户-服务器组织架构在分布式系统中很重要。在这一章节，我们将仔细研究客户端和服务端的典型组织。我们将注意到服务端的典型设计，包括那些在object-based分布式系统中使用的类型。一个应用很广泛的Web服务是Apache，我们将单独了解。服务端的集群组织也很重要，特别是他们需要合作得像是单一系统在工作一样。我们将介绍类似的例子，包括广域服务PlanetLab。

在广域分布式系统中，一个重要的课题是机器间的晋城移动。进程转移或者代码转移可以帮助实现可伸缩性，或者帮助动态配置客户端和服务端。关于代码转移还有其如何实现的，我们将在这一章节讨论。

## 3.1线程

虽然进程组成了分布式系统的基本单元，但实践表明，分布式系统以进程作为最小粒度并不足够充分（就是还有线程呗）。作为替代，有进程控制的多个线程组成了更小的粒度，这能更好的构建分布式应用并获得更好的性能。在这一节中，我们将关注在分布式系统中的线程是如何发挥作用的并解释它为什么这么重要。关于线程以及如何通过线程构建应用可以参考

[Lewis and Berg, 1998; Stevens, 1999; Robbins and Robbins, 2003]。这些很好的从概念上诠释了多线程并发程序。

**线程的介绍**

为了理解分布式系统中的线程，我们需要首先理解进程是什么，并且进程和线程是什么关系。为了运行一个程序，操作系统会创造一系列虚拟进程，每一个会运行不同的程序。为了跟踪这些进程，操作系统有一个进程对照表，包括存储CPU的寄存器只，内存映射，打开文件，计数器，privileges，等。这些条目构成了进程的上下文。

进程上下文可以看作是硬件处理器上下文的软件模拟。硬件处理器上下文包括硬件处理终端，以及存储CPU处理现场的最小信息。这个进程上下文包括程序计数器，或者其他的寄存器值比如栈指针。

进程通常定义为一个运行的程序，也即正在操作系统虚拟处理器上运行的程序。一个关键问题是，操作系统会很小心的保证独立的进程之间不会相互干扰的彼此。换句话说，多个进程共享同一个cpu以及其他硬件资源这件事，是透明的。通常，操作系统需要硬件的支持来完成这种分割操作。

这种透明性是有代价的。比如，当一个进程被创建时，操作系统必须创建一个完全独立的地址空间。这种分配可以意味着初始化内存段，例如，将数据段归零，将相关程序复制到文本段，并为临时数据建立堆栈。这样，在两个进程间切换CPU需要消耗性能。除了在寄存器临时存放数据外（包括程序寄存器和栈指针），操作系统同样会修改内存管理单元（MMU）的寄存器和使地址转换缓存无效，例如在转换后备缓冲区(TLB)中。另外，如果操作系统需要支持超过他内存大小的进程数量，那么需要在内存和磁盘间进行内存交换。

像进程一样，线程也会执行一段代码，并独立于其他线程。然而，不同于进程，如果实现并发透明性会导致性能下降，则不会尝试实现高度的并发透明性。因此，线程系统通常保持最小的信息来允许CPU可以被其他线程共享。特别是，一个**线程上下文**通常和进程上下文的组成没有区别，除了一些关于线程管理的信息。例如，线程系统可以跟踪一个线程当前在互斥量变量上被阻塞的事实，以便不选择它执行。关于其他线程的不必要的信息是不必理会的。因此，保护数据不受单个进程内线程的不适当访问完全留给应用程序开发人员。因为我们发现一个处理器的上下文包含在线程上下文中，而线程上下文包含在进程上下文中。

正如我们刚才概述的，部署线程有两个重要含义。首先，多线程应用的系统性能很难比同样的单线程系统性能差。实际上，多线程可以获得更多的性能。其次，由于线程不能自动的像进程一样保护自己不受其他线程干扰，多线程应用的独立需要更多的智慧和努力。合适的设计和事情的简单化是有帮助的选择。不幸的是，从目前的实践来看，这个原则并不好理解。

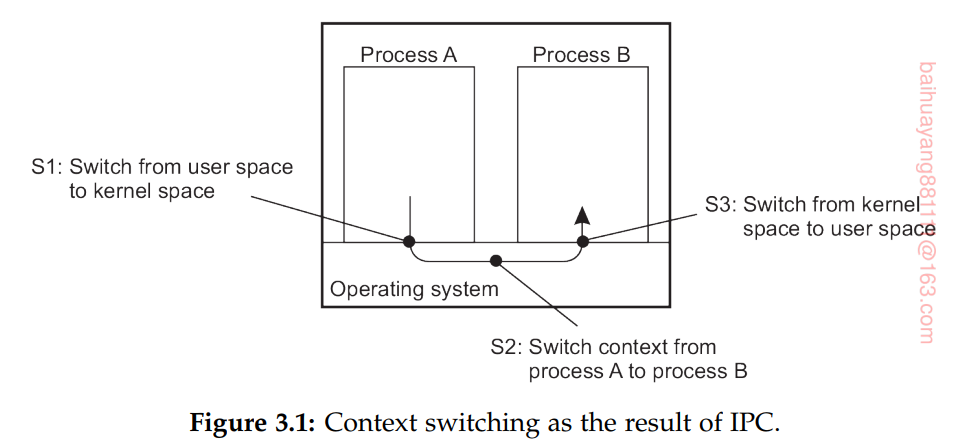
**在非分布式系统中的线程使用**

在讨论分布式系统中线程的功能之前，我们首先讨论在非分布式系统中线程的功能。使用多线程系统有很多好处。

最重要的一个优势是，作为但线程进程，当一个阻塞方法被调用时，进程将会被阻塞。为了说明这一点，考虑一个应用程序，例如电子表格程序，并假设用户持续地、交互地想要更改值（举个访问数据库的栗子更好，比如需要向数据库刷10分钟的数据，进程将阻塞10分钟）。电子表格程序的一个重要特性是，它维护不同单元格之间的功能依赖关系，这些单元格通常来自不同的电子表格。因此当一个单元更改时，其他依赖的单元会自动更新。当用户更改一个单元时，这种更改会出发一系列计算。如果只有一个进程进行控制，在程序等待输入之前，计算是不能够完成的。像这样，当依赖的计算完成之前，提供数据是不容易的。一个简单的处理方式是，获取两个线程：一个负责处理用户输入，另一个负责更新。同时，第三个线程将数据备份。

另一个多线程的优点是当在一个多处理器或者多核系统上运行程序是，可以进行并行计算。当数据都被存储在共享的内存时，每一个线程都被分配给了不同的CPU和处理器。如果涉及得当，这种并行计算可以是透明的：这个进程同样可以在单核系统中运行，尽管这样运行会比较慢。在多核和多处理器价格变得相对便宜的情况下，多线程的并行处理变得很重要了。这种计算机系统通常发生在客户-服务端系统的服务端，但目前也广泛应用于智能手机等设备。

多线程同样在大型应用的上下文中很有作用。这些应用通常由一些相互合作的程序，即独立执行的进程组成。这个方式是典型的Unix 环境的方式。程序之间的合作通过**interprocess communication（IPC）**来实现。对于Unix系统，这些机制包括，管道，消息队列和共享内存等。这种IPC机制的主要缺点是上下文切换很消耗资源，图3.1介绍了这些机制。



除了使用进程，一个应用还可以通过不同的独立线程来构成。他们之间的通信完全通过共享数据来完成。线程切换有时可以只在用户态完成，在一些其他的实现中，内核也会参与线程并进行调度。这样会极大的提高性能。最终，这里有一个使用线程的纯粹的软件工程的原因：许多由线程构成的应用会变得更简单。考虑需要执行多个(或多或少独立的)任务的应用程序，如前面讨论的电子表格示例。

**线程实现**

线程通常以线程包的形式提供。这种包包括线程的创建和摧毁操作，同样包括线程同步相关的锁和条件变量等。这里主要有两种方法来实现线程包。第一种方法是构成完全在用户控件执行的一个线程库。第二种方法是，使用内核感知这些线程并调度线程。

用户级别的线程库有一系列好处。第一，创建和摧毁线程都不是很好资源。因为所有的线程权限都在用户地址区域，创建线程唯一值得考虑的就是在栈空间分配内存。同样的，摧毁线程主要包括在栈空间释放内存，但现在已经不再使用这种方式了。这两种方式都很廉价。

第二个使用用户级别线程的优点是上下文切换可以在几个指令下完成。基本上，只有CPU寄存器的值需要保存并且接下来当线程切换时重新加载之前存储的值。这里不需要改变内存映射，冲刷TLB，或做CPU计数等。然而，如Note3.1讨论的，上下文切换的大部分开销是由扰乱内存缓存造成的。

用户级线程的一个主要缺点是使用了**多对一的线程模型**：即多个线程映射为一个调度实体。作为一个结果，调用阻塞系统调用将立即阻塞线程所属的整个进程，从而阻塞该进程中的所有其他线程。正如我们解释的，线程通常对于构建由许多在逻辑上同时执行的小组成部分构成的大型应用是很有帮助的。在这种情况下，在同一时刻，IO阻塞不应该阻止其他部分的执行。对于这种应用，用户级别的线程是不会起到作用的。（注意这里是用户级别的，我们平常用的是内核级别的）

这个问题可以通过内核级别的实现来解决，也即使用**一对一线程模型**，即一个线程为一个调度实体。当然，每个线程的操作都是由操作系统内核负责，每一次都需要系统调用。这里线程上下文切换看起来会比进程的上下文切换更耗资源。然而，考虑到上下文切换的性能通常是由内存缓存的无效使用决定的，而不是由多对一或一对一线程模型之间的区别决定的，许多操作系统现在提供了后者模型，即使只是为了简单。

最后一点需要注意的是，必须认识到使用线程是组织应用程序中同时执行和并发执行的一种方式。在实践中，我们经常看到应用程序被构造为一组并发进程，共同使用操作系统提供的进程的设施。一个很好的例子就是ApacheWeb服务器，它是通过一系列进程来处理请求的。每个进程都组成一个单一的线程实例，但是都能通过标注你的方法与其他实例通信。

正如Srinivasan讨论的，使用进程代替线程对于将数据空间分离有很大的好处：每个线程都在独立的自己的数据区工作，并在操作系统层面与其他的进程隔离。这种分离的好处不应低估：线程编程被认为是出了名的困难，因为开发人员完全负责管理对共享数据的并发访问。使用进程，数据空间最终受到硬件支持的保护。如果一个进程访问了它自己之外的内存，硬件将会抛出一个异常，这个异常将在之后被操作系统处理。在线程则没有这种待遇。

**分布式系统的线程**

线程的一个重要性质是，它能够提供方便的手段使得在系统中的阻塞调用，不影响该线程所属进程的运行。这个性质使得线程在分布式系统中的使用变得很有吸引力，因为它使得同时维持多个逻辑通信的连接变得容易。为了说明这个观点，我们将分别观察多线程的客户-服务端系统。

**多线程客户端**

为了建立高度透明的分布式系统，在广域网中运行的分布式系统可能需要隐藏较长的进程间消息传播时间。广域网中的往返延迟很容易达到几百毫秒，有时甚至是几秒钟。

隐藏通信延迟的通常方法是启动通信并立即处理其他事情。这在Web浏览器中很常见。在很多情况下，一个Web文档包含很多个HTML文件如图片，标签等。为了获取这些元素，浏览器需要建立TCP/IP连接，读取传入的数据，并将它发送到需要展示的区域。建立连接或者读入数据通常都是阻塞操作。在处理长途通信时，我们也有一个缺点，每个操作完成的时间可能相对较长。

一个Web浏览器通常都是抓取HTML页面然后展示它。为了尽量隐藏通信延迟，一些浏览器开始会一边接受数据，一边进行展示。当文本对用户可用，或者使用滚动工具等，浏览器将继续获取组成页面的其他文件，图像等。这些文件和图像将被展示。用户不用等到页面所有组件都被获取才能看到页面。

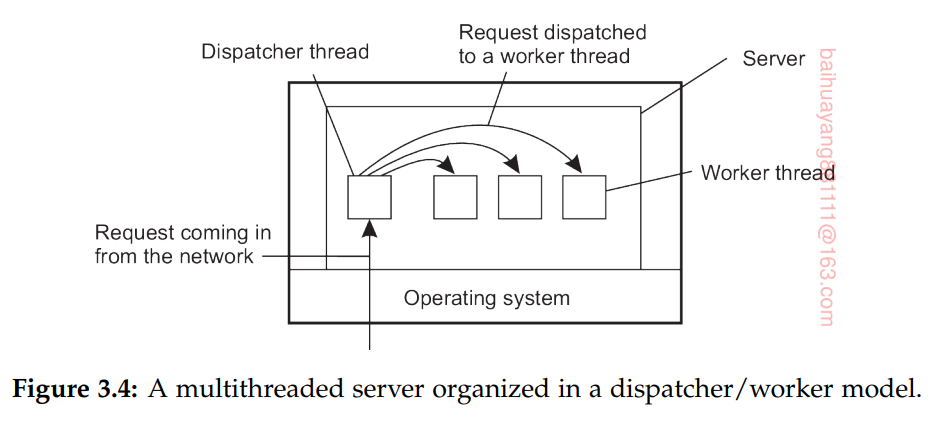
实际上，Web浏览器在同时做一系列任务。事实证明，将浏览器开发为多线程客户机大大简化了工作。当主要的HTML文件被获取时，其他的线程会被激活并获取其他部分。每个线程都单独和服务器建立连接并抓取数据。建立连接和读取数据可以使用标准系统调用，这里假定一个阻塞调用不会暂停所有的进程。同样Stevens提到，每个线程的代码都很简单。与此同时，用户只注意到图像显示的延迟等等，但是可以浏览文档。Web浏览器使用多线程的另一个好处是可以同时打开多个连接。在之前的例子中，和服务端建立了多个连接。如果这个服务器超负荷了，那么这种获取方式将不会提高性能。

然而，在许多例子中，Web服务器会复制到多个机器，每个服务器提供相同的Web文档。这些复制的服务器位于同一个站点，并且都叫同一个名字。当一个Web请求到来时，这个请求会来到其中的某一个服务器，这通常会使用轮询或者其他负载均衡的技术。当使用多线程客户端时，连接会在多个副本上建立，允许数据并行传输，这与不适用复制服务器相比，整个Web文档的传输要快很多。只有当客户机能够处理真正并行的传入数据流时，这种方法才有可能。线程是实现此目的的理想方法。

**多线程服务器**

虽然多线程在客户端具有很好的优势，但更主要的应用是放在了服务端。实践表明，多线程不仅大大简化了服务器代码，而且使开发利用并行性获得高性能的服务器变得更加容易，甚至在单处理器系统上也是如此。然而，随着现代多核处理器的发展，多线程并行成为了一种主要的实现方式。

为了了解服务端多线程的有点，考虑一个偶尔会因为硬盘等待儿阻塞的文件系统。这个文件服务器通常会等待一个输入请求，进而获取请求，然后返回一个应答。如图3.4所示。这里一个线程，或者说是**分配器**，读取进入的请求。紧接着请求被发送到服务器。检查这个请求后，服务器挑选一个空闲的**工作线程**来处理请求。



这个工作线程会执行一个本地读取操作，这是阻塞的，可能会使得获取数据的线程暂停。如果线程暂停，其他的线程将会继续执行。这样，分配线程可以执行更多的工作。作为替代，另一个工作线程会被选择来继续运行。

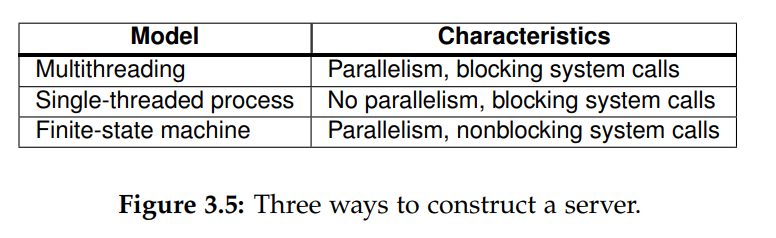
现在考虑服务没有使用线程的情况。一个可能就是讲服务器处理为单线程了。主要的循环即为获取请求，检测，处理返回，然后再进行下一个请求。当在硬盘上等待时，服务器是空闲的并且没有处理其他的请求。通常，其他客户端的请求不会被处理。此外，如果文件服务器运行在专用机器上(通常是这种情况)，那么在文件服务器等待磁盘时，CPU只是处于空闲状态。这样处理的结果是更少的请求会被执行。线程虽然获得了可观的性能，但是每个线程都是顺序执行的。

到目前为止，我们看到两种设计：多线程文件服务和单线程文件服务。第三种选择是将服务器作为大型单线程有限状态机运行。当请求到达时，唯一的线程将会检测它。如果可以从内存缓存中得到满足，那很好，但是如果不能，线程必须访问磁盘。然而，和磁盘阻塞不同的是，线程会调度一种异步磁盘操作，即之后会被操作系统中断。为了使它工作，线程会记录请求的状态，并持续观察之后是否会有其他的请求。（不是多个线程，而是单一线程接收请求，系统调用一个神奇的异步操作访问硬盘，之后发出中断来通知线程。）

当一个磁盘操作完成时，一个操作系统将会通知线程，并查找关联的请求状态并进一步执行它。最终，一个响应将会被一个非阻塞网络调用返回给客户端。

在这个设计中，我们在前两种情况下拥有的“顺序过程”模型丢失了。每次线程需要执行阻塞操作时，它都需要准确地记录它在处理请求时的位置，还可能存储其他状态。一旦完成这些工作，就可以开始操作并继续进行其他工作。其他工作意味着处理新到达的请求，或者处理先前启动的操作已经完成的请求。当然，如果没有其他工作，线程将会阻塞。实际上，我们正在艰难地模拟多线程及其各自堆栈的行为。该流程作为一个有限状态机进行操作，该有限状态机获取一个事件，然后根据其中的内容对其做出反应。

现在我们能明白线程提供了什么好处了。它可以同时实现如同进程一样的顺序执行或者执行并行。阻塞系统调用使编程变得更容易，因为它们看起来就像普通的过程调用一样。另外，多线程允许并发执行并提高性能。单线程服务器保留了阻塞系统调用的易用性和简洁性，但就每个时间单元可以处理的请求数量而言，它可能严重阻碍性能。有限状态机方法通过并行实现高性能，但使用非阻塞调用，这通常很难编程，因此很难维护。这些模型的总结如图3.5.



同样，我们也可以使用多进程来代替多线程。它的优势是操作系统可以提供进程维度的数据隔离。而，如果进程需要进行大量通信，那么与使用线程相比，我们可能会看到对性能的明显负面影响。