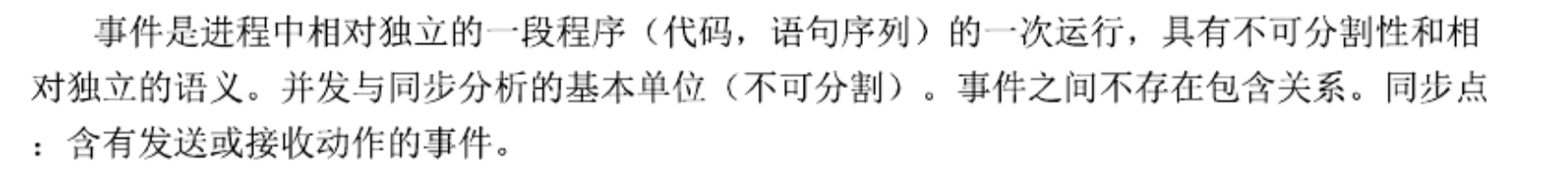
一、事件和时钟

1. 形式化定义
2. 为啥设计和实验中要用到这个东西
3. 设计理念，评判

事件定义：假设进程是算法，而事件是它们执行过程中表现的特定行为。

时钟的定义：时钟在一般意义上指的是一个计算机的物理时间，每个计算机都会包括他们自己的物理时钟，不同的计算机的物理可能会不同。时钟仅是赋予数字给一个事件的方法。这个数字被认为是事件发生的时间。更精确一点，我们将每个进程Pi的时钟Ci定义成一个函数。这个函数赋予数字Ci(a)给进程Pi的事件a。

为啥设计和实验中要用到这个东西

现实生活中物理时间有统一的标准，而分布式系统中每个节点记录的时间并不一样，即使设置了 [NTP](http://www.zhihu.com/question/24960940) 时间同步节点间也存在毫秒级别的偏差[1][2]。因而分布式系统需要有另外的方法记录事件顺序关系，这就是逻辑时钟(logical clock)。

[Leslie](https://en.wikipedia.org/wiki/Leslie_Cheung) [Lamport](https://en.wikipedia.org/wiki/Leslie_Lamport) 在1978年提出逻辑时钟的概念，并描述了一种逻辑时钟的表示方法，这个方法被称为Lamport时间戳。

钟分为逻辑时钟与向量时钟

向量时钟方法在分布式系统中用于保证操作的有序性和数据的一致性。

相对于其他方法，向量时钟的主要优势在于：

节点之间不需要同步时钟，即不需要全局时钟。

不需要在所有节点上存储、维护一段数据的版本数。

设计理念，评判

有的时候计算机上的物理时间往往不同，但是可以自己定义一套逻辑时间的规范，来记录消息事件的有序性。典型的方法就是向量时钟表示法，(V[1], V[2], V[3],…)，V[i]代表的是第Pi进程的已发生事件数。通过比较向量值的大小得出时间event的顺序，只有每个V[i]<V[j]才能定义前者小于后者，此算法的缺点是太占存储空间了，每个时间上存储了所有的进程的时间数的表示。

二、系统设计的评价指标

A吞吐量B延时（什么情况下会延时）C资源利用率D吞吐率（I/O吞吐率）

1. 它们的数学定义是什么

2. 定义，它们相互之间有什么关系

3. 还有什么模式提供系统服务：a面向事件b 非阻断c异步，它们的服务机制是什么，工作原理是什么（给出示例代码）

吞吐量：吞吐量是指系统在单位时间内处理请求的数量。对于无并发的应用系统而言，吞吐量与响应时间成严格的反比关系，实际上此时吞吐量就是响应时间的倒数。前面已经说过，对于单用户的系统，响应时间（或者系统响应时间和应用延迟时间）可以很好地度量系统的性能，但对于并发系统，通常需要用吞吐量作为性能指标。

延时（什么情况下会延时）：延时是一个时间概念，指比原来自然状态下延长了一段时间，在不同的时间中具体表现不同。延时又可称为响应时间，响应时间是指系统对请求作出响应的时间。由于一个系统通常会提供许多功能，而不同功能的处理逻辑也千差万别，因而不同功能的响应时间也不尽相同，甚至同一功能在不同输入数据的情况下响应时间也不相同。

资源利用率：资源利用率反映的是在一段时间内资源平均被占用的情况。对于数量为1的资源，资源利用率可以表示为被占用的时间与整段时间的比值；对于数量不为1的资源，资源利用率可以表示为在该段时间内平均被占用的资源数与总资源数的比值。

吞吐率：吞吐率是一种关于计算机或数据通信系统(如网桥、路由器、网关或广域网连接等)数据传输率的测度。吞吐率通常是对一个系统和它的部件处理传输数据请求能力的总体评价。简单的说吞吐率就是指在一指定时间内由一处传输到另一处或被处理的数据量。以太网吞吐率的单位为“兆比特每秒”或“Mb/s”。

2. 定义，它们相互之间有什么关系

响应时间和吞吐量之间的关系是什么?

吞吐量图显示的是虚拟用户每秒钟从服务器接收到的字节数。当和[响应时间](https://www.baidu.com/s?wd=%E5%93%8D%E5%BA%94%E6%97%B6%E9%97%B4&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1dhnH79nH0dm1DduWRzn1bY0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3ErH6znHmsnWDLPW63rjTzPHn4)比较时，可以发现随着吞吐量的降低，[响应时间](https://www.baidu.com/s?wd=%E5%93%8D%E5%BA%94%E6%97%B6%E9%97%B4&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1dhnH79nH0dm1DduWRzn1bY0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3ErH6znHmsnWDLPW63rjTzPHn4)也降低，同样的，吞吐量的峰值和最大[响应时间](https://www.baidu.com/s?wd=%E5%93%8D%E5%BA%94%E6%97%B6%E9%97%B4&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1dhnH79nH0dm1DduWRzn1bY0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3ErH6znHmsnWDLPW63rjTzPHn4)差不多在同时出现。

本来吞吐量和时延是两个完全不同的概念，似乎它们应当是彼此无关的。然而，吞吐量和时延却是密切相关的。当网络的吞吐量增大时，分组在[路由器](https://www.baidu.com/s?wd=%E8%B7%AF%E7%94%B1%E5%99%A8&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1YYmhRLPjFbnHPBuHRdrym0IAYqnWm3PW64rj0d0AP8IA3qPjfsn1bkrjKxmLKz0ZNzUjdCIZwsrBtEXh9GuA7EQhF9pywdQhPEUiqkIyN1IA-EUBtvnW61PjDzPHRLnjRvnWmYnjf)中等待转换时就会经常处在更长的队列中，因而增加了排队的时间。这样，时延就会增大。当吞吐量进一步增加时，还可能产生网络的拥塞（见教材的5.3节）。这时整个网络的时延将大大增加。可见吞吐量与时延的关系是非常密切的。

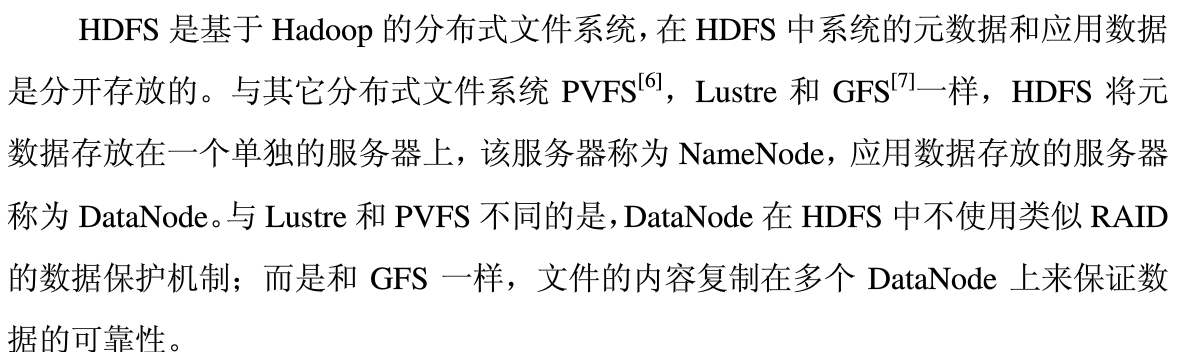
3. 还有什么模式提供系统服务：a面向事件b 非阻断c异步，它们的服务机制是什么，工作原理是什么（给出示例代码）

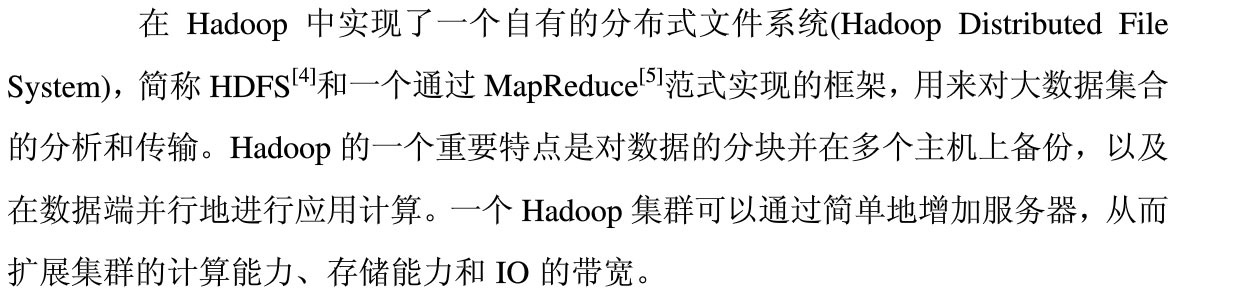
　系统服务表示开机后apache和mysql会以系统服务的形式自动生效，非服务模式需要打开phpstudy才可以生效。

[安全模式](https://www.baidu.com/s?wd=%E5%AE%89%E5%85%A8%E6%A8%A1%E5%BC%8F&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1YkPyNbryR4PHf3nWcvPv7h0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPHmLnHm1Pjfz)是[Windows操作系统](https://www.baidu.com/s?wd=Windows%E6%93%8D%E4%BD%9C%E7%B3%BB%E7%BB%9F&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1YkPyNbryR4PHf3nWcvPv7h0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPHmLnHm1Pjfz)中的一种特殊模式，经常使用电脑的朋友肯定不会感到陌生，在安全模式下用户可以轻松地修复系统的一些错误，起到事半功倍的效果。安全模式的工作原理是在不加载第三方设备驱动程序的情况下启动电脑，使电脑运行在系统最小模式，这样用户就可以方便地检测与修复计算机系统的错误。

三、大数据热门技术：Hadoop的HDFS文件系统 和 RAID传统的文件备份机制

1. 可靠性之间有什么区别

HDFS可靠性：

RAID备份可靠性：

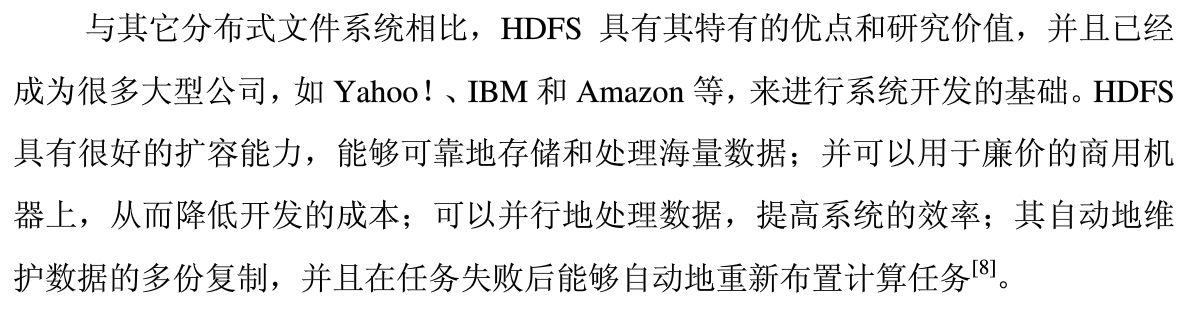
我们知道，Raid技术能够通过纠错编解码来实现文件存储的可靠性。一个磁盘的数据发生损坏，可以通过纠错编解码来恢复该数据。虽然它也会产生多余的校验码parity，但是原数据本身并不会存成多份。作为另外一种提高可靠性的理念，Raid技术也被引入到了分布式文件系统中，Google新一代文件系统Colossus便采用了Raid技术(Reed-Solomon纠错编解码)实现了更经济的可靠性，

RAID是一种把多块独立的硬盘（物理硬盘）按不同的方式组合起来形成一个硬盘组（逻辑硬盘），从而提供比单个硬盘更高的存储机能和提供数据备份技术。组成磁盘阵列的不同方式成为RAID级别（RAID Levels）。在系统中，RAID被看作是一个逻辑分区，但是它是由多个硬盘组成的（最少两块）。它通过在多个硬盘上同时存储和读取数据来大幅提高存储系统的数据吞吐量（Throughput），而且在很多RAID模式中都有较为完备的相互校验/恢复的措施，甚至是直接相互的镜像备份，从而大大提高了RAID系统的容错度，提高了系统的稳定冗余性，这也是Redundant一词的由来。数据备份的功能是在用户数据一旦发生损坏后，利用备份信息可以使损坏数据得以恢复，从而保障了用户数据的安全性。

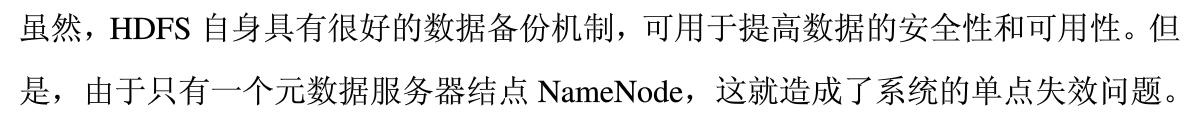
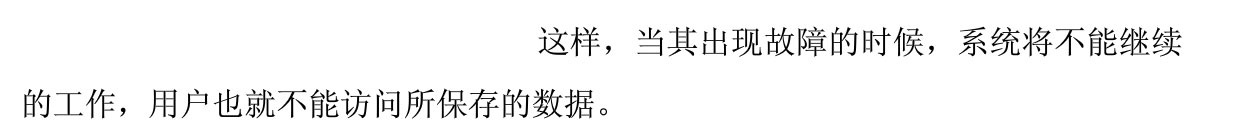
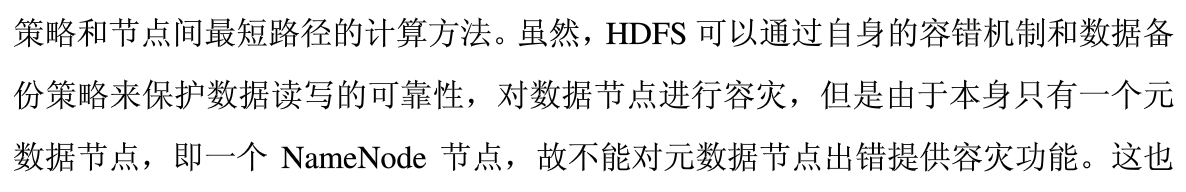
RAID如何实现数据存储的高稳定性呢？我们不妨来看一下它的工作原理。RAID按照实现原理的不同分为不同的级别，不同的级别之间工作模式是有区别的。整个的RAID结构是一些磁盘结构，通过对磁盘进行组合达到提高效率，减少错误的目的，

RAID技术分为几种不同的等级，分别可以提供不同的速度，安全性和性价比。根据实际情况选择适当的RAID级别可以满足用户对存储系统可用性、性能和容量的要求。常用的RAID级别有以下几种:NRAID，JBOD，RAID0，RAID1，RAID0+1，RAID3，RAID5等。目前经常使用的是RAID5和RAID(0+1)。

2. 分别的优缺点

Hadoop的HDFS文件系统优点

Hadoop的HDFS文件系统缺点：



RAID备份优点：

RAID技术的两大特点：一是速度、二是安全

最初开发RAID的主要目的是节省成本，当时几块小容量硬盘的价格总和要低于大容量的硬盘。目前来看RAID在节省成本方面的作用并不明显，但是RAID可以充分发挥出多块硬盘的优势，实现远远超出任何一块单独硬盘的速度和吞吐量。除了性能上的提高之外，RAID还可以提供良好的容错能力，在任何一块硬盘出现问题的情况下都可以继续工作，不会受到损坏硬盘的影响。

1 、成本低，功耗小，传输速率高。

2 、可以提供容错功能。

3、 RAID的另一特征是具备数据校验(Parity)功能，校验可被描述为用于RAID级别2，3，4，5的额外的信息，当磁盘失效的情况发生时，校验功能结合完好磁盘中的数据，可以重建失效磁盘上的数据。

4 、RAID比起传统的大直径磁盘驱动器来，在同样的容量下，价格要低许多。

RAID备份缺点：

RAID能够让你的磁盘系统更快，容量更大，但是只是在你其中的一块磁盘数据丢失后会起到保护的作用，他并不是一个到位的备份系统，

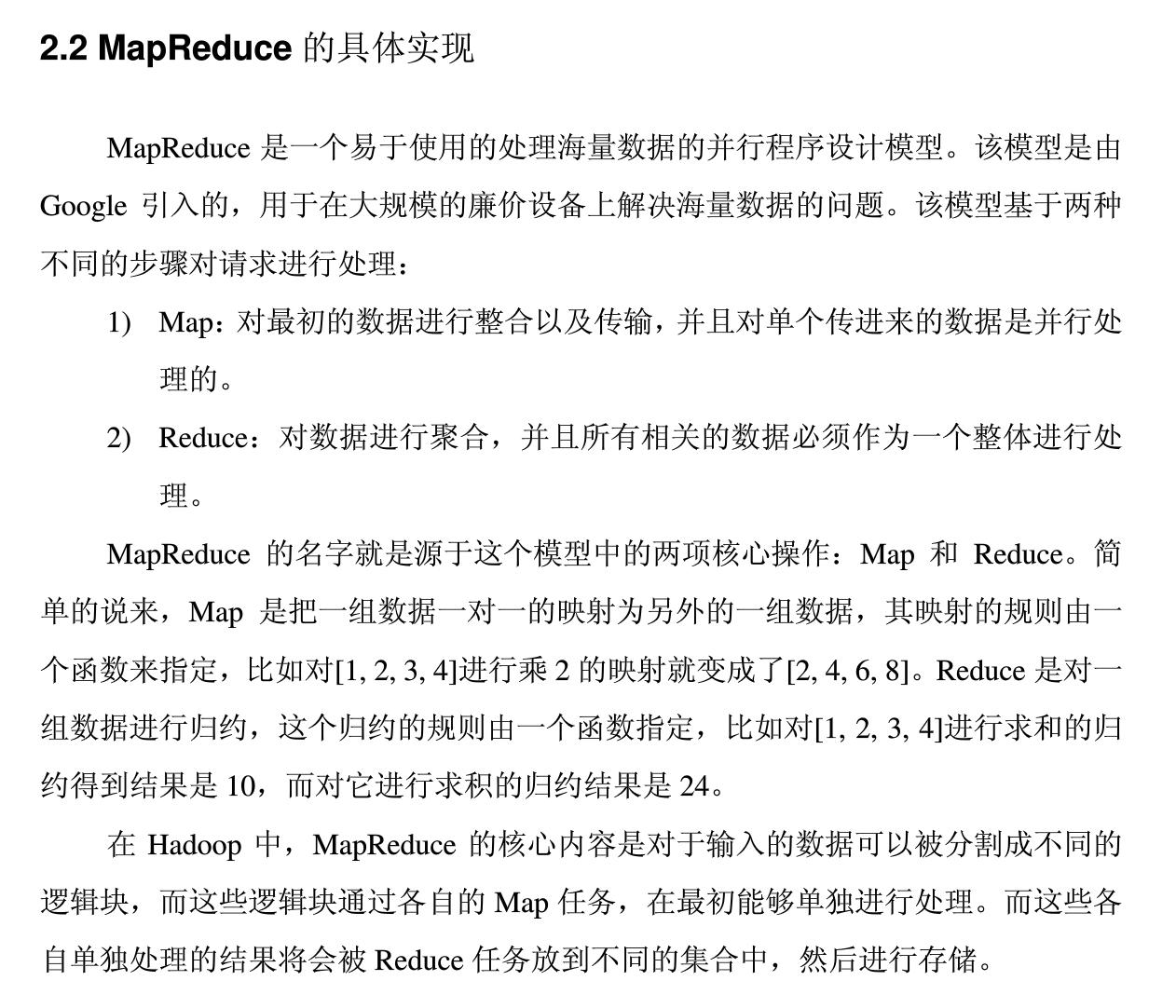
1、RAID0没有冗余功能,如果一个磁盘(物理)损坏,则所有的数据都无法使用.

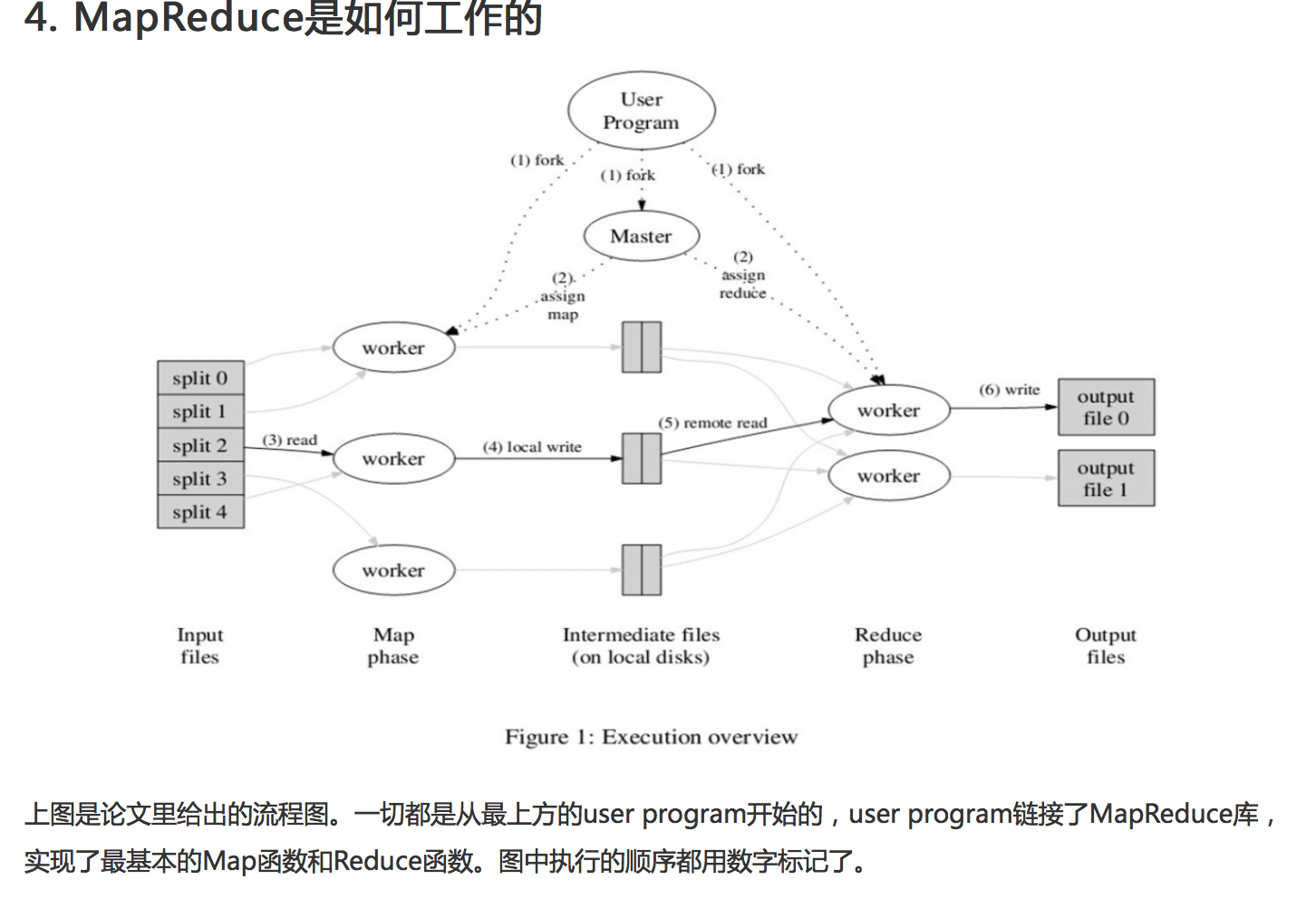
2、RAID1磁盘的利用率最高只能达到50%(使用两块盘的情况下),是所有RAID级别中最低的.

3、RAID0+1以理解为是RAID 0和RAID 1的折中方案.RAID 0+1可以为系统提供数据安全保障,但保障程度要比 Mirror低而磁盘空间利用率要比Mirror高.

四、计算模式 Map Reduce 和 对应的Open API

1. 有什么区别（用代码结构来说明）

Map Reduce：





Open API：

Open API即开放API,也称开放平台。 所谓的开放API（OpenAPI）是服务型网站常见的一种应用，网站的服务商将自己的网站服务封装成一系列API（Application Programming Interface，应用编程接口）开放出去，供第三方开发者使用，这种行为就叫做开放网站的API，所开放的API就被称作OpenAPI（开放API）。

根据开放层级划分，开放平台主要分为四层，即“硬件”层、系统层、业务层和应用层。其中，业务层开放和应用层开放是更加贴近互联网层面的开放方式。

根据开放技术划分，可将开放平台划分为五种类型：OpenAPI 型开放平台、插件式开放平台、综合型开放平台、应用超市型开放平台和基础服务型开放平台（云计算平台）。

五、数据库管理

No SQL 和 关系型数据库

1. 定义
2. 比较优缺点
3. 适用于什么场景（什么时候用）

No SQL 定义：

  NoSQL，泛指非关系型的数据库。随着互联网[web2.0](http://baike.baidu.com/view/733.htm)网站的兴起，传统的关系数据库在应付web2.0网站，特别是超大规模和高并发的[SNS](http://baike.baidu.com/subview/8258/5896174.htm)类型的web2.0纯[动态网](http://baike.baidu.com/view/528572.htm)站已经显得力不从心，暴露了很多难以克服的问题，而非关系型的数据库则由于其本身的特点得到了非常迅速的发展。NoSQL数据库的产生就是为了解决大规模数据集合多重数据种类带来的挑战，尤其是大数据应用难题。

关系型数据库 定义：

关系数据库，是建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据。现实世界中的各种实体以及实体之间的各种联系均用关系模型来表示。现如今虽然对此模型有一些批评意见，但它还是[数据存储](https://www.baidu.com/s?wd=%E6%95%B0%E6%8D%AE%E5%AD%98%E5%82%A8&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1YznW03nWc3mWwBPHm4Pvcd0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPHTYrH0krHfv)的传统标准。标准数据查询语言SQL就是一种基于关系数据库的语言，这种语言执行对关系数据库中数据的检索和操作。 关系模型由关系数据结构、关系操作集合、关系完整性约束三部分组成。

请列举关系型数据库相对于其它类型的数据库的优缺点？

  答：关系数据库相比于非关系型数据库在数据中添加了逻辑结构关系和约束，并且结构单一，数据操作方便，更容易被用户接受，应用也比较广泛。缺点：数据类型表达能力差，复杂查询功能差，支持长事务能力差，环境应变能力差

NoSQL 适用场景：

NoSQL数据库在以下的这几种情况下比较适用：1、数据模型比较简单；2、需要灵活性更强的IT系统；3、对数据库性能要求较高；4、不需要高度的数据一致性；5、对于给定key，比较容易映射复杂值的环境。

关系型数据库 适用场景：

   主机操作系统为window，主要用于web网站的建设，承载中小型web后台数据。

    在租赁的虚拟主机中一般会预安装SQL Server作为数据库软件。