### 镜像容灾技术方案

一、    容灾技术的选择
（一）基于SAN城域网的镜像技术
目前实现容灾的技术方案看起来很多，不过目前比较通用的和实用的容灾技术，主要有两类，一种是基于主机的跨阵列数据镜像容灾方案，一种是基于阵列系统的数据复制容灾方案。
而随着光纤存储网络技术的成熟和在距离上的拓展，光纤城域存储网络的实现已经趋于成熟，这就使得我们今天可以不再需要基于阵列系统的数据复制技术，就可以实现同城容灾了。新的容灾方案所利用的是最为传统的基于主机的跨阵列数据镜像技术，也就是说我们可以利用基于城域SAN存储网上的镜像技术，轻松实现数据容灾，然后在此基础上，我们可以利用先进的集群软件，构建应用级的容灾系统。
1）为什么我们建议选择基于主机的镜像技术容灾方案？
λ    零停机、零切换: 基于镜像的原理，镜像中的任何一个磁盘阵列出现问题停顿时，都不会导致应用中断。从而使容灾的效果能够达到无缝的数据高可用性，对业务系统的高可用性提供了保障。当然没有停机，也就没有切换操作，也就不会造成切换带来的一系列麻烦。
λ    成熟性：基于主机镜像技术是从磁盘容错技术中最成熟、历史最悠久、可靠性最高的数据保护技术而来。
λ    简单性：镜像技术是实现最为简单的数据容错技术。
λ    异构性：基于主机镜像技术不仅可以在磁盘阵列内部实现，也可以跨越不同磁盘系统来实现，镜像技术完全不依赖于磁盘系统的品牌和型号。所以，利用镜像技术实现容灾，我们就可以保留在任何时候自由、灵活的选择磁盘系统的权利。
2）为什么以前比较少用主机镜像做容灾，而用阵列镜像或复制技术容灾？
λ    在SCSI技术时代，SCSI只能支持25米的距离，所以虽然基于阵列系统间的复制被广泛采用，但无法实现远距离容灾
λ    复制技术可以很好的解决远距离和低带宽带来的问题，当然在解决低带宽问题时，普遍采用的技术的是异步复制方案，因为同步复制带来的I/O延时是任何应用都无法接受的。所以一直以来，是容灾方案的主要选择。
λ    FC光纤存储出现以后，为远距离镜像提供了可能，随着同城光纤存储网络的普及，利用镜像技术构建同城容灾系统，将是一个既简单、又实用的容灾解决方案。
3）为什么现在不建议采用基于阵列系统的复制来构建容灾系统？
λ    数据复制技术所能解决的距离和带宽的问题，当前已经不再是问题了。
λ    基于复制的容灾会机制，在灾难发生时将导致停机、切换甚至是数据一致性风险，从容灾效果上来说不很理想。
λ    利用数据复制技术实现容灾，由于其技术本身的局限性，必然导致容灾系统的结构要比没有容灾系统的时候复杂得多。
λ    同时，这样的容灾系统的故障环节不可避免的会增加，导致维护难度、维护成本的增加。
λ    基于磁盘系统的数据复制技术，要求生产中心和灾备中心选用同一个厂商的磁盘系统，甚至要求选用同一系列、同一型号的磁盘系统，用户完全失去讨价还价的余地，完全失去了在任何时候自由、灵活地选择合作厂商和磁盘系统的权利，必然给用户造成被动和经济上的损失。
（二）传统网络传输与SAN 网络传输的主要区别
λ    数据传输的延时，SAN网络与传统网络有着数量及的差别，举例来说，如果传统网络的延时为毫秒级，SAN网络的延时则为纳秒级。由于SAN网络中的延时远远小于传统网络，因而其数据传输速度，远远快于传统网络。
λ    不同协议的封包结构，导致传输效率的差别。传统网络协议如TCP/IP协议的数据传送的封包结构，比SAN 网络上的协议如FCP要复杂，也导致了数据传输速度，远远慢于SAN网络。
λ    因此，我们可以简单的得出结论，即便是在同样的物理带宽下，采用SAN网络方式的数据传输，其速度，稳定性都要远远优于传统的网络。
综上所述，既然可以采用基于光纤的城域SAN网络，那么，我们就没有必要再采用过时的基于磁盘系统数据复制的容灾方案，采用基于主机镜像技术来实现容灾方案，是一个更好的选择。
（三）构建容灾系统的技术要求
基于磁盘镜像技术构建容灾系统，在数据系统容错实现的逻辑上，和本地磁盘系统间的镜像没有任何区别，所以我们大可以放心的去构建这样的系统。
但是，在实际应用环境中，我们还是发现了一个不同的地方，是我们在考虑容灾方案的时候必须解决的，那就是距离。不是说我们已经解决了距离的问题吗？是的，我们在技术上已经解决了远距离数据镜像的问题，但是，有几个问题是任何技术都必须考虑的，不管是镜像技术还是复制技术，包括：
λ    距离决定了链路需要更高的可靠性。
我们在短距离（比如一个机房内）可以忽略不计的链路故障（可以通过链路冗余解决）在长距离上就无法忽略了，比如，我们无法保证我们的光纤不会被野蛮施工挖断。
λ    距离会影响我们管理一个系统的复杂性和现场感
远距离灾备系统永远都是生产中心的附属品，如果我们不能将其用直观的方式纳入到日常的管理体系中，很快我们就会发现我们对生产系统和灾备系统的管理会脱节，所以，容灾系统肯定不仅仅是镜像那么简单，他需要一个完整的、直观的管理界面去管理，才能形成一个行之有效有的灾备系统。
所以，我们在构建容灾系统的时候，必须考虑由距离引起的问题给我们的技术方案造成的各种负面影响，并要很好的解决它。有关问题和解决方法，我们会在下一节中详细论述，这里我们先简单的把要求提出来，供参考。
λ    对磁盘镜像容灾管理的要求：可视化磁盘管理和高度可控的容灾管理。
λ    应对链路故障引起的各种问题：实现增量镜像和避免Split-Brain(脑裂)等
二、    易腾数信基于镜像技术的容灾方案及其实现
（一）数据同城容灾方案原理分析
1）易腾数信远程镜像数据容灾原理
易腾数信建议利用EterneMirrorCluster系列软件的磁盘镜像技术，来构建容灾方案。利用EterneMirrorCluster的镜像技术构建容灾系统是非常简单的，它只有一个条件，就是将生产中心和灾备中心之间的SAN存储区域网络通过光纤连接起来，建立城域SAN存储网络。然后，我们就可以通过EterneMirrorCluster提供的非常成熟的跨阵列磁盘镜像技术来实现同城容灾了，容灾方案的结构如下图所示：
 

从镜像原理上讲，在城域SAN存储网络上的两套磁盘系统之间的镜像，和在一个机房内的SAN上的两个磁盘系统之间的镜像并没有任何区别。
利用裸光纤将生产中心和灾备中心的SAN网络连接起来，构成城域SAN网络以后，利用 EterneMirrorCluster的先进的镜像磁盘卷管理功能，EterneMirrorCluster镜像功能把这两个磁盘阵列的盘组合为一个EterneDisk盘，形成磁盘对，确保两个盘数据完全一致（类似于RAID1），这样我们就可以非常方便的实现生产中心磁盘系统和灾备中心磁盘系统之间的镜像了。如下图所示。
 

我们可以看到，利用EterneMirrorCluster，我们可以创建任意一个镜像磁盘卷（EtDisk）供业务主机使用，镜像磁盘卷（EtDisk）实际上是由两个完全对等的，容量相等的磁盘片或分区构成的，通过镜像磁盘卷保证两个磁盘片上的数据完全一样，业务主机对该镜像磁盘卷的任意修改，都将同时被写到位于生产中心和灾备中心的两个磁盘系统上。
采用这种方式，生产中心的磁盘阵列与同城容灾中心的磁盘阵列对于两地的主机而言是完全同等的。利用城域SAN存储网络和EterneMirrorCluster镜像功能，我们可以非常轻松的实现数据系统的异地容灾。并且消除了复制技术（无论是同步还是异步）的切换的动作，从而保证零停机时间，零数据损失的实现。
并且除了数据的同步镜像外，业务应用也可利用EterneMirrorCluster实现应用在生产中心和容灾中心间的高可用。也就是说，生产中心的业务服务器如果发生宕机或业务程序中断，应用可以自动切换到生产中心的备机，或者切换到容灾中心的服务器上。EterneMirrorCluster可高效的保证了数据和应用的安全。
2）磁盘阵列同步复制原理及局限性
在同城的容灾方案中，也可利用阵列提供的软件功能实现基于磁盘阵列的同步复制方案，这也是较为流行的一种。其具体原理如下：


以上的原理图清楚地指出了基于磁盘阵列的同步复制方案的原理，其原理适用于所有类型的磁盘阵列间的同步复制原理。因此，我们可以看到，当生产中心的阵列发生问题时，会产生如下一系列的问题：
1． 主机必须手工的将I/O 路径切换到容灾中心的阵列上，这将导致无法避免的停机时间和用户的业务停顿。
2． 阵列的切换操作，是直接在磁盘阵列上进行的，在发生磁盘阵列问题，也就是在用户业务处于停顿状态的时候，找什么人来做这个操作（此操作磁盘阵列厂商通常建议由厂商或是由资质的代理来做。），或是用户自己来做此操作，都是对操作人员是非常严峻的考验。无形中也就增加了用户的风险。
3． 当整个过程进行到第4或第5步时，生产中心磁盘阵列突然发生故障，主机将认为数据没有写入磁盘阵列，而实际上数据已经写入容灾中心磁盘阵列，一旦启用容灾中心磁盘阵列，将导致数据不一致风险。
4． 如果用户选择的是在不同品牌的磁盘阵列之间实现镜像，基于磁盘阵列的同步复制方案更是无能为力了。
5.  如果用户有业务高可用的需求，此方案无法满足生产中心和容灾中心直接的应用高可用自动切换。
因此，无论从容灾效果，还是从易操作，易管理的角度，基于磁盘阵列的同步复制方案，都不是同城容灾（基于SAN）的最佳选择。
3）操作系统镜像命令作用解释
我们都知道，操作系统如windows其本身有如软RAID方式来实现镜像。同样是镜像，为什么不能作为同城数据容灾的方案呢？
1． OS 镜像还有一个无法解决的问题就是，一旦镜像打破，再修复镜像时，其重新同步，需要重新做全同步，对系统影响极大。在实际经验中我们发现，重新同步的时间，跟系统停机的时间几乎没有什么区别。
2． 对于异构的磁盘阵列环境，OS不支持对不同品牌的多路径（主机到磁盘阵列的光纤路径）管理。因为不同阵列厂商提供各自的多路径管理软件，而不同的多路经管理软件是不能在同一主机上正常工作的，因此无法真正意义上实现异构磁盘阵列间的镜像。

（二）易腾数信远程镜像数据容灾系统故障和灾难的响应
一个完整的灾备系统，除了在数据灾难发生时，能够完成灾备的使命，需要考虑灾备系统本身的可维护性和可操作性，以及对系统尽可能快的恢复。下面，我们来看一下，数据系统故障和灾难情况下，易腾数信的远程镜像容灾系统是怎样响应的。
1）当生产中心数据系统故障
生产中心数据系统故障意味着灾难，磁盘故障，链路故障，或者数据系统的计划内停机时间，也就是一切导致主机无法访问生产中心数据系统的情况。我们来看一下我们推荐的容灾方案是如何响应的，见下图：
 

当生产中心的磁盘系统发生故障（灾难）时，由于同城容灾中心的磁盘是它的镜像，所以镜像磁盘(EtDisk）会自动隔离生产中心的磁盘，转而对容灾中心的数据进行访问。而对操作系统而言没有任何的变化。从上图我们看到，业务系统可以通过城域SAN网络利用镜像磁盘(EtDisk)直接访问灾备中心的磁盘系统的数据，而不需要有任何针对业务系统的动作。也就是说，生产中心磁盘系统的灾难，对业务系统是透明的，应用和数据库不会因为生产中心磁盘系统的故障而停止；更重要的是，因为应用和数据库不会因为灾难而异常中止，从而避免了发生数据库损坏（数据一致性风险）的可能。
值得注意的是：整个过程对应用完全透明，不需要也不会中断业务系统的正常运行。这是基于磁盘系统间复制技术构建的容灾系统无法实现的。
2）灾备中心数据系统故障以及生产中心和灾备中心SAN链路故障
灾备中心数据系统故障，以及生产中心到灾备中心的链路故障，我们都可以把其看成是容灾部分的故障，其原理和后果与生产中的数据系统故障相同。都是导致了镜像的破坏。而后，系统将自动的只与状态健康的磁盘阵列继续工作。整个过程对应用和操作系统完全透明。
3）故障修复后的恢复（远程镜像快速恢复）
磁盘系统故障修复之后，我们需要尽可能快的将远程镜像系统恢复起来，以确保容灾的功能继续得以实现，同时，在整个镜像恢复的过程中，势必会对应用造成影响。因为磁盘数据的同步，一定会造成I/O的极度繁忙而导致应用性能下降，如果镜像恢复无法快速完成，其后果跟系统应用停机也非常接近了。因此，如何快速有效的实现镜像的重新同步，同样是一个容灾方案是否成功的关键因素。
传统的镜像技术（如OS的镜像技术），在镜像链路被中断以后，中断的镜像会被认为完全作废，在链路恢复以后，我们不得不将数据完整地从生产中心拷贝一份到容灾中心。这种方式，对于用户的应用是无法接受的。链路方面的故障如果经常发生，我们就需要不断的重复将生产中心的数据全部同步到灾备中心的磁盘系统上，实际上，这种方案不具有可实施性和可维护性，是不现实的。这也是为什么主机厂商虽然也有类似镜像功能，但不会用于容灾的的根本原因。
为了解决这个问题，EterneMirrorCluster提供了UUID+BitMap+FastSync技术，其中UUID+BitMap是一种针对镜像磁盘块的日志记录技术，该技术允许EterneMirrorCluster在镜像链路中断后记录磁盘的数据变化情况，以便在镜像链路恢复后，由FastSync技术实现数据的增量恢复。所谓FastSync意思就是“镜像的快速同步”，实现数据的快速增量同步，只同步利用EterneMirrorCluster 的UUID+BitMap技术记录的数据变动的数据块，我们现在可以不用再担心容灾系统本身的可维护性了。利用UUID+BitMap+FastSync，我们的应对步骤如下：
1．    当任何故障发生，导致镜像被破坏。
2．    生产中心的EterneMirrorCluster利用 UUID+BitMap日志记录因业务数据的变化而变化的数据块。
3．    一旦故障被修复，EterneMirrorCluster的FastSync功能模块，会根据 UUID+BitMap日志记录的情况，将链路中断后更新的业务数据（变化量）同步到灾难端实现增量更新。
4．    镜像快速同步的过程中，用户的应用始终可以正常工作。
整个过程的发起是自动完成，或者手动只需要执行一条命令即刻完成。整个过程的速度，由于只是同步增量，时间远远小于整个数据系统的完全同步。从而大大减小对用户应用的影响，这也是传统镜像技术如OS镜像所以不具备的。

（三）易腾数信远程镜像数据容灾方案的技术优势
和其他容灾方案相比，易腾数信容灾方案具有明显的优势，这些优势不仅仅表现在技术实现方面，还表现在开放性、可维护性等各个方面。
1）零停机时间和零数据损失
由于EterneMirrorCluster 采用的是跨异构阵列的镜像技术，就决定了在这种方式下，无论是哪一边的磁盘阵列由于物理故障停顿，都不会影响数据的可用性而造成数据的损失，这从根本上实现了在物理故障的情况下，数据的高度可用性。
2）故障修复后的快速重新同步
EterneMirrorCluster 提供的镜像技术，是基于日志的镜像技术，无论由于主机发生故障，还是由于镜像中的链路或是硬盘发生故障导致的镜像被破坏的情况，都可以通过镜像日至得以快速恢复。这使得镜像恢复过程对系统的性能影响微乎其微。
3）应用的高可用性
EterneMirrorCluster 可完善的提供应用的高可用保护方案，从而提供用户关键业务的24小时在线保障。
4）数据同步过程高度透明性
EterneMirrorCluster提供完整的图形化管理，在数据同步的过程中，可以随时得知同步的进度和状态。
5）节约硬件采购成本
EterneMirrorCluster 支持对不同品牌，不同型号的异构阵列做镜像，这可以帮助用户方便的选择适应自身业务需求的硬件，从而获得最好的价格性能比。

（四）易腾数信远程镜像数据容灾方案的工程优势
λ    结构简单
基于镜像的容灾方案较任何一种容灾方案为简单。比如基于磁盘系统复制技术的容灾方案，必须在磁盘系统内部专门配置相应的接口卡，使用该磁盘系统专有的复制软件，才可以构建容灾系统。明显的，该方法增加了一个在非容灾系统中完全不需要的环节，不仅增加了故障源，同时也增加了维护难度。事实上，基于主机的复制技术存在同样的问题。
在一个使用镜像磁盘卷管理的应用系统中，我们只是利用了其中最常用、最成熟的镜像功能，就可以实现容灾，而不必像其他容灾系统那样增加很多不必要的环节。
λ    技术成熟
镜像技术是比任何数据复制技术更早使用于高可用系统的成熟的功能，这种功能已经广泛应用于包括IBM Mainframe、AS400、RS6000、HPUX、Digital Unix、SUN Solaris、Linux、Windows等在内的所有服务器系统上，同时也被广泛用于磁盘系统内部作为企业级解决方案，象EMC、HDS、HP、IBM、SUN、Compaq等众多存储设备生产厂商，都将镜像技术内置于其磁盘系统内部，用于对数据可用性要求最高的用户群。
λ    存储开放性
利用EterneMirrorCluster的镜像技术，我们构建容灾方案的同时，还构建起一个支持异构磁盘阵列存储虚拟化的平台。不再要求生产系统和灾备系统的存储系统必须是同一个品牌的，对具体型号更没有任何要求，体现了存储平台选择的开放性。由此，用户可以获得在存储平台选择上的主动权，避免被存储厂商“绑架”的尴尬。
λ    技术的完整性
EterneMirrorCluster拥有一个容灾方案必须的所有技术特性，它不仅可以提供可靠的镜像技术，实现跨磁盘系统的数据容灾，同时，我们可以利用EterneMirrorCluster 的UUID+BitMap+FastSync技术，保证该容灾系统的可维护性。
λ    容灾系统的可视化管理
一个灾备系统和生产系统是一个整体，而不是两个孤立的系统，一个系统的可视化管理工具，有利于管理者将两个系统有机的结合起来，而不是分别处理两个独立的系统。
EterneMirrorCluster 提供完善的管理GUI视图，通过任何一台安装有EterneMirrorCluster界面端的系统，我们就可以查看和管理镜像磁盘卷，这意味着，我们可以实现对生产中心和灾备中心存储的统一管理。EterneMirrorCluster管理界面丰富，下图为其中一个管理界面，仅供参考。
 
λ    应对灾难，应用不中断，数据不丢失
易腾数信建议的容灾方案，不会因为生产中心（或者灾备中心）磁盘系统的故障和灾难而导致应用和数据库的异常中止，这不仅避免了对应用系统正常运作的影响，还避免了发生数据库损坏的可能。
相比较而言，如果采用数据复制的方式（无论是基于磁盘系统的硬件复制方式还是基于软件的数据复制方式），都需要在生产中心故障时对数据系统进行切换操作，反而造成业务的停顿。另外，由于在灾难发生时，数据库系统的复制是即刻停止的，数据库系统没有经过正常的Shutdown，所以不仅不可避免的导致部分交易的损失，甚至还有可能导致数据库的损坏。
λ    I/O效率最佳
从性能上来分析，在操作系统一级进行镜像，数据会在同一时间写入到两地的磁盘。数据可写入过程是两组并行的操作，即本地写和完成信号返回+异地写和完成信号返回。
而相比较而言，数据复制技术需要通过以下4个步骤才算写操作完成：
1．    数据先有操作系统写入本地磁盘系统
2．    磁盘系统将数据通过链路复制到异地系统
3．    异地磁盘系统完成写操作，返回信号给本地磁盘系统
4．    本地磁盘系统返回信号给操作系统
明显的，这个过程和直接数据镜像相比，把原先并行的2组步骤变成了纯串行的4个步骤，并且需要SCSI到复制协议的转换过程，无论在流程上和反应时间上都会比直接镜像造成更多的延时，对应用系统有更大的影响。
另外，在镜像磁盘卷这个层面，我们可以根据需要对镜像的数据灵活设置，我们不需将所有磁盘进行镜像，而只需镜像必要的系统分区即可。这种灵活性在实际使用过程中将大大减少数据远程复制的数量，从而避免对系统不必要的影响。
（五）易腾数信远程镜像数据容灾方案对系统性能的影响
大家都比较关心容灾系统建立以后对原有业务系统性能的影响，考察容灾系统对业务系统性能的影响，主要从两个方面衡量，一是CPU资源的消耗，二是I/O，特别是写操作的延迟效应。
λ    CPU资源消耗
采用主机端的软件镜像技术，对CPU资源的损耗，实际上是微乎其微的，但很多时候被磁盘系统厂商人为的夸大了。具体的事实我们可以通过简单的测试得到，我们可以设置这样一个测试，就一目了然了：
1）在测试系统上，往一个没有镜像的磁盘上Copy一个大文件，察看CPU使用率；
2）在测试系统上，往一个有镜像的磁盘上Copy一个大文件，察看CPU使用率。
事实上，处理镜像需要的CPU时间是非常非常小的，原因是磁盘I/O操作的速度是毫秒（ms）级的，磁盘系统Cache I/O的速度是受限于光纤通道的100-200MB（8bit\*10ns）带宽和距离（15公里 == 0.1ms）的，而相反的，高端主机总线的宽度一般是64-128Byte，甚至更高，主机CPU的处理速度更是在千兆的水平（ns级），所以I/O对主机CPU的消耗往往都是可以忽略不计的，如果说需要关心的话，也主要针对象RAID-5这样的技术（需要一定量计算，从而消耗主机的CPU资源），而像镜像这样的技术，是几乎不需要消耗CPU时间的。
λ    I/O的延迟效应（特别是写操作的延迟效应）
采用EterneMirrorCluster的镜像技术构建容灾系统，其对系统 I/O的延迟效应要小于任何一种数据复制技术，不管是基于磁盘系统的硬件数据复制技术，还是基于主机软件的数据复制技术。
这里我想要补充的是在整个容灾系统中，对业务系统的性能的影响最大的不是任何一种技术所产生的负面作用，而是“距离”，正如前面提到的，在Cache命中率较高的系统中，距离对写操作的影响较大，这和光的传播速度有关，光在150公里距离上的一个来回需要1ms，在15KM距离上一个来回需要0.1ms，我们列出一个对照表，供大家参考。本对照表不包含设备协议转换和光在光纤中的折射等因素。同时，我们知道，100MB光纤对应的速度是ns级的。

|  |  |  |  |
| --- | --- | --- | --- |
| **距离** | **光传输距离** | **传输时间** | **说明** |
| 15 km | 30  km | 100us | **同城容灾中心的距离级别** |
| 150 m | 300 m | 1us |  |
| 15 m | 30 m | 100ns | **一个机房的内部距离** |
| 1.5 m | 3 m | 10ns | **一个主板的内部距离** |
| 本地Cache写的时间 | ns~us级 |  |
| **本地磁盘写的时间** | **us~ms级** |  |