**第十四章 RAID磁盘阵列的原理与搭建**

**本节所讲内容：**

**14.1 RAID概念-企业级RAID 0, 1,5,10的工作原理**

**14.2 RAID-0-1-5-10搭建及使用-删除RAID及注意事项**

**14.3 实战：企业中硬件raid5的配置**

# 14.1 RAID概念

**磁盘阵列（Redundant Arrays of Independent Disks，RAID），有“独立磁盘构成的具有冗余能力的阵列”之意。 磁盘阵列是由很多价格较便宜的磁盘，以硬件（RAID卡）或软件（MDADM）形式组合成一个容量巨大的磁盘组，利用多个磁盘组合在一起，提升整个磁盘系统效能。利用这项技术，将数据切割成许多区段，分别存放在各个硬盘上。 磁盘阵列还能利用同位检查（Parity Check）的观念，在数组中任意一个硬盘故障时，仍可读出数据，在数据重构时，将数据经计算后重新置入新硬盘中**

**注：RAID可以预防数据丢失，但是它并不能完全保证你的数据不会丢失，所以大家使用RAID的同时还是注意备份重要的数据**

**RAID的创建有两种方式：软RAID（通过操作系统软件来实现）和硬RAID（使用硬件阵列卡）；在企业中用的最多的是：raid1、raid5和raid10。不过随着云的高速发展，供应商一般可以把硬件问题解决掉。**

## **14.1.1 RAID几种常见的类型**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RAID类型** | | **最低磁盘个数** | **空间利用率** | **各自的优缺点** |
| **级 别** | **说 明** |
| **RAID0** | **条带卷** | **2+** | **100%** | **读写速度快，不容错** |
| **RAID1** | **镜像卷** | **2** | **50%** | **读写速度一般，容错** |
| **RAID5** | **带奇偶校验的条带卷** | **3+** | **(n-1)/n** | **读写速度快，容错，允许坏一块盘** |
| **RAID6** | **带奇偶校验的条带集，双校验** | **4+** | **(n-2)/n** | **读写快，容错，允许坏两块盘** |
| **RAID10** | **RAID1的安全+RAID0的高速** | **4** | **50%** | **读写速度快，容错** |
| **RAID50** | **RAID5的安全+RAID0的高速** | **6** | **(n-2)/n** | **读写速度快，容错** |

**RAID基本思想：把好几块硬盘通过一定组合方式把它组合起来，成为一个新的硬盘阵列组，从而使它能够达到高性能硬盘的要求**

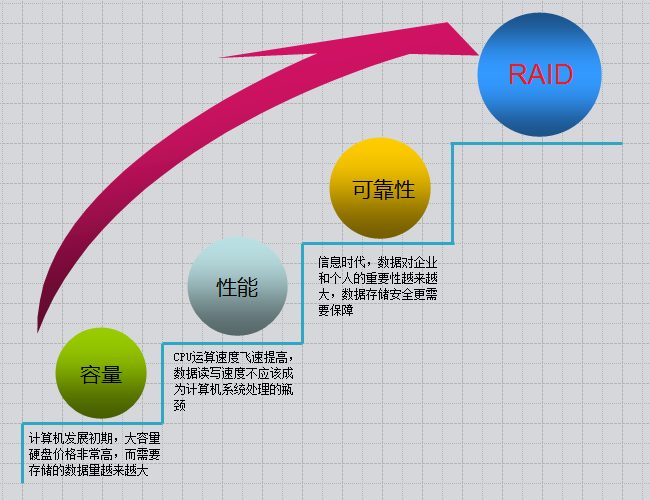
**RAID有三个关键技术：**

**镜像：提供了数据的安全性；**

**chunk条带（块大小也可以说是条带的粒度），它的存在的就是为了提高I/O，提供了数据并发性**

**数据的校验：提供了数据的安全**

**Raid相对于单个磁盘优点：**

****

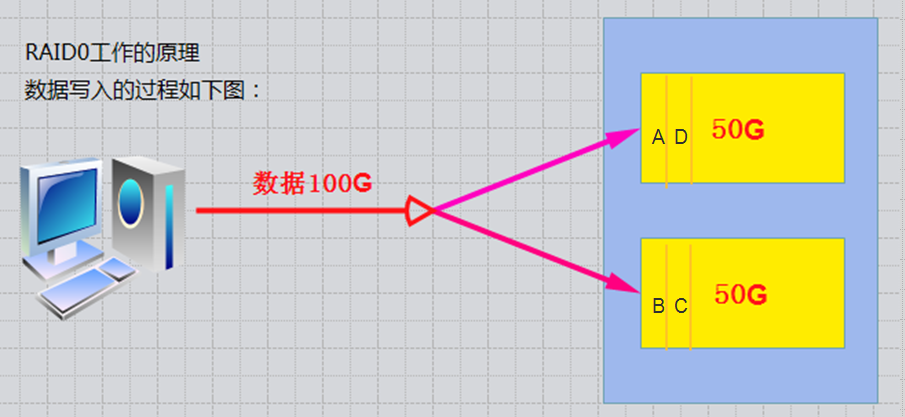
## **14.1.2 RAID-0的工作原理**

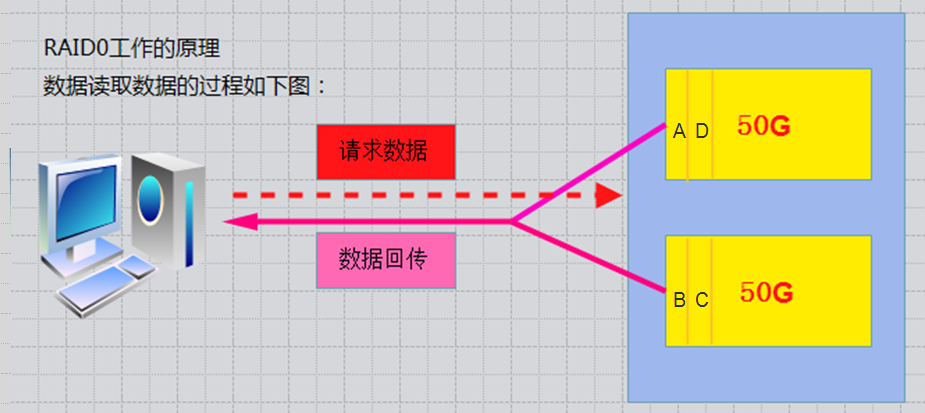
**条带 （strping），也是我们最早出现的RAID模式**

**需磁盘数量:2块以上(大小最好相同)，是组建磁盘阵列中最简单的一种形式，只需要2块以上的硬盘即可.**

**特点:成本低，可以提高整个磁盘的性能和吞吐量。RAID 0没有提供冗余或错误修复能力，速度快.**

**任何一个磁盘的损坏将损坏全部数据；磁盘利用率为100%。**

****

****

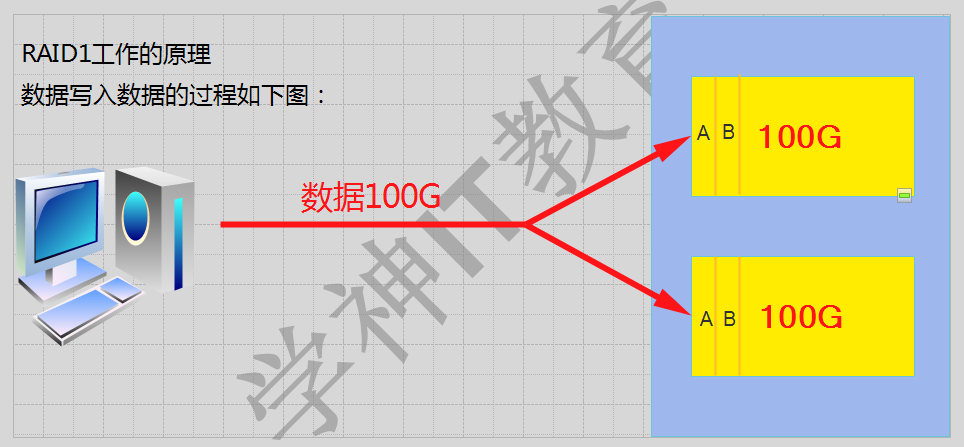
## **14.1.3 RAID-1**

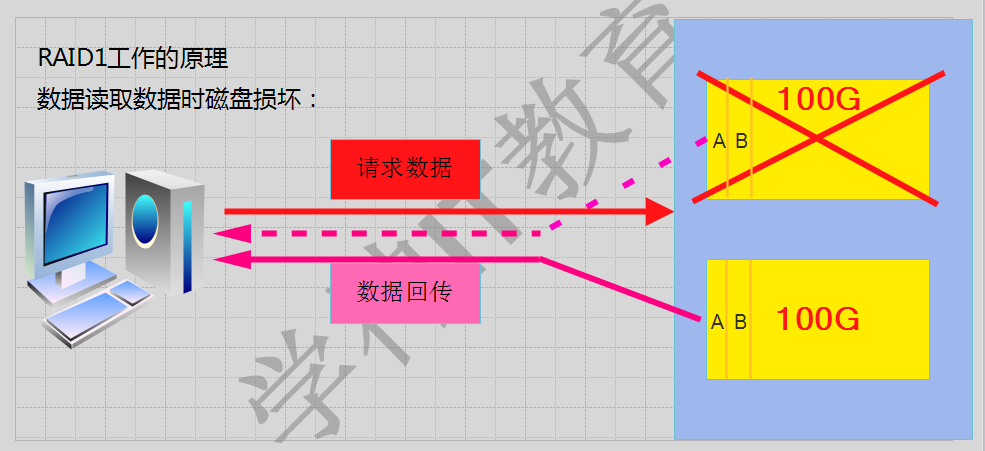
**mirroring（镜像卷），需要磁盘两块以上**

**原理:是把一个磁盘的数据镜像到另一个磁盘上，也就是说数据在写入一块磁盘的同时，会在另一块闲置的磁盘上生成镜像文件，(同步)**

**RAID 1 mirroring（镜像卷），至少需要两块硬盘，raid大小等于两个raid分区中最小的容量（最好将分区大小分为一样），数据有冗余，在存储时同时写入两块硬盘，实现了数据备份；**

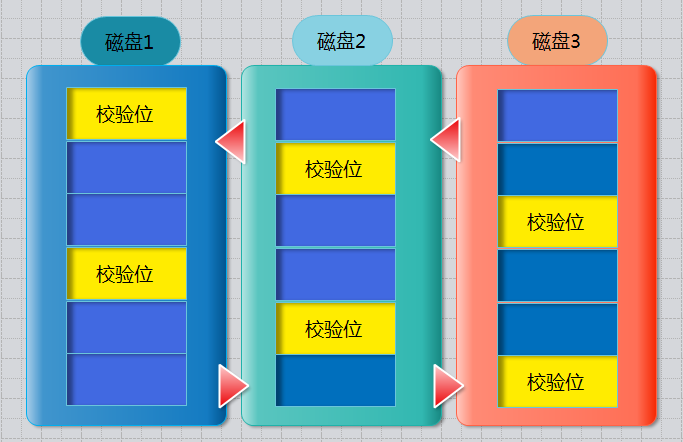
**磁盘利用率为50%，即2块100G的磁盘构成RAID1只能提供100G的可用空间。如下图**

****

****

## **14.1.4 RAID-5**

**需要三块或以上硬盘，可以提供热备盘实现故障的恢复；只损坏一块，没有问题。但如果同时损坏两块磁盘，则数据将都会损坏。 空间利用率： (n-1)/n 2/3 如下图所示**

****

**奇偶校验信息的作用:**

**当RAID5的一个磁盘数据发生损坏后，利用剩下的数据和相应的奇偶校验信息去恢复被损坏的数据。**

**扩展：异或运算**

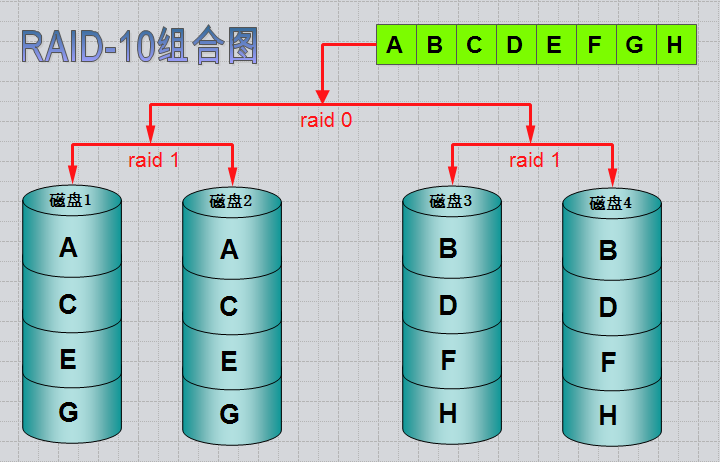
**是用相对简单的异或逻辑运算（相同为0，相异为1）**

|  |  |  |
| --- | --- | --- |
| **A值** | **B值** | **Xor结果** |
| **0** | **0** | **0** |
| **1** | **0** | **1** |
| **0** | **1** | **1** |
| **1** | **1** | **0** |

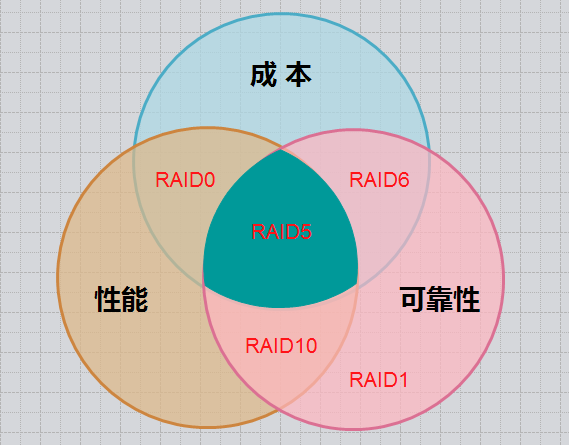
## **14.1.5 嵌套RAID级别**

**RAID-10镜像+条带**

**RAID 10是将镜像和条带进行两级组合的RAID级别，第一级是RAID1镜像对，第二级为RAID 0。比如我们有8块盘，它是先两两做镜像，形成了新的4块盘，然后对这4块盘做RAID0；当RAID10有一个硬盘受损其余硬盘会继续工作，这个时候受影响的硬盘只有2块**

****

**几个方案对比下来， RAID5是最适合的，如下图：**

****

## **14.1.6 RAID硬盘失效处理**

**一般两种处理方法：热备和热插拔**

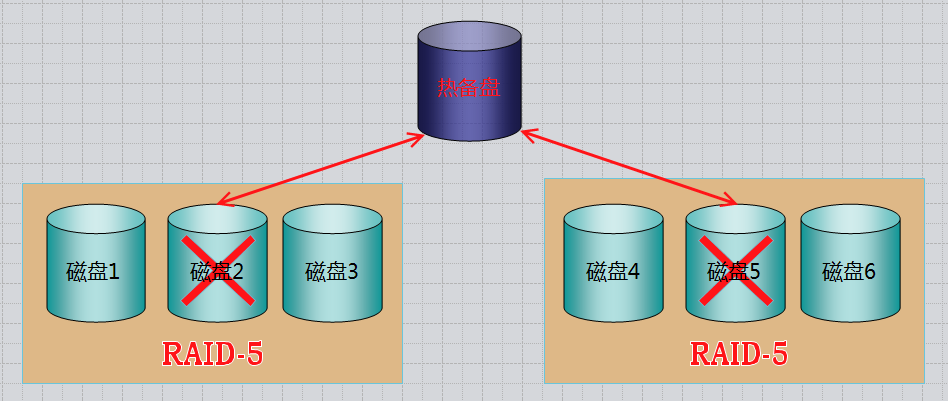
**热备：HotSpare**

**定义：当冗余的RAID组中某个硬盘失效时，在不干扰当前RAID系统的正常使用的情况下，用RAID系统中另外一个正常的备用硬盘自动顶替失效硬盘，及时保证RAID系统的冗余性**

**全局式：备用硬盘为系统中所有的冗余RAID组共享**

**专用式：备用硬盘为系统中某一组冗余RAID组专用**

**如下图所示：是一个全局热备的示例，该热备盘由系统中两个RAID组共享，可自动顶替任何一个RAID中的一个失效硬盘**

****

**热插拔：HotSwap**

**定义：在不影响系统正常运转的情况下，用正常的物理硬盘替换RAID系统中失效硬盘。**

# 14.2 RAID-0-1-5-10搭建及使用-删除RAID及注意事项

## **14.2.1 RAID的实现方式**

**互动：我们做硬件RAID，是在装系统前还是之后？**

**答：先做阵列才装系统 ，一般服务器启动时，有显示进入配置Riad的提示，比如：按下CTRL+L/H/M进入配置raid界面**

**硬RAID：需要RAID卡，我们的磁盘是接在RAID卡的，由它统一管理和控制。数据也由它来进行分配和维护；它有自己的cpu，处理速度快**

**操作视频链接：http://pan.baidu.com/s/1jIJqJp8 密码：mfcb 视频没有声音**

**软RAID：通过操作系统实现**

**Linux内核中有一个md(multiple devices)模块在底层管理RAID设备，它会在应用层给我们提供一个应用程序的工具mdadm ，mdadm是linux下用于创建和管理软件RAID的命令。**

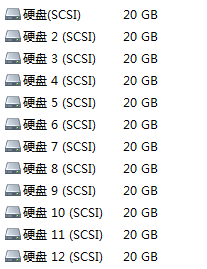
**mdadm命令常见参数解释：**

|  |  |  |  |
| --- | --- | --- | --- |
| **-C或--creat** | **建立一个新阵列** | **-r** | **移除设备** |
| **-A** | **激活磁盘阵列** | **-l 或--level=** | **设定磁盘阵列的级别** |
| **-D或--detail** | **打印阵列设备的详细信息** | **-n或--raid-devices=** | **指定阵列成员（分区/磁盘）的数量** |
| **-s或--scan** | **扫描配置文件或/proc/mdstat得到阵列缺失信息** | **-x或--spare-devicds=** | **指定阵列中备用盘的数量** |
| **-f** | **将设备状态定为故障** | **-c或--chunk=** | **设定阵列的块chunk块大小 ，单位为KB** |
| **-a或--add** | **添加设备到阵列** | **-G或--grow** | **改变阵列大小或形态** |
| **-v** | **--verbose 显示详细信息** | **-S** | **停止阵列** |

**互动： raid5需要3块硬盘。 那么使用4块硬盘，可以做raid5吗？**

**可以的**

**实验环境：新添加11块硬盘，每块磁盘的作用如下：**



**互动：磁盘达到sdz以后，名字应该如何排？**

**sdaa 、 sdab 。。。**



**实验环境：**

|  |  |  |
| --- | --- | --- |
| **raid种类** | **磁盘** | **热备盘** |
| **raid0** | **sdb、sdc** |  |
| **raid1** | **sdd、sde、** | **sdf** |
| **raid5** | **sdg、sdh、sdi** | **sdj** |
| **raid10** | **分区：sdk1,sdk2,sdk3.sdk4** |  |

**注：工作中正常做raid全部是使用独立的磁盘来做的。为了节约资源，raid10以一块磁盘上多个分区来代替多个独立的磁盘做raid，但是这样做出来的raid没有备份数据的作用，因为一块磁盘坏了，这个磁盘上所做的raid也就都坏了。**

## **14.2.2 创建RAID0**

**实验环境：**

|  |  |  |
| --- | --- | --- |
| **raid种类** | **磁盘** | **热备盘** |
| **raid0** | **sdb、sdc** |  |

### 1、创建raid0

### [root@xuegod63 ~]# mdadm -C -v /dev/md0 -l 0 -n 2 /dev/sdb /dev/sdc

### mdadm: chunk size defaults to 512K

### mdadm: Defaulting to version 1.2 metadata

### mdadm: array /dev/md0 started.

### [root@xuegod63 ~]# mdadm -Ds

### 2、 查看阵列信息

**[root@xuegod63 ~]# mdadm -Ds**

**ARRAY /dev/md0 metadata=1.2 name=xuegod63.cn:0 UUID=cadf4f55:226ef97d:565eaba5:3a3c7da4**

**[root@xuegod63 ~]# mdadm -D /dev/md0**

**/dev/md0:**

**Version : 1.2**

**Creation Time : Thu May 17 15:59:16 2018**

**Raid Level : raid0**

**Array Size : 41910272 (39.97 GiB 42.92 GB)**

**Raid Devices : 2**

**Total Devices : 2**

**Persistence : Superblock is persistent**

**Update Time : Thu May 17 15:59:16 2018**

**State : clean**

**Active Devices : 2**

**Working Devices : 2**

**Failed Devices : 0**

**Spare Devices : 0**

**Chunk Size : 512K #chunk是raid中最小的存储单位**

**Consistency Policy : none**

**Name : xuegod63.cn:0 (local to host xuegod63.cn)**

**UUID : cadf4f55:226ef97d:565eaba5:3a3c7da4**

**Events : 0**

**Number Major Minor RaidDevice State**

**0 8 16 0 active sync /dev/sdb**

**1 8 32 1 active sync /dev/sdc**

**[root@xuegod63 ~]# mdadm -Ds > /etc/mdadm.conf #生成配置文件**

### 3、 对创建的RAID0进行文件系统创建并挂载

**[root@xuegod63 ~]# mkfs.xfs /dev/md0**

**[root@xuegod63 ~]# mkdir /raid0**

**[root@xuegod63 ~]# mount /dev/md0 /raid0/**

**[root@xuegod63 ~]# df -Th /raid0/**

**文件系统 类型 容量 已用 可用 已用% 挂载点**

**/dev/md0 xfs 40G 33M 40G 1% /raid0**

**[root@xuegod63 ~]# echo 324 > /raid0/a.txt**

### 4、 开机自动挂载

**[root@xuegod63 ~]# blkid /dev/md0**

**/dev/md0: UUID="3bf9c260-dc7b-4e37-a865-a8caa21ddf2c" TYPE="xfs"**

**[root@xuegod63 ~]# echo "UUID=5bba0862-c4a2-44ad-a78f-367f387ad001 /raid0 xfs defaults 0 0" >> /etc/fstab**

## **14.2.3 创建RAID1**

**实验内容如下：**

|  |  |  |
| --- | --- | --- |
| **raid种类** | **磁盘** | **热备盘** |
| **raid1** | **sdd、sde、** | **sdf** |

**1）创建RAID1**

**2）添加1个热备盘**

**3）模拟磁盘故障，自动顶替故障盘**

**4）从raid1中移出故障盘**

**[root@xuegod63 ~]# mdadm -C -v /dev/md1 -l 1 -n 2 -x 1 /dev/sd[d,e,f]**

### 将RADI信息保存到配置文件

**[root@xuegod63 ~]# mdadm -Dsv > /etc/mdadm.conf**

### 查看 RAID 阵列信息：

**[root@xuegod63 ~]# mdadm -D /dev/md1**

**Raid Level : raid1**

**Array Size : 20955136 (19.98 GiB 21.46 GB)**

**。。。**

**Number Major Minor RaidDevice State**

**0 8 48 0 active sync /dev/sdd**

**1 8 64 1 active sync /dev/sde**

**2 8 80 - spare /dev/sdf**

### 在RAID设备上创建文件系统

**[root@xuegod63 ~]# mkfs.xfs /dev/md1**

**[root@xuegod63 ~]# mkdir /raid1**

**[root@xuegod63 ~]# mount /dev/md1 /raid1/**

### 准备测试文件

**[root@xuegod63 ~]# cp /etc/passwd /raid1/**

### 模拟损坏

**下面模拟RAID1中数据盘/dev/sde出现故障，观察/dev/sdf备用盘能否自动顶替故障盘**

**[root@xuegod63 ~]# mdadm /dev/md1 -f /dev/sde**

**查看一下阵列状态信息**

**[root@xuegod63 ~]# mdadm -D /dev/md1**

**...**

**Number Major Minor RaidDevice State**

**0 8 96 0 active sync /dev/sdg**

**2 8 128 1 spare rebuilding /dev/sdi #热备盘已经在同步数据**

**1 8 112 - faulty /dev/sdh**

**更新配置文件**

**[root@xuegod63 ~]# mdadm -Dsv > /etc/mdadm.conf**

**查看数据是否丢失**

**[root@xuegod63 ~]# ls /raid1/ #数据正常，没有丢失**

**重要的数据如：数据库 ； 系统盘 （把系统安装到raid1的md1设备上，可以对md1做分区）**

**移除损坏的设备：**

**[root@xuegod63 ~]# mdadm -r /dev/md1 /dev/sde**

**mdadm: hot removed /dev/sde from /dev/md1**

**查看信息：**

**[root@xuegod63 ~]# mdadm -D /dev/md1**

**Number Major Minor RaidDevice State**

**0 8 96 0 active sync /dev/sdd**

**2 8 128 1 active sync /dev/sdf #已经没有热备盘了**

### 添加一块新热备盘

**[root@xuegod63 ~]# mdadm -a /dev/md1 /dev/sde**

**mdadm: added /dev/sde**

## **14.2.4 创建RAID5**

**实验环境：**

|  |  |  |
| --- | --- | --- |
| **raid种类** | **磁盘** | **热备盘** |
| **raid5** | **sdg、sdh、sdi** | **sdj** |

**1）创建RAID5, 添加1个热备盘，指定chunk大小为32K**

**-x或--spare-devicds= 指定阵列中备用盘的数量**

**-c或--chunk= 设定阵列的块chunk块大小 ，单位为KB**

**2）停止阵列，重新激活阵列**

**3）使用热备盘，扩展阵列容量，从3个磁盘扩展到4个**

### 14.2.4.1 创建RAID-5

**[root@xuegod63 ~]# mdadm -C -v /dev/md5 -l 5 -n 3 -x 1 -c32 /dev/sd{g,h,i,j}**

**[root@xuegod63 ~]# mdadm -D /dev/md5**

**/dev/md5:**

**Version : 1.2**

**Creation Time : Thu May 17 18:54:20 2018**

**Raid Level : raid5**

**Array Size : 41910272 (39.97 GiB 42.92 GB)**

**Used Dev Size : 20955136 (19.98 GiB 21.46 GB)**

**Raid Devices : 3**

**Total Devices : 4**

**Persistence : Superblock is persistent**

**Update Time : Thu May 17 18:54:31 2018**

**State : clean, degraded, recovering**

**Active Devices : 2**

**Working Devices : 4**

**Failed Devices : 0**

**Spare Devices : 2**

**Layout : left-symmetric**

**Chunk Size : 32K**

**Consistency Policy : resync**

**Rebuild Status : 7% complete**

**Name : xuegod63.cn:5 (local to host xuegod63.cn)**

**UUID : fa685cea:38778d6a:0eb2c670:07ec5797**

**Events : 2**

**Number Major Minor RaidDevice State**

**0 8 96 0 active sync /dev/sdg**

**1 8 112 1 active sync /dev/sdh**

**4 8 128 2 spare rebuilding /dev/sdi**

**3 8 144 - spare /dev/sdj #热备盘**

### 14.2.4.2 停止MD5阵列

**[root@xuegod63 ~]# mdadm -Dsv > /etc/mdadm.conf #停止前，一定要先保存配置文件**

**[root@xuegod63 ~]# mdadm -D /dev/md5 ##停止前，请确认数据已经同步完**

**Consistency Policy : resync #数据已经同步完**

**[root@xuegod63 ~]# mdadm -S /dev/md5**

**mdadm: stopped /dev/md5**

### 14.2.4.3 激活MD5阵列

**[root@xuegod63 ~]# mdadm -As**

**mdadm: /dev/md5 has been started with 3 drives and 1 spare.**

### 14.2.4.4 扩展RAID5磁盘阵列

**将热备盘增加到md5中，使用md5中可以使用的磁盘数量为4块**

**[root@xuegod63 /]# mdadm -G /dev/md5 -n 4 -c 32**

**#-G或--grow 改变阵列大小或形态**

**[root@xuegod63 ~]# mdadm -Dsv > /etc/mdadm.conf #保存配置文件**

**备注：阵列只有在正常状态下，才能扩容，降级及重构时不允许扩容。对于raid5来说，只能增加成员盘，不能减少。而对于raid1来说，可以增加成员盘，也可以减少。**

**[root@xuegod63 ~]# mdadm -D /dev/md5 #查看状态**

**。。。**

**Array Size : 41910272 (39.97 GiB 42.92 GB) #发现新增加硬盘后空间没有变大，为什么？**

**Used Dev Size : 20955136 (19.98 GiB 21.46 GB)**

**。。。**

**Reshape Status : 3% complete #重塑状态：3%完成 ，等到100%， 数据才同步完，同步完后会变成成：Consistency Policy : resync #一致性策略：再同步，表示已经同步完**

**。。。**

**Number Major Minor RaidDevice State**

**0 8 96 0 active sync /dev/sdg**

**1 8 112 1 active sync /dev/sdh**

**4 8 128 2 active sync /dev/sdi**

**3 8 144 3 active sync /dev/sdj**

**等一会，等所有数据同步完成后，查看md5空间大小：**

**Array Size : 62865408 (59.95 GiB 64.37 GB) #空间已经变大**

**Used Dev Size : 20955136 (19.98 GiB 21.46 GB)**

## **14.2.5 创建RAID10**

**实验环境：**

|  |  |  |
| --- | --- | --- |
| **raid种类** | **磁盘** | **热备盘** |
| **raid10** | **分区：sdk1,sdk2,sdk3.sdk4** |  |

**[root@xuegod63 ~]# fdisk /dev/sdk #分4个主分区，每个分区1G大小**

**[root@xuegod63 ~]# mdadm -C -v /dev/md10 -l 10 -n 4 /dev/sdk[1-4]**

**[root@xuegod63 ~]# cat /proc/mdstat**

## **14.2.6删除RAID所有信息及注意事项**

**[root@xuegod63 ~]# umount /dev/md0 /raid0 #如果你已经挂载raid，就先卸载。**

**[root@xuegod63 ~]# mdadm -Ss #停止raid设备**

**[root@xuegod63 ~]# rm -rf /etc/mdadm.conf #删除raid配置文件**

**[root@xuegod63 ~]# mdadm --zero-superblock /dev/sdb #清除物理磁盘中的raid标识**

**[root@xuegod63 ~]# mdadm --zero-superblock /dev/sdc #清除物理磁盘中的raid标识**

**参数：--zero-superblock : erase the MD superblock from a device. #擦除设备中的MD超级块**

# 14.3 实战：企业中硬件raid5的配置

**联想（ThinkServer） RD650做Raid**

**联想（ThinkServer） RD650（640升级）2U机架式 服务器2.5‘’盘位 2\*E5-2609V4/双电源 升级至32G内存3个300G硬盘**

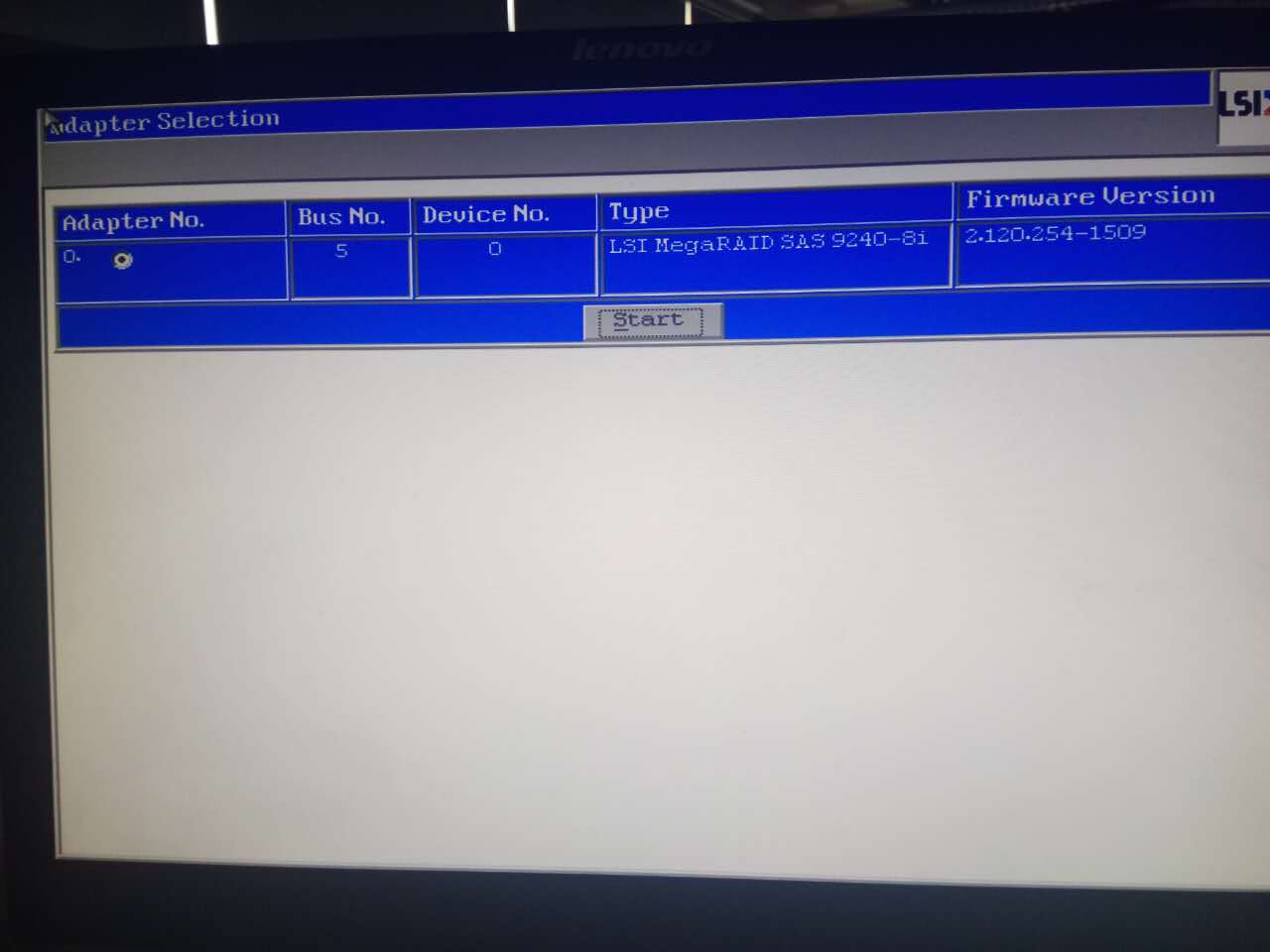
**https://item.jd.com/10502926632.html?jd\_pop=eef9047a-999c-421a-8f04-fd4678c9cd4c#crumb-wrap**



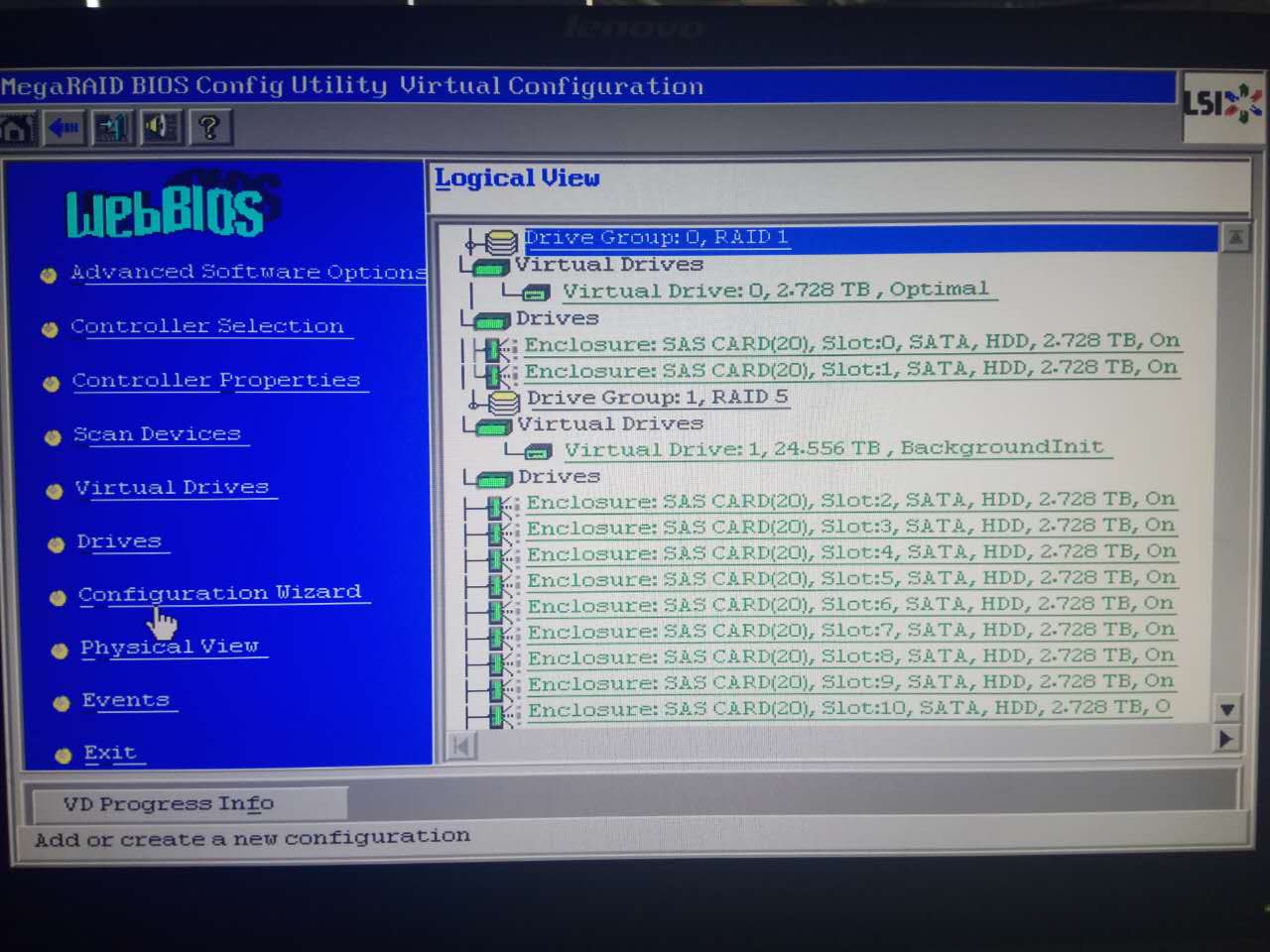
**操作步骤：**

**开机后，进入raid配置界面有提示，按ctrl +h进入raid配置界面：**

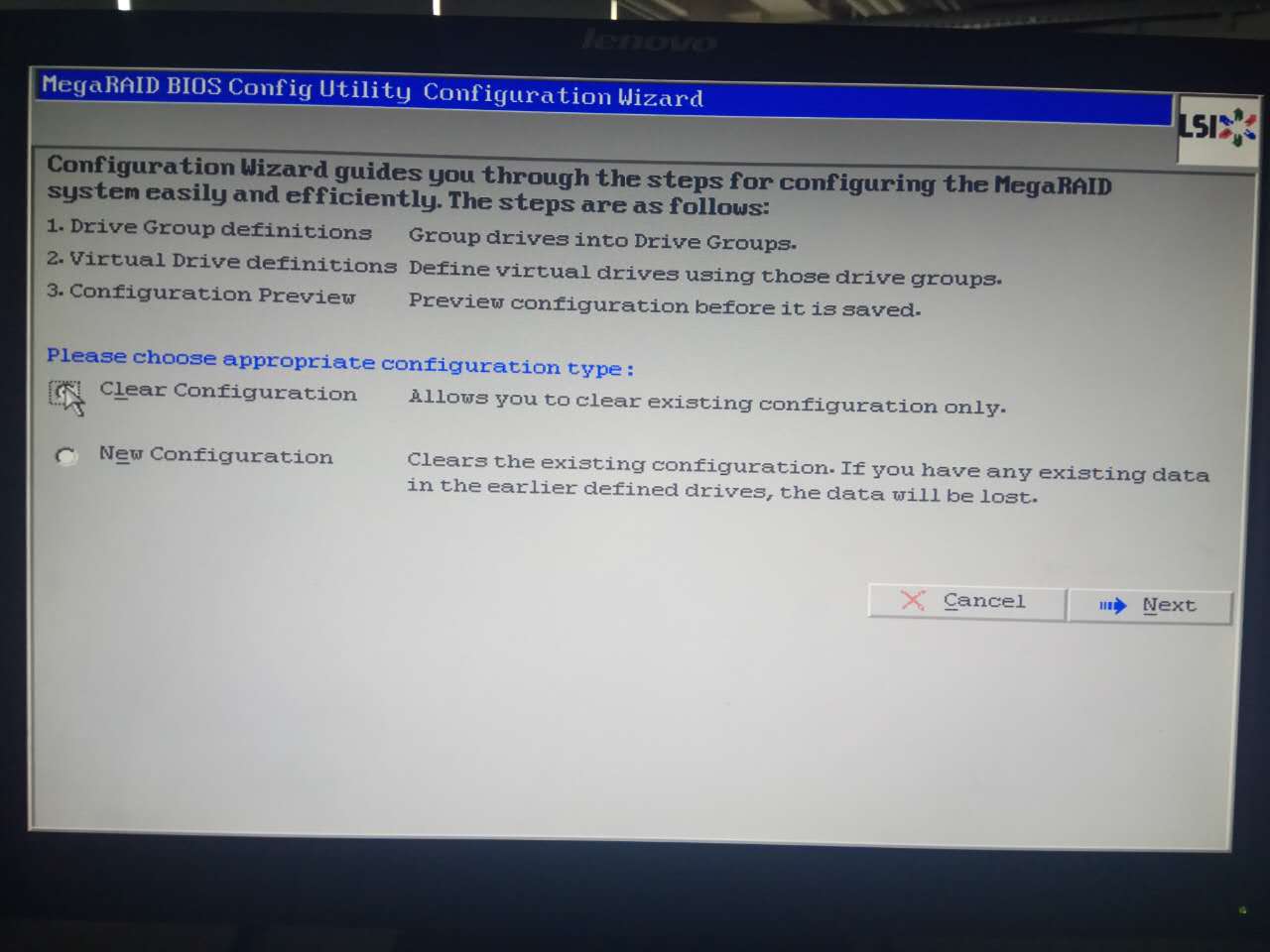
**连接服务器以后，显示以下界面,单击start进入配置页面：**

****

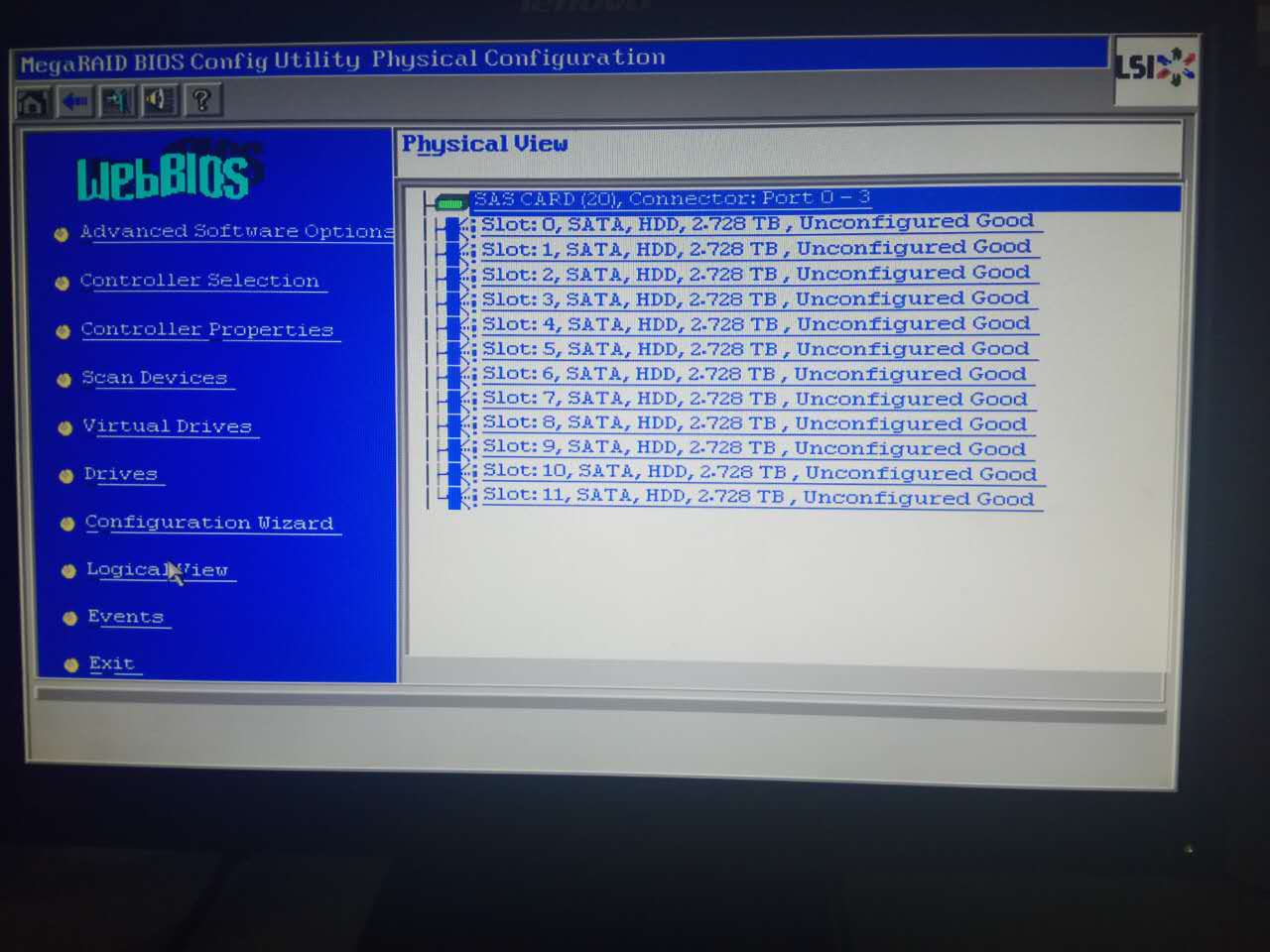
**单机Configuration Wizard （配置向导）进行配置：**

****

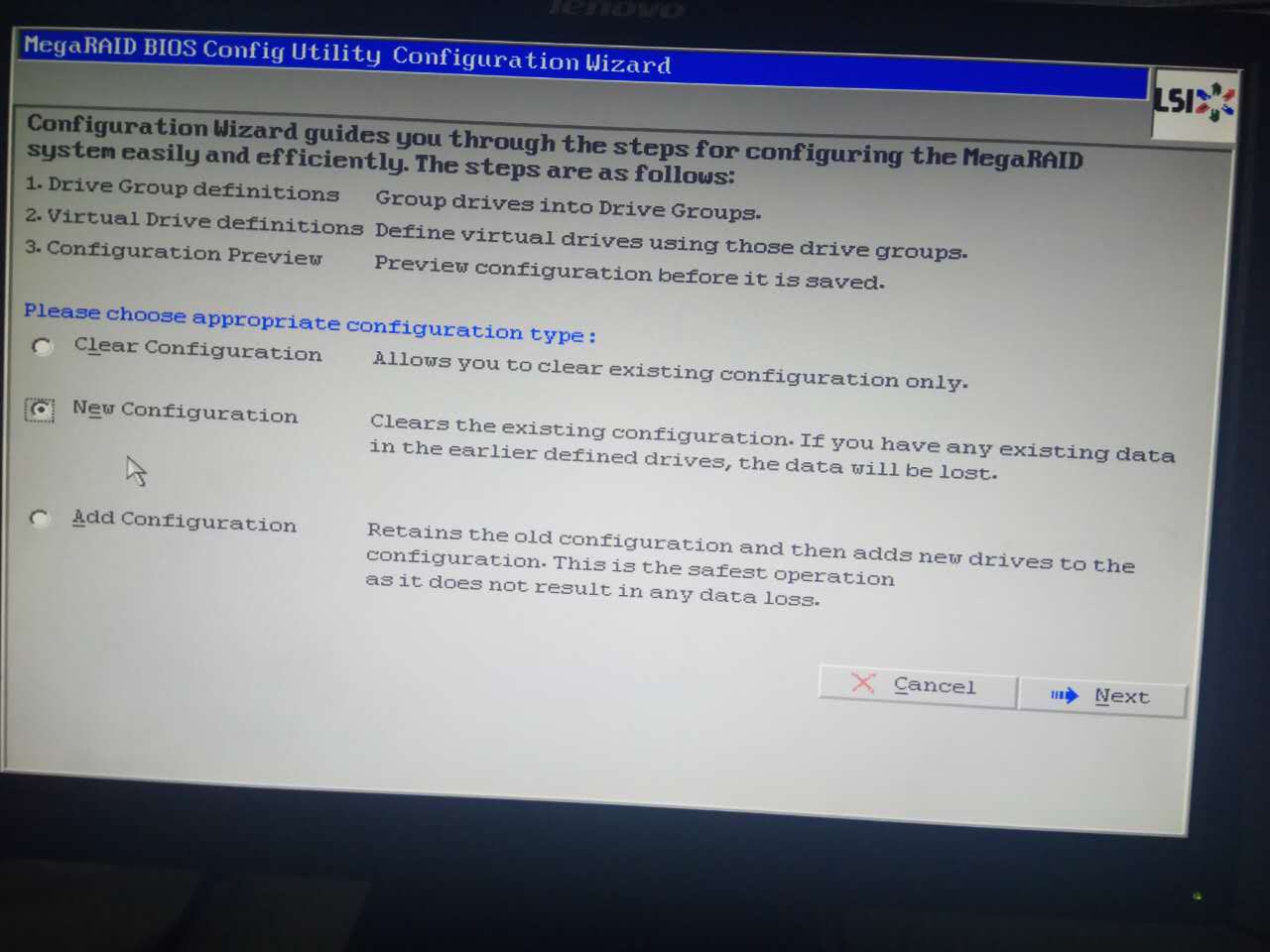
**单机Clear Configuration(清除配置)清除旧的配置：**

****

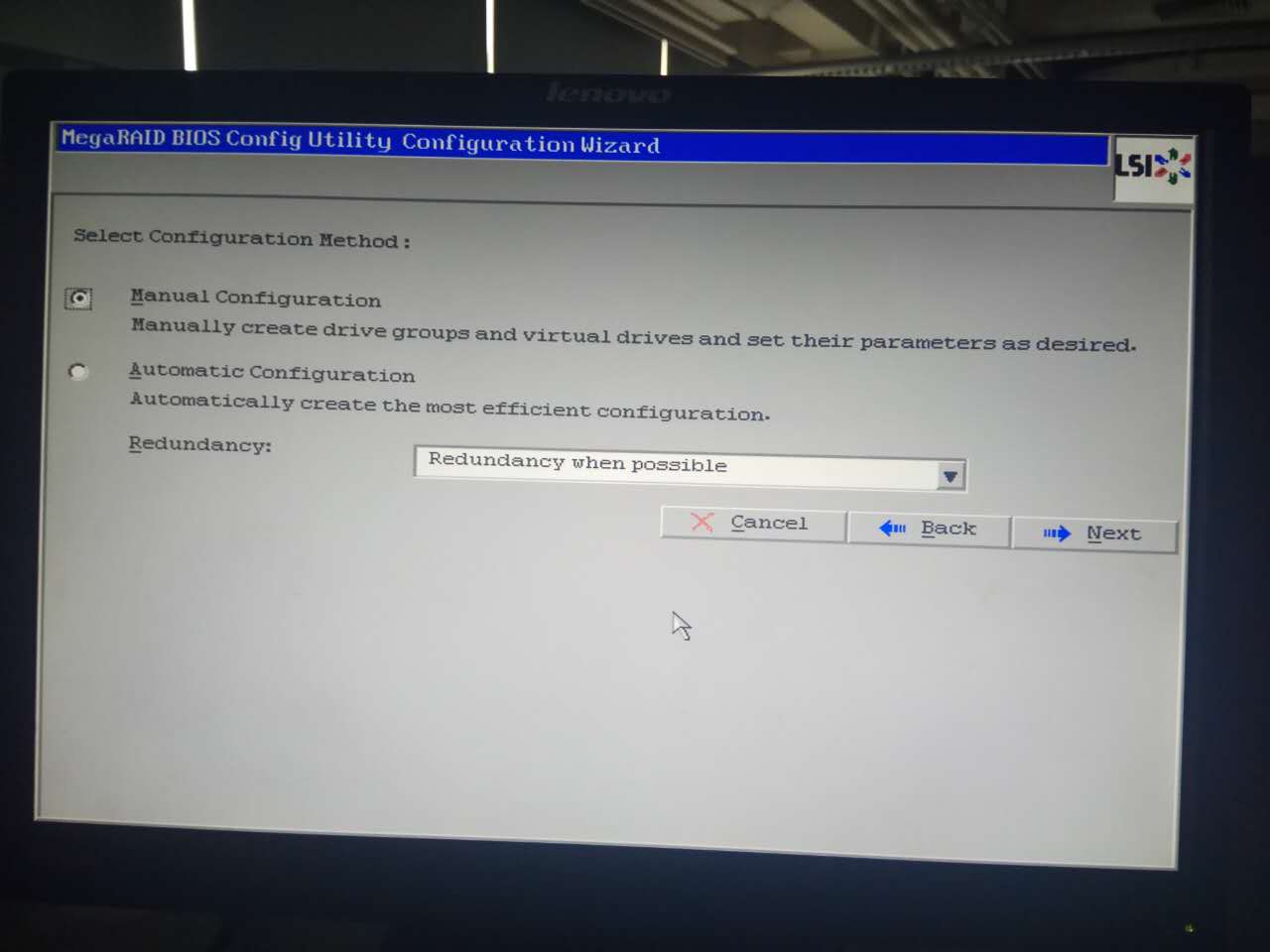
**清除以后，显示如下图，再次单机Configuration Wizard进行配置,：**

****

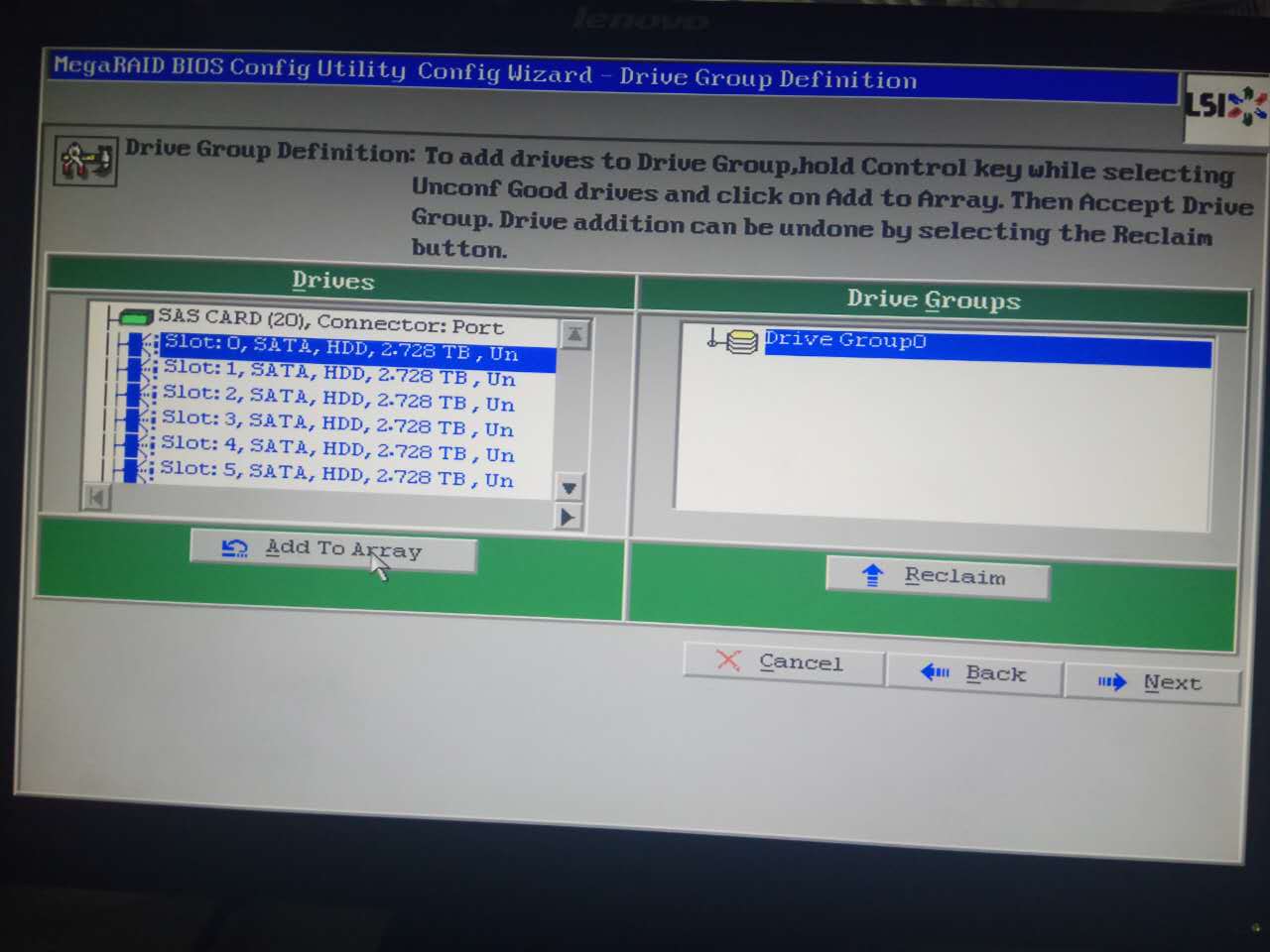
**单机new Configuration 进行新的配置：**

****

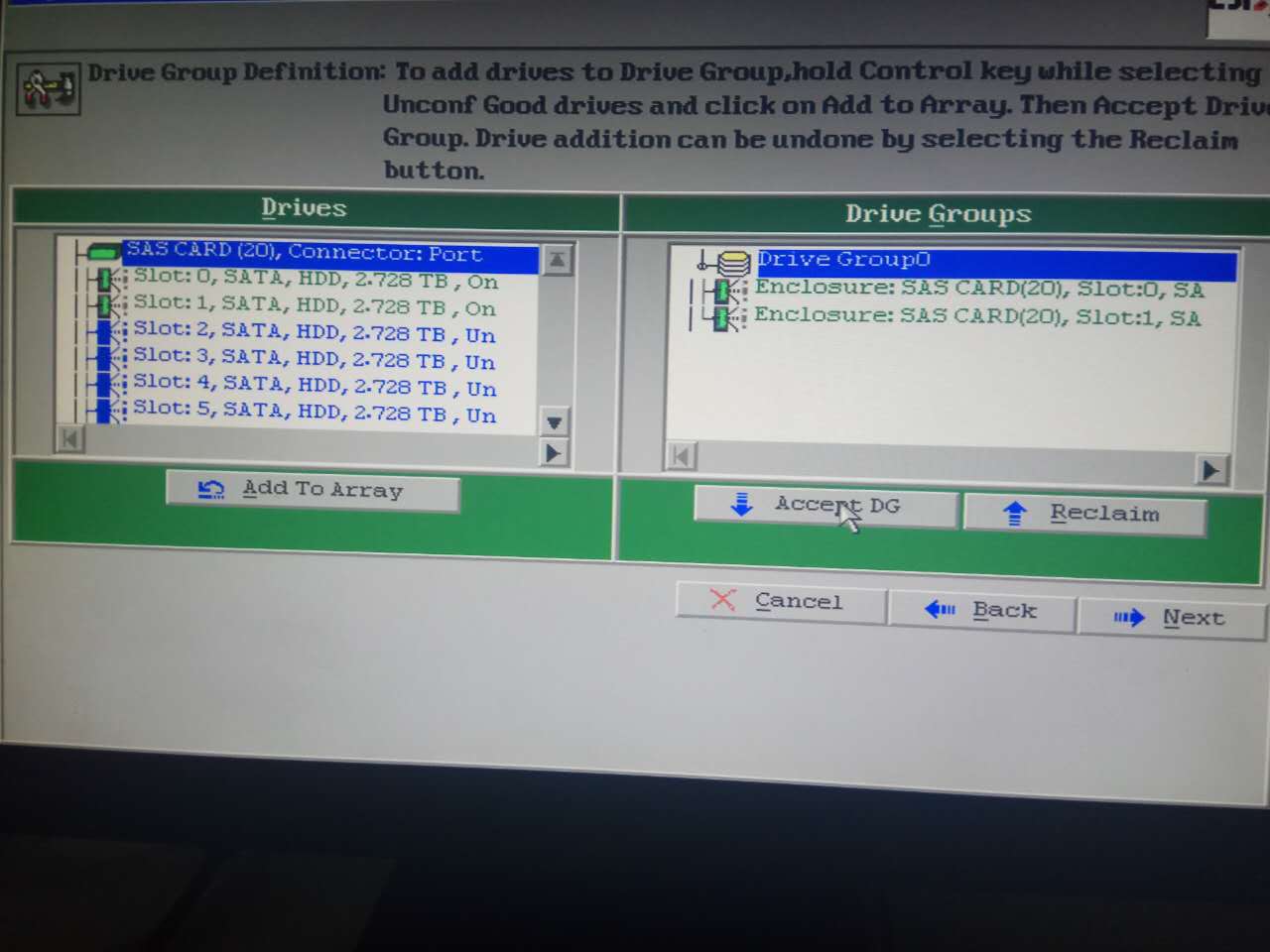
**进入如下页面，单击Manual Configuration（配置手册）：**

****

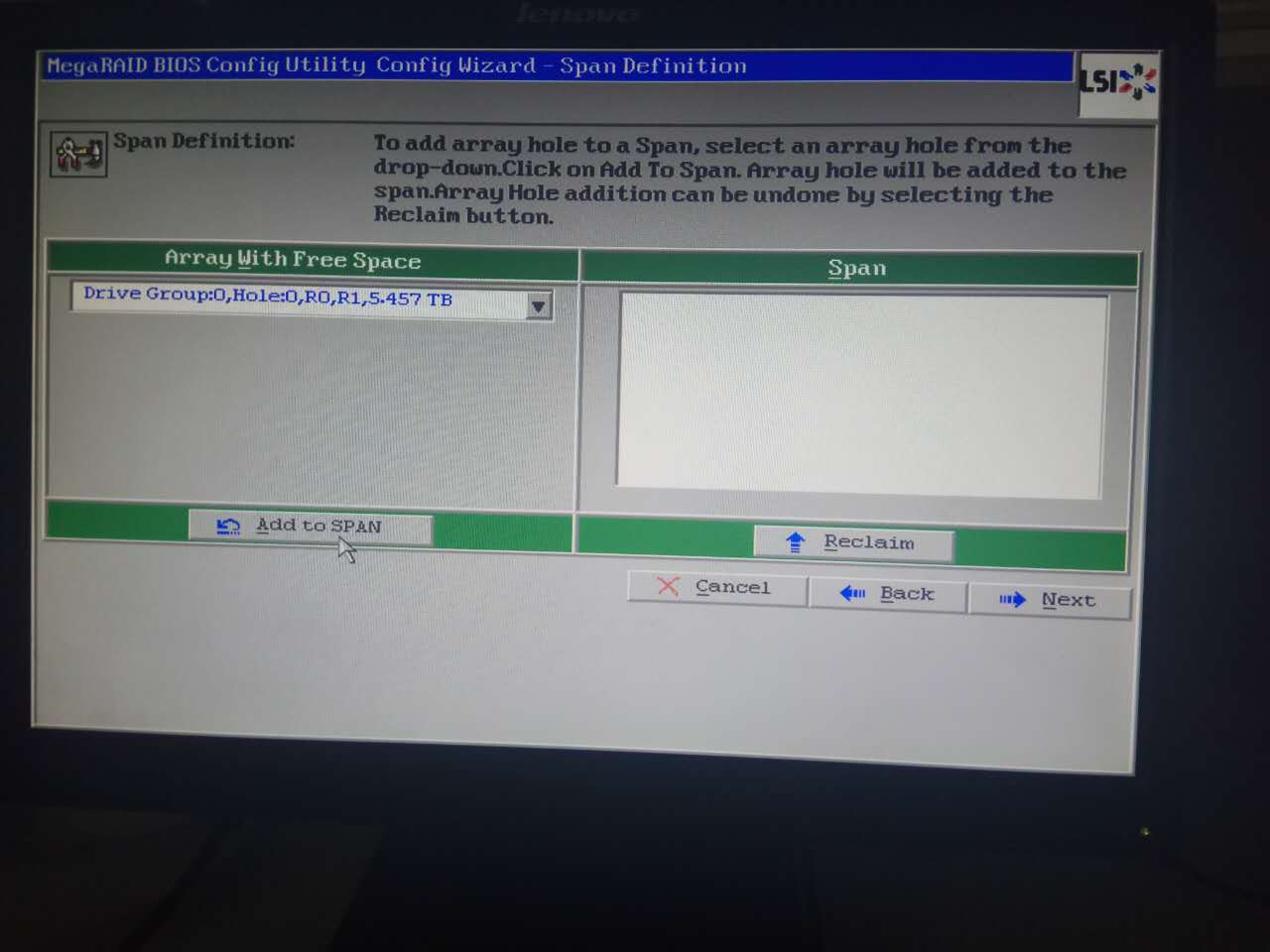
**选择左侧两块磁盘，做个raid1，单机 Add To Array(加入阵列)：**

****

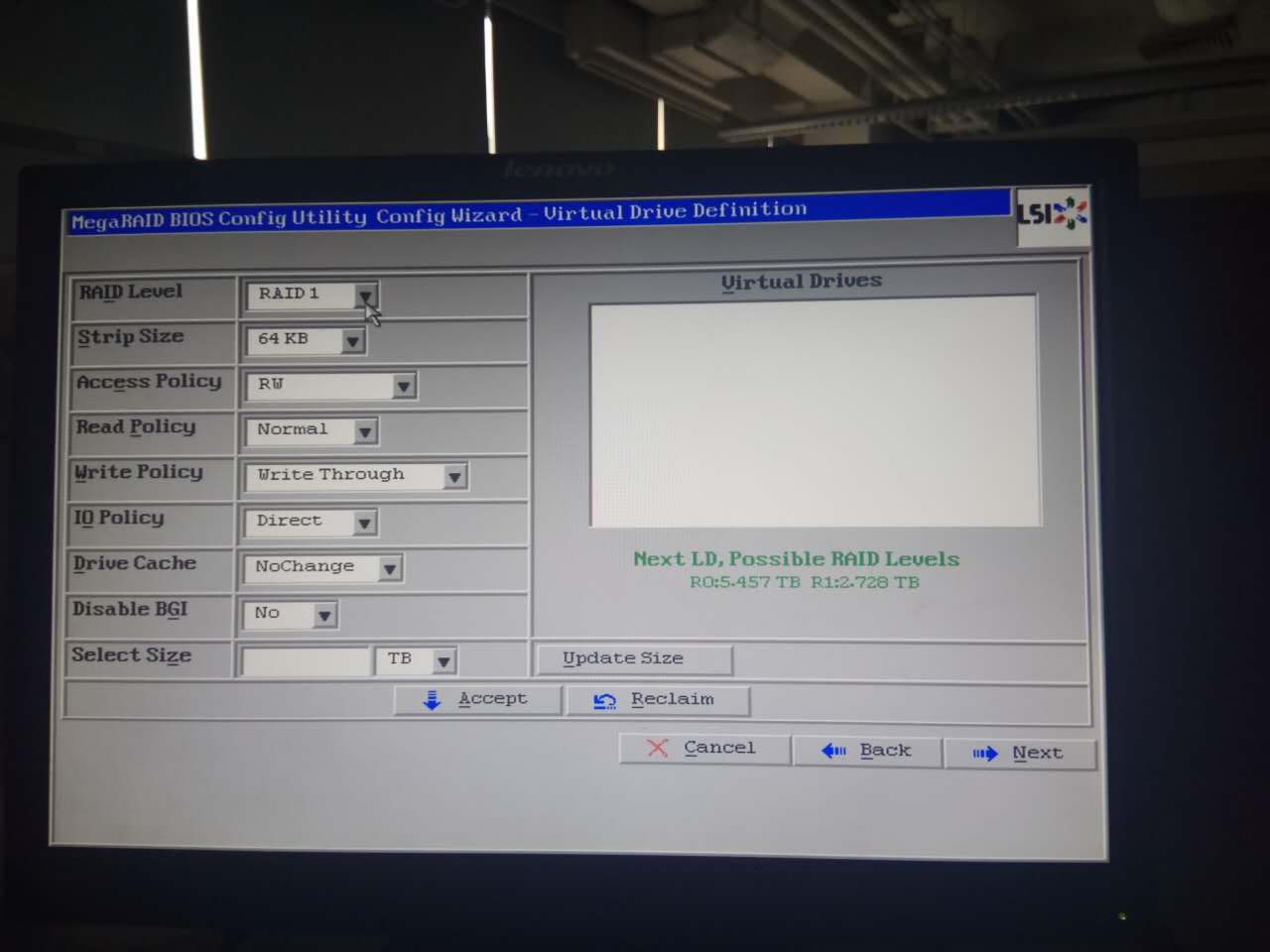
**此处我们把两块盘做raid1，单机Accept DG（接受磁盘组，DG为disk groups的缩写），：**

****

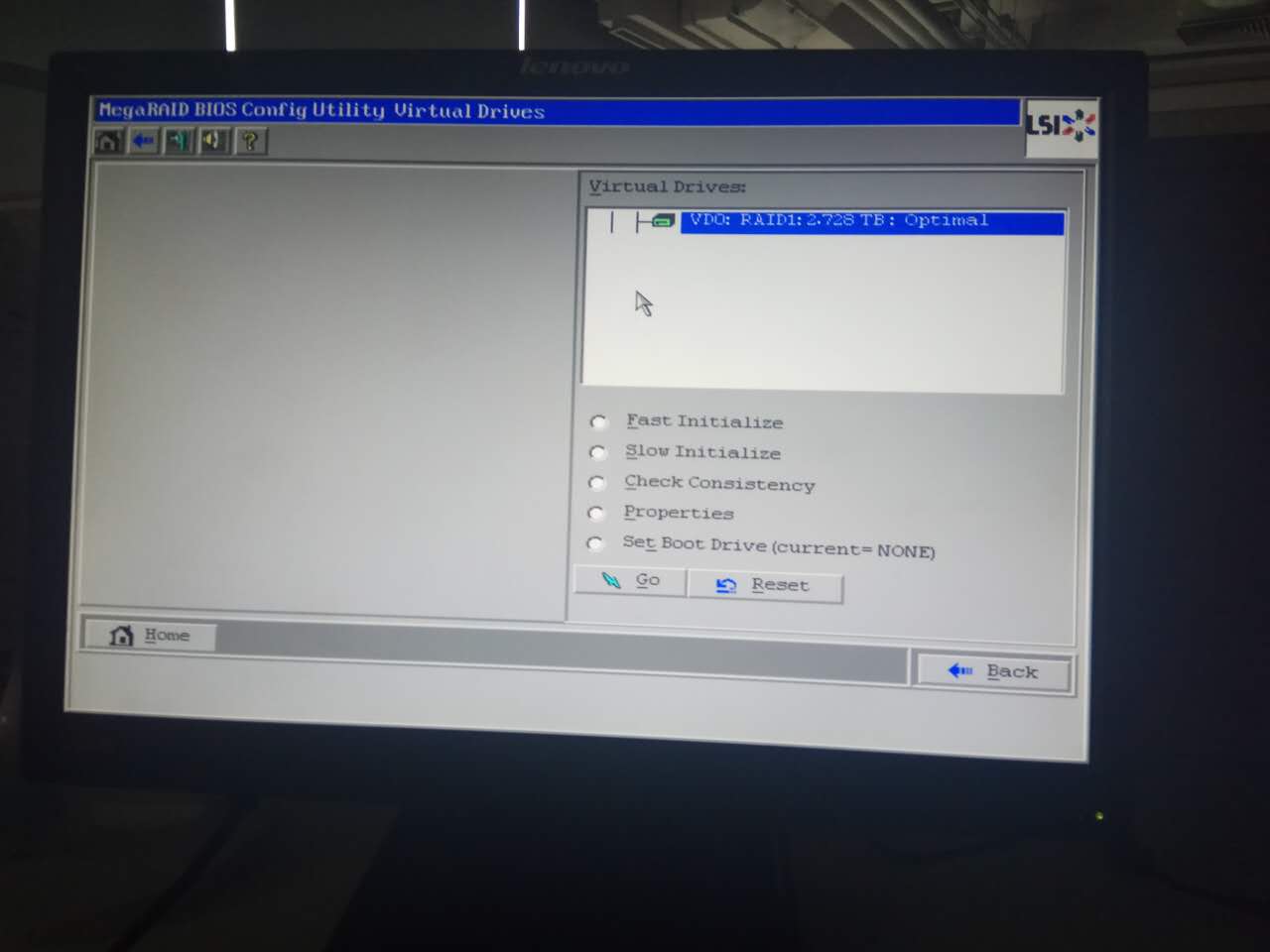
**然后单机next，会进入如下页面，单机Add to SPAN(缚住或扎牢的意思，理解为将两块盘捆绑到一起)，单机next：**

****

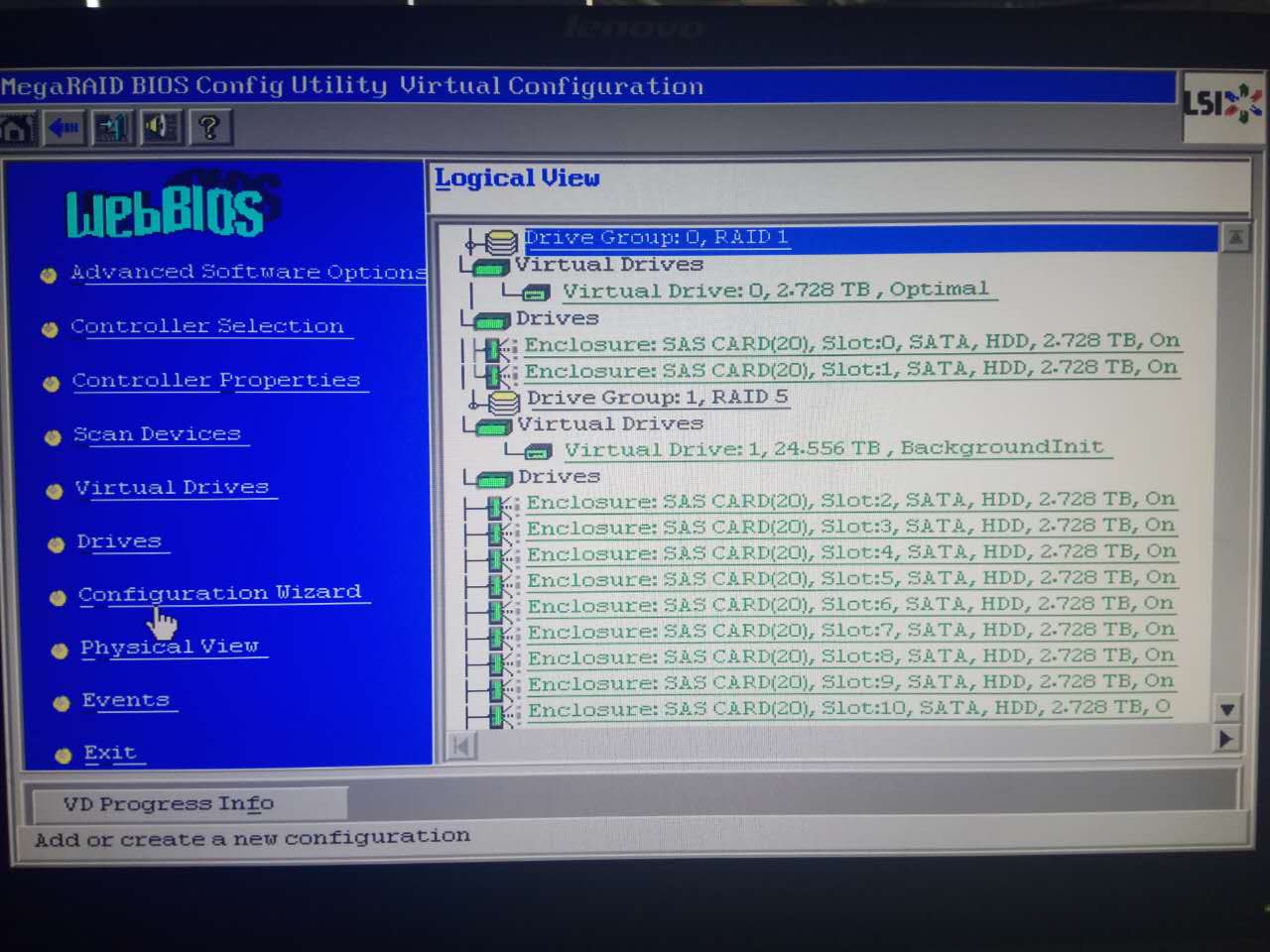
**进入如下页面，可以选择raid（我们做的而是raid1），然后单机Update Size，accept，直接next就可以：**

****

**后面全部选next或者yes，当碰到下面这步骤时，可以忽略，直接点back：**

****

**会回到之前的Configuration Wizard配置向导页面, 后面选择add Configuration(添加配置)，后面做raid5的10块盘操作步骤和之前相同。最后配置完成的结果如下：**

****

**互动：为什么先把两块磁盘做raid1，然后把后面的磁盘都做成raid5**

**raid1是镜像卷，安装系统用，一块坏了，不影响系统运行。 raid5存数据**

**如果服务器主板不支持硬raid ， 可以用raid卡**

**扩展：常见raid卡：**

**戴尔（DELL） 服务器RAID阵列卡 H730P 大卡 2G缓存+后备电池保障数据读写不受断电影响**

**https://item.jd.com/13179864635.html ￥1990.00元 H730P 小卡是1G缓存**







**硬raid如果阵列卡坏了 怎么办？**

**如果更换的RAID卡，与原卡规格型号相同，则不会有什么问题。若两卡规格型号不同，需要重新安装新卡驱动程序。这时就会有不确定的可能性，如驱动程序差别不大时，RAID盘组合顺序正常，硬盘原数据可正确读写；若驱动程序结构差别较大，则可能发生读写错误。**

**因此，要尽量选购与原卡规格相同的产品，以确保数据安全。**

**总结：**

**14.1 RAID概念-企业级RAID 0, 1,5,10的工作原理**

**14.2 RAID-0-1-5-10搭建及使用-删除RAID及注意事项**

**14.3 实战：企业中硬件raid5的配置**