ldc压入栈后是否在heap中创建对象？

TODO：static 局部的static和类static存储位置。

# 第一篇 J2SE

## Base Data Type

### String

TODO1:hostspot 加载string后是否在持久区创建对象？

## Collection

### 1.2.1 概述



**Collection VS Array**

数组为定长，Collection均可动态增长。

### 1.2.2 List

元素保持特定的顺序

ArrayList

基于数组实现，可动态的扩展和缩减。

1. **private** **transient** Object[] elementData;

排序：Comparator

### 1.2.3 Set

HashSet

元素不重复，无顺序。根据对象HashCode()返回值决定存储位置

LinkedHashSet

维护插入顺序，性能略低于HashSet

### 1.2.3 Map

每个元素为key-value形式

HashMap

底层是一个数组结构，数组中的每一项又是一个链表。非线程安全

碰撞冲突

hashMap数组每一项是一个链表，添加到对应位置链表最后一项的next。

HashMap冲突很厉害，最差性能，你会怎么解决?从O（n）提升到log（n）咯，用二叉排序树的思路说了一通

ConcurrentHashMap

线程安全且效率高

**实现原理**

HashTable: 每个方法加synchronized

ConcurrentHashMap: 引入了一个“分段锁”的概念,就是把Map分成了N个Segment，put和get的时候，都是现根据key.hashCode()算出放到哪个Segment中：

LinkedHashMap

### Queue

队列，保证先进先出。

### Tree

TODO1:TreeMap 内部红黑树排序

## IO

### 1.3.1 简介

NIO

NIO是一种同步非阻塞IO模型，基于IO多路复用（事件驱动），解决高并发环境的大量连接，有较高吞吐量。

优势

NIO的优势在于单个Thread可以同时处理多个网络连接的IO，避免创建大量的线程和线程切换造成资源的浪费。（原理：NIO优化了IO的读写, 最大化压榨CPU，线程一个时间片可以处理更多IO）

IO多路复用

多路指N个连接，复用指复用一个或少量线程。

同步 VS 异步

按照POSIX标准来划分只分为两类：同步IO和异步IO。主语是IO的操作。同步和异步说的是消息的通知机制，阻塞非阻塞说的是线程的状态。

**同步**：NIO调用后没有返回，在调用没有得到结果之前，并未返回。主动等待这个调用结果，一旦调用返回，就得到返回值了。

**异步**：调用后直接返回，由被调用者通知来通知调用者，或通过回调函数处理这个调用。

### 1.3.2 IO优化

BIO

特点：JDK1.4以前的唯一选择，程序直观简单。

原理：同步阻塞。

使用场景：连接数少且固定的架构，这种方式对服务器资源要求比较高，并发局限于应用中。

NIO

对于大部分的IO场景, 它都能适应. 但是, 由于它的阻塞性, 每一个流的读写都需要占用一个线程. 这意味着, 流IO的可伸缩性很差. NIO就是IO Multiplexing在Java中的实现。IO Multiplexing在系统级语言如C/C++中应用了很长时间. 使用IO Multiplexing, IO的伸缩性大大提高, 使用单个线程, 就可以处理大量的IO对象.

特点：JDK1.4开始支持.**有响应实时性的优势**

原理：同步非阻塞。需一个线程去轮询所有IO操作的状态改变。

使用场景：连接数目多且每个连接只发送少量数据，例如一个聊天服务器。编程比较复杂，

AIO

特点：编程比较复杂，JDK7开始支持.

原理：异步非阻塞。无需轮询，充分调用OS参与并发操作，由系统通知对应的线程来处理。

使用场景：连接数目多。比如相册服务器

TODO: 通常来说，异步操作会让性能的吞吐率有很大提升（Throughput），但是会牺牲系统的响应时间（latency）。这需要业务上支持。

理想状态CPU，内存，IO外设都满负荷工作，大部分情况瓶颈都在IO的读写部分，NIO是优化了IO的读写。NIO最大化压榨CPU，把时间片都更好利用起来。如果系统的瓶颈不在这里，那么BIO和NIO在性能上没有区别。

### 1.3.3 实践

TODO1 用netty+spring VS servlet +tomcat 比较

Netty&Mina

对JDK网络编程的封装，屏蔽了繁杂的编程细节，提供便于用户开发网络应用程序的api,让开发者可以更加专注于业务逻辑的实现。

NoSql/redies todo3 如何使用NIO的

TODO3 Reactor模式

TODO2:tomcat能实现多大的并发

## 1.4 Thread

### 1.5.1 多线程

多线程本质上是利用闲置的CUP资源，加快的处理速度。

#### 线程 VS 进程

进程是资源(CPU时间，内存等)分配的基本单位，线程是CPU调度的基本单位。

#### ****多线程 VS 单线程****

**多线程不一定就比单线程程序跑的快，**取决于应用场景及程序的设计，有时候可能引入多线程带来的性能提升抵不过多线程而引入的开销。

多线程的代价

* 线程**创建和销毁**所花时间，以及系统资源开销(线程池)
* 线程调度, **上下文切换** ： **上下文切换过程并不廉价**
* **内存同步**的开销
* 设计更复杂

#### 适用CASE

理想状态CPU，内存，IO外设都满负荷工作，大部分情况瓶颈都在IO的读写部分，多线程适合计算性case。

### 1.5.2 线程同步

#### 线程状态

[Reference](http://blog.csdn.net/dongnan591172113/article/details/51967711)

New->Runable(start)->Running->Blocked

#### 线程同步

#### 同步和互斥

同步：多个线程并发访问共享资源时，保证同一时刻只有一个（信号量可以多个）线程使用。

实现同步的方法：

1）临界区（CriticalSection）：通过对多线程的串行化来访问公共资源或一段代码，速度快，适合控制数据访问。进程内可用。

2）互斥量：**互斥量用于线程的互斥。只能为0/1。**一个互斥量只能用于**一个资源**的互斥访问，可跨进程使用。

3）信号量：**信号线用于线程的同步。可以为非负整数，**可实现**多个同类资源**的**多线程**互斥和同步。当信号量为单值信号量是，也可以完成一个资源的互斥访问。可跨进程使用。

4）事件：用来通知线程有一些事件已发生，从而启动后继任务的开始,可跨进程使用。

synchronized的底层实现就用到了临界区和互斥锁（重量级锁的情况下）这两个概念。

同步指多线程环境，线程之间的协同和配合。线程同步通过加锁机制实现线程安全。

实现同步方法

* ThreadLocal
* sychronized
* Lock
* Atomic

#### 线程安全

[Reference](http://www.jasongj.com/java/thread_safe/)

多线程环境，如果每次运行结果和单线程运行的结果是一样的，就是线程安全的。

线程安全问题都是由**共享变量**(全局变量及静态变量)引起的。

线程不安全，会造成所得到的数据不一致或者数据污染，而出现脏数据。

提到线程安全,我们需要同时关注原子性、可见性和有序性和问题。

****原子性****

保证**串行**的顺序。

[Reference](http://blog.psjay.com/posts/summary-of-java-concurrency-two-synchronized-and-atomicity/)

原子操作：不可中断的一个或一系列操作。

An atomic operation is one that **cannot be interrupted** by the scheduler; if the operation begins, and then it will run to **completion** before the possibility of a context switch.

这里的原子性指java规范中定义的原子操作，指原始类型或引用类型的读取/赋值。

Synchronized/Lock

要实现更大范围操作的原子性，可以通过synchronized和Lock来实现。这里的原子性是互斥。可以中断，但即使中断了，组中的其他方法也不能被执行，保证是串行的顺序。

**在java多线程里指的原子性即为互斥，串行执行。**

原子性操作



只有语句1是原子性操作

x++和 x = x+1包括3个操作：读取x的值，进行加1操作，写入新的值。



x是共享变量，线程公有的。test方法读取变量到线程私有的栈中，x在内存中有两份copy，需要同步。和x的内存位置不同。

非原子操作：

**只要是非原子操作，操作就可能被中断或阻塞，需要同步来保证线程安全。**

****可见性****

在一些情况下即便是原子操作也可能会引发一些错误，Java提供了volatile关键字来保证可见性。

volatile ([Reference](http://www.cnblogs.com/dolphin0520/p/3920373.html))

使用恰当的话，它比synchronized的使用和执行成本会更低，因为它不会引起线程上下文的切换和调度。

实现原理

加入volatile关键字关键字时所生成的汇编代码（**非字节码**），会多出一个lock前缀指令

lock前缀指令，相当于一个内存屏障（也称内存栅栏）：

* 禁止指令重排。时不会把其后面的指令排到内存屏障之前的位置，也不会把前面的指令排到内存屏障的后面。
* 强制将对cache的修改立即写入主存并使其他CPU中对应的缓存行无效；

使用场景1

在一些情况下即便是原子操作（flag的赋值）也可能会引发一些错误



线程访问对象的值时，将**堆内存中实例变量的值**copy到**工作内存**。Java HotSpot Server版VM并不会将修改的值写回内存。

使用场景2

当使用volatile修饰某个变量时，它会保证对该变量的修改会立即被更新到内存中，并且将其它缓存中对该变量的缓存设置成无效，因此其它线程需要读取该值时必须从主内存中读取，从而得到最新的值。**这里更新内存的时间点具有不确定性只能用在上述场景，不能用于下面的自增操作，**



保证操作是原子性操作，才能保证使用volatile关键字的程序在并发时能够正确执行



自增并非原子操作,不同保证互斥串行操作。



volatile还可以禁止指令重排，保证有序性



如果语句2会在语句1之前执行，线程2中就使用未初始化的context，导致程序出错

****有序性****

Java中可通过volatile在一定程序上保证顺序性，另外还可以通过synchronized和锁来保证顺序性。

调用栈

理解JVM方法调用机制：



方法method1中currentThread也是Rt

#### ThreadGroup

[Reference](http://www.cnblogs.com/0616--ataozhijia/p/3710571.html)

Code: ThreadGroup100

使用线程组的好处是可以对这一组的线程进行整体操作。

Thread.currentThread ().getThreadGroup ()

main默认group为main

### 1.5.3 Sychronized

实现线程的互斥串行操作。保证接口的原子性和可见性。

#### 实现原理

在HotSpot虚拟机中，synchronized通过对象内部的监视器锁（monitor）来实现。

[Reference](http://www.cnblogs.com/dennyzhangdd/p/6734638.html)

**修饰代码段:**synchronized在JVM被编译为monitorenter、monitorexit指令来获取和释放互斥锁.。

**修饰方法：**方法flags标记为ACC\_SYNCHRONIZED

**Hotspot实现**

加锁值存在对象头部，即mark中。cas操作实现轻量级锁,但仅用于交替执行同步块的情况。对于同一Thread可用偏向锁，减少cas操作。多个线程在同一时刻进入临界区膨胀为重量级锁。依赖OS层面实现。

### 1.5.4 线程锁

加锁是保护共享资源逻辑一致性的通用方案：在并发环境下，通过锁使得在某个时刻只有一个线程或进程可以访问共享的资源

#### 重量级锁

即悲观锁。假定会出现并发问题，读取数据时即将数据锁定。悲观锁需要操作系统参与，会引起线程阻塞和上下文切换。相对于乐观锁，系统资源开销大。

实现原理：monitor依赖OS的Mutex Lock来实现的。操作系统实现线程之间的切换需要从用户态转换到核心态，这个成本非常高，需要相对比较长的时间。

#### 轻量级锁

轻量级锁即乐观锁。假定不会出现并发问题，每次不加锁，冲突失败就重试，直到成功为止。

原理：

**借助CPU提供的CAS（compare and swap）指令。实现乐观锁。在冲突不激烈的多线程环境，会有比较好的性能。**

实现原理





ABA问题

潜在问题:如果需要关注A值变化过程，是会漏掉一段时间窗口的监控 .

CAS指令

CPU的CAS指令是原子性操作，执行完成前不可中断。本质上采用了缓存一致性协议（比如MESI），执行过程加了排他锁，不会出现线程安全问题。

注意：CAS操作只是实现了原子性,还需考虑可见性。

加锁优先级

无锁数据结构>锁区块>方法>对象锁>类锁

说明：尽可能使加锁的代码块工作量尽可能的小，避免在锁代码块中调用RPC方法。

#### 可重入锁

可重入锁，也叫做**递归锁**，指同一线程外层函数获得锁后，内层函数仍可获取该锁。

**特点：发生在同一线程内**

Code: ReentrantTest101



执行输出“method2”

可重入锁的最大作用就是可以避免死锁。

这里讲的是广义上的可重入锁，而不是单指JAVA下的ReentrantLock。ReentrantLock 和synchronized 都是可重入锁

实现原理

state不为0，

#### 锁优化

高效并发是从JDK 1.5到JDK 1.6的一个重要改进，HotSpot虚拟机开发团队在这个版本上  
花费了大量的精力去实现各种锁优化技术，这些技术都是为了在线程之间更高效地共享数据，减少获得锁和释放锁所带来的性能消耗，以及解决竞争问题，从而提高程序的执行效率。

**锁粗化**:将多个连续的锁扩展成一个范围更大的锁，用以减少频繁互斥同步导致的性能损耗。

**锁消除:** JIT通过逃逸分析，去除堆上不会被其他线程访问到的代码锁。。

**轻量级锁：**轻量级锁即乐观锁。多个线程交替进入临界区，只需要依靠一条CAS原子指令就可以完成锁的获取及释放。

**偏向锁：**对轻量级锁进一步优化。使用CAS指令置换对象头ThreadID，下一次同一个线程进入则偏向该线程，无需任何同步操作。当有竞争，需解除偏向锁，进入轻量级锁

**适应性自旋:**为了避免线程频繁挂起、恢复的状态切换消耗。产生了忙循环（循环时间固定），即自旋。JDK1.6引入了自适应自旋。自旋时间根据之前锁自旋时间和线程状态，动态变化，用以期望能减少阻塞的时间。

#### 死锁

[Reference](https://juejin.im/post/5aaf6ee76fb9a028d3753534)

多进程在执行过程中，由于竞争资源或者通信而造成的一种阻塞的现象。

竞争的资源，可以是线程池里的线程、网络连接池的连接，数据库中数据引擎提供的锁，等等一切可以被称作竞争资源的东西。

如何排查？

执行jstack命令，查看线程是否一致处于WAITING状态。



### 1.5.5 Concurrent包

JDK5将Doug Lea的并发库引入到Java标准库中，提供新的启动(submit)、调度、管理线程的一大堆API了。增强了并发功能，简化了代码。主要特性：

* ReentrantLock
* 线程池
* 新增并发集合类ConcurrentHashMap，ConcurrentLinkedQueue等
* Semaphore

#### 1.5.5.1 线程池

四种线程池

* **newCachedThreadPool** 根据处理需要，可灵活回收和新建线程。

maximumPoolSize = Integer.MAX\_VALUE(Integer是4位，-2^31~2^31-1)

* **newFixedThreadPool** 定长线程池，超出的线程会在队列中等待。
* **newScheduledThreadPool** 定长线程池，支持定时及周期性任务执行。
* **newSingleThreadExecutor** 单线程化的线程池，保证所有任务按照指定顺序(FIFO, LIFO, 优先级)执行。

原理

复用执行前面任务的线程，add过程受阻将添加到workQueue

TODO1 如何复用

线程池作用

* 减少在创建和销毁线程上花的时间
* 控制系统资源的开销。如果不使用线程池，有可能造成系统创建大量同类线程而导致消耗完内存或者“过度切换”的问题

**DB Pool源码分析**

[Reference](https://blog.csdn.net/shuaihj/article/details/14223015)

TODO4

TODO1:如何检查线程池内线程运行情况，busy？ study监控API

#### 1.5.5.2 ReentrantLock

实现互斥串行操作的

ReentrantLock VS synchronized

本质上synchronized的lock升级到重量级锁依赖OS的mutex lock，有线程切换，会引起线程阻塞和上下文切换。相对于乐观锁，系统资源开销大。

* Synchronized的锁是JVM层面上实现的,是原生语法层面。ReentrantLock是API层面的互斥锁（用起来会复杂一些，需要写控制锁的代码，加锁和解锁都需要显式写出）。
* ReentrantLock更灵活、更强大，增加了锁轮训、超时(tryLock[time])、中断(lockInterruptibly)等高级功能。
* synchronized采用悲观锁?，而Lock用的是乐观锁方式。

#### 1.5.5.3 ReadWriteLock

Compare with synchronized

 synchronized的修饰区域每次只有一个线程可以访问。这种强烈的互斥性使得每次不管是读数据还是写数据都只能有一个线程可以操作。**在有多个读线程可以并行执行**的情况下，它并不是一个理想的选择。

Compare with volatile

使用读写锁的方式在单个写线程加多个读线程的情况下，其实差别不大。在使用volatile变量的时候，因为每次操作修改的结果对于全局都是可见的。那么在只有一个写线程的情况下，只有这个线程可以唯一修改数据。不会存在有几个写线程而产生的竞争条件。对于有**多个写线程的情况下，volatile变量就不能保证数据的一致性了**。而ReentrantReadWriteLock却可以有锁的机制保证互斥。它同时也尽可能保证了足够大的并行性。

#### 1.5.5.5 Semaphore

基于java同步器AQS

#### 1.5.5.6 Atomic

Atom包下的原子操作类实现了一种乐观锁，其本质是使用了CPU级别的CAS指令。

AtomicReference

对象引用变量读和写都是原子性的

原理：它是通过"volatile"和"Unsafe提供的CAS函数实现"原子操作

compareAndSet(expect,update)

LockSupport

比较底层，应用开发还是用Lock

park VS wait

#### LimitLatch

### 1.5.6 fork-join

[Reference](https://www.liaoxuefeng.com/article/001493522711597674607c7f4f346628a76145477e2ff82000)

TODO1

## 1.5 JVM

**谈谈你对JVM的理解？**

Java语言的一个非常重要的特点就是与平台的无关性。而使用Java虚拟机是实现这一特点的关键。Java源文件经编译成字节码程序，通过JVM将每一条指令翻译成不同平台机器码，通过特定平台运行。

JVM执行程序过程 ：加载class文件 🡪管理并分配内存(生成对象) 🡪执行垃圾收集

### 1.5.2 类加载机制

#### 类加载过程

Class对象自身的实例化过程。

注意：这里的加载指类的初始化过程，和类是否实例化无关。

[Reference](https://blog.csdn.net/ns_code/article/details/17881581)

类加载分了三个阶段，loading，linking和initializing。

1 )加载

1.加载Class编译后的二进制字节流到方法区

2.//方法区初始化静态变量

3.在堆中创建一个 Class 类**对象**。通过该对象访问方法区中的这些数据。

hotspot源码显示最后调用allocate\_permanent方法在PermGen分配内存，所以instanceKlass对象保存在永久代区域；

2 )链接

验证--准备--解析

**准备**

* 为类变量**分配内存**（方法区）空间并设置默认值。
* 不包括实例变量。实例变量会在对象实例化时随着对象一块分配在堆中。

**解析**

* 解析阶段是虚拟机将常量池中的符号引用转化为直接引用的过程。
* 解析可能在Initializing之前，也可能在之后这是为了支持 Java 语言的运行时绑定（也成为动态绑定或晚期绑定）

3 )初始化

 初始化是类加载过程的最后一步，到了此阶段，才真正开始执行类中定义的代码**。**

从另一个角度来表达：初始化阶段是执行类构造器<clinit>()方法的过程。

**<clinit>**

Code: ClassClinit100

编译器自动收集，类变量的赋值语句和静态语句块中的语句合并产生。收集顺序是由源文件中出现的顺序所决定的。如果一个类中没有静态语句块，也没有对类变量的赋值操作，那么编译器可以不为这个类生成此方法。



**<clinit> VS <init>**

两种都是编译器生成的方法。

**<clinit>：**是类（或接口）构造器，在类加载的初始化阶段对类进行初始化，类变量内存分配在方法区。

**<init>：**实例构造器。

<init>加入到构造器第一句执行



#### 触发条件

* JVM执行指令new/getstatic/putstatic/invokestatic/。
* 类反射。
* 当初始化一个类的时候，如果其父类还没有进行初始化，则需要先触发其父类的初始化。
* 当虚拟机启动时，用户需要指定一个要执行的主类，虚拟机会先执行该主类。

具体Case分析

1.类的初始化后并未类的实例生成，比如getstatic….



2.单例中类Singleton初始化阶段会生成对象instance

这里生成对象instance和Singleton的初始化顺序？

Code: Singleton101



几个阶段是按顺序**开始**，而不一定按顺序**完成**，通常都是交叉进行的。这里<clinit>执行过程激活new的执行。

#### 模块化

（jboss modules、osgi、jigsaw）

### 对象内存管理

#### 运行时数据区



JVM区域总体分两类，heap区和非heap区。

Heap

堆是可供各条线程共享的运行时内存区域，也是供所有类实例和数组对象分配内存的区。

存储对象实例，原则上所有的对象都在堆区上分配内存。



Eden Space（伊甸园）&Survivor Space(幸存者区)&Old Gen（老年代）。

非heap

Code Cache(代码缓存区)&Perm Gen（永久代）&Jvm Stack&Local Method Statck

#### 内存分配和回收

[Reference](http://www.cnblogs.com/zhguang/p/3257367.html#access)

Java内存分配和回收的机制概括的说就是：分代分配，分代回收。

根据对象存活的时间分为：年轻代、年老代、永久代(方法区）,详细见见下文。。。

new过程：JVM加载类🡺为对象分配内存🡪 调用构造函数初始化成员变量(<init>)

* Eden区：伊甸(yī diàn)园(传说夏娃和亚当偷吃禁果的地方，用来表示对象内存首次分配的区域)。绝大多数刚创建的对象会被分配在Eden区（大对象可以直接被创建在老年代）。Eden是连续的内存空间，因此在其上分配内存极快。
* Survivor: Eden区分配的多数对象很快就会消亡，Eden区满的时候，执行Young GC，将消亡的对象清理掉，并将剩余的对象复制到空闲的Survivor。回收过程中，总是有一块 Survivor 区域是空闲的。
* Old区：Young GC过程中, 在两个Survivor区多次切换（HotSpot虚拟机默认15次）的长生命周期对象会将被复制到老年代。

**Q:java虚拟机堆何时分配? 存放哪些数据? 何时回收？**

**Q:jvm 如何分配直接内存， new对象如何不分配在堆而是栈上，常量池解析?**

#### 内存溢出&泄漏

**内存溢出**

申请了10个字节的空间，写入11或以上字节的数据，就会溢出。

java.lang.OutOfMemoryError，一般都是程序问题或者JVM参数配置问题引起

* Java heap space:
* PermGen space ：1）应用中很多class，web服务器对JSP进行pre compile时2）webapp加载大量的第三方jar
* unable to create new native thread

Memory leak

[Reference](http://blog.csdn.net/dongnan591172113/article/details/79044678)

对象申请的内存不再使用，因为仍被引用着，GC无法移除，导致内存泄漏。

内存泄漏可能导致内存溢出。

**1. 长生命周期的对象持有短生命周期对象的引用**

静态集合类的生命周期跟应用程序一样长，有可能会发生内存泄漏，看下面代码：



集合类如果是局部变量，在方法栈退出后就没有引用了会被jvm正常回收，不会造成内存泄露。而如果是全局性的变量提供这样的删除机制或者定期清除策略非常必要。

Java集合类已经提供了删除机制



**2.集合里的对象属性被改变**



**3.监听器**

在Java中，我们经常会使用到监听器，如对某个控件添加单击监听器addOnClickListener()，但往往释放对象的时候会忘记删除监听器

**4.各种连接**

Java中的连接包括数据库连接、网络连接和io连接，如果没有显式调用其close()方法，是不会自动关闭的，这些连接就不能被GC回收而导致内存泄漏。一般情况下，在try代码块里创建连接，在finally里释放连接，就能够避免此类内存泄漏。

**5.外部模块的引用**

调用外部模块的时候，方法传入对象，也应该注意防止内存泄漏。如：

public void register(Object A)

这个方法传入对象A模块的引用，这时需要B模块提供了去除引用的方法，如unregister()。这种情况容易忽略，而且发生了内存泄漏的话，比较难察觉，应该在编写代码过程中就应该注意此类问题。

**6.单例模式**

因为单例对象初始化后将在JVM的整个生命周期内存在，如果它持有一个外部对象（生命周期比较短）的引用，那么这个外部对象就不能被回收，而导致内存泄漏。如果这个外部对象还持有其它对象的引用，那么内存泄漏会更严重，因此需要特别注意此类情况。

StackOverflowError（堆栈溢出）

请求的栈的深度大于虚拟机所允许的深度

同步性

类加载由JVM自动同步，而执行对象的构造器可以中断的并发原子操作

### 1.6.3 GC

[Reference](http://www.cnblogs.com/ityouknow/p/5614961.html)

概括地说，GC对对象的内存进行标记，并确定哪些内存需要回收。根据一定的回收策略，自动的回收。保证JVM中的内存空间，防止出现内存泄露和溢出问题。

学习Java GC机制，可以帮助我们在日常工作中排查各种内存溢出或泄露问题，解决性能瓶颈，达到更高的并发量，写出更高效的程序。

Java 中的堆也是 GC 收集垃圾的主要区域。

#### 垃圾收集器

* Serial收集器:串行收集器是最古老，最稳定以及效率高的收集器，可能会产生较长的停顿，只使用一个线程去回收。新生代、老年代使用串行回收；新生代复制算法、老年代标记-压缩；垃圾收集的过程中会Stop The World（服务暂停）

参数控制：-XX:+UseSerialGC 串行收集器

* ParNew收集器:ParNew收集器其实就是Serial收集器的多线程版本。新生代并行，老年代串行；新生代复制算法、老年代标记-压缩

参数控制：-XX:+UseParNewGC ParNew收集器

-XX:ParallelGCThreads 限制线程数量

#### 对象存活判定

**Q:谁会被GC，什么时候GC,如果想不被GC怎么办?**

目前虚拟机基本都是采用可达性算法

**引用计数器**

每个对象有一个引用计数器，当对象被引用一次则计数器加1，当对象引用失效一次则计数器减1，对于计数器为0的对象意味着是垃圾对象，可以被GC回收。

* 优点：1.实时回收 2.回收应用无需挂起3.回收具有区域性，不用扫描全部对象。
* 缺点 ：1.无法解决循环引用问题 2.浪费资源，即使内存够用，仍然在计数统计

**可达性算法**

GC Roots可达的对象便都是存活对象，不可到达的对象便是GC回收的对象。

可作为GC Roots的对象：

* 栈的局部变量表所引用的对象；
* 本地方法栈的JNI所引用的对象；
* 方法区的静态变量和常量所引用的对象；



对象实例1、2、4、6都具有GC Roots可达性，也就是存活对象，不能被GC回收的对象，而对于对象实例3、5直接虽然连通，但并没有任何一个GC Roots与之相连，这便是GC Roots不可达的对象，这就是GC需要回收的垃圾对象。

#### 垃圾回收算法

GC 分为Partial GC（Young GC）和Full GC。

**GC回收策略**

Mark-Sweep

标记-清除回收算法

1. 标记和清除过程的效率都不高 2.空间碎片太多。常用可达性算法。



Copying

复制收集算法

这种算法的代价是将内存缩小为原来的一半，持续复制长生存期的对象则导致效率降低



**Young GC**

* GC触发条件**：** Eden区分配满的时候触发。
* 回收过程： 在Eden区创建对象 🡪Eden区处理消亡对象 🡪复制剩余对象到Survivor 🡪 清理出一块空闲的Survivor 🡪复制长周期对象到老年代

**Old GC**

只收集old gen的GC。只有CMS的concurrent collection是这个模式。

**Full GC**

* 触发条件： old gen剩余的空间不够用。
* 回收范围：对整个新生代、老生代、元空间（metaspace）的全局范围的GC；。

4. 介绍GC 和GC Root不正常引用。

7. 数组多大放在 JVM 老年代（不只是设置 PretenureSizeThreshold ，问通常多大，没做过一问便知）

 老年代中数组的访问方式

9. GC 算法，永久代对象如何 GC ， GC 有环怎么处理

11. 12. 如果想在 GC 中生存 1 次怎么办

深入分析了Classloader，双亲委派机制

### 1.6.5 VM 指令集

[Reference](http://blog.csdn.net/dongnan591172113/article/details/51832690)

常量🡪栈顶。

|  |  |  |
| --- | --- | --- |
| iconst\_1 | 将int型的1推送至栈顶 | 4: iconst\_1 |
| fconst\_1 | 将float型的1推送至栈顶 |  |
| bipush | 将单字节(push byte)的常量值(-128~127)推送至栈顶 | 9: bipush 10 |
| sipush | 将一个短整型常量值(-32768~32767)推送至栈顶 |  |
| ldc | 超过32767。或者String常量 | 15: ldc #5 |

栈顶 ⬄ 本地变量

|  |  |  |
| --- | --- | --- |
| istore\_1 | 将栈顶int型数值存入第二个本地变量 |  |
| astore | 将栈顶**引用型数值**存入指定本地变量 |  |
| astore\_0 | 将栈顶引用型数值存入第一个本地变量 |  |
| astore | 将指定的引用类型本地变量推送至栈顶 |  |
| iload\_1 | 将第二个int型本地变量推送至栈顶 |  |
| aload\_0 | 将第一个引用类型本地变量推送至栈顶 |  |

pop

对栈顶进行操作

|  |  |  |
| --- | --- | --- |
| pop | 将栈顶数值弹出 |  |
| dup | 复制栈顶数值并将复制值压入栈顶 |  |

**invoke**

|  |  |  |
| --- | --- | --- |
| invokevirtual | 调用实例方法 |  |
| invokestatic | 调用静态方法 |  |
| invokeinterface | 调用接口方法 | Map.put |
| invokespecial | 调用超类构造方法/实例初始化方法/私有方法 |  |

5: invokespecial #3 // Method "<init>":(I)V

4: invokespecial #3 // Method "<init>":()V

域操作

成员变量

|  |  |  |
| --- | --- | --- |
| getstatic | 获取静态域，并将其值压入栈顶 |  |
| putstatic | 用栈顶的值为指定的类的静态域赋值 |  |
| getfiled | 获取实例域，并将其值压入栈顶 |  |
| putfileld | 用栈顶的值为指定实例域赋值 |  |

运算

|  |  |  |
| --- | --- | --- |
| iadd | 将栈顶两int型数值相加并将结果压入栈顶 |  |
| ishl | 将int型数值左移位指定位数并将结果压入栈顶 |  |

new及数组

|  |  |  |
| --- | --- | --- |
| new | 创建一个对象，并将其引用值压入栈顶 |  |
|  | 将int型数值左移位指定位数并将结果压入栈顶 |  |

javap

[Reference](http://blog.csdn.net/qq_24489717/article/details/53837493)

* -c 输出编译后的JVM指令



* -verbose 输出常量池，栈大小，方法参数。
* -l 输出行和变量的表。
* -public 只输出public方法和域
* -protected 只输出public和protected类和成员
* -package 只输出包，public和protected类和成员，这是默认的
* -p -private 输出所有类和成员
* -s 输出内部类型签名
* -constants 输出静态final常量
* 不加参数:只显示方法名
* -help 帮助

实例：

**int** a1=10;**int** a2=10;**int** a3=a1+a2;编译后的字节码



可见

变量赋值过程：数值通过命令的操作数push到栈顶->store到变量

Integer a1=10;



这里astore传的是引用！

局部变量。bipush/ldc等。

[Reference](https://blog.csdn.net/zhangjg_blog/article/details/22432599)

包含LineNumberTable和LocalVariableTable



LineNumberTable：描述了一条字节码和源码行号的对应关系

LocalVariableTable：局部变量与源代码中的局部变量之间的对应关系

## 1.6 Feature

本章主要讲述各特性的原理。

JDK

TODO3JDK&JRE&JVM

和Google的Android关系https://www.zhihu.com/question/270271649

### Java发展

JDK 1.1

1997年发布。

* 反射
* JDBC；
* 支持内部类；
* 引入Java Bean；
* 引入RMI（Remote Method Invocation）；

JDK 1.2

1998年发布

* Collection
* 引入Java 插件；
* 对字符串常量做内存映射；
* 引入JIT（Just In Time）编译器；
* 引入对打包的Java文件进行数字签名；
* 引入控制授权访问系统资源的策略工具；
* 引入JFC（Java Foundation Classes），包括Swing 1.0、拖放和Java 2D类库；

JDK 1.4

2002年发布

* NIO
* XML处理；
* Java打印服务；
* 引入Logging API；
* 引入Java Web Start；
* 引入断言Assert；

引入Preferences API；

* 引入链式异常处理；
* 支持IPv6；
* 支持正则表达式；
* 引入Image I/O slot machine API。

JDK 5

2004年发布

[Reference](https://sakuraffy.github.io/java_jdk5.0/)

* 泛型
* concurrent包
* 枚举
* 元数据（Metadata）
* foreach循环
* 自动拆箱装箱
* 静态导入
* 可变参数 add(int... arr)
* 内省

JDK6

2006年发布

* 支持脚本语言

JDK7

2011年发布

* 支持动态语言
* 异步IO（NIO 2）
* Lambda的实现

JDK 8

2014年发布

* Lambda表达式
* Optional防NPE
* Pipelines和Streams
* Date和Time API
* Default方法
* Type注解
* Nashhorn JavaScript引擎
* 并发计数器
* Parallel操作
* 移除PermGen Error
* TLS SNIJava 8的内存分代改进

### 泛型

TODO3

### 反射

Class.forName和classloader的区别

Class.forName(xxx);的作用是要求JVM查找并加载指定的类。

反射效率低。需要读取字节码到内存然后分配内存生成类实例最后执行。

内省（Introspector）

### Assert

TODO3

### transient

TODO2

### Lambda

面向对象编程是对数据进行抽象，而函数式编程是对行为进行抽象，本质是java的语法糖，让开发更加专注于业务实现。

代码简洁，可读性好。比如事件处理，集合处理，回调。可以传递函数。

### Optional

TODO2:

### Pipelines&Streams

TODO1:JDK8,9新特性

JDK

# 第二篇 Java Web

## 2.1 Base

### 2.1.1 网络技术

### 2.1.2 Servlet规范

Servlet接口是一套处理网络请求的规范，J2EE规范之一，主要为了扩展java的web功能。应用开发人员根据规范编写业务程序，servlet容器实现HTTP服务(面向接口编程)。

## 2.2 框架

### 2.3 Spring

#### 2.3.1 IoC

**依赖倒置**：高层对抽象接口编程，**低层依赖于高层模块的需求抽象。**

**好处：**通过IoC容器管理对象之间的依赖关系，降低了资源双方的耦合度。Spring容器管理资源对象和对象之间的依赖关系，实现资源的可配置和易管理。

#### 2.3.2 AOP

方便进行面向切面的编程，很多传统OOP不容易实现的功能可以通过AOP轻松应对。

声明式事务

通过声明的方式灵活地进行事务管理，开发不用写事务管理代码（提高开发效率和质量）。

日志

#### 2.3.3 Spring MVC

基于Servlet

Request🡪 DispatcherServlet (web.xml)🡪 HandlerMapping(All Controller Map)🡪 Controller

在HandlerMapping中找到对应服务方法并调用。这里是反射call，CglibAopProxy静态代理。

MVC分层好处

1. 有利于分工
2. 复用
3. 好扩展，好维护

除了上述特性，使用spring，方便集成各种优秀的框架，降低各种框架的使用难度。降低Java EE API的使用难度等。

### 2.4 Hibernate

#### 缓存机制

* Session的缓存: 应用场景太单一，系统中大量的列表式查询缓存起不到作用。系统中通过ThreadLocal在线程中重用Session，每个线程可能需要大量处理不用的业务逻辑，缓存命中率很低。
* SessionFactory应用级缓存。EHCache是Hibernate中的二级缓存插件，使用Hibernate的系统可以直接使用EHCache缓存。

#### API

##### SessionFactory

openSession打开一个新的session

getCurrentSession获取当前线程里的session，如果没有才打开新的

##### Session

[Reference](http://coolvinson.iteye.com/blog/520610)

get

load 返回代理对象，延迟加载

#### 丢失更新

略

实现方案

处理的数据变化引起其他潜在问题,比如营业额增加，银行转账。

Mysql解决的方法：

* 乐观锁：对数据version化，update前检查数据一致性。Hibernate框架已经实现。
* 悲观锁(for update)，在DB层面实现。
* 设置为Serializable

(Oracle可以把隔离级别设置成read-only。?)

Hibernate LockMode

LockMode.NONE ：有缓存用缓存，没缓存则从数据库读   
LockMode.READ ：直接从数据库读，不使用缓存数据   
LockMode.WRITE ：在insert update数据的时候，HIBERNATE内部使用的。   
以上3种均为HIBERNATE级别的锁，也就是缓存级别的锁。   
下面2种为数据库级别的锁：   
LockMode.UPGRADE：相当于SQL语句select for update，被select的数据都被数据库锁住了，不能被其他事务修改。   
LockMode. UPGRADE\_NOWAIT ：是ORACLE数据库特有的select for update nowait

只读事务

[Reference](https://www.cnblogs.com/hackem/p/3890656.html)

设置了readonly后，connection都会被赋予readonly，效果取决于数据库的实现



Mysql：true的时候会读到最新值

##### 加锁策略

并发修改同一记录时，避免更新丢失，需要加锁。要么在应用层加锁，要么在缓存加锁，要么在数据库层使用乐观锁，使用version作为更新依据。

说明：如果每次访问冲突概率小于20%，推荐使用乐观锁，否则使用悲观锁。乐观锁的重试次数不得小于3次。

#### DB连接池

建立连接是一个费时的操作，频繁的进行数据库连接操作势必占用很多的系统资源，影响网站的响应速度。

如果查询完数据库后，不关闭连接，而是暂时存放起来，当别人使用时，把这个连接给他们使用。就避免了一次建立数据库连接和断开的操作时间消耗。

基本思想

为数据库连接建立一个“缓冲池”。预先在缓冲池中放入一定数量的连接，当需要建立数据库连接时，只需从“缓冲池”中取出一个，使用完毕之后再放回去

TODO3:C3P0源码

## 2.3 中间件

[Reference](https://kb.cnblogs.com/page/196448/)

一般情况，中间件应用于分布式环境，主要解决异构网络环境下模块的互连与互操作问题，提供标准接口、协议，屏蔽实现细节（将具体业务和底层逻辑解耦的组件），提高应用系统移植性。

分类：数据访问中间件，远程调用中间件，消息中间件，交易中间件，对象中间件。

### 消息中间件

引入消息中间件是用来降低应用之间的耦合的。

对于上游应用来说，它往往会依赖非常多的下游应用，如果全部用同步调用的方式，那么上游应用必须等到所有下游应用全部执行成功后，才能返回成功，这种等待往往是不可忍受的。为此，需要将同步异步化，而消息中间件可以实现这种异步化。

使用消息中间件后，上游应用只需要向消息服务器发送消息就可以了，它不用等待所有应用执行完，它可以提前返回执行成功，后续，消息服务器会去向这些下游应用发消息，通知它们执行。

[Reference](http://blog.csdn.net/dongnan591172113/article/details/8771441)

**使用场景**

1. 系统整合：

系统整合：分布式环境中,用于同构/异构系统的整合,而且系统间耦合度较低，

1)AB系统,A宕机

2)性能：系统差不干扰B

3) 系统整合模块交互点统一，变化集中在一点，不用逐个模块修改,垂直切分容易

2.分解系统: 通过消息分解系统。

3. 异步:1.推消息 2.流量削峰(据同步:web应用->缓存，搜索，db)

### Memcached

并发的处理

1.单个命令是原子操作。命令将被串行化、先后执行。

2.命令序列不是原子的。在并发的情况下，get and set可能覆写其他进程set的item。Memcached 1.2.5以及更高版本，提供了gets和cas命令，可解决。

JAVA客户端

1. Memcached client for java (code: MemCache120)

官方提供, 运行比较稳定，使用阻塞IO，不支持CAS操作。

1. XMemcached

**分布式缓存**

## 2.4 应用服务器

TODO1:为什么大多数的服务都运行在Linux

TODO2如何做压力测试出负荷值

### 2.6.1 Tomcat

遵循J2EE Servlet规范，实现Servlet和JSP的支持，又叫Servlet容器。在中小型系统和并发访问用户不是很多的场合下被普遍使用。

Apache VS tomcat

apache是web服务器，专门提供HTTP服务。处理静态内容，比如HTML/图片。

tomcat是java应用服务器，servlet容器，是apache的扩展

#### 5.2.1 性能优化

Tomcat的默认配置作为生产环境，尤其是内存和线程的配置，默认都很低，容易成为性能瓶颈。 可以从三方面入手，内存/线程/IO

### 5.2 Jetty

### 5.2 Apache

### 5.3 Nginx

**优点**

* 异步非阻塞，高并发环境消耗更少的资源
* 配置简洁,

#### Nginx与apache

Nginx性能高，而apache是阻塞的。但比较稳定。

1.反向代理，缓存静态文件 2.负载均衡

## 2.5 分布式应用

TODO2 JTA分布式事务等

分布式系统往往是把应用拆分成多个应用，每个团队维护一个应用，应用与应用通过远程过程调用或者消息中间件通信。这种系统的优点是能够做到高内聚低耦合，可以支撑业务的快速发展。

对于分布式系统中的某个应用来说，它的入口往往有

**三个入口：**远程调用，消息中间件，web访问

**出口则有很多**

* 远程过程调用：
* 消息中间件
* 访问分布式缓存
* 访问分布式日志
* 访问数据库
* 执行调度任务

分布式 VS 集群

分布式：一个业务分拆多出个子业务，部署在不同的服务器上

集群：同一个业务，部署在多个服务器上

通过接口分解系统解耦性，不同子系统通过相同的接口描述进行服务调用。

涉及的技术

* 远程服务框架

远程服务框架是分布式系统中不可或缺的，对于远程服务框架来说，关键的几个点是：

服务消费者、服务提供者、协议（java序列号、反序列化、hessian协议等等）

网络（NIO、多路复用等等）

地址列表

目前业界比较有名的开源框架有：Apache Thrift（跨语言）、Dubbo等等。

* 消息中间件

引入消息中间件是用来降低应用之间的耦合的。对于上游应用来说，它往往会依赖非常多的下游应用，如果全部用同步调用的方式，那么上游应用必须等到所有下游应用全部执行成功后，才能返回成功，这种等待往往是不可忍受的。为此，需要将同步异步化，而消息中间件可以实现这种异步化。

使用消息中间件后，上游应用只需要向消息服务器发送消息就可以了，它不用等待所有应用执行完，消息服务器会去向这些下游应用发消息，通知它们执行。

总结:发送异步消息，使上游应用不必等下游应用执行完，降低应用之间的耦合的。

* wrapper层

wrapper层主要是上面两个入口的包装层，进行一些异常管理，日志管理和AOP拦截

* 服务管理和配置管理

业务逻辑层：流程引擎和规则引擎，配合spring来治理极易变动的业务逻辑，后续有空单独写个流程引擎

负载均衡器单点故障

实时管理集群

分布式事务

### RPC

### Restful

规范

接口授权OAuth

TODO1

### SOA

传统分布式按功能模块把应用垂直拆分成多个应用。缺点：当垂直应用越来越多，模块间相互交叉调用越来越复杂，模块耦合度太高，扩展性差，不好维护和部署

SOA对系统水平拆分，将核心业务抽取出来，作为独立的服务,降低模块之间耦合度，好扩展，方便部署。

总结：当垂直应用越来越多，应用之间交互不可避免，将核心业务抽取出来，作为独立的服务，逐渐形成稳定的服务中心，

1.扩展性 2提高开发和维护效率 3.伸缩性?

### Web service

WebService，顾名思义就是基于Web的服务。基于HTTP传输协议的程序。接收和响应外部系统的某种请求。从而实现远程调用。

### CAP理论

### 分布式事务

### 微服务

微服务 vs SOA

SOA和微服务有很多概念是相通的，并且微服务是由SOA演化过来的，SOA更加专注服务之间的治理，微服务考虑的更全面一些，微服务更独立一些，最明显微服务要求有自己独立的库，soa只是强调服务的独立。

TODO1

# 第三篇 架构设计

## 4.1 OO设计准则

基本思想：尽可能的抽象，将变化控制到最小，做到高内聚、低耦合。

实现方法：使用对象、类、继承、封装、多态等基本概念来进行程序设计。

### 3.1.1 Object-Oriented

Object-Oriented是一种程序设计方法，从现实世界中客观存在的事物出发来构造软件系统，尽可能运用人类的自然思维方式。

基本思想：使用对象、类、继承、封装、多态等基本概念来进行程序设计。

**OO VS Procedure-oriented**

都是模块化编程，Procedure-oriented把问题分割成以功能函数为单位的模块中，

Object-Oriented把问题抽象描述成以对象为单位的模块中。

注意：本章主要讲述各特性的原理。

#### 封装

封装数据结构和函数，隐藏细节(private/public),避免被外界破坏。

好处1.符合人的自然思维2.从单一职责的角度来说？。

#### 多态

实现原理

方法的动态绑定，实现运行时的类型决定对象的行为。多态的表现形式是父类指针或引用指向子类对象，在这个指针上调用的方法使用子类的实现版本。多态是IOC、模板模式实现的关键

TODO4: invokevirtual执行过程

[Reference](https://blog.csdn.net/qq_35614059/article/details/79141866)

C++多态是通过虚函数来实现的。

#### 接口&抽象

接口优于抽象类：接口更加抽象，所以更稳定。

Interface vs Abstract

1. abstract class和interface最大的区别在于设计理念不同。其实abstract class表示的是"is-a"关系，是一种继承关系。interface表示的是"like-a"关系。一个类只能使用一次继承关系而一个类却可以实现多个interface。

2.语法上的区别，abstract class 可以有自己的数据成员，也可以有非abstract的成员方法

### 3.1.2 低耦合&高内聚

模块独立性指每个模块只完成系统要求的独立子功能，并且与其他模块的联系最少且接口简单，两个定性的度量标准――耦合性和内聚性。

#### 耦合性

指模块间相互联系紧密程度。为提高模块的独立性，模块间尽可能松散的耦合。

耦合性分类(低🡪高):

* 无直接耦合: 无调用关系，通过主模块的控制和调用产生联系。比如controller调用的service。

下面几种模块之间存在调用关系：

* 数据耦合: 传递简单的数据值，即值传递。
* 标记耦合: 传递的是数据结构，如数组名、类。
* 控制耦合: 传递的是控制变量（如开关、标志等）。
* 公共耦合: 访问同一个全局数据结构（包括单个变量），
* 内容耦合: 可能在汇编语言中出现。大多数高级语言都已设计成不允许出现。

1.高耦合度导致系统难维护，牵一发动全身，又会引入新的BUG。2.且不好拓展。

#### 内聚性

模块内各元素（变量，语句、、、）联系的越紧密，则它的内聚性就越高。在模块划分时，遵循“一个模块，一个功能”的原则，尽可能使模块达到功能内聚。

内聚性分类(低🡪高)

* 偶然内聚: 模块内各元素之间没有任何联系。
* 逻辑内聚: 模块把几种相关的功能组合在一起，每次被调用时，由传送给模块参数来确定该模块应完成哪一种功能 。
* 时间内聚: 把需要同时执行的动作组合在一起。(无顺序)
* 过程内聚：必须以特定次序执行。（有顺序）
* 通信内聚:处理使用相同的输入数据或者产生相同的输出数据。
* 顺序内聚: 顺序+同一功能。前一功能元素输出就是下一功能元素的输入。（）
* 功能内聚: 顺序+同一功能 +不可再分

好处见单一职责模式。设计类或者模块尽量做到“高内聚，低耦合”。

### 3.1.3 设计准则

[Reference](https://blog.csdn.net/dongnan591172113/article/details/51996421)

需求总是变化，所以就需要根据准则设计软件，变化满足需求的同时还能保持软件内部的封装体系稳定，不被需求的变化影响。且使我们的代码稳健，易于复用，易于拓展，灵活优雅。

设计准则 VS 设计模式

不同的设计模式对应不同的需求，而设计原则则代表永恒的灵魂，需要在实践中时时刻刻地遵守。

在实际编程中，根据设计准则，总结如下几点：

* 低层模块尽量都要有抽象类或接口，或者两者都有。
* 变量的声明类型尽量是抽象类或接口。
* 使用继承时遵循里氏替换原则。

#### 3.1.4.1 单一职责

一个类(模块)，一个功能。职责过多，可能引起它变化的因素就越多，增加系统错误风险。

#### 3.1.4.2 开放封闭原则

核心思想

对修改封闭的，对扩展开放。

具体实现

**封闭：**对抽象编程，而不对具体编程。**因为抽象相对稳定**。让类依赖于固定的抽象，所以修改就是封闭的。

**开放：**通过面向对象的继承和多态机制，重写/重载方法来改变固有行为，实现拓展，所以就是开放的。

#### 3.1.4.3 接口隔离

核心思想

使用多个小的专门的接口，而不要使用一个大的总接口。接口应该是内聚的，应该避免“胖”接口。

一个类对另外一个类的依赖应该建立在最小的接口上，不要强迫依赖不用的方法，这是一种接口污染。  
    胖接口存在明显的弊端，会导致实现的类型必须完全实现接口的所有方法、属性等；而某些时候，实现类型并非需要所有的接口定义，而且对胖接口的修改将导致一连串的客户端程序需要修改

具体实现

分离的手段主要有以下两种：

1、委托分离，通过增加一个新的类型来委托客户的请求，隔离客户和接口的直接依赖，但是会增加系统的开销。

2、多重继承分离，通过接口多继承来实现客户的需求，这种方式是较好的。

分离后，每个类对应一个功能方法，将接口组装到类中，有个性化需求时，只要跟换新的实现类即可。

#### 3.1.4.4替换原则

核心思想：子类必须能够替换其基类。这一思想体现为对继承机制的约束规范

好处：保证继承复用是可靠地，违反必然导致违反开放封闭原则。

实现：面向接口编程

#### 3.1.4.5 依赖倒置

**核心思想**

**面向接口编程**。**上层模块不应该依赖底层模块，它们都应该依赖于抽象，低层依赖于高层模块的需求抽象。**

面向接口

依赖存在于类与类、模块与模块之间。当两个模块之间存在紧密的耦合关系时，最好的方法就是分离接口和实现，来有效控制耦合关系，达到依赖于抽象的设计目标。  
    抽象的稳定性决定了系统的稳定，依赖于抽象是面向对象设计的精髓，也是依赖倒置原则的核心。  
传递依赖：1.接口传递 2.构造方法3.setter

**倒置**



（图一） （图二）

图1依赖关系为高层依赖底层，图2依赖关系被颠倒（反转），从而使得低层次模块依赖于高层次模块的需求抽象。

## 4.2 设计技巧

### 3.2.1 继承&封装&多态

面向接口

### 3.2.2 内部类

TODO1

**什么是java序列化，如何实现java序列化?**

序列化就是一种用来处理对象流的机制，所谓对象流也就是将对象的内容进行流化。可以对流化后的对象进行读写操作，也可将流化后的对象传输于网络之间。序列化是为了解决在对对象流进行读写操作时所引发的问题。

### 3.2.3 异常

[Reference](https://www.jianshu.com/p/503901e5be2b)

#### 3.2.2.1 异常分类



**Error**

编译和系统的错误，不允许捕获。

Checked异常

可以被处理的异常，必须显示处理。比如ClassNotFoundException / IOException

RuntimeException

产生频繁，处理麻烦，若显示申明或者捕获将会对程序的可读性和运行效率影响很大。所以由系统自动检测并将它们交给缺省的异常处理程序。

#### 3.2.2.1 使用技巧

1处理好异常能保证系统的健壮性,安全性

2.重要的调试手段，根据抛出的异常能判断发生的位置和出了什么出错，为什么出错。

3.代码可读性，干掉if else，减少嵌套?

仅异常情况下使用异常

异常不要用来做流程控制，条件控制。异常设计的**初衷**是解决程序运行中的各种意外情况，且异常的处理**效率**比条件判断方式要低很多。



（虽然这里的异常可以用return Result，来减少嵌套并携带错误信息）

异常带上参数信息

记录日志时，尽可能带上参数信息，便于调试

不要捕捉所有异常

1. **try** {
2. // do someting that might throw IOException
3. // do someting that might throw CertificateException
4. } **catch** (Exception e) {  }

不推荐的写法，应尽量避免,这样些会可能带来一些问题：

* 调用的接口的潜在Bug被忽略
* 不同类型的异常可能需要不同的处理，写在一起内聚性比较差

上面的代码可以更改为：

1. **try** {
2. // do someting that might throw IOException
3. // do someting that might throw CertificateException
4. } **catch** (IOException e) {
5. // Do something handle IOException
6. } **catch** (CertificateException e) {
7. // Do something handle CertificateException
8. }

如果所有异常处理方式一致，且JDK版本高于1.7，可以简化如下：

1. **try** {
2. // do someting that might throw IOException
3. // do someting that might throw CertificateException
4. } **catch** (IOException | CertificateException e) {
5. // Do something handle these exception
6. }

抛异常 VS 返回错误码

开放接口层要将异常处理成错误码和错误信息方式返回。页面要跳转到友好错误页面， 加上用户容易理解的错误提示信息。

应用内部推荐异常抛出；

跨应用间使用Result(Success,Code,message).

开放接口必须使用“错误码”

{"errcode":40018,"errmsg":"invalid name"}

说明：关于RPC方法返回方式使用Result方式的理由： 1）如果使用抛异常返回方式，调用方如果没有捕获到就会产生运行时错误。 2）如果不加栈信息，只是new自定义异常，加入自己的理解的error message，对于调用端解决问题的帮助不会太多。如果加了栈信息，在频繁调用出错的情况下，数据序列化和传输的性能损耗也是问题。

#### 3.2.2.3 异常设计

异常框架要 1方便定位到错误 2使代码整洁 3给用户友好的错误提示。

一般一个项目都会有一套自己的异常框架。如下几种设计

1）设计的很细

异常太多，太麻烦。比如User表格字段校验异常。

2）AppException

只使用一个AppException，在这个异常类中添加一个type属性，通过type来区分不同的异常。type可以用一个枚举来维护，这个type值可以是上面提到的参数格式、业务校验、服务。也可以只用一个AppException，所有的异常都封装成这个，不再按type细分。

3）按层定义

DaoException/ServiceException/AppException。使用的时候会在服务层把DaoException转化为ServiceException，再在应用层把ServiceException封装成AppException。最外层有一个拦截全局异常，拦截AppException就可以了。

4）分类

ArgumentException: 参数格式异常，在所有的参数格式校验出错的时候使用。

BusinessException: 业务校验异常

ServiceException: 服务异常

### 3.2.4 接口设计

TODO1

## 4.3 设计模式

TODO:32中设计模式的好处从5大准则的角度来答。

### 单例模式

[单例模式](http://blog.csdn.net/dongnan591172113/article/details/51027374)

**非同步**



**延迟加载**



**Double Checked locking**



* 1处volatile：new Singleton并非原子操作，instance为类变量，准备阶段



1.new给instance分配内存

2.invokespecial初始化对象

3.aload将instance指向分配的内存

JVM 的即时编译器中存在指令重排序的优化。最终的执行顺序可能是 1-2-3 也可能是 1-3-2

2处为什么判空?

防止阻塞Thread恢复后会多次new

### 观察者模式



观察者 VS 事件驱动

TODO3

## 4.4 软件架构分析

**软件架构优劣的衡量要素：**1.性能2.可用性 3.伸缩性4.扩展性5.安全性

### 3.3.1 性能

#### 性能指标

* 并发数：Concurrency。能同时处理的用户数。高并发
* 响应时间：发出请求到收到相应数据的时间
* 吞吐量：系统单位处理的请求数。QPS（query per second）
* 性能计数器：包括System Load、对象和线程数、内存、CPU、磁盘、网络IO

#### 性能测试方法

* 性能测试：性能预期
* 负荷测试：找到性能安全临界线
* 压力测试：超过安全符合继续施压，获得最大压力承受能力

### 3.3.2 伸缩性

随着不断上升的并发访问压力和数据增长，需要不断向集群加入服务器的手段来缓解。伸缩性衡量是否方便构建集群，是否容易添加服务器。

#### 架构设计

**不同功能物理分离实现伸缩**

分割的粒度可以非常小，甚至可以一个关键网页部署一个独立服务，比如对电商非常重要的商品详情页面，商铺页面，搜索列表页面，每个页面都可以独立部署，专门维护

**单一功能通过集群规模实现伸缩**

应用服务器，搜索，缓存，数据库

#### 应用服务器

#### DB集群伸缩性

关系数据库

1）主从分离

2）分表分库

NoSql

#### 缓存设计

### 3.3.3 扩展性

指现有系统影响最小的情况下，系统扩展或提升的能力。

**纵向拆分**：按功能模块拆分为多个应用系统，新增业务可以设计成独立的web系统。

**横向拆分：**拆分出复用的基础公用业务，独立部署为分布式服务。

分布式系统集成方式：1）分布式消息队列2）分布式服务

消息中间件

通过消息分解系统解耦性，不同子系统处理同一消息

分布式服务

通过接口分解系统解耦性，不同子系统通过相同的接口描述进行服务调用。

### 可用性

实现高可用架构的手段是数据冗余备份和服务失效转移。

服务器：由无状态服务器和有状态的Session服务器组成。

## 4.5 软件工程

### 软件类型

这个世界变化实在是快，各种新名词层出不穷，让人眼花缭乱（云，大数据，AR,AI）。其实很多东西只不过新瓶装老酒而已。

单机类型

最开始的那些不需要互联网的单机软件。

C/S

通过安装的客户端访问服务。早期多为局域网。



B/S

浏览器/服务器架构。软件云端部署，用户通过通过浏览器来使用。



SaaS

软件即服务，SaaS不是一种软件架构，而是一种软件销售方式。本质上还是采用B/S架构。

但通过对每个使用者收取年租费或月租费来销售

优势：云端办公和较低资费。

目标用户：中小企业用户，因为他们要求不高，不像大企业那般“挑食”，通用版开发难度也小。

ERP型SaaS

互联网

### 云计算

**服务模式**

SaaS（软件即服务）、PaaS（平台即服务）和IaaS（基础设施即服务）



**部署方式**

私有云、社区云、公有云和混合云

**基本特征**

按需自助服务：按需使用，按需付费

广泛的网络访问、资源共享、快速的可伸缩性和可度量的服务。

企业级云

### 企业级软件

面向的用户为企业，政府，业务比较复杂。按功能划分为财务会计、ERP（企业资源规划）、CRM（客户关系管理）、SCM（供应链管理）、HRM（人力资源管理）、BI（商务智能）、CMS（内容管理系统）和企业通信工具等。也可以按行业划分为制造业、零售业、医药业等解决方案。

ERP

Enterprise Resource Planning，是一种供应链管理思想。包括采购、销售、制造、财务等功能。

#### 复杂度高

政府学校和企业的采购模式、利益诉求不尽相同；不同生产、管理环节对软件特性、服务模式等的需求大相径庭；不同企业对同种软件也总会有个性化的要求；同一个软件在同一个企业内的同一个功能，也可能因部门间的厉害纠葛，导致需求不可理喻地分裂。

#### 国内现状

### 互联网公司

广义上来讲，企业规划、业务的开展都是基于互联网的前提之下施行的，都可以叫互联网公司。企业直接面向真实用户，即TO C。

特点：

追求高效，扁平化管理，去层级化，人本主义。

TODO3：和传统企业有什么本质区别？

2B VS 2C

2C，用户即购买者，更注重用户体验。2B，购买者为决策层，更注重提供生产和管理

国内现状：

用户有限，一天就是 24 小时，除去睡觉、工作，留给互联网的时间非常少，却有成千上万的产品在争夺这少得可怜的时间。而 B 端用户在工作时间使用产品，粘性大，高频刚需，还有可观的服务费收入。

### 开发模式

#### 敏捷开发

把产品开发引向了快速迭代、小步快跑的路线上

#### 瀑布模型

将功能的实现与设计分开，便于分工协作，将软件生命周期划分为制定计划、需求分析、软件设计、程序编写、软件测试和运行维护等六个基本活动，并且规定了它们自上而下、相互衔接的固定次序，如同瀑布流水，逐级下落

**缺点**

项目各个阶段之间极少有反馈，不适应用户需求的变化

我们的系统也从第一代平台开始到现在第四代平台更换中，对这四代平台做一个简单的介绍： 第一代平台，主要是集中式，以快速上线为目的；

第二代平台主要是分布式改造，缓解各服务压力；

第三代平台主要做服务端SOA治理，后台统一账户中心；

第四代微服务化改造，已达到灰度上线、动态部署集中管理的目的。

我也从负责Java端，到负责整个技术团队，慢慢的在领导的信任下测试交给了我，再后来分公司独立后将运维也交给了我，于是成了整个分公司的技术负责人。这就是我的故事。未来仍然有更多的挑战，感谢我们团队的兄弟姐妹，感谢工作中遇到的所有同事和领导。

### GPL协议

## 4.6 Design Tools

### ER图

TODO4

### UML图

1）实现(implements)



2）继承



3）依赖和关联

依赖和关联都是使用到另一个类，关联是一种更强的依赖关系。

* **依赖**：局部变量(方法里new对象)、方法的参数或者是静态方法的调用;
* **关联**：成员变量，关联分为组合/聚合



关联分为组合/聚合，区别在于存在的生命周期不同。

**聚合(aggregation):**



**组合(Composition)**

Room是house组成部分，有相同的生命周期。



# 第四篇 DB

## 4.1 DB设计

### 4.1.1 数据类型

* 更小的通常更好：占用更少的磁盘、内存和CPU缓存，CUP处理周期
* 简单就好：简单类型CPU处理快。例如整形比字符操作代价低
* 尽量避免NULL：NULL使索引更复杂，需要更多存储空间

### 4.4.2 范式

[Reference](http://blog.csdn.net/dongnan591172113/article/details/7910988)

范式和反范式

范式需要联表，代价昂贵。还可能使一些索引策略无效，比如覆盖索引。

## 4.2 事务

### 4.2.1事务特性

**Atomic**（原子性）

要么全部成功，要么全部失败。

**Consistency**（一致性）

数据不会因为事务的执行而遭受破坏.只有合法的数据可以被写入数据库，否则事务应该将其回滚到最初状态。

实现: 完整性约束：域约束，基本表约束和断言

**Isolation（隔离性）**

事务允许多个用户对同一个数据进行并发访问，而不破坏数据的正确性和完整性。

实现：事务的锁来进行控制。

**Durability（持久性）**

### 4.2.2 隔离级别

[Reference](http://blog.csdn.net/dongnan591172113/article/details/47948935)

* Read uncommitted 出现脏读。
* Read committed Problem:其他提交事务所做的修改或删除，出现不可重复读
* Repeatable read 其他事务update or delete，事务依然可以Repeatable read。read的数据可能过期，所以可能出现丢失更新(P1)。 P2:其他事务插入，此时发生幻像读。
* Serializable 序列化。表锁。

丢失更新

[Reference](http://weifuwu.io/2015/12/30/mysql-second-kind-lose-update/)

|  |  |  |
| --- | --- | --- |
| **时间** | **trx1** | **trx2** |
| T1 | select，100 |  |
| T2 | update to 200.lock X |  |
| T3 |  | select，100元 |
| T4 |  | update to 200元 |
| T5 |  | [此时commit,会阻塞到超时] |
| T6 | commit |  |
| T7 |  | select，100 |
| T8 |  | commit |

解决思路：

* 加悲观锁: 。trx1和trx2在select都for update
* 乐观锁。对数据version化，update前检查数据一致性，此时业务上就可以返回失败或者重新select再计算。数据库自动加排他锁，update本身是原子操作。

实际开发中，如更新数据基于原始数据，比如库存数，要注意lost update问题。

### 4.2.3 事务锁

DB提供的锁机制有行级锁....锁定的资源包括行、表、页、簇和数据库。为了最小化锁的成本，尽量锁定比较小的对象

#### 行级锁

加锁粒度最小，但加锁的开销也最大。行级锁分为[共享锁](http://www.hollischuang.com/archives/923)和[排他锁](http://www.hollischuang.com/archives/923)。

共享锁(Share Lock)

事务T1对数据A加共享锁后，Mysql会对查询结果中的**每行都加共享锁。事务T2只能对A再加共享锁，不能加排他锁。**

用法

1. **SELECT** ... LOCK IN SHARE MODE;

排他锁（eXclusive Lock）

排他锁又称写锁，如果事务T1对数据A加上排他锁后，**T2不能再对A加任任何类型的锁，但是可以读**。

用法

1. **SELECT** ... **FOR** **UPDATE**;

innoDB对update、insert、delete会**自动加排他锁（**Select不加**）**。

意向锁

意向锁是表级锁，其设计目的主要是为了在一个事务中揭示下一行将要被请求锁的类型。InnoDB中的两个表锁：

意向共享锁（IS）：表示事务准备给数据行加入共享锁，也就是说一个数据行加共享锁前必须先取得该表的IS锁

意向排他锁（IX）：类似上面，表示事务准备给数据行加入排他锁，说明事务在一个数据行加排他锁前必须先取得该表的IX锁。

意向锁是InnoDB自动加的，不需要用户干预。

在 Mysql 中，行级锁并不是直接锁记录，而是**锁索引**。索引分为主键索引和非主键索引两种，如果一条sql 语句操作了主键索引，Mysql 就会锁定这条主键索引；如果一条语句操作了非主键索引，MySQL会先锁定该非主键索引，再锁定相关的主键索引。

InnoDB 行锁是通过给索引项加锁实现的，如果没有索引，InnoDB 会通过隐藏的聚簇索引来对记录加锁。也就是说：**如果不通过索引条件检索数据，那么InnoDB将对表中所有数据加锁，实际效果跟表锁一样。**因为没有了索引，找到某一条记录就得扫描全表，要扫描全表，就得锁定表。

#### 表级锁

#### 页级锁

#### 悲观锁&乐观锁

不是数据库中真正存在的锁，是并发控制采用的技术手段。其实不仅仅是关系型数据库系统中有乐观锁和悲观锁的概念，像memcache、hibernate、tair等都有类似的概念。

悲观锁

实现：在mysql层常用的悲观锁实现方式是加一个排他锁。

悲观并发控制实际上是“先取锁再访问”的策略，

不足：增加额外的开销还会降低了并行性 。

乐观锁

乐观并发控制直到提交时锁定，跟新锁定的数据是原子串行操作，所以不会产生任何锁和死锁。注意要加version控制。有点类似线程的cas操作，cas是原始数据和target数据比较。

### 4.2.4 Tools

#### Transaction

开启事务 start transaction;

select \* FROM information\_schema.innodb\_trx\G;



trx\_id

trx\_state: RUNNING/LOCK WAIT

trx\_mysql\_thread\_id

trx\_stared 事务开始时间

trx\_wait\_stated:block事务wait stated

trx\_query:

trx\_rows\_locked 锁定row行数

trx\_tables\_locked 锁定table数量

Autocommit



show variables like 'autocommit';

set autocommit = 0;仅对当前连接生效

autocommit VS start transaction

[Reference](https://blog.csdn.net/qq_26941173/article/details/77872038)

set autocommit=0 所有语句自动开启事务，commit结束事务。  
start transaction 启动一个新事务。语句"挂起"自动提交模式

ReadOnly

查看 show global variables like "%read\_only%";

设置

1.查看InnoDB存储引擎 系统级的隔离级别和会话级的隔离级别

select @@global.tx\_isolation,@@tx\_isolation;

#### Lock

Syntax: show status like '%Innodb\_row\_lock%';



Innodb\_row\_lock\_current\_waits 当前阻塞线程数

Innodb\_row\_lock\_waits 总共wait次数

2.select \* from information\_schema.innodb\_locks;

查看lock具体数据



lock\_trx\_id找到事务

lock\_table+lock\_data找到具体锁定的数据。

Lock\_mode: X/S

lock\_type： RECORD行锁。

Timeout

show variables like 'innodb\_lock\_wait\_timeout'; 等待获取锁

## 4.3 MySQL

### 4.3.1发展

MySQL 1.0

1996年，MySQL 1.0发布 ，只有CRUD基本操作

1999-2000年，有一家公司在瑞典成立了，叫MySQL AB。 开发出了 Berkeley DB引擎, MySQL从此开始支持事务处理了

2000 年 MySQL 公布了自己的源代码，并采用GPL许可协议，正式进入开源世界。

MySQL 3.23（2001）

实现了 SQL 查询的系统，MyISAM 代替了老旧而且有诸多限制的 ISAM 引擎。InnoDB 引擎也已经可以使用，Mysql真正“诞生”的时刻

MySQL 4.0（2003）

不支持存储过程、触发程序、服务器端指针或视图，这时候的MySQL还不是一个企业级数据库。

MySQL 5.0（2006）

出现了一些“企业级”特性：视图、触发器、存储过程和存储函数。同时 Oracle 收购的 InnoDB Oy发布了 InnoDB plugin。

版本5.1（2008）

Sun 收购 MySQL AB 以后发布的首个版本，研发时间长达五年。

2008年 MySQL被Sun公司收购。

2009年Oracle收购Sun 公司，MySQL 转入Oracle 门下。

Oracle 对MySQL版本重新进行了划分，分成了社区版和企业版，企业版

MySQL Community Server

MySQL Cluster 集群版，开源免费。可将几个MySQL Server封装成一个Server。

MySQL 5.5（2010）

默认InnoDB plugin引擎。具有提交、回滚和crash恢复功能、ACID兼容。

行级锁

### 4.3.2 逻辑架构

mysql分为 server 和 storage engine 两个组件, server 负责 sql的parse, 执行; storage engine 去真正的 做 数据/index的 读取/写入. 以前是这样: server 命令 storage engine 按 index 把相应的 数据 从 数据表读出, 传给server, server来按 where条件 做选择; 现在 ICP则是在 可能的情况下, 让storage engine 根据index 做判断, 如果不符合 条件 则无须 读 数据表. 这样节省了disk IO.



每个虚线框为一层，总共三层。

第一层，服务层(为客户端服务):为请求做连接处理，授权认证，安全等。

第二层，核心层:查询解析，分析，优化，缓存，提供内建函数;存储过程，触发器，视图。

第三层，存储引擎层，data的save和query，事务处理。

### 4.3.3 索引

#### 4.3.3.1 Base

在数据之外，数据库系统还维护着满足特定查找算法的数据结构，这些数据结构以某种方式引用（指向）数据，这样就可以在这些数据结构上实现高级查找算法。这种数据结构，就是索引。索引优化了查询。本质是空间换时间。

查找算法

顺序查找：复杂度O(n)，缺点：的算法在数据量很大时显然是糟糕的

二分查找（binary search）：要求被检索数据有序。

二叉树查找（binary tree search）等。，只能应用于二叉查找树上

##### 优点

1）索引大大减少了服务器需要扫描的数据量

2）索引可以帮助服务器避免排序和临时表

3）索引可以将随机I/O变成顺序I/O

哪些字段要索引

Where column, order by

##### 判断标准

##### 索引和页的关系

索引通常十分庞大，需要以文件形式存储在磁盘，索引查找过程中就要产生磁盘I/O消耗，相对于内存存取，I/O存取的消耗要高几个数量级，磁盘IO的存取次数就是评价一个数据库索引优劣的关键性指标。

TODO 页大小为什么是4K

InnoDB的页面分为

叶子节点：B树层次为0的页面，存储记录的所有内容

非叶子节点：B树层次大于0的页面，只存储索引键和页面指针。

#### 4.3.3.2 索引类型

普通索引

Unique Index: 1. 提高访问速度 2.保证数据唯一性

Primary Index

组合索引（联合索引、复合索引）

Hash索引

全文索引(Full Text)

有时还可以看到**多列索引**、**短索引（前缀索引）、覆盖索引（索引覆盖）、聚簇索引（聚集索引）和非聚簇索引**，这些索引其实都不能将单独它归为一个索引类型。

B-Tree

[Reference](https://blog.csdn.net/Roy_70/article/details/76581028)

即Balanced Tree, 有别于二叉查找树（Binary Search Tree）。二叉搜索树的查询时间复杂度是O(log(n))，从算法逻辑上来讲，无论查找速度还是比较次数都是最小的，但在数据库的索引通常十分庞大，需要以文件形式存储，而磁盘IO的存取次数就是评价一个数据库索引优劣的关键性指标。

B-Tree优化了磁盘IO

#### 4.3.3.3 索引策略

##### Multiple-Column Indexes

这里指在多个列上分别建索引。建索引的一个误区是在where后面条件每个列建一个索引，这样最好的情况也只能用到一个索引。

生效原则：

Or：index merge。耗费性能做数据合并排序。特别是返回数据量较大时

And：

索引顺序:

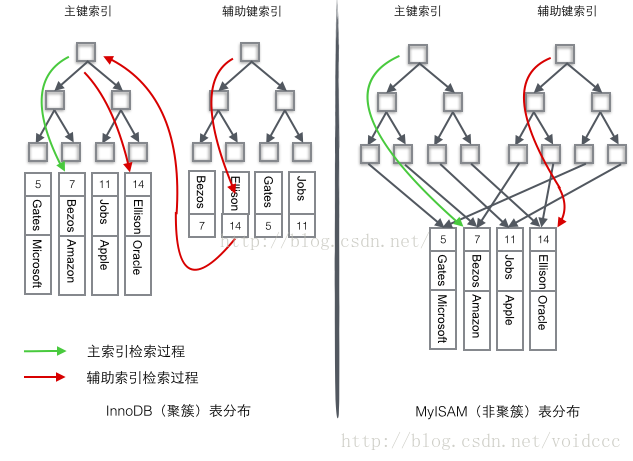
##### Clustered Index

聚簇索引

聚簇索引是数据存储方式，聚簇是指数据行和键值存在一起。InnoDB是聚簇索引，行数据就储存在叶子节点上。辅助索引的叶节点是主键，通过辅助键索引需要两步定位到数据。

非聚簇索引

即二级索引, Secondary index 。MyISAM是非聚簇索引，叶子节点上的data是数据的地址。主索引和辅助索引没啥区别。索引和表数据存储在独立的地方。



优点

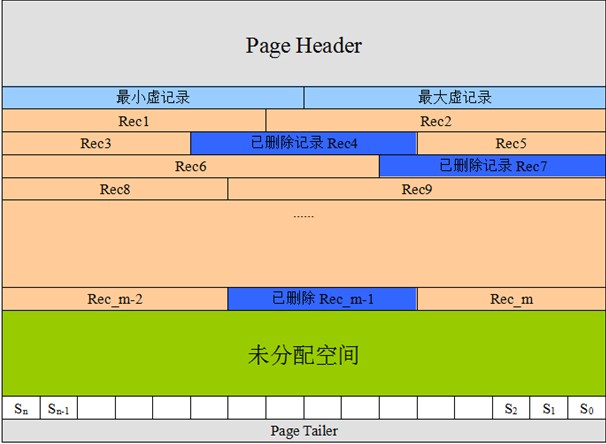
* 聚集相关数据，比如userId做主键，读相关邮件只需读少量数据页，否则每一每一封邮件可能导致一次磁盘I/O。
* 使用覆盖索引扫描，可直接使用叶节点的主键值。

缺点

* 基于聚集索引插入，可能导致页裂变，
* 全表扫描慢，尤其是行比较稀疏，或也分裂导致存储不连续。
* 二级索引包含主键值，可能更大，索引时需要两次查找，而不是一次。

页面格式

[Reference](https://www.cnblogs.com/vinchen/archive/2012/09/10/2679478.html)



页内的有序性

按序存储：插入导致记录移动

链表：查找需要逐个比较

slot和每个record对应，只需保证slot按序存储，record链表存储即可。

**slot区**：slot是一些页面有效记录的指针，每个slot占两个字节，存储了记录相对页面首地址的偏移。如果页面有n条有效记录，那么slot的数量就在n/8+2~n/4+2之间。下一节详细介绍slot区，它是记录页面有序和二分查找的关键。

##### 覆盖索引

覆盖索引是建立在联合索引之上的，指索引的叶子节点已包含所有要查询的列，因此不需要访问表数据。

判断标准：eplain的extra显示using index。

##### 组合索引

### 4.3.4 存储引擎

InnoDB VS MyISAM

一是 InnoDB 支持**事务**。二是 InnoDB 采用了**行级锁**。

影响引擎性能的因素：数据大小，I/O请求量，主键还是二级索引

InnoDB

优先选择InnoDB，除非用到InnoDB不具备的特性，没有其他办法替代。

1.事务2.热备份3.崩溃恢复4.聚簇索引

MyISAM

一般读写性能高，但随着应用压力上升，各种锁争用，崩溃数据丢失问题

MyISAM引擎缓存索引，而不缓存数据

不支持事务、外键和行级锁,适合小项目

优点:1.读多写少的项目，效率高2.没有磁盘碎片，节省空间

|  |  |  |
| --- | --- | --- |
| **时间** | **Tran1** | **Tran 2** |
| T1 | start transaction | start transaction |
| T2 | select，100 | select，100元 |
| T3 |  | update to 200元 |
| T4 | update to 200元 |  |
| T5 |  | commit |
| T6 |  |  |
| T7 | commit |  |
| T8 |  |  |

1.T4处update或select..for update将阻塞到报错



### 4.3.5 主从分离架构



## 4.4 ORACLE

### PLSQL

**优势**

1 SQL语言只是访问、操作数据库的语言。PLSQL具备高级语言的特性，比如模块化，流程控制，异常.

2.集成在数据库中，调用更快。减少了网络的交互，从而减少网络传输时间。

## 4.5 NOSQL

现代互联网应用，关系数据库的很多主要特性却往往无用武之地

1.数据库事务一致性需求

2.数据库的写实时性和读实时性需求

3.对复杂的SQL查询，特别是多表关联查询的需求

去掉关系型数据库的两大重要基础：以关系代数为基础的结构化查询语句（SQL）和事务一致性保证（ACID）

1. 高并发读写2)海量数据的高效率存储和访问3)高可用性和可伸缩性

### 4.6.1 Redis

Redis是一款开源的高性能key-value数据库，支持更丰富的数据结构，例如hashes, lists, sets等。除此之外，Redis还提供一些类数据库的特性，比如事务，HA，主从库。与一般数据库不同，redis是使用内存作为主存，周期性的将数据写到硬盘上。因为读写数据都使用内存，所以它的速度是非常快的，很适合我们来存放一些临时性的数据。

Redis vs Memcached

1.使用简单的key-value存储的话，Memcached的内存利用率更高

2. 由于Redis只使用单核，而Memcached可以使用多核， Memcached相对性能更高

3.拥有更多的数据结构和并支持更丰富的数据操作。

4.集群管理的不同

缓存雪崩

缓存击穿

**管理Session**

## 4.6 DB优化

### 4.6.1 库表结构优化

选择合适的数据类型

* 更小的通常更好：占用更少的磁盘、内存和CPU缓存，CUP处理周期
* 简单就好：简单类型CPU处理快。例如整形比字符操作代价低
* 尽量避免NULL：NULL使索引更复杂，需要更多存储空间

### 4.6.2 索引优化

索引增加写数据的额外消耗，频繁写的表不宜建索引。

三星系统

第一颗星：与查询相关的索引行是相邻的，即where后面的等值谓词，可以匹配索引列顺序

第二颗星：索引行的顺序与查询语句需求一致，也就是order by 中的排序和索引顺序是否一致

第三颗星：索引行包含查询语句中所有的列

### 4.6.3 查询优化

查询的生命周期

客户端🡪服务器🡪解析🡪生成执行计划🡪执行🡪客户端

执行**：**调用storage engine和调用后的数据处理，包括排序、分组等。

每个过程都花费时间

#### 4.6.3.1 优化数据访问

查询性能低下最基本原因是访问数据太多，有些情况不可避免，但大部分性能低下都可以通过减少访问的数据量来优化。常规步骤：

1.确认应用服务检索必要的数据列或和数据行。

2.确认MySQL服务层分析必要的行。

取不需要的行

记得分页

取全部列

* 额外的I/O，内存和CUP消耗
* 覆盖索引失效

select \* 虽然简化开发和提高代码的复用性，但要考虑性能影响，必要时缓存。

MySQL是否扫描额外的记录

explain分析执行计划，避免全表扫描。 对于group by 或count/一些 会引起全表扫描，必要时使用单独的汇总表或者放Memcached。

type 没建索引导致全表扫描

rows理想情况扫描行数=返回函数，越小越好。

#### 4.6.3.2 重构查询方式

复杂查询 Or 简单查询？

大的语句影响

锁住很多数据，占满整个事务日志，耗尽系统资源，阻塞很多小而重要查询。

关联查询

分解关联查询

传统实现总是强调db完成尽可能的工作，原因在于以前网络通信，每个SQL的查询解析和优化代价高。对Mysql并不适用。有时候把一个大查询分解为多个小查询很有必要。

用单表还是联表，看具体业务场景。没到阀值前，都可以，有的业务也许永远到达这个阀值。在高并发高读环境，遇到性能瓶颈时,分解关联查询好处：

* 减少锁竞争
* 将联表分解成单表冗余字段的形式，建覆盖索引。
* 汇总表
* 缓存。缓存更高效。且dao层的复用率更好，更好维护，后期好扩展。数据库好拆分。

1.关联查询优化

需要联表查询

* 连表时，保证被关联的字段需要有索引，一般在连表顺序第二个create index。
* 数据类型必须绝对一致
* 考虑对相对应的查询条件做索引。加快查询速度。
* 一般联表消耗更多系统资源，还可能使索引失效，超过三个表禁止join。

4）缓存

* 缓存重复查询和获取比较慢的数据。

5）搜索

页面搜索严禁左模糊或者全模糊，如果需要走搜索引擎来解决。

说明：索引文件具有 B-Tree 的最左前缀匹配特性，如果左边的值未确定，那么无法使用此索

引。

1.3.1 strategy2

根据不同业务场景有不同的侧重，优化的策略不同，比如、读写比例、并发量、数据量和数据增长数据。

* 读写比例🡺根据具体业务读写比例，降低主库压力
* 高并发🡺集群
* 数据是否需要强一致🡺NOSQL
* 数据量，数据增长快。

1.针对数量大的表进行历史表分离（如交易流水表）。

2.做分表分库，单表行数超过 500 万行或者单表容量超过 2GB，才推荐进行分库分表。说明：如果预计三年后的数据量根本达不到这个级别，不要在创建表时就分库分表

#### 4.6.3.3 Case

##### 分页

Mysql的分页查询十分简单，但是当数据量大的时候一般的分页就吃不消了。

原理

先读取前面n条记录，然后抛弃前n条，读后面m条想要的，所以n越大，偏移量越大，性能就越差。

## 附录1 Explain执行计划

**id**：id越大则优先级越高，越先会被执行。如果id相同，则执行顺序从上至下

**possible\_keys** :基于SQL的访问列和where中使用的比较符(=><)

**Key：**The index actually chosen

**key\_len**：使用索引最大byte

**ref：**The ref column shows which columns or constants are compared to the index named in the key column to select rows from the table.

**rows**：估算读取的行数

### select\_type

simple: 不使用UNION或子查询等

primary: 最外层的select

subquery，

derived:里层的子查询



union，

unionresult

### Type

all < index < range < ref < eq\_ref < const < system < NULL

* **ALL**：即full table scan，（到存储引擎）遍历全表
* **index**：为full index scan，遍历全部的索引树。可避免排序。



* **range**： 1.使用索引 2.只检索给定范围的行



* **ref**： 索引查找 1.使用索引 2.匹配条件用于查找索引列上的值



* **eq\_ref:** 类似ref，最多返回一条记录。区别:索引列是primary key或者unique index
* **const，system**: mysql能对查询部分优化转成变量，比如id在where子句，mysql将查询转成常量，然后就可高效将表从联表执行中移除。



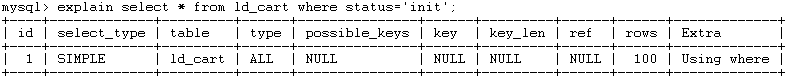
* **NULL：**表示在执行语句中，不用查表或索引

### Extra

Additional information about how MySQL resolves the query

* using index：使用覆盖索引。区Type中index。
* Using where: A where clause is used to restrict which rows to match against the next table or send to the client。在server里过滤。
* Using index condition： v5.6加入。可以在store engine过滤。
* using temporay表示用临时表来存储结果集，常见于排序和分组查询
* using filesort，mysql中无法用索引完成的排序成为文件排序。

**Using where**



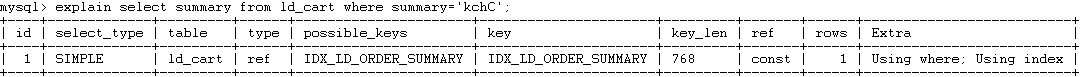
full table scan，并根据where条件在server层面过滤。

**Using index**



full index scan。没有过滤。

Using where;Using index;

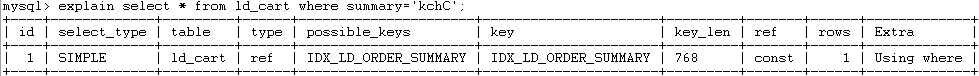


这里的Using where**指在server层面过滤index。一次IO。**

Using index condition

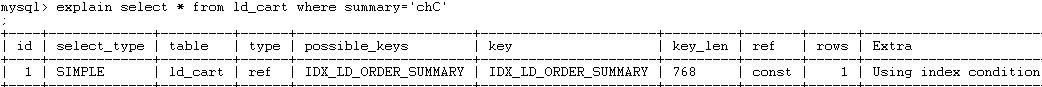
Tables are read by accessing index tuples and testing them first to determine whether to read full table rows. In this way, index information is used to defer (“push down”) reading full table rows unless it is necessary

先看5.6之前版本，

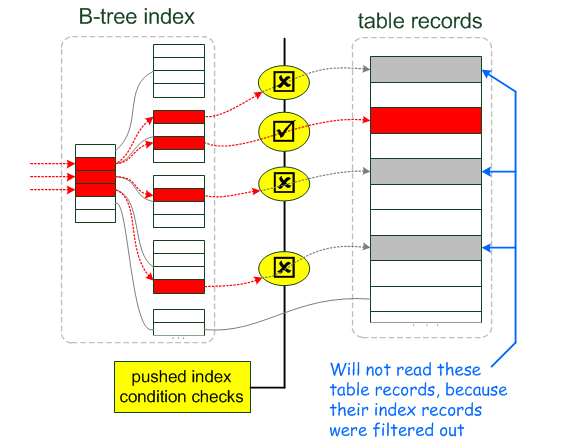


server先读到index在server层面过滤，再到store engine读到full rows。没有Using index因为index没有覆盖。

v5.6



storage engine 根据index 做判断, 再read full rows. 这样节省一次disk IO.



Using union

1. mysql> explain **select** \* **from** ld\_cart **where** summary='kchC' or member\_id=13\G;
2. \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* 1. row \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*
3. id: 1
4. select\_type: SIMPLE
5. **table**: ld\_cart
6. type: index\_merge
7. possible\_keys: IDX\_LD\_ORDER\_SUMMARY,idx\_ld\_cart\_member\_id
8. **key**: IDX\_LD\_ORDER\_SUMMARY,idx\_ld\_cart\_member\_id
9. key\_len: 768,9
10. ref: NULL
11. **rows**: 2
12. Extra: Using **union**(idx\_summary,idx\_member\_id); Using **where**

同时扫描summary和member\_id两个列索引，并将结果合并。

Where

Mysql应用where条件记录，从好到坏依次为

1）存储索引层，在索引中使用where条件过滤(Using index condition)。过滤覆盖索引。

2) 在mysql服务器层, 扫描索引过滤并返回记录。无需在回表查询记录(Using where，Using index)。**这里过滤的是索引**。

3) 在mysql服务器层，先从数据表读出记录然后过滤(Using Where)

1 2 有待验证

## 附录 2 CMD

**连接：**mysql –uroot [-h slc11fsp] –p

**版本：**mysql –V；或者select version();

### Mgr

**Status**: show [**global**/session] status like '%xxx%';

**Variables:** show variables like '%max\_connections%';

**Innodb:** show [**global**] status like 'Innodb\_%'

### information\_schema

##### indexs

**Show**: show **index** **from** fund\_history\G;

**Create**: create **index** idx\_fund\_name on fund(name);

##### tables

show databases/tables;

**Show DDl:** show create table ld\_cart;

### performance\_schema

##### Load

查看mysql占用系统资源

top –p pis

pidstat |grep mysql

##### Connections

show full processlist;



每一条代表一个Thread，一个Tcp连接

Threads\_connected 当前连接数。只有数量值，show processlist有更详细的信息;

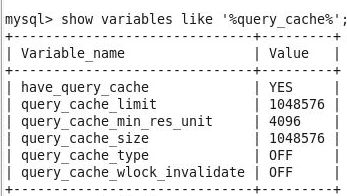
connections 试图连接mySQL服务器的次数

Max\_used\_connections

##### Cache

Query

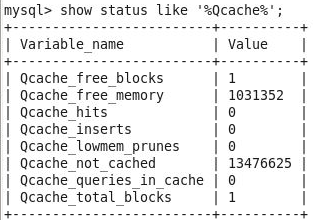
show variables like '%query\_cache%'



query\_cache\_limit 缓冲区大小，缺省为1M；

query\_cache\_min\_res\_unit缓存块大小，默认4KB，设置值大对大数据查询有好处，但如果你的查询都是小数据 查询，就容易造成内存碎片和浪费；

Cache Status



Qcache\_free\_blocks：缓存中相邻内存块的个数。数目大说明可能有碎片。FLUSH QUERY CACHE会对缓存中的碎片进行整理，从而得到一个空闲块。

Qcache\_free\_memory：缓存中的空闲内存。

Qcache\_hits：每次查询在缓存中命中时就增大

Qcache\_inserts：每次插入一个查询时就增大。命中次数除以插入次数就是不中比率。

Qcache\_lowmem\_prunes：缓存出现内存不足并且必须要进行清理以便为更多查询提供空间的次数。这个数字最好长时间来看;如果这个 数字在不断增长，就表示可能碎片非常严重，或者内存很少。(上面的 free\_blocks和free\_memory可以告诉您属于哪种情况)

Qcache\_not\_cached：不适合进行缓存的查询的数量，通常是由于这些查询不是 SELECT 语句或者用了now()之类的函数。

Qcache\_queries\_in\_cache：当前缓存的查询(和响应)的数量。

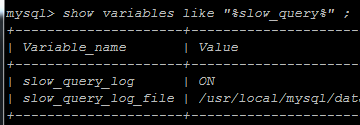
Qcache\_total\_blocks：缓存中块的数量。

##### 日志

show variables like 'log%';

show query

show variables like "%slow\_query%" ;



开启日志

set global slow\_query\_log=ON;

分析

mysqldumpslow /usr/local/mysql/data/mysql/slc11fsp-slow.log

#### 数据大小

select data\_length,index\_length from information\_schema.tables where table\_schema='bigdata' and table\_name = 'ld\_cart';

默认单位是byte，优化显示

select concat(round(sum(data\_length/1024/1024),2),'MB') as data\_length\_MB,concat(round(sum(index\_length/1024/1024),2),'MB') as index\_length\_MB from information\_schema.tables where table\_schema='bigdata' and table\_name = 'fund\_history';

行数select count(1) from ld\_cart;

idex

show index from bigdata.fund;

#### 统计

#### Questions

Slow\_queries： 超过long\_query\_time数量。（show variables like "%long\_query\_time%"）;

Questions ： 发往服务器的查询的数量。统计开始时间点为service start。

DDL统计

Innodb\_rows\_read/inserted/deleted/updated

Innodb\_data\_read/writes

buffer

Threads\_running 不在睡眠的线程数量。   
Uptime 服务器工作了多少秒。

# 第五篇 Linux

## 3.1 系统资源分析

### top

–p pid 监控某个进程

-H show thread, top –Hp pid

4 users：当前登录user数

Tasks:进程数。-

Cpus(s)

us/sy — 用户/内核 占用CPU的百分比。

ni — 改变过优先级的进程占用CPU的百分比

id — 空闲CPU百分比

wa — IO等待占用CPU的百分比

hi — 硬中断（Hardware IRQ）占用CPU的百分比

si — 软中断（Software Interrupts）占用CPU的百分比

Title

PR 优先级

VIRT 进程使用的虚拟内存总量，单位kb。VIRT=SWAP+RES

RES 进程使用的、未被换出的物理内存大小，单位kb。RES=CODE+DATA

SHR 共享内存大小，单位kb

%CPU  CPU时间占用百分比

%MEM  进程使用的物理内存百分比

TIME+  进程使用的CPU时间总计，单位1/100秒  
CMD 进程名称（命令名/命令行）

**快捷键**

P:CPU排序。默认按cpu占用量排序。

M:内存排序

H: show thread

1:show cups 列出单个cpu的负荷情况

### ps

查看线程。Process Status

a 显示现行终端机下的所有程序，包括其他用户的程序

f 用ASCII字符显示树状结构，表达程序间的相互关系

e 列出程序时，显示程序所使用的环境变量

-A 显示所有进程

-u 以用户为主的格式来显示程序状况

-x 显示所有程序，不以终端机来区分

常用

ef



USER 用户名

UID User ID

PID Process ID（进程），当然jps命令也可以查到进行ID

PPID parent pid

TIME 进程使用的总cpu时间

COMMAND 正在执行的命令行命令

aux



SID 会话ID（Session id）

%CPU **进程的cpu占用率，单个CPU，双核的最大值为200%**

%MEM 进程的内存占用率

VSZ 进程所使用的虚存的大小（Virtual Size）

RSS 进程使用的驻留集大小或者是实际内存的大小，Kbytes字节。

TTY 与进程关联的终端（tty）

STAT 进程的状态：进程状态使用字符表示的（STAT的状态码）

START 进程启动时间和日期

### netstat

系统所有的网络套接字连接情况，包括 tcp, udp 以及 unix 套接字。

Active Internet connections，称为有源TCP连接

Active UNIX domain sockets，称为有源Unix域套接口(和网络套接字一样，但是只能用于本机通信，性能可以提高一倍)

Param

-a (all)显示所有选项，默认不显示LISTEN相关

-t (tcp)仅显示tcp相关选项

-n 拒绝显示别名(域名)，能显示数字的全部转化成数字(ip)。

-u (udp)仅显示udp相关选项  
-p 显示建立相关链接的程序名

-l 仅列出有在 Listen (监听) 的服務状态

-x**只列出所有监听 UNIX 端口**

提示：LISTEN和LISTENING的状态只有用-a或者-l才能看到

-p 显示建立相关链接的程序名  
-r 显示路由信息，路由表  
-e 显示扩展信息，例如uid等  
-s 按各个协议进行统计  
-c 每隔一个固定时间，执行该netstat命令

**常用**

1)netstat -at|grep slc01boa

2)统计80端口连接数

netstat -nat|grep -i "8080"|wc –l

3)Linux下查看tomcat连接数

netstat -na | grep ESTABLISHED | grep 8080 | wc -l

1. 统计httpd协议连接数  
   ps -ef|grep httpd|wc –l
2. 统计已连接上的，状态为“established

netstat -na|grep ESTABLISHED|wc -l

3)找出程序运行的端口

netstat -ap | grep java

**4)**[根据端口查找进程](http://www.cnblogs.com/paul8339/p/6638370.html)

netstat -apn | grep 8989

netstat -n | awk '/^tcp/ {++S[$NF]} END {for(a in S) print a, S[a]}'

windows

netstat -ano -p tcp -n 1|find "8080"

-n 1间隔1s自动刷新



PID是进程

统计

netstat -an|find "ESTABLISHED" /c

title

local address本机的地址

foreign address 外部电脑电脑有联系的ip地址

### free

[Reference](http://www.cnblogs.com/coldplayerest/archive/2010/02/20/1669949.html)



-/+ buffers/cache: 应用程序认为系统被用掉多少内存

设置缓存大小：

### df

Disk File

-i 查看inodes

df –hl

df -h /home/nadong

### du

Disk usage of each File

du –sh

分区

### jps

显示java进程的pid



## 3.2 java

#### jar

-c 创建一个jar包  
-t 显示jar中的内容列表  
-x 解压jar包  
-u 添加文件到jar包中  
-f 指定jar包的文件名  
-v生成详细的报造，并输出至标准设备

1）解压

Jar -xvf test.ear

2）创建

Jar -cvf test.jar ./\*

3）更新

jar -uvf RESTWebService.ear APP-INF/lib/AdfHcmUsersRestModel.jar

4）查看

jar tvf test.jar

5）搜索：

jar tvf ZhcLib.jar | grep 'pattern'

搜索整个目录下的所有jar文件：

find . -name "\*.jar"|awk '{print "jar -tvf "$1}' | sh | grep ".\*.xml"

## 3.2 文件

### 文件属性

文件颜色

绿色文件---------- 可执行文件，可执行的程序

红色文件-----------压缩文件或者包文件

蓝色文件----------目录

白色文件----------一般性文件，如文本文件，配置文件，[源码](https://www.2cto.com/ym/)文件等

浅蓝色文件----------链接文件，主要是使用ln命令建立的文件



链接文件

分配不同权限

缩写路径

软链接 VS 硬链接

软链接和Windows的快捷方式类似

Chmod

chmod u+w test

chmod ugo+r test;

### 文件搜索

find

find [dir] -name filename

1.在当前及子文件夹查找

find –name common.jar

2.在etc及子文件夹查找

find /etc -name httpd.conf

3.模糊查找

find –name ‘httpd\*'

Grep

-n ：输出行号

-r ：在某一目录下递归[所有子目录]查找某一字串:

grep -r ‘str\*' . (注意此处有点)

grep -r a /scratch

Whereis

whereis useradd

### 文件查看及编辑

more

Enter:一行

Space:翻页

less

less 的用法比起 more 更加的有弹性。在 more 的时候，我们并没有办法向前面翻， 只能往后面看

Space:翻页

G:倒序查看

vim

**全部删除：**按esc后，然后dG

lsof

### 6.3 用户权限

#### 用户组

**1)groups**:  查看当前登录用户的组内成员

**2)/etc/group**

group\_name:passwd:GID:user\_list

**3) groupadd**

group add group1

**4) groupdel**

删除群组。先删除group的用户，方能删除群组

**usermod**

usermod -a -G root nadong

将一个用户添加到用户组中c

**chown**

将指定文件的拥有者改为指定的用户或组

chown -R mysql: mysql ./

#### 用户

**useradd**

adduser moon

新建用户::useradd -d /home/moon -m moon

修改用户的密码: passwd tes

**userdel**

userdel  moon

1. 将新用户添加到管理组gpasswd -a test admin

3、给 test 用户创建自己的目录：

cd /home

mkdir test

chown test /home/test

4、重新启动，

reboot 然后用 test 登录，

**su**

5. :Su switch use /sudo输入当前用户的密码。

**sudo**

1.需要root赋予sudo权限

su root🡪 hadoop ALL=(ALL) ALL

## VNC

* Kill VNC: vncserver -kill :1
* Create VNC: vncserver -geometry 1910x1050
* VNC set copy: vncconfig -nowin &
* List: vncserver -list

## Nginx

1)启动操作  
nginx -c /usr/local/nginx/conf/nginx.conf  
2)停止操作  
从容停止Nginx  
快速停止Nginx：  
kill -TERM 主进程号  
  
强制停止Nginx：  
kill -9 主进程号  
3)  
nginx -s reload

## mysql

#启动服务 service mysqld start

查看状态：sudo service mysqld start

#连接

./mysql/bin/mysql –uroot

mysql –uroot [-h slc11fsp] -p

root是db的账号

DDL

show databases/tables  
desc table\_name;

GRANT ALL PRIVILEGES ON \*.\* TO 'root'@'%' IDENTIFIED BY '123456' WITH GRANT OPTION;

### redis

1)启动： redis-server.exe redis.windows.conf

开启memcache，并连接测试：  
以守护进程模式启动memcache  
memcached-1.5.2/bin/memcached -d -l slc11fsp.us.oracle.com -p 11211 -m 2048 -u root

## 6.5 SHELL

变量

$# 是传给脚本的参数个数

$0 是脚本本身的名字

$1 是传递给该shell脚本的第一个参数

$2 是传递给该shell脚本的第二个参数

$@ 是传给脚本的所有参数的列表

$\* 是以一个单字符串显示所有向脚本传递的参数，与位置变量不同，参数可超过9个

$$ 是脚本运行的当前进程ID号

$? 是显示最后命令的退出状态，0 = success, 其他 = error

${var}等价于$var（简写）

if/else

[Reference](http://www.cnblogs.com/kangyoung/p/3556173.html)

比较

if ([ condition ] || [condition])

then

cmd

else

cmd

fi

if method()

method return 0=true ,1=false

文件表达式

-e filename 如果 filename存在，则为真  
-d filename 如果 filename为目录，则为真   
-f filename 如果 filename为常规文件，则为真  
-L filename 如果 filename为符号链接，则为真  
-r filename 如果 filename可读，则为真   
-w filename 如果 filename可写，则为真   
-x filename 如果 filename可执行，则为真  
-s filename 如果文件长度不为0，则为真  
-h filename 如果文件是软链接，则为真

整数变量

eq:=

ne:!=

gt:>

ge:>=

lt: <

le <=

函数

LaunchJDeveloper(){

DisplayBanner

}

### expr

手工命令行计数器



模式匹配

# 第六篇 Improve

## 7.1 Summary

### 7.1.1 前端优化

* 合并css、js，减少Http请求
* 启用压缩，减少数据传输
* 使用浏览器缓存。生成新的js时，只需更新Html中引用。
* CSN加速：缓存静态资源，如图片、文件、CSS、JavaScript脚本、静态网页等。
* 反向代理

### 7.1.2 应用服务器

#### 系统架构

* 缓存：加快数据读取，降低DB访问压力。网站遇到性能瓶颈，第一个想到的解决方案
* 异步操作：加快请求响应及实现削峰。
* 使用集群：使用负载均衡，有更好的响应延迟，提高吞吐能力。

#### 代码优化

* Thread：多线程,最大发挥CPU和处理IO阻塞。
* IO：使用NIO(非阻塞，基于事件驱动)
* GC：减少垃圾回收。
* 锁优化：尽量不使用锁,减小锁的力度。
* 资源复用，比如单例和线程池。
* 合理的数据结构

#### Protocol

HTTP 1.1

**最大并发数限制**

并发指建立的TCP连接数。可以一个连接发送多个请求。Server端按发送顺序处理。

### 7.1.3 DB优化

#### 7.1.3.1 库表结构优化

选择合适的数据类型

* 更小的通常更好：占用更少的磁盘、内存和CPU缓存，CUP处理周期
* 简单就好：简单类型CPU处理快。例如整形比字符操作代价低
* 尽量避免NULL：NULL使索引更复杂，需要更多存储空间

#### 7.1.3.2 索引优化

索引增加写数据的额外消耗，频繁写的表不宜建索引。

三星系统

第一颗星：与查询相关的索引行是相邻的，即where后面的等值谓词，可以匹配索引列顺序

第二颗星：索引行的顺序与查询语句需求一致，也就是order by 中的排序和索引顺序是否一致

第三颗星：索引行包含查询语句中所有的列

#### 7.1.3.3 查询优化

查询的生命周期

客户端🡪服务器🡪解析🡪生成执行计划🡪执行🡪客户端

执行**：**调用storage engine和调用后的数据处理，包括排序、分组等。

每个过程都花费时间

查询性能低下最基本原因是访问数据太多，确认应用服务检索必要的数据列或和数据行。

1.关联查询优化

需要联表查询

* 连表时，保证被关联的字段需要有索引，一般在连表顺序第二个create index。
* 数据类型必须绝对一致
* 考虑对相对应的查询条件做索引。加快查询速度。
* 一般联表消耗更多系统资源，还可能使索引失效，超过三个表禁止join。

分解关联查询

用单表还是联表，看具体业务场景。没到阀值前，都可以，有的业务也许永远到达这个阀值。(好处),在高并发高读环境，遇到性能瓶颈时，将联表分解成单表冗余字段的形式，这样做有几个好处

* 好做缓存
* 建覆盖索引，把所有需要的列都放到索引中，这样存储引擎无须返回表获取对应行就可以返回结果了。
* dao层的复用率更好，更好维护，后期好扩展。

2.分析执行计划

Explain 分析服务器执行计划**，**确认MySQL服务器是否在分析大量超过需要的数据行，避免全表扫描。 对于group by 或count/一些 会引起全表扫描，必要时使用单独的汇总表或者放Memcached。

4）缓存

* 缓存重复查询和获取比较慢的数据。

5）搜索

页面搜索严禁左模糊或者全模糊，如果需要走搜索引擎来解决。

说明：索引文件具有 B-Tree 的最左前缀匹配特性，如果左边的值未确定，那么无法使用此索

引。

扫描行数和访问类型

ALL-index-range-ref

1.3.1 strategy2

根据不同业务场景有不同的侧重，优化的策略不同，比如、读写比例、并发量、数据量和数据增长数据。

* 读写比例🡺根据具体业务读写比例，降低主库压力
* 高并发🡺集群
* 数据是否需要强一致🡺NOSQL
* 数据量，数据增长快。

1.针对数量大的表进行历史表分离（如交易流水表）。

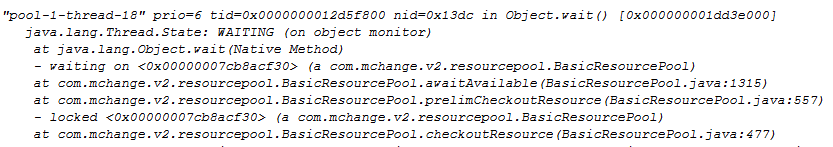
2.做分表分库，单表行数超过 500 万行或者单表容量超过 2GB，才推荐进行分库分表。说明：如果预计三年后的数据量根本达不到这个级别，不要在创建表时就分库分表

#### 7.1.3.2 Case

insert时间长 (远程)

每个请求占用一个request/缓解mysql压力，将数据存入redies。然后批量插入mysql

2．未关闭事务，导致系统无响应。



WAITING等待连接池连接。

##### 分页

Mysql的分页查询十分简单，但是当数据量大的时候一般的分页就吃不消了。

原理

先读取前面n条记录，然后抛弃前n条，读后面m条想要的，所以n越大，偏移量越大，性能就越差。

优化

##### 联表

## 7. 2 性能优化

### 7.2.1 Sumary

#### 7.2.1.1 性能指标

PV QPS TPS

##### Concurrency

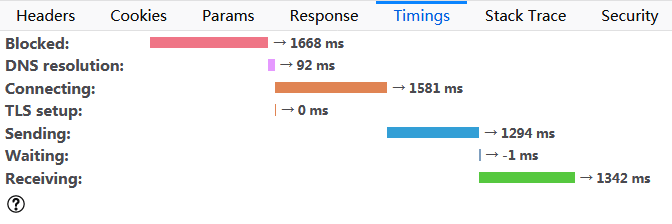
并发数。服务器能同时处理的并发TCP连接数。

##### Response

响应时间。发出请求到收到相应数据的时间。

响应时间没有一致的规律或公式。受诸多因素影响。比如锁，多线程，缓存，硬件响应等，所以响应时间可能是问题的结果也可能是问题的原因。需要自行判断是否合理。能不能满足需求。

FireFox



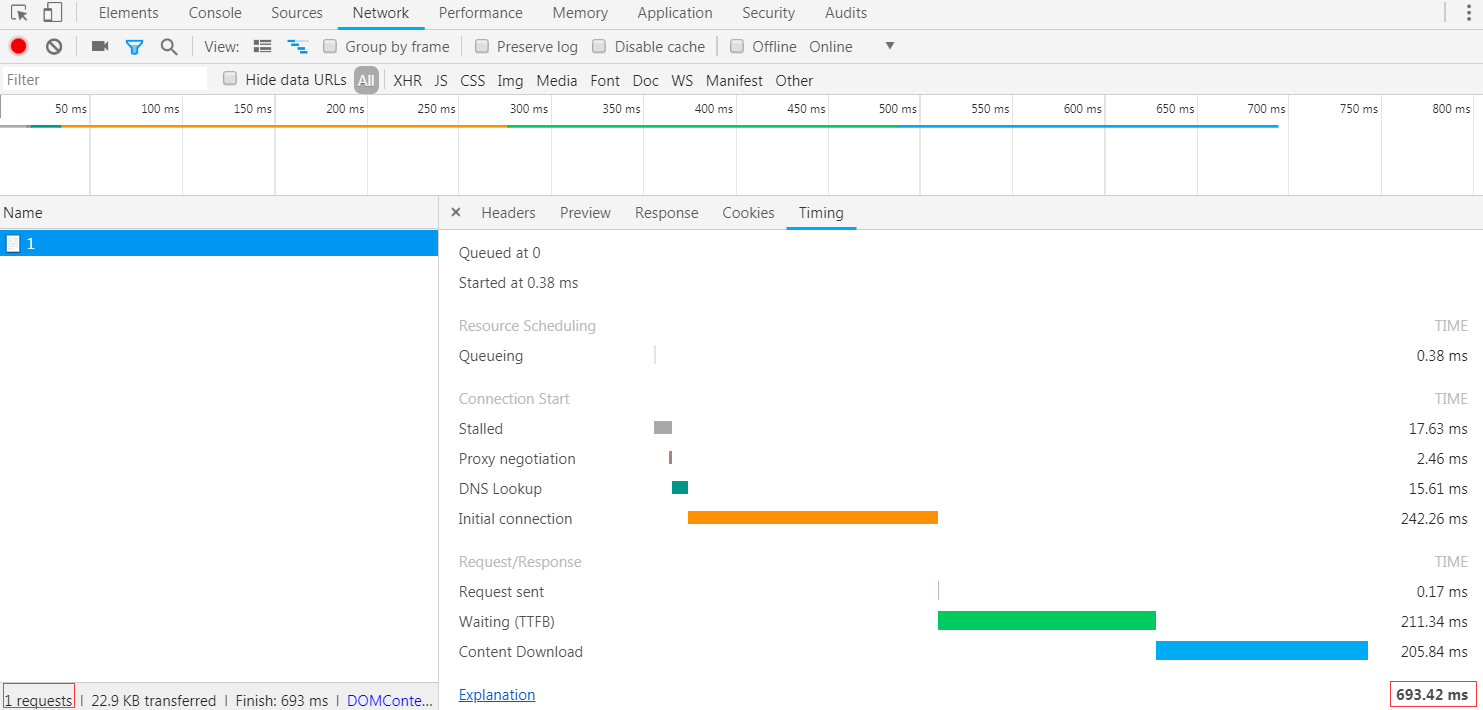
Connecting:TCP建立连接时间

Waiting : Waiting for a response from the server.

Receiving :read entire response from server

Chrome

[Reference](https://developers.google.com/web/tools/chrome-devtools/network-performance/understanding-resource-timing)



Queuing：阻塞的原因

Finish： timeline里所有request总时间

DOMContentLoaded ： 加载html dom内容，不包括js，css等

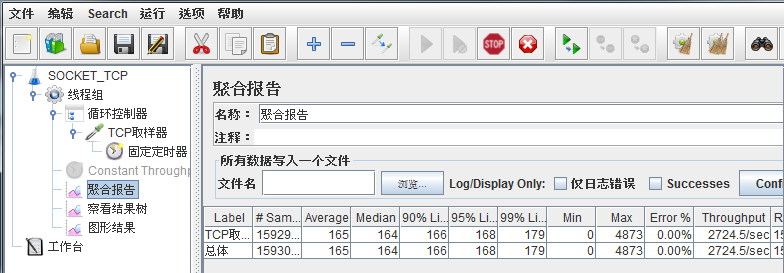
Load： 整个页面的时间

##### 吞吐量

系统单位处理的请求数（QPS）。一个TCP连接可同时发出多个请求，吞吐量是衡量服务器性能的指标。

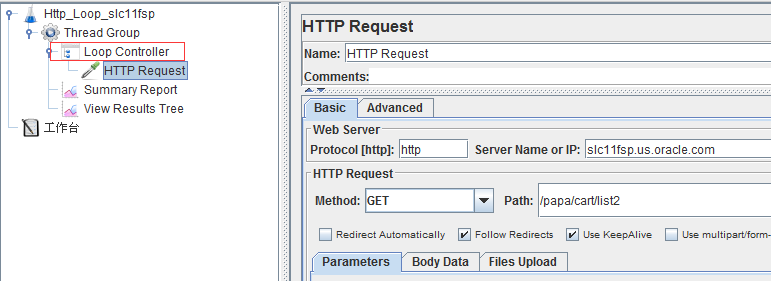
JMeter

TCP



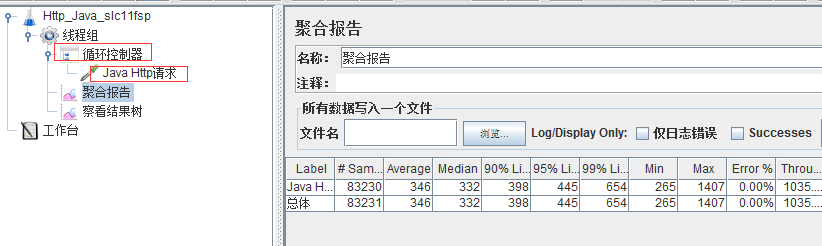
HTTP

短连接



此处是自带的HTTP Request，虽然Loop，但每个request都会重新建立TCP连接，使用完close。

长连接



java中不要close连接

性能计数器

包括System Load、对象和线程数、内存、CPU、磁盘、网络IO

#### 7.2.1.2 性能测试方法

* 性能测试：性能预期
* 负荷测试：找到性能安全临界线
* 压力测试：超过安全符合继续施压，获得最大压力承受能力

压力测试

<https://blog.csdn.net/wyingquan/article/details/414608>

#### 7.2.1.3 测试流程

很多时候在线上出现问题时，往往会伴随着某些指标的异常。大部分情况下，在问题发生之前，某些指标就会提前有异常显示。

step 1 概括

top查看整个系统负荷，正常情况单核负荷在0.7，四核就是3一下。

### 7.2.2 系统卡顿

系统卡顿，响应慢，往往会伴随一些系统指标异常，第一步应找到系统瓶颈。可能在CPU，IO，也可能在内存。一些常见引起卡顿的原因

* 操作数据库 %idle值高
* 大量计算 %idle值低，
* 大量IO。%idle值高，%iowait值高，详细信息看iostat
* 死锁：大量线程执行同一方法？方法可能有request请求
* 大量频繁线程切换
* 内存：内存分配

定位线程

ps或jps找java进程pid🡪top -Hp $pid H 查看耗cpu时间最多的几个线程🡪线程ID转换成16进制🡪jstack thread dump

#### 场景分析

CPU使用率高，响应慢

CPU持续飙高，定位到占用CPU最大的线程分析。可能的原因，

CPU处理能力相对较低/计算量大。

CPU使用率不高，响应慢

特征：%idle值高，dump线程，分析线程stuck原因。

**IO**：%iowait值高。分析IO，详细信息看iostat

**数据库**

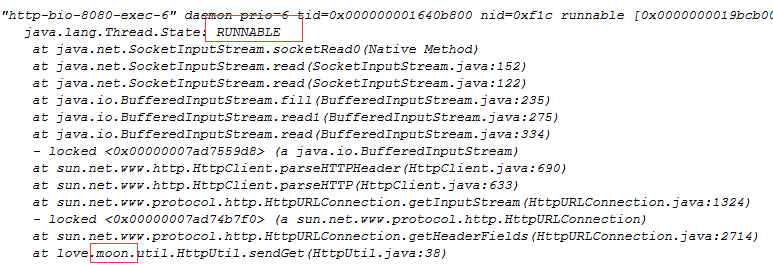
**内存**：内存分配

线程切换：

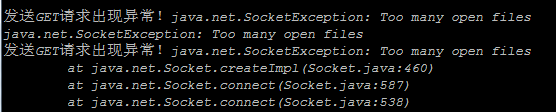
请求无响应

各项指标正常。是否所有的Runnable线程都一直在执行相同的方法，如果是的，恭喜你，锁住了！

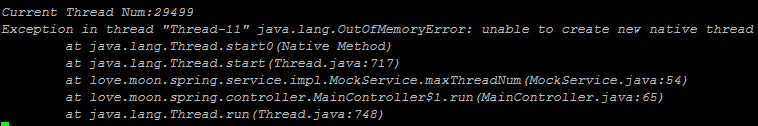
TODO线程池 大量runnable



请求第三方，如果不返回，线程一直RUNNABLE



超过4000个连接后



### 7.2.2 线程优化

系统负载高原因

1.线程切换

理想状态CPU，内存，IO外设都满负荷工作，大部分情况瓶颈都在IO的读写部分，多线程适合计算性case。

CPU主频

CPU内核工作的时钟频率，单位之间产生脉冲个数。用于同步指令执行。数值不代表运算速度(和运算速度相关：cache，指令集 ，CPU位数)，但对提高运算速度至关重要。

cat /boot/config-`uname -r` | grep 'CONFIG\_HZ='

cat /proc/interrupts

中断

TODO 操作系统时间片和中断、主频的关系？

#### 7.2.2.1 线程优化

线程切换开销

[Reference](http://ifeve.com/java-context-switch/)

上下文切换会带来直接和间接两种性能开销

**直接消耗**: CPU寄存器需要保存和加载, 系统调度器的代码需要执行, 刷新TLB(TLB：translation lookaside buffer，快表，可以理解为页表缓冲，地址变换高速缓存), CPU 的pipeline需要刷掉;

**间接消耗**：多核的cache之间得共享数据,对于程序的影响要看线程工作区操作数据的大小。

切换时间实测

线程数量

**1) Max Number of threads:**

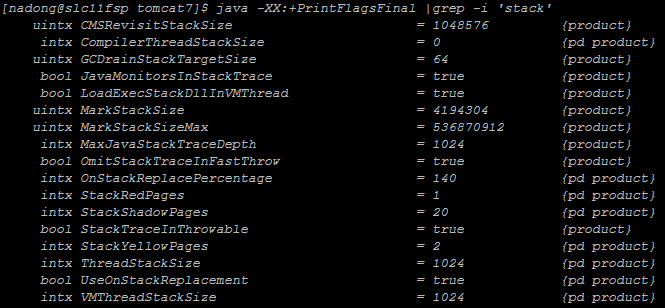
(MaxProcessMemory - JVMMemory - ReservedOsMemory) / ThreadStackSize

* MaxProcessMemory 指的是一个进程的最大内存。在32位的 windows下是 2G
* JVMMemory         JVM内存
* ReservedOsMemory  保留的操作系统内存
* ThreadStackSize      线程栈的大小(-Xss)

2)线程数量

((CUP时间+CUP等待时间)/CUP时间)\*CUP数量

ThreadStackSize



java -XX:+PrintFlagsFinal | grep ThreadStackSize

#### Thread Dump

tomcat

dump分析发现大量WAITING

#### 7.2.2.2 TOMCAT

tomcat的线程优化，本质上就是java 的线程池优化。结合socket特点来优化。实现线程资源和任务的一种平衡。

Server

超过最大连接数，响应时间变长

最大连接数越大，cup load约高

线程数越大，超过500，响应时间变长。线程切换时间？

吞吐量是否应该在服务端测，排除传输时间干扰。

##### 线程池

**thread busy**: jstack pid | grep " RUNNABLE" | wc –l

**waitting**: jstack pid | grep "WAITING" | wc –l

如果tomcat线程池minSpareThreads设置较大，大量线程将parking

##### maxThreads

Tomcat7默认值200。

实际应用中线程真正用于计算的时间可能很少，大多数线程都在等待或阻塞，如等待数据库返回数据、等待硬盘读写数据等。maxThreads应设大点，能支持更改的并发，最大化利用cpu闲置资源。但并不是越大越好，过大CPU会花费大量的时间用于线程的切换，整体效率会降低。

平衡点

只能做性能测试？

##### 设置策略

默认有12个系统线程负责。。。大量的线程用于处理sock，由于web one thread per one requst，看作一个线程和一个sock对应。线程池设置策略看作针对sock的设置

常设的参数

maxThreads：最大的线程数。看作最大连接数。

maxConnections

maxConnections的设置与Tomcat的运行模式有关。如果tomcat使用的是BIO，那么maxConnections的值应该与maxThreads一致；如果tomcat使用的是NIO，那么类似于Tomcat的默认值，maxConnections值应该远大于maxThreads

acceptCount

所有可用的线程数都被使用时，可以放到处理队列中的请求数。默认值10。

通过前面的介绍可以知道，虽然tomcat同时可以处理的连接数目是maxConnections，但服务器中可以同时接收的连接数为maxConnections+acceptCount 。

acceptCount的设置，与应用在连接过高情况下希望做出什么反应有关系。如果设置过大，后面进入的请求等待时间会很长；如果设置过小，后面进入的请求立马返回connection refused

KeepAlive

maxKeepAliveRequests： 保持请求数量，默认值100。

keepAliveTimeout 默认值60000

keepAliveTimeout VS connnectionTimeout

**connnectionTimeout**： 网络连接超时，默认值60000，单位：毫秒。通常可设置为30000毫秒。设置为0表示永不超时，这样设置有隐患的

minSpareThreads：初始创建数。默认25。

maxSpareThreads： max，超过将关闭socket。默认值75。

VS maxThread最大创建数

**acceptorThreadCount**： acceptor数量

bufferSize： 输入流缓冲大小，默认值2048 bytes。

compression： 压缩传输，取值on/off/force，默认值off。 其中和最大连接数相关的参数为

enableLookups： 是否反查域名，默认值为true。为了提高处理能力，应设置为false

server.xml

1. **<Connector** port="8080" protocol="HTTP/1.1"
2. connectionTimeout="20000" maxThreads="1000" minSpareThreads="60"
3. maxSpareThreads="600" acceptCount="120"  redirectPort="8443" URIEncoding="utf-8"**/>**

### 7.2.3 内存优化

JVM的编译优化

指令重排序，内存栅栏等

#### 查看内存

PrintFlagsFinal

java -XX:+PrintFlagsFinal

**PrintCommandLineFlags**

已经设置参数

jmap –heap pid

1. Heap Configuration:
2. MinHeapFreeRatio         = 0
3. MaxHeapFreeRatio         = 100
4. MaxHeapSize              = 1073741824 (1024.0MB)
5. NewSize                  = 314572800 (300.0MB)
6. MaxNewSize               = 314572800 (300.0MB)
7. OldSize                  = 759169024 (724.0MB)
8. NewRatio                 = 2
9. SurvivorRatio            = 8
10. MetaspaceSize            = 21807104 (20.796875MB)
11. CompressedClassSpaceSize = 1073741824 (1024.0MB)
12. MaxMetaspaceSize         = 17592186044415 MB
13. G1HeapRegionSize         = 0 (0.0MB)

jstat –gcutil pid 查看内存变化。

#### JAVA OPS优化

**方法区Perm**

* -XX:PermSize=5M -XX:MaxPermSize=7M

**32G 内存配置示例：**

-Xms10g -Xmx10g -Xmn1024m -Xshare:off

**jvm性能调优**

* 控制GC的行为. 频繁GC导致性能下降,因此经常会根据系统运行的程序的特性来更改GC行为
* 调整JVM堆栈大小.xmx xxx
* 控制JVM线程的内存分配.xss

**JVM调试方法：**

1.Jconsole查看内存各部分变化

2.gc 用jstat

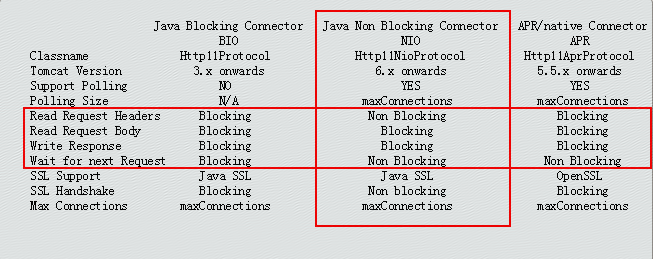
#### Heap Dump

jmap -dump:format=b,file=/scratch/tools/heamdump.out 22204

### 7.2.4 IO优化

配置文件server.xml，略。

IO模型支持



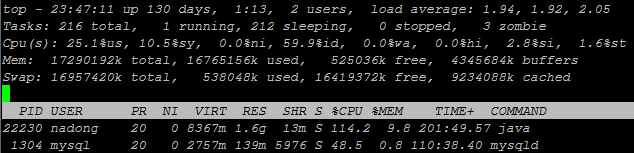
tomcat要实现servlet规范所以不能最大发挥NIO的特性，servlet3.0之前完全是同步阻塞模型，Read http body 以及 response还是同步阻塞，因为servlet规范规定的就是这样。

### 7.2.5 监控Tools

#### 7.2.5.1 Linux

##### top

动态实时显示。关注%CPU %MEM



**Cpu(s)**

%us user CPU time，这里指整个CPU计算能力的百分比。标题中%CPU指单核的。

%ni：改变过优先级的进程，%id：idle，%wa ：I/O waiting

%hi：硬中断，%si：软中断，%st: steal time

**Mem**

used：程序使用的+buffers. centos 7不包括缓存。

free：未分配

**S**：R=运行 S=睡眠 D=不可中断的睡眠状态T=跟踪/停止 Z=僵尸进程

**快捷键：**P:CPU排序。M:内存排序，H: show thread，1:show cups 列出每个cpu负荷

Title:

VIRT 进程使用的虚拟内存总量，单位kb。VIRT=SWAP+RES

**Load Average**

一段时间 (1 分钟、5分钟、15分钟) 内平均 Load

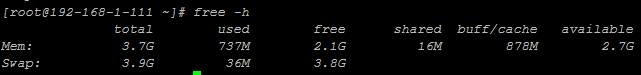
系统负荷正常范围：

* 大于0.7，必须开始调查了，问题出在哪里，防止情况恶化。
* 持续大于1.0，必须动手寻找解决办法，把这个值降下来。
* 当系统负荷达到5.0，就表明系统有很严重的问题，长时间没有响应，或者接近死机了。

以上指标基于单CPU的，四核机器的负载最好保持在3(4\*0.7 = 2.8)以下。

##### strace

##### free



used 程序使用的

free：未分配

avaliable：free+buff/cache-不可回收部分（tmpfs）

##### netstat

1.Linux下查看并发连接数，一个浏览器可能建立几个连接，非并发用户数

netstat -na | grep ESTABLISHED | grep 8080 | wc -l

#### 7.2.5.1 Sysstat

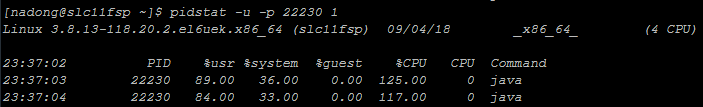
##### pidstat

–p pid.

-u/r/d cpu,内存, io. 默认显示CPU信息.

pidstat 2 10 2s为间隔显示10次。

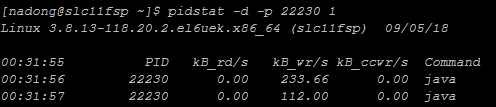
-u



CPU:在那个核运行

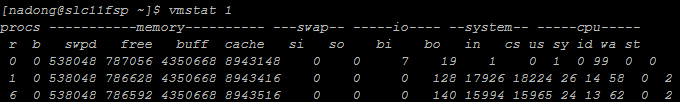
针对单个的线程stat，和top类似

-d



##### vmstat

[Reference](https://linuxtools-rst.readthedocs.io/zh_CN/latest/tool/vmstat.html)



**proces**: r 分配到CPU进程数，b阻塞进程

**memory**: swpd虚拟内存已用大小。

**Swap**: si 每秒从交换区写到内存的大小，so，每秒写入交换区的内存大小

io

单位：blocks/s，Linux块的大小为1M。从内存角度，bi 从block device 输入, bo 输出到device,

system

in中断数/每秒，cs 上下文切换/s。context指process context

**cpu**: us,sy,id,wa

##### iostat

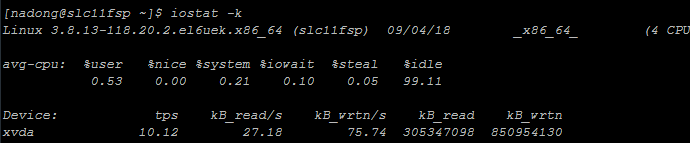
统计CPU和IO。都是针对整个系统的，不能对单个线程统计，区别于top/pidstat。

[Reference](https://linuxtools-rst.readthedocs.io/zh_CN/latest/tool/iostat.html)

-C/d 显示CPU/磁盘 使用情况

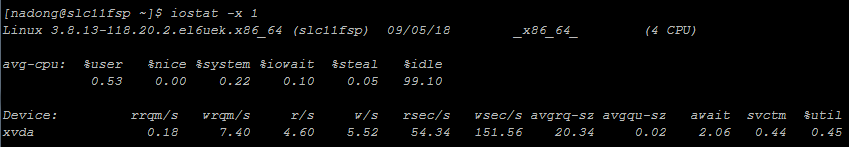
-k/m 以kb/M为单位显示

-d



TPS 每秒I/O请求。可理解为吞吐量

-x 详细信息。



r/s: 每秒完成的读 I/O 设备次数。即 rio/s

w/s: 每秒完成的写 I/O 设备次数。即 wio/s

avgqu-sz: 平均I/O队列长度。

await: 平均每次设备I/O操作的等待时间 (毫秒)。

svctm: 平均每次设备I/O操作的服务时间 (毫秒)。

%io消耗的cpu百分比

##### sar

#### 7.2.5.2 Java

##### jstack

生成线程快照。

[Reference](http://www.hollischuang.com/archives/110)

**用法**：jstack <pid>

**主要目的**：定位线程长时间停顿原因，如线程间死锁、请求外部资源导致的长时间等。

Thread Dump分析

prio：默认值为5，最大值为10，最小值为1 os\_prio ，tid：线程id，nid：native id，操作系统记录的id，thread state

**RUNNABLE**

locked <地址> (目标)。进入synchronized方法, 申请对象锁成功才会出现此句



BLOCKED(on object monitor)

waiting to lock <地址> (目标)：使用synchronized申请对象锁未成功,在迚入区等待。

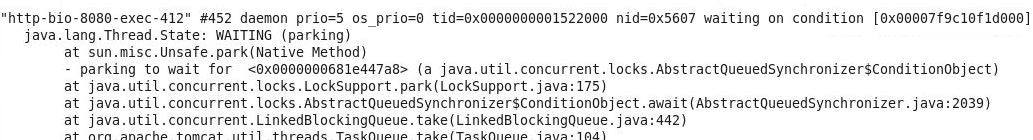


WATTING(on object monitor)

waiting on <地址> (目标)。synchronized修饰的方法调用wait()。Thread释放锁后会**等待锁**



WATTING(parking)

- parking to wait for <地址> (Object)

TIMED\_WAITING(parking)



这里tomcat线程池minSpareThreads设置较大，大量线程将parking

TIMED\_WAITING(sleeping)/ TIMED\_WAITING(on object monitor)

at java.lang.Thread.sleep

**Thread State**

Wait on condition

死锁

http://www.hollischuang.com/archives/2409

http://www.hollischuang.com/archives/110

##### jstat

jstat -option pic < interval >

**gcutil**



S0：年轻代中第一个survivor（幸存区）已使用的占当前容量百分比   
S1：年轻代中第二个survivor（幸存区）已使用的占当前容量百分比   
E：年轻代中Eden（伊甸园）已使用的占当前容量百分比   
O：old代已使用的占当前容量百分比

P：perm代已使用的占当前容量百分比   
YGC：从应用程序启动到采样时年轻代中gc次数   
YGCT：从应用程序启动到采样时年轻代中gc所用时间(s)

FGC：从应用程序启动到采样时old代(全gc)gc次数   
FGCT：从应用程序启动到采样时old代(全gc)gc所用时间(s)   
GCT：从应用程序启动到采样时gc用的总时间(s)   
**gc**



 S0C：年轻代中第一个survivor（幸存区）的容量 (byte)   
 S1C：年轻代中第二个survivor（幸存区）的容量 (byte)   
 S0U：年轻代中第一个survivor（幸存区）目前已使用空间 (byte)   
 S1U：年轻代中第二个survivor（幸存区）目前已使用空间 (byte)   
 EC：年轻代中Eden（伊甸园）的容量 (byte)   
 EU：年轻代中Eden（伊甸园）目前已使用空间 (byte)   
OC：Old代的容量 (byte)   
OU：Old代目前已使用空间 (byte)   
PC：Perm(持久代)的容量 (byte)   
PU：Perm(持久代)目前已使用空间 (字节)

**gcnew**

年轻代对象的信息



**gcold**

 old代对象的信息



**gcoldcapacity**

old代对象的信息及其占用量



OGCMN：old代中初始化(最小)的大小 (字节)   
OGCMX：old代的最大容量 (字节)   
OGC：old代当前新生成的容量 (字节)   
OC：Old代的容量 (byte)

**gcnewcapacity:**



 NGCMN：young初始化(最小)的大小 (字节)   
 NGCMX：年轻代(young)的最大容量 (字节)   
 NGC：年轻代(young)中当前的容量 (字节)

S0CMX：年轻代中第一个survivor（幸存区）的最大容量 (字节)   
S1CMX ：年轻代中第二个survivor（幸存区）的最大容量 (字节)

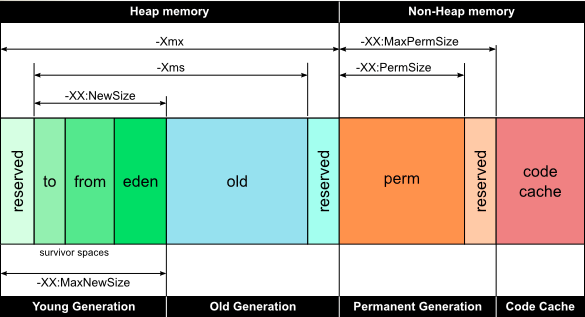
ECMX：年轻代中Eden（伊甸园）的最大容量 (字节)   
**gcpermcapacity**

 PGCMN：perm代中初始化(最小)的大小 (字节)   
  PGCMX：perm代的最大容量 (字节)     
  PGC：perm代当前新生成的容量 (字节)   
PC：Perm(持久代)的容量 (byte)

##### jmap

heap

NewSize: young size



Heap

* -Xms heap的最小内存；-Xmx heap的最大内存；
* -Xmn young的大小，一般设置为Xmx的3、4分之一
* -XX:NewSize: young的大小
* –XX:MaxNewSize
* –XX:NewRatio: young:old,2表示1:2,即young/heap=1/5
* -XX:SurvivorRatio:2\*Survivor:eden,8代表1:1:8
* -server 启用JDK 的 server 版；一定要作为第一个参数，在多个CPU时性能佳

No Heap

* -XX:+UseParNewGC ：缩短minor收集的时间
* -XX:+UseConcMarkSweepGC ：缩短major收集的时间。此选项在Heap Size 比较大而且Major收集时间较长的情况下使用更合适

dump

jmap -dump:format=b,file=/scratch/tools/heamdump.out 22204

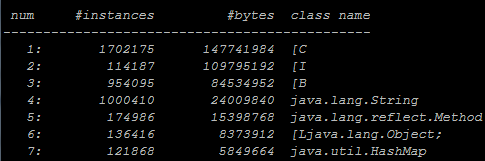
live：dump only live objects; if not specified,all objects in the heap are dumped.

format=b：binary format

-histo

jmap –histo[:live] pid

堆对象信息



##### Tomcat Mgr

[Reference](https://tomcat.apache.org/tomcat-7.0-doc/manager-howto.html)

Server Status

JVM

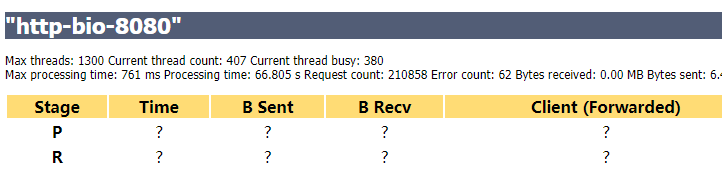
Free:可用内存

Total：当前分配。包括Free和未使用

Max：最大:

Thread Pool

<http://localhost:8080/manager/status>



Current thread busy: Runnable状态的线程

**Stage**

P: Parse and prepare request 正在处理请求

R: Ready. 线程sleep状态

S: Service

F: Finishing

K: Keepalive

TODO2 linux 命令 tomcat 每次只能accept一个request？

Status

<http://localhost:8080/manager/status?XML=true>

Thread Dump

<http://localhost:8080/manager/text/threaddump>

VM Info

<http://localhost:8080/manager/text/vminfo>

<http://localhost:8080/manager/status?XML=true>

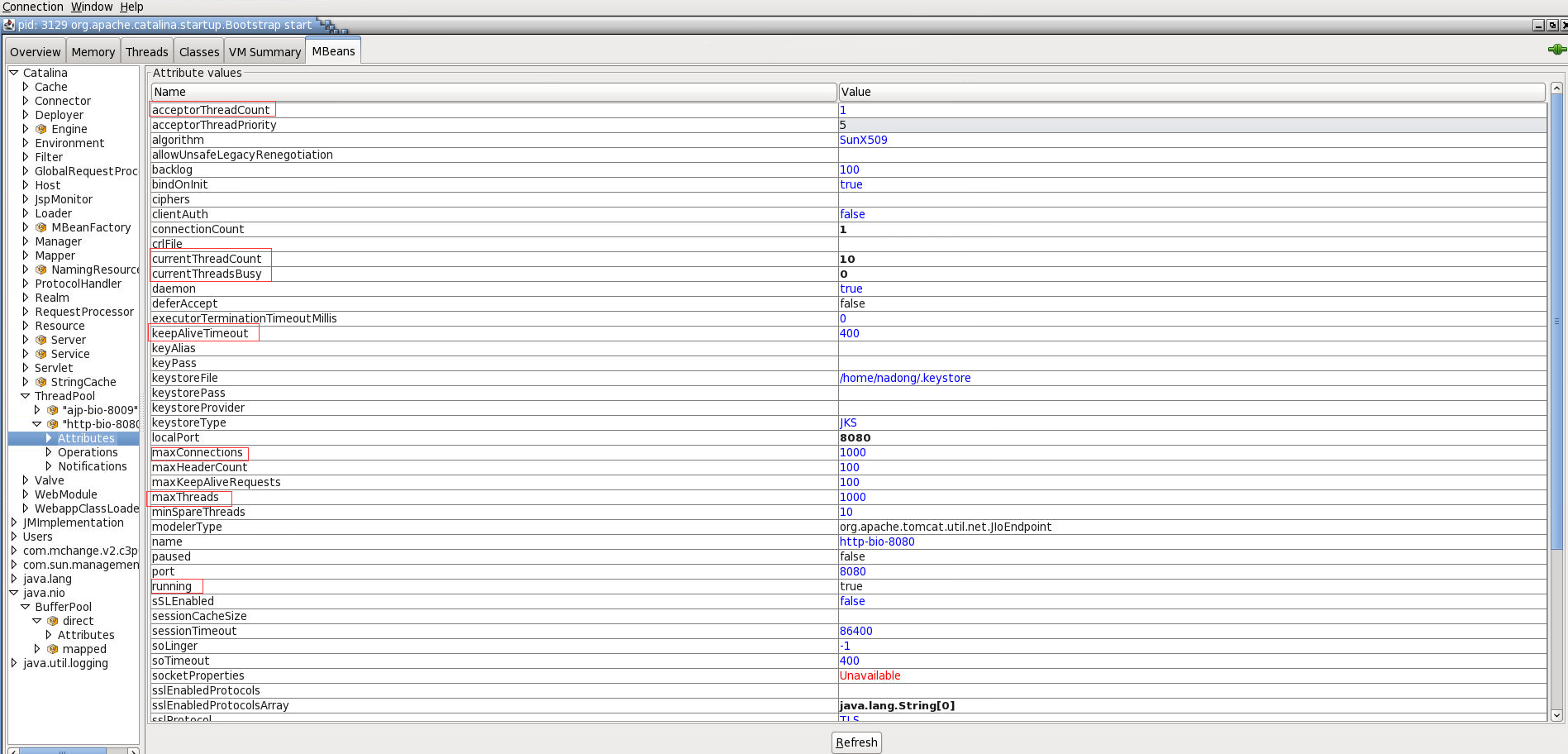
##### jconsole

配置

JAVA\_OPTS

-Djava.rmi.server.hostname=127.0.0.1 -Dcom.sun.management.jmxremote.port=8090 -Dcom.sun.management.jmxremote.authenticate=false -Dcom.sun.management.jmxremote.ssl=false

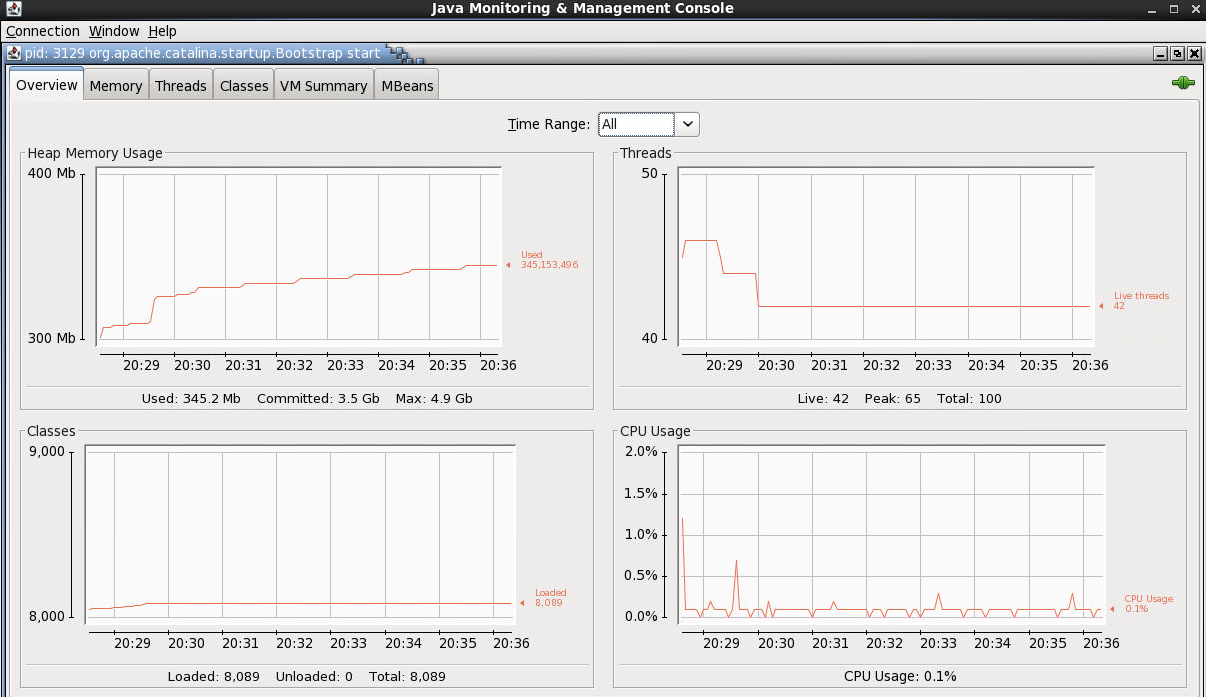
查询Acceptor属性



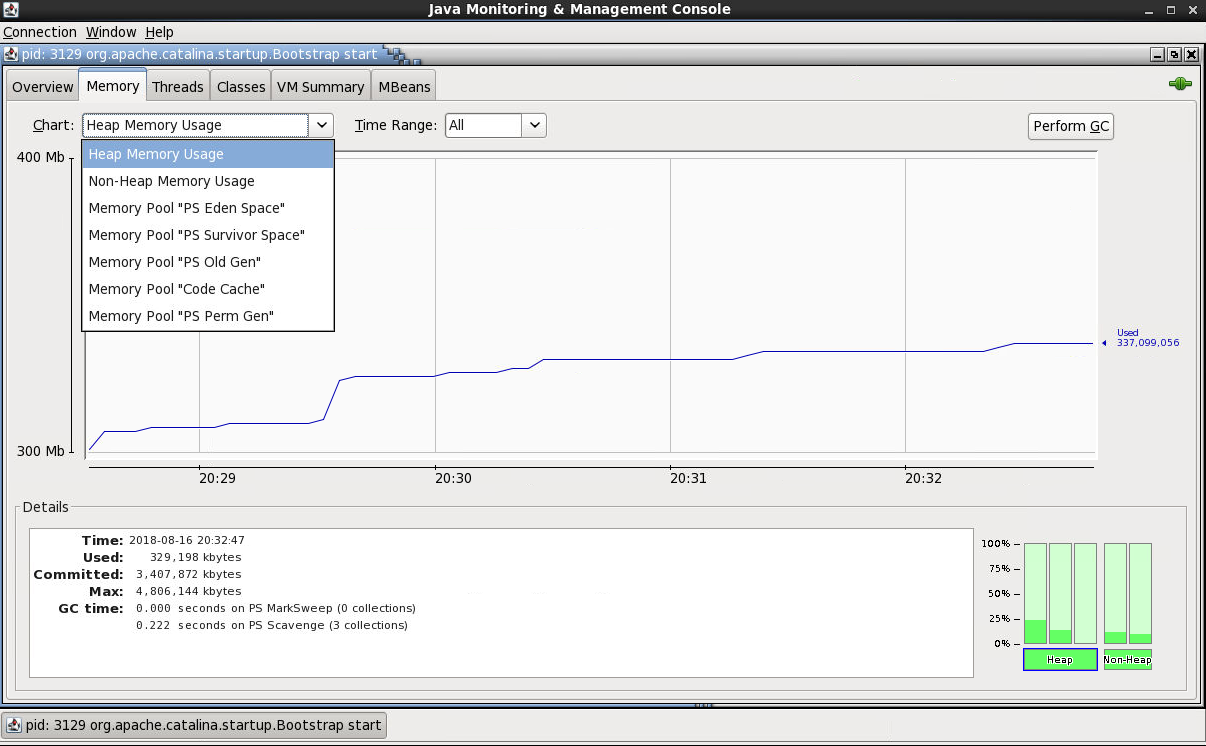
currentThreadCount

currentThreadBusy 线程池

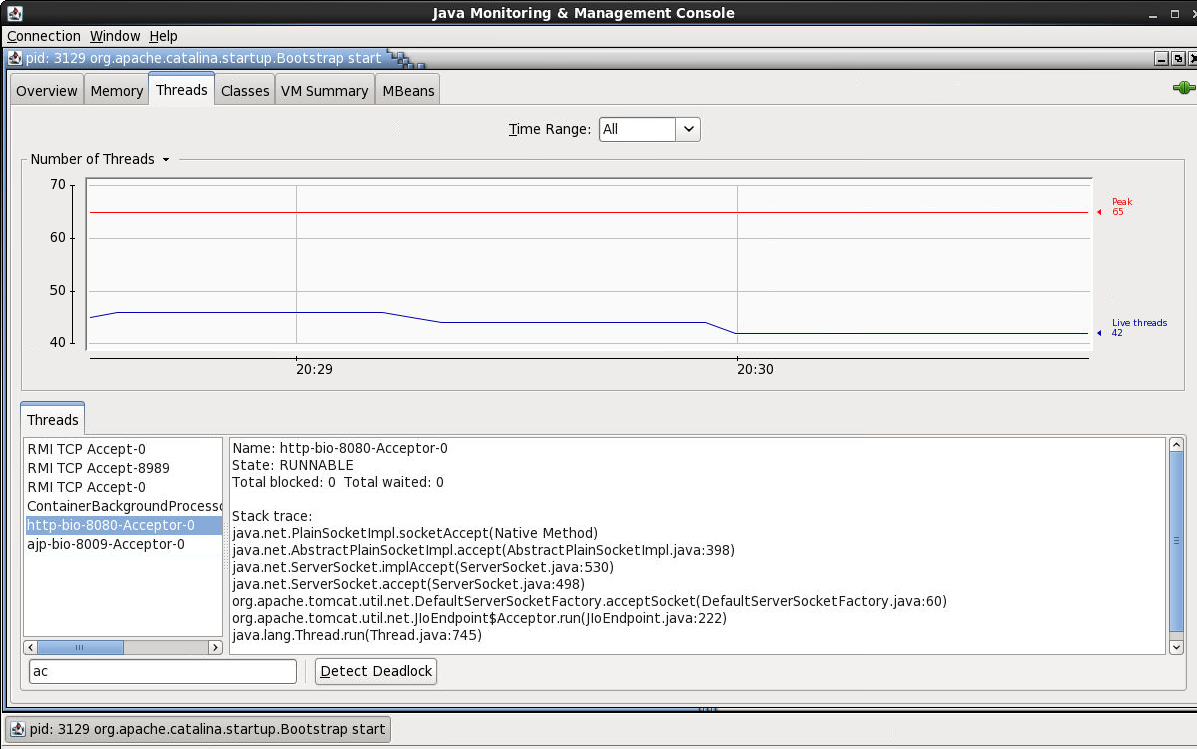
Overview



Memory



Thread



Total blocked

Total waited:

**state**

Wait on condition

##### Jvisual

JDK自带性能检测工具



**Running**: thread is still running.   
**Sleeping**: thread is sleeping (method yield () was called on the thread object)  
**Wait**: thread was blocked by a mutex or a barrier, and is waiting for another thread to release the lock

**Park**: parked threads are suspended until they are given a permit. Unparking a thread is usually done by calling method unpark () on the thread object  
**Monitor**: threads are waiting on a condition to become true to resume execution

## 7.3 MYSQL

#### MySQL Workbench



Key Efficiency

<https://gxnotes.com/article/72479.html>