## 信息管理技术

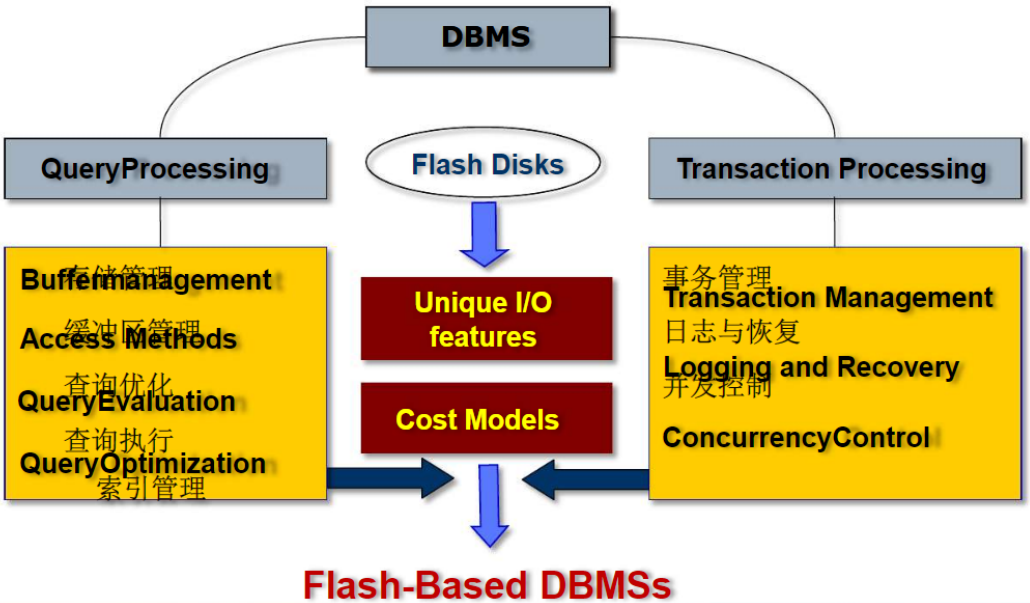
### 1、Flash-base database

基于 flash 的数据库系统（Flash-base database，也称为闪存数据库系统）：在了解基于 flash 的数据库系统之前，关于 flash，我们的了解或许停留在浏览网页时候的插件。通常，flash 用于前端的现实和交互，而大多数情况下数据的收集处理借助其他语言（如 java、C++、php）的帮助。

其实随着闪存技术的不断发展，闪存已经应用十分广泛了，比如我们的手机、电脑等等。过去，我们的数据管理通常基于磁盘存储，这是由特定的技术限制决定的。闪存是一种新的数据存储方式。闪存相对于磁盘，期读操作的销毁在磁盘读消耗的 2%左右，而写消耗也在 30%左右，而在效率上，这个比值则是数百。我们可以发现，flash 的存储操作特性中读取的速度较快而随机写的速度较慢，这主要是由于擦除的操作导致的，如何平衡这种性能提升的差异来形成和谐的数据库处理性能环境也成为基于 flash 的数据库系统的挑战之一。而由于闪存技术的发热，其读写的高效特性，也引发人们对已有的存储管理作出改变。其中，数据库管理也不断发生着技术改变。

随着闪存技术的出现，数据库系统也对应带来了升级。我们对于现有的数据库管理系统要进行改造，例如：日志存录结构和事物处理的相关变化，索引结构的调整和查询优化的要求蜂拥而至。如果我们直接使用闪存代替原始的磁盘，那么，可能不能很好地发挥闪存的物理特性，不能最大化发挥闪存的优良特性。有相关研究表明，一些数据库的事务处理能力在固态硬盘上的存储速度是传统磁盘的十倍，但是，实际上，闪存的读写速度相比磁盘应该远超这个倍数值，因而，如何更好地发挥闪存的提速提性能效果十分重要，对于提升数据库处理效率发挥着重要作用。

为了改善数据库的性能，需要利用独特的 flash I/O 特性。下图是本人从互联网上寻找的一种闪存数据库的示意图。我们可以看到，数据库分为查询处理和事务处理两种处理内容。其中查询处理分为：存储管理、缓冲区管理、查询优化和索引管理。而事务处理分为事务管理、日志与恢复和开发控制。而为了改善性能，闪存磁盘利用自身的 I/O 特性来加速查询和事务处理。对于数据库来说，这是十分必要的。相应地，要利用好闪存的特性，也需要相应的算法调整，来利用好闪存特性。




### 2、GPU 的应用

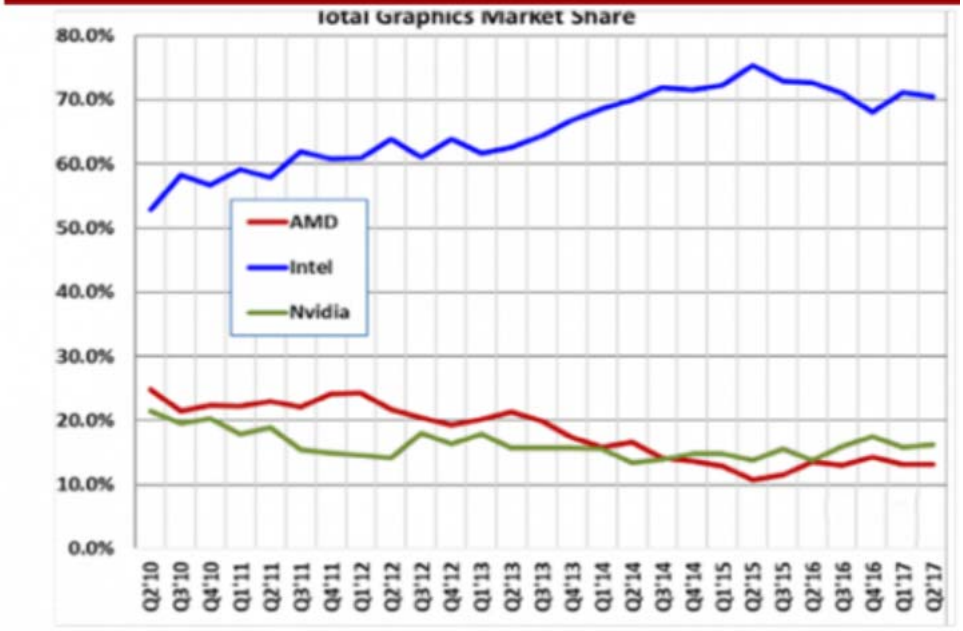
GPU 在面向大规模数据计算的领域中具有广阔的应用前景，主要应用与图形渲染、字符串匹

配、矩阵乘法、数据挖掘和 OLAP 计算等等方面。GPU 的主要技术思路有两点，其一是面向并行的计算分解，其二是数据传输的瓶颈。所谓 GPU 就是 Graphics Processing Unit 图形处理器，是一种为计算机设备进行图形处理的微处理器设备。GPU 减轻了 CPU 关于图形处理，3D 图像效果增强和图像纹理等等与显示有关的的工作。图形处理器包括了我们经常选电脑的时候提到的显卡。我们在购买电脑之前，首先会看看它的显卡配置，一些廉价的电脑并没有独立显卡，可能在显示图像上的效果相应较差，cpu 的负担相对较重，不适合进行需要复杂图像处理的客户。而拥有独立显卡的电脑相对来说显示性能更好。处理显卡以外，GPU 还包括显示缓冲存储器和 RAMD/A 转换器（负责吧数字信号转换为模拟信号）。

当今的 GPU 是以传统 Z-buffer 算法为基础的，但是它已经不能满足新的应用需求，随着人们对于实时画面图像的要求提高，一些更加复杂的图像算法不得不被开发而加以利用，从而提升 GPU 性能。

根据 2017 年中国报告网关于《2017 年我国 GPU 的应用分析及发展趋势预测》，PCU 在图形处理上广泛应用，而报告也显示，PC 端的 FPU 份额中 Intel 独占鳌头，剩下的小部分市场被瓜分。报告还指出，在智能手机时代，移动端的 GPU 适应性要求更高，可是市场却呈现更加集中的态势。

图：PC 端 GPU 市场份额




### 3、MapReduce 计算流程

实质上是一个 Map 和 Reduce 两个思想的结合，Map 代表映射，而 Reduce 代表归约。MapReduce 实质上为不会分布式系统编程的程序员将程序运行到分布式系统上提供了方便。MapReduce 基于并行计算，我们都知道并行程序的编写是十分复杂繁琐的，在保证一致性和复杂性上常常让人费解头疼。MapReduce 通过对大规模数据的各个节点分配可靠性信息的方式来化简操作。从目前的发展趋势来看，并行计算的发展还没有很长的历史，也并没有暴热。目前的并行计算对于编程人员的要求仍然过高，MapReduce 这样的并行编程简化计算流程从时间维度来看，还有很大的发展空间。

### 4、海量网络数据管理

在我们的网络环境中，存在许多实体和关系。这些网络包括技术网络、生物网络、社会网络和概念网络。其中技术网络包括因特网、万维网、电网和传感器网络等等，而生物网络包括蛋白质相互左营网络、基因调控网络和代谢网络等等，社会网络包括社会中个人、组织

之间的社会关系网络、国际经济体系中全球贸易网和在线社交网络等等，概念网络包括实体关系网络、语义链接网络和 LinkedData 等等。我们可以看到，网络无处不在，无论是在日常生活中还是网络技术中，网络都不可或缺，交织这复杂的关系。面对纷繁复杂的关系实体，我们需要进行抽象，形成概念化的网络，进行组织管理和信息提取，从而获得对于我们有用的但是在网络中分散存在或者需要归纳提取的内容判断。

老师提到了我们复旦中文只是图谱平台（km.fudan.edu.cn）。这个平台包含 2000 万实体，5000 万关系，这个平台部分领域准确率，比如军事上，准确率达到 95%。这个平台自身包含一个海量的关系网络，由一个标签牵引出多个标签，不断扩张形成一个巨大的网络。

目前，大数据正火。一个原因在于许多企业认为海量的数据存在潜在的未来价值。几个月前，我在粤港澳大湾区调研的时候就发现许多企业都在花费一定的金钱存储海量数据。一些公司甚至不知道这些海量的数据应该怎么使用，但是却表示相信其未来的价值，并表示如果自己现在不关注海量大数据，未来可能就会落伍，就会被淘汰。从目前的情况看，许多高校的大数据专业都开设不久，整个学科体系也还不完善，未来的发展空间还很大。我们了解到，广州互联网法院还在积极探索区块链和大数据对于互联网法院的价值。可见，计算机知识正在飞速侵入各个行业各个领域，即使以严肃庄严为底色的法律界也不例外。大数据的未来前景比较广阔。

**5、数据库替换：ForkBase**

ForkBase 的数据库有旨在支持需要数据版本控制、分叉和防篡改等功能的应用。现有的索引结构并不适合数据中途分叉，且后续多个分支可以合并的环境。将 Merkle 树该用面向模式的拆分树，它支持下列属性：

快速查找和更新、快速确定两棵树之间的差异，然后合并、高校的重复数据删除和防止篡改（保留使用 Hash 校验数据功能）。

其实，ForkBase 的一个例子是区块链。区块链（Blockchain），是比特币的一个重要概念，它本质上是一个去中心化的数据库，同时作为比特币的底层技术，是一串使用密码学方法相关联产生的数据块。我之前在出于兴趣在各个网站也查询了一下区块链技术，并结合他人的博客，总结了自己的理解：

**区块链：**

首先我们建立一个去中心化的借贷系统。如果这个系统里面 A 借了 B100 元，那么没有一个可信赖的中间人，A 以后不承认借了 B100 元怎么办？在这个系统里面，B 就可以高喊“看，A 在 xxxx 年 xx 月 xx 日借了我 100 元”。于是群众甲乙丙丁等人就会听到这次喊话，如何拿小本本记录下来。如果有一天[i]A 不承认借了 B100 元，那么这时候群众就可以跳出来说：“不对，我本本上记着你某一天从 B 那里借了 100 元。”

那么问题来了，要是有人动了心眼，明明整个系统只有 1000 元，有个人只有 100 元却说：“我有 1000 元”，那样岂不是可以伪造数据了？为了反之这种情况，我们就要给每次喊话记上编号，这样逐层被记录之后就可以解决伪造的问题。

可是大家凭什么把你的话记录到小本本上呢？那么就需要给第一个记录这句话的人一笔酬劳￥￥￥，之后的人便放弃记录这句话。这之后，第一个记录的人给这条记录编号（例如 001），然后在把这句话和记录编号一起喊出来给下一个人记录。

**区块链技术有优势：**

①去除中间机构，运行成本降低

由于区块链技术去中心化的特点，不需要中间机构介入，交易过程经手着更少，而是通过整个体系规则自动运转，这降低了运行成本。

②公开透明，无需专人核查，提升效率

区块链中人们拥有相同的账本，而一些错误会通过程序本身受到隔离，不需要专人核查。而

无中间商和核查过程都提升了系统效率
③数据安全
在区块链技术中，除非有人恶意掌握 50%以上的节点，否则就不会出现错误。而要掌握 50%以上节点是几乎不可能的。以比特币为例，没有黑客能在比特币运行系统内得手。
但目前区块链还面临着性能问题有待突破、隐私保护有待加强以及升级修复机制有待探索等问题。

---

[i] 【参考资料】
https://baike.baidu.com/item/%E5%8C%BA%E5%9D%97%E9%93%BE/13465666?fr=aladdin
https://www.zhihu.com/question/37290469/answer/107612456