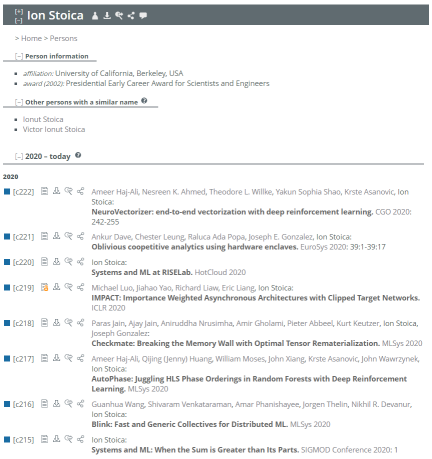
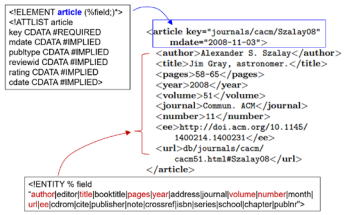
编程作业 3: 分布式 DBLP 数据查询系统 P3: Distributed DBLP Data Query System

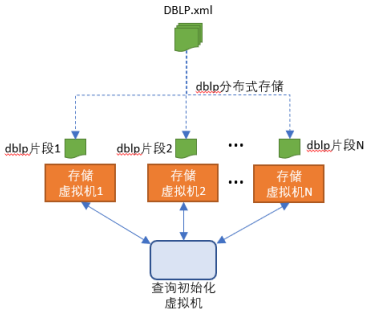
1. **DBLP 数据集介绍**. DBLP (Digital Bibliography & Library Project)是计算机领域内，其 url 链接为（http://dblp.uni-trier.de），DBLP 研究的成果以作者为核心的一个计算机类英文文献的集成数据库系统，按年代列出了作者的科研成果，包括国际期刊，会议，专著等公开发表的论文。这个项目是德国特里尔大学的 Michael Ley 开发和维护，它提供计算机领域科学文献的搜索服务，但只储存这些文献的相关元数据，如标题，作者，发表日期等，并没有包括文献的全文文件（如 pdf 文件）。下图为加州大学伯克利分校教授 Ion Stoica 的 DBLP 论文列表（对应的 url 为：https://dblp.uni-trier.de/pid/s/IonStoica.html）



DBLP 使用 XML 存储元数据（http://dblp.uni-trier.de/xml/），对应的 dblp.xml.gz 文件大小为 553 MB。DBLP 在学术界声誉很高，很多论文及实验基于 DBLP 开发。所收录的期刊和会议论文质量较高，也比较全面，文献更新速度很快，能很好地反应了国外学术研究的前沿方向。如下图为一个 XML 文件格式示例（见 http://dblp.uni-trier.de/xml/docu/dblpxml.pdf）



2. 给定一个查询初始化虚拟机和 N（如 6）个存储虚拟机，其中 N 个虚拟机负责分布式存储上述 dblp.xml 片段文件，查询初始化虚拟机负责发送一个查询请求，协调 N 个虚拟机，完成如下查询功能需求：



2.1 给出输入条件：作者名字 author=[Ion Stoica]，返回作者的 DBLP 发表论文总数。

2.2 给定上述作者信息和年份区间(如: year>2010，或者 2010< year <2022，或者 year=2010)返回该作者在对应年份的 DBLP 发表论文总数

3. 非功能性需求：

a) 存储负载均衡：每个虚拟机随机存储 dblp.xml 内容数据块，要求每个虚拟机所存储的内容空间将可能均衡；
b) 查询容错机制：这里假定查询发起客户端虚拟机不会出现故障、而 N 个存储虚拟机存在故障的可能性，要求该查询在大于 1、至多 N/2 个存储虚拟机出现故障的情况下还可以正确的查询结果；
c) 最小化端到端的查询时间：在满足上述 a)和 b)两个非功能需求的情况下，尽可能的降低在提交查询之后和返回最终查询结果之间的查询时间。

4. 提示：
   a) 通过 socket 方式实现 client/server 网络通信模式，在每个存储虚拟机实现一个 server 程序，在查询初始化虚拟机实现一个 client 程序，初始化查询请求发送至每个 server；server 程序接收到查询请求后，负责查询本地 dbml 片段文件，返回查询结果给 client 端，client 端则需要汇聚 N 个查询返回结果，最终得到汇总后的查询结果。
   b) dblp.xml 的分布式存储：简化起见，有关 dblp 论文的片段分区准则，可以考虑将 dblp.xml 随机切分，然后将一个 dblp 分片随机存储在 N 个存储虚拟机之一，以达到负载均衡目的；为了保证容错机制，每个 dblp 分片额外存储另外一份副本，使得每个 dblp 分片最多有两个副本，并且这两个副本分别存储在不同的虚拟机。在进行查询处理的时候，避免由于两个副本的缘故导致查询结果重复的问题，因此需要对容错副本进行区分。
   c) 查询时间的优化：简化起见，考虑到本次作业仅涉及到查询 author 和 year 信息，只须采用类似于 grep 或者 wc 命令统计论文的频次；可以在每个存储虚拟机统计包含输入统计论文的频次，然后存储虚拟机将本地的论文频次返回至查询初始化虚拟机，然后由存储虚拟机进行汇总，最终得到查询结果。

5. 查询初始化虚拟机最终生成 1 个 log 文件：学号-hw2-q1.log，该 log 日志文件最后一行为所用查询的时间。

6. 评分准则（共计 20 分）：
   a) 查询 Demo 演示：查询结果正确（4 分）、存储负载均衡（3 分）、查询容错（3 分）、在规定时间（如 1 分钟）内返回结果（4 分）
   b) 作业报告：2 页，作业报告包括：组成员算法设计、如何支持可扩展性、低延迟、消息格式、作业 2 如何与本次作业集成、网络带宽开销（节点 join、leave、failure-rejoin）、消息丢包时组成员列表的错误比例。（6 分）
   c) [**加分项**] 可通过一个构建一个本地索引结构，例如哈希表结构，对 author 和 year 通过论文频次进行统计，从而避免对本地 xml 数据块进行全局扫描，并对比使用本地索引结构和 grep/wc 全局扫描的查询时间。（5 分）

编程作业 4: 分布式组成员服务 P4: Distributed Membership Service（25 分）

1. 在上述编程作业 1 中包括了 1 个查询初始化虚拟机和 N 个存储虚拟机，本次作业在此基础之上，要求设计和实现一个分布式组成员服务(distributed group membership service)，可以使得N个存储虚拟机形成一个分布式组系统，每个存储虚拟机通过一个后台进程(daemon)，维护其他**组成员列表**（membership list），该列表维护所有在线(live)且连接到该组的节点成员 ID。

   在如下情况，需要动态更新上述组成员列表：

   a) 任何一个虚拟机（或及其对应的 daemon）进入到该组
   b) 该组的任何一个虚拟机（或及其对应的 daemon）自愿主动离开该组
   c) 该组的任何一个虚拟机（或及其对应的 daemon）发生 crash 宕机、被动离开该组（或假定该虚拟机在规定的较长时间内没有从宕机故障得以恢复）

   在该组服务中，任意时间存在仅一个组服务，不允许多个组服务；此外，由于该组服务是采用 crash/fail-stop 模型，故在一个虚拟机在发生故障之后重新加入到该组服务之中，虚拟机 ID 编号除了包含对应的 IP 地址之外还必须包含当前的时间戳 timestamp，使得在一个虚拟机在发生故障之后出现重复加入的情况下，组成员列表仅维护最新时间戳的虚拟机 ID。

   低延迟需求：在虚拟机发生宕机故障之后，在 2 秒内（假定采用同步时钟）至少在一个其他虚拟机所维护的组成员列表中反应出其故障情况—即所谓的时间约束的完整性（time bounded completeness）。虚拟机的宕机故障、加入和（主动）离开须在 6 秒内、所有虚拟机所维护的组成员列表中反应该情况（这里假定虚拟机之间的通信延迟较低）。

   简化起见，最多 1 个虚拟机同时发生故障，并且在 1 个虚拟机发生故障之后，要求至少在 20 秒之后才会再次发生其他虚拟机发生故障情况，避免多个虚拟机同时发生故障。

   可扩展性需求：此外，该组服务必须具有较好的可扩展性，可支持较多数量的虚拟机组成员服务，典型情况下要求 N 在 3-8 个虚拟机。

   Gossip 协议：在实现该组服务的故障检测和节点主动离开，不能使用简单的 master/leader 维护全局的组成员列表，而是通过诸如 gossip 协议实现组服务；不过，为了便于存储虚拟机节点加入 join 到该组服务操作，可以通过某个固定的节点机器（或称之为引荐人 introducer），假设所有的节点（包括新加入的节点）均知晓该 introducer 机器。通过该 introducer，新加入的节点可获取该 introducer 当前所负责的组成员列表。当 introducer 发生故障的时候，新节点则不能加入该组；但是当前的组成员还是可以正常工作，包括检测节点故障、节点主动和被动离开。

   通过 heart-heating 和 gossip 这两个机制进行故障检测，在进行故障检测的进程中，可以考虑这些虚拟机组成员所形成的拓扑结构。在实现该组服务，要求带宽开销较小，比如避免使用多到多 N*N 的 ping 消息来侦测节点失效。为了确保带宽有效性，必须通过 UDP 网络协议来完成。

   消息格式：确保平台相关的字段（如 int）需经过封装为平台无关的消息格式，例如 Google Protocol Buffers，不过在虚拟机平台没有安装 Google Protocol Buffers 包的情况下，建议使用自定义的消息格式，并要求在实验报告中给出该消息格式说明。

2. 每个存储虚拟机节点生成组服务的日志文件，利用作业 1 的分布式查询技术进行调试 debug，
   a) 每次在组成员列表发生变化（包括 join、leave、failure）的时候，生成对应的日志信息；
   b) 每次检测到节点故障或者节点之间的通信，记录对应的日志信息；

   然后利用作业 1 的分布式查询功能，给定输入关键字输出包含该关键字的日志记录。因此，本次作业要求尽可能将作业 1 的代码进行集成。

3. 评分准则（共计 25 分）：
   a) 组服务 Demo 演示：支持节点 join、leave、failure-rejoin（9 分）；动态显式组成员列表（5 分）；组服务容错机制，即：消息丢包比率为 3%、10%、30%（5 分）；利用作业 2 查询日志关键字（5 分）
   b) 作业报告：2 页，作业报告包括：组成员算法设计、如何支持可扩展性、低延迟、消息格式、作业 2 如何与本次作业集成、网络带宽开销（节点 join、leave、failure-rejoin）、消息丢包时组成员列表的错误比例。（6 分）
   c) [**加分项**]上述内容假设最多仅有一个虚拟机同时发生，先若允许多余 1 个、少于 N/2 虚拟机同时发生故障，则要求通过加锁机制控制共享变量(组成员结构)，在此功能基础之上，统计在变化虚拟机同时发生故障的数量时，对比组服务的性能指标（如：所有组成员列表收敛一致所需的时间）（5 分）