|  |
| --- |
| **YARNsim：模拟Hadoop YARN**    **YARNsim: Simulating Hadoop YARN** |

|  |  |
| --- | --- |
| 院（系）名 称 ： | 计算机学院 |
| 专 业 名 称 ： | 计算机技术 |
| 学 生 姓 名 ： | 柯晨曦 |
| 学号： | 2017282110254 |

摘要

Apache Hadoops系统非常流行是一个非常流行的系统，但是它在集中式作业/任务管理以及对除MapReduce以外的其他编程模型的支持等问题上仍有许多不足，下一代Hadoop（Apache Hadoop YARN）旨在解决这些问题。在本文中，我们提出了YARNsim，一个Hadoop YARN的仿真系统。 YARNsim基于并行离散事件模拟，提供模拟YARN关键组件的协议级精度。YARNsim提供了一个虚拟平台，系统架构师可以在这个平台上评估Hadoop YARN系统的设计和实现。 此外，应用程序开发人员可以调整工作绩效，了解不同配置之间的权衡，Hadoop YARN系统供应商可以在有限预算下评估系统效率。

为了验证YARNsim的有效性，我们用它来模拟两个真实系统，并比较YARNsim和实际系统的实验结果。 实验包括标准Hadoop基准测试，合成工作负载和生物信息学应用程序。 结果表明，大多数测试用例的错误率在10％以内。 实验证明，YARNsim能够为系统设计人员提供假设分析，与实际系统上的测试和评估相比，能够以最低的成本及时进行。

关键词：Hadoop;MapReduce; YARN; 并行离散事件仿真；HDFS

**目录**

[1 概述 1](#_Toc468699222)

[1.1 论文来源 1](#_Toc468699223)

[1.2 研究背景 1](#_Toc468699224)

[1.3 研究现状及相关工作 1](#_Toc468699225)

[2 YARNSIM的设计与实现 3](#_Toc468699229)

[2.1 Hadoop模块 3](#_Toc468699230)

[2.2 YARNsim的分析模型 3](#_Toc468699231)

3 [实验评估 5](#_Toc468699229)

[3.1 实验环境 5](#_Toc468699230)

[3.2 Hadoop I/0基准的测试 5](#_Toc468699231)

[3.3 Hadoop综合基准的测试 6](#_Toc468699230)

[3.4 生物信息学的测试 8](#_Toc468699231)

[4 总结和展望 9](#_Toc468699229)

[5 阅读体会 10](#_Toc468699232)

1、绪论

1.1论文来源

本文来自于ACM国际集群研讨会（IEEE/ACM International Symposium on Cluster）的论文‘YARNsim: Simulating Hadoop YARN’，这篇论文的作者是芝加哥伊利诺理工学院的Ning Liu等学者。

1.2研究背景

Apache Hadoop 是MapReduce 框架最流行的开源实现。 但是，第一代Hadoop基于集中式模型，JobTracker管理资源分配和作业处理。 这种设计方便了实现和管理，但是随着Hadoop进入数百万并发任务和数万个节点的系统的工作时代，其内在限制成为问题。 主JobTracker是一个单一的故障点，这意味着在JobTracker中的一个简单的故障将影响整个系统。这也是可扩展性的瓶颈，因为所有的工作和任务都是通过这一点来管理的。在HDFS 中，NameNode的功能类似于JobTracker。它管理所有的文件，关联的块和它们的位置。

为了最佳地量化这些设计限制并帮助设计和实现Hadoop系统，许多研究人员都使用了仿真器或仿真器等工具。 存在很多Hadoop性能模型和仿真系统。 在仿真技术方面，Mumak ，SimMR ，MRperf ，MRSG 和MRsim 是基于离散事件建模和仿真的，Starfish 基于分析建模和仿真。 在系统设计方面，Mumak和SLS 是以MapReduce作业调度和资源分配为核心的Hadoop调度模拟器， MRperf，MRSG和MRsim专注于模拟MapReduce作业/任务执行引擎。

YARNsim基于两个并行的仿真软件包：ROSS和CODES的协同设计。

1.3研究背景和相关工作

模拟已经为Hadoop系统的性能提供了有用的见解。最重要的是，仿真为设计和验证下一代系统提供了一个及时且具有成本效益的方法。但是，这些仿真系统是具有限制性的，原因如下：

（1）大多数离散事件仿真系统都是基于顺序事件处理引擎，限制了可能在合理时间内运行的仿真的保真度; （2）基于分析模型的仿真系统无法对复杂系统进行精确建模，而且由于必要的简化，往往牺牲精度。 （3）没有一个全面的仿真系统来模拟包含软件栈和硬件组件的Hadoop。例如，由于缺乏详细的网络模型，Mumak和SimMR不能捕获资源争夺。 MRperf基于网络仿真系统ns-2 ，但缺乏HDFS和本地文件系统的详细模型。事实上，大多数仿真系统都不包含HDFS的模型。

Hadoop YARN 最近被提议作为下一代Hadoop系统，解决第一代的问题并提供新的功能。 Hadoop YARN的核心思想是为管理资源和支持各种编程模型（包括MapReduce和MPI）提供一个通用的平台。 许多研究人员已经将注意力集中在Hadoop YARN的性能优化上。 然而，社区仍然缺乏一个全面的Hadoop YARN仿真系统，可以提供一个沙箱，用于探索设计方案和理解YARN行为。 在本文中，我们提出YARNsim：Hadoop YARN的模拟系统。 我们的贡献如下：

（1）据我们所知，YARNsim是第一个全面的Hadoop YARN仿真系统。

（2）YARNsim并行运行，因此能够对大规模场景进行建模和模拟。

（3）YARNsim包含详细的协议级精确HDFS模型，可提高仿真保真度。 大多数现有的MapReduce仿真系统都没有涉及到详细的HDFS模型。

（4）我们开发生物信息学应用模型，并通过比较YARNsim和真实系统的性能结果来验证其准确性。

本文的第二节介绍了YARNsim的设计与实现，第三节进行了实验评估，最后一节进行了总结和展望未来工作。

2、YARSIM的设计与实现

Apache Hadoop 是MapReduce 框架最流行的开源实现。 但是，第一代Hadoop基于集中式模型，JobTracker管理资源分配和作业处理。 这种设计方便了实现和管理，但是随着Hadoop进入数百万并发任务和数万个节点的系统的工作时代，其内在限制成为问题。 主JobTracker是一个单一的故障点，这意味着在JobTracker中的一个简单的故障将导致整个系统。 这也是可扩展性的瓶颈，因为所有的工作和任务都是通过这一点来管理的。 在HDFS 中，NameNode的功能类似于JobTracker。 它管理所有的文件，关联的块和它们的位置。

2.1 Hadoop模块

从守护进程的角度YARNsim中建模的所有hadoop模块如图1所示。

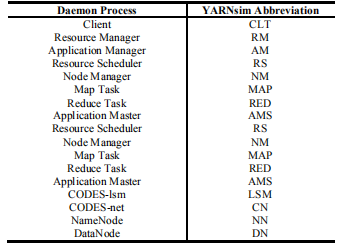


图1. 模型化的系统进程

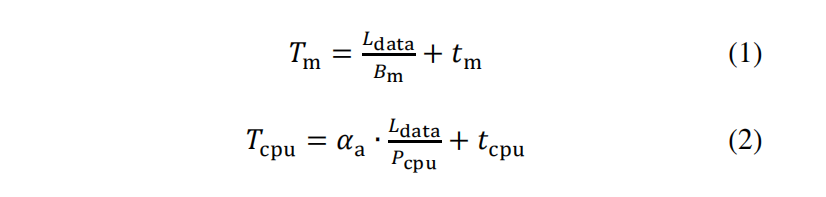
前一部分表示的进程模块，后一部分是进程模块名称的缩写，在本文的后面用的都是缩写。

2.2YRAMSIM的分析模型

在YARNsim中，我们将内部通信分类为元数据流和数据流。这两个流程都可以对整个系统性能产生重大影响。元数据流的模型遵循YARN中使用的基本协议。具体而言，YARN内部通信遵循客户端 - 服务器模型，并基于定制的RPC库。 Hadoop RPC服务器由一个监听器模块和一个callQueue模块组成。在不同守护进程之间传递的消息本质上是RPC调用。在YARNsim通信系统的设计中，我们遵循客户 - 服务器模型。具体而言，客户端和服务器被建模为两个不同的LP，并且在客户端和服务器之间传递的消息被建模为在LP之间发送和接收的事件。我们的模型不包括线程级细节，但是服务器模拟消息排队效应。我们后来表明，这个抽象层次在绩效评估中提供了合理的准确性。

硬件是影响Hadoop系统和应用程序性能的关键因素。 一般来说，在YARNsim中考虑的硬件模型是CPU，内存，网络和磁盘。 我们使用CPU和内存的分析模型，因为它们是在数据移动方面建模较不复杂的场景的理想选择。

等式1和2描述了YARNsim中使用的分析模型来计算存储器中的数据移动成本和CPU中的数据处理成本。



硬件相关参数可通过用户定义的输入配置文件进行配置。 初始化后，用户可以定制Hadoop硬件系统并获得一个精确的硬件平台。 YARNsim的软件栈模型如图1所示。主要组件是应用程序模块，MapReduce模块，YARN模块和HDFS模块。 椭圆是用来说明在YARNsim中建模的守护进程。 每个椭圆代表一种守护进程。 硬件相关参数可通过用户定义的输入配置文件进行配置。 初始化后，用户可以定制Hadoop硬件系统并获得一个精确的硬件平台。

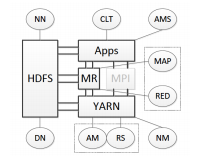


图2. YARNsim结构

YARNsim的软件栈模型如图1所示。主要组件是应用程序模块，MapReduce模块，YARN模块和HDFS模块。 椭圆是用来说明在YARNsim中建模的守护进程。 每个椭圆代表一种守护进程。

3、实验评估

3.1 实验环境

我们的实验是在伊利诺理工学院Scalabe计算机软件实验室的两个小型集群Craysun和HEC上进行的。 Craysun有16个节点，每个节点都配备了一个四核英特尔®至强®CPU，处理器频率为3.40 GHz。但是，每个节点目前只有1 GB的内存。对于这项研究，我们想要模拟在商品级设备的低端群集上构建MapReduce框架的情况。HEC是基于Sun Fire Linux的51节点集群，具有一个头节点和50个计算节点。头节点是Sun Fire X4240，配备双2.7 GHz Opteron四核处理器，8GB内存和配置为RAID5磁盘阵列的12500 GB 7200RPM SATA-II驱动器。计算节点为Sun Fire X2200服务器，每个节点配备双2.3GHz Opteron四核处理器，8GB内存和250GB 7200RPM SATA硬盘。所有51个节点通过千兆以太网连接。就本文而言，最新的Hadoop YARN是版本2.5.0，这个版本是我们用于所有实验的版本。

3.2 Hadoop I/O基准实验

我们选择了Teragen作为I / O基准测试，专注于HDFS模型验证。 此外，生成的数据被用作实际系统中以下基准测试的输入。 在Craysun中，我们共使用了16个节点，并将生成的数据从128 MB变为16 GB。 块大小被配置为32 MB，映射任务的数量被配置为匹配块的数量，减少任务的数量为默认值1.在YARNsim中，我们也配置了上述系统。 为了比较Craysun和YARNsim的结果，我们在YARNsim中为Teragen建立了一个模型来模拟数据生成过程。

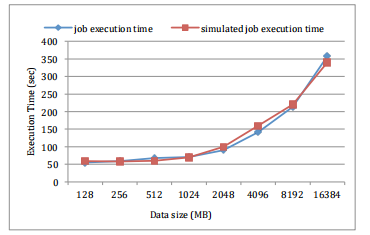


图3. HEC和YARNsim在Teragen基准测试中的性能比较

比较结果如图3所示。在所有的测试案例中，仿真结果和实际系统结果之间的误差为10％。 由于数据是平衡到每个计算节点，我们可以在测试中获得一个平衡的MapReduce任务执行。

3.3 Hadoop 综合基准测试

我们选择了Terasort和Wordcount作为我们的实验，因为这两个基准都被广泛接受，可以代表一类Hadoop应用程序。为了进一步分析应用程序性能，假设映射阶段和缩小阶段包含合并排序操作，我们将每个作业分解为三个阶段 - map, shuffle, 和 reduce。在Craysun和HEC中，我们使用Teragen基准测试产生的16个节点上的数据，并将输入数据大小从128 MB改为16 GB。为了准确地记录每个阶段在实际系统中的性能，我们利用Hadoop提供的工作历史服务，报告每个阶段的详细性能。我们收集这些数字，并将其与从YARNsim系统收集的数字进行比较。在YARNsim中，我们使用与Craysun和HEC相同的配置来配置仿真的集群。我们还为Terasort和Wordcount基准构建了模型，并在两个不同的模拟群集上分别运行它们。

我们比较了每个阶段的模拟工作和实际工作之间的表现，实验结果如图4-7所示。

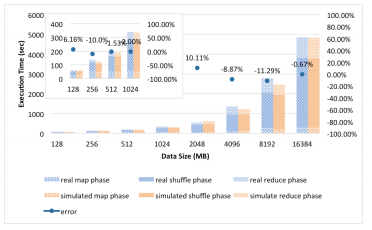


图4. Craysun和YARNsim在Terasort基准测试中的性能比较

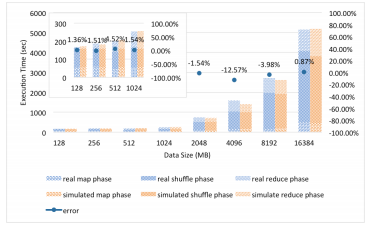


图5. Craysun与YARNsim在Wordcount基准测试中的性能比较

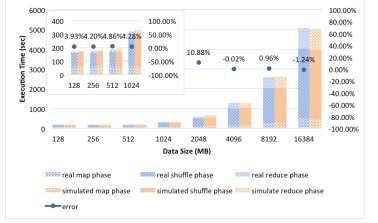


图6. HEC和YARNsim在Terasort基准测试中的性能比较

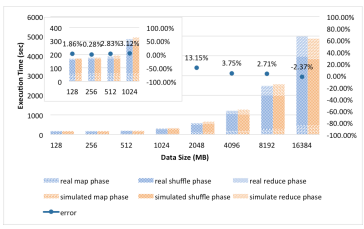


图7. HEC和YARNsim在Wordcount基准测试中的性能比较

我们在图4的子图中的小数据集上报告了Hadoop YARN的性能。这些测试的目的是验证Hadoop系统开销。 正如我们所看到的，YARNsim在轻量级实验中精确捕获这些开销。 在图5中，我们使用Wordcount基准，最大错误率是12.57％。 我们注意到，对于一些测试用例，reduce阶段会产生相对较大的错误。然而，在这组实验中，reduce阶段时间仅占整个执行时间的一小部分。 因此大多数测试用例的整体错误率仍然可以接受。

3.4生物信息学应用

在生物信息学领域，大数据集群是一个具有挑战性的问题。许多生物科学家诉诸于Hadoop MapReduce的大规模和并行处理解决方案。例如，特拉华大学的研究人员开发了一种基于八叉树的聚类算法来分类蛋白质 - 配体结合几何形状。该方法在Hadoop MapReduce中实现，分为两阶段MapReduce工作流程。几何缩减和密钥生成构成了第一阶段MapReduce作业，其中大型数据集由map任务读取。第一阶段的输出是第二阶段MapReduce作业的输入。在这里，基于八叉树的迭代聚类算法被实现为一系列MapReduce作业，指示搜索已经迭代到搜索树的深层。在第一阶段，输出数据大小约为输入数据大小的1％。因此MapReduce作业大部分时间都花在了map和shuffle阶段。

我们使用这个应用程序作为YARNsim的示例。为了建模应用程序，我们确定了每个阶段中所有数据块的大小和位置，并将它们用作建模的MapReduce作业的输入。我们将蛋白质几何数据的输入文件从948MB变为60GB，并使用16个节点在HEC上进行实验。我们还为这个集群应用程序构建了一个模型，并使用不同的配置在YARNsim上运行它。 HEC和YARNsim的性能如图8所示。

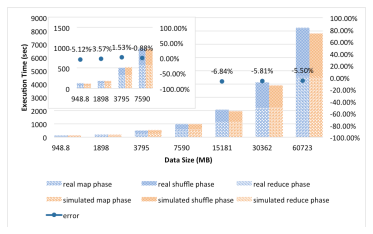


图8. HEC与YARNsim在生物信息学应用测试中的性能比较

如图8所示，基于YARNsim的模型可以帮助捕捉应用程序的性能，从而为进一步优化应用程序及其环境提供潜力。本文使用的聚类应用具有鲜明的特点。例如，由于reduce阶段的数据大小只是map阶段输入数据大小的1％，因此其map阶段比reduce阶段长得多。 YARNsim忠实地捕捉这个特征，并且可以预测不同输入数据大小下的应用性能。最大误差是6.84％。

4、总结和展望

在本文中，我们展示了Hadoop下一代仿真系统YARNsim。 YARNsim由四个核心模块组成：一个应用程序模块，一个YARN模块，一个MapReduce模块和一个HDFS模块。 YARNsim还包含一组硬件模型。 不同的模块在功能上是独立的，但在操作上连接; 它们一起捕获YARN中不同的软件和硬件层，并能以令人满意的粒度重放YARN系统行为。 YARNsim为用户提供了一个全面的仿真平台，系统架构师可以评估各种设计点，应用程序开发人员可以测试/调整应用程序的性能。

我们通过一系列全面的Hadoop基准测试来验证YARNsim的性能，包括Terasort，Teragen和Wordcount在仿真系统和真实集群上。 我们通过生物信息学应用程序进一步验证YARNsim。 实验结果显示大多数测试用例的错误率不到10％。

YARNsim不是没有限制。 YARNsim的当前版本不支持容错模型。 因此，用户将无法模拟失败时的作业执行。 另外，HDFS模块还不能建模辅助NameNode。 我们计划采用一种机制来支持系统组件故障模型和仿真。 YARNsim也只包含FIFO调度算法。 我们计划在不久的将来开发容量调度程序，公平调度程序和其他高级调度算法，以支持复杂作业和任务执行的模拟。

5、阅读体会

本文的主要问题是第一代Hadoop有一些局限性，单点故障、集中式作业/任务管理以及对除MapReduce以外的其他编程模型的支持等问题成了限制它发展的因素，Hadoop YARN 最近被提议作为下一代Hadoop系统，解决第一代的问题并提供新的功能。 而为了最佳地量化这些设计限制并帮助设计和实现Hadoop系统，许多研究人员都使用了仿真器或仿真器等工具。但是现在的仿真工具都有局限性，所有作者设计和实现了一个新的仿真工具YARNsim。然后作者做了大量的实验，包括Hadoop的I/O基准测试，Hadoop的综合基准测试和生物信息学应用测试。通过比较仿真工具和实际环境的实验结果，作者得出YARNsim的误差基本都没有超过10%。也就是说这是这个仿真工具的设计是较合理的，但是也还有一些问题没有解决，例如当前版本不支持容错等。