**大数据系统应用可靠性基准测试**

当前Hadoop/Spark等分布式处理框架被广泛应用于大数据的处理分析，如Web索引建立、日志挖掘、机器学习、大规模图分析等。然而应用在处理大数据时，经常出现内存溢出、IO异常、任务超时等运行时错误，这些错误会直接造成应用执行失败。已有的研究工作表明这些错误的原因多种多样，既包括系统缺陷（如设计缺陷和实现bugs），也包括数据异常和应用缺陷（如参数配置不当、代码缺陷等）。由于运行时错误的多样性和复杂性，当前还没有一个通用的静态错误检测方法能够提前发现系统、应用和数据的潜在问题。因此，我们的总体目标是去构造一个可靠性基准测试框架，通过构造典型应用、生成异常数据、异常配置和高负载来对大数据系统应用进行压力测试，从而提前暴露系统和应用存在的问题和缺陷。主要创新点包括：(1) 提出针对大数据应用的异常数据生成方法。该方法通过分析应用程序特征，来生成相应的异常数据（包括异常分布、高维度等特征）。(2) 提出一种针对大数据应用的异常参数生成方法。该方法通过组合系统与应用参数，并通过分析参数相关性来削减组合测试空间。(3) 提出针对流式大数据应用的高负载生成方法。该方法通过构建多个负载发生客户端，并通过并行划分数据流量、流速的方法，来生成高负载的应用数据。最后，框架将实现这些方法并完成自动化测试。该项目的研究成果将对开源社区、学术界和工业界产生重要影响。可靠性基准测试框架可以用于开源社区更新Hadoop/Spark版本时的可靠性测试，提前发现系统缺陷，修复bugs。同时，也可以用于学术界、工业界开发新应用时的可靠性测试，提前发现应用潜在的问题，节省成本，避免损失。另外，在前期研究中，我们已经分析了大数据处理框架及应用的内存溢出错误原因、诊断方法与修复方法，具有较好的研究基础。

**流式图计算引擎**

由于图结构能够很好表达数据之间的复杂依赖关系和聚集情况，因此图计算引擎被广泛应用于大规模图数据挖掘，如社交网络意见领袖挖掘、商品购买推荐、舆论检测、信用卡欺诈检测等。然而，当前的图计算引擎如Google Pregel, Spark GraphX，GraphLab等只支持静态图计算，而现实世界中的图数据又是动态变化、实时更新的，因此我们的研究目标是构造一个可以在动态图上进行实时计算的流式图计算引擎。该引擎的主要构成和主要创新点包括：(1) 基于状态半异步更新的动态图计算模型。该模型将动态图计算表达为增量计算（状态＋更新），并通过异步状态更新和全局同步相结合的半异步方式来解决状态更新和更新冲突的问题。(2) 状态更新消息传播控制方法。该方法针对不同应用的增量计算特点，通过控制状态更新消息传播范围（如邻接点、连同子图、全图）来减少计算量。(3) 状态更新与计算并行化方法。该方法改进现有的基于顶点划分和基于边划分方法，考虑新增顶点和边的分布，对图进行动态划分，减少计算代价和通信代价。由于学术界目前还没有对动态图进行分布计算的通用方法，工业届也没有成熟的动态图计算系统，因此本研究成果将对学术界和工业界产生重要影响。我们前期已经分析了Pregel，Spark GraphX等系统设计和实现，具有良好的研究基础。