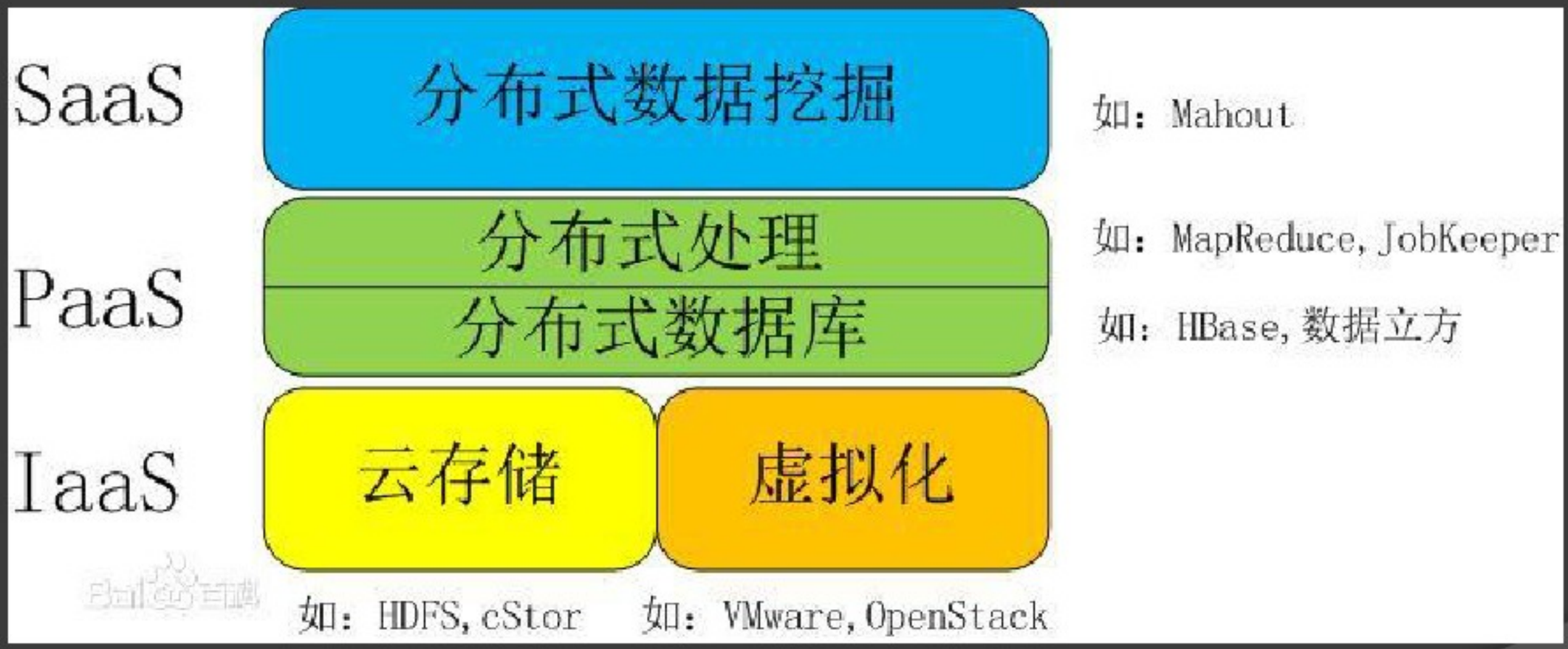
STORM技术分享

大数据

- 巨量资料，指所涉及资料量规模巨大（Volume、Variety）无法透过目前主流软件工具合理时间(Velocity)内达撷取、管理、处理、并整理成帮助企业经营决策（Variety）更积极目的的资讯

大数据与云计算


SaaS
分布式数据挖掘
如：Mahout
PaaS
分布式处理
如：MapReduce, JobKeeper
分布式数据库
如：HBase, 数据立方
IaaS
云存储
虚拟化
如：HDFS, cStor
如：VMware, OpenStack

分布式实时流式处理系统

流式处理系统（规范且稳定的结构）：
http://video.sina.com.cn/v/b/59974476-1213608837.html

分布式系统（网格，云计算，快速的部署能力和容灾性，方便的扩展）

实时系统（快速的处理能力）

典型的场景

- 日志统计系统：

传统的解决方案：

queue+worker实时系统： 云统计

问题：

1、部署维护消息队列

2、自动容错机制，进程、机器挂掉自动处理

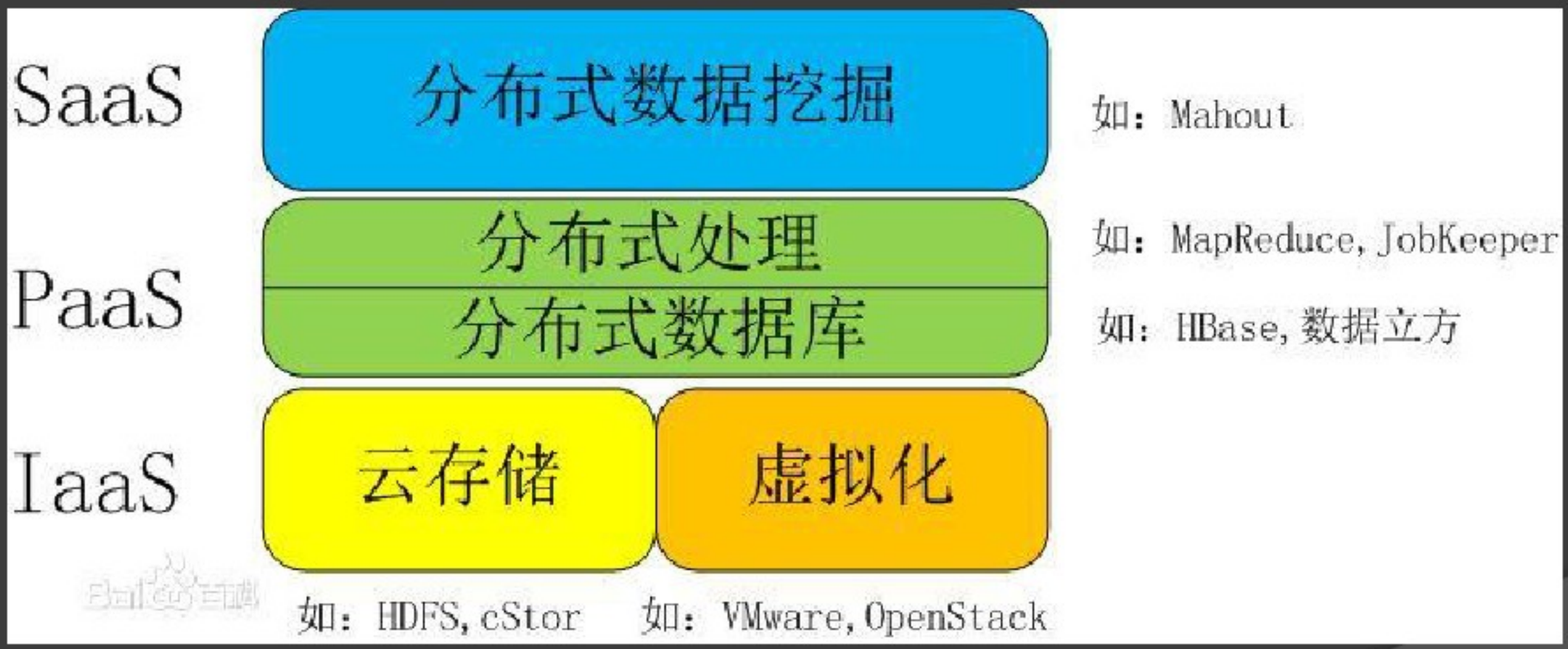
3、功能扩展性

Storm是什么

分布式实时流式处理系统

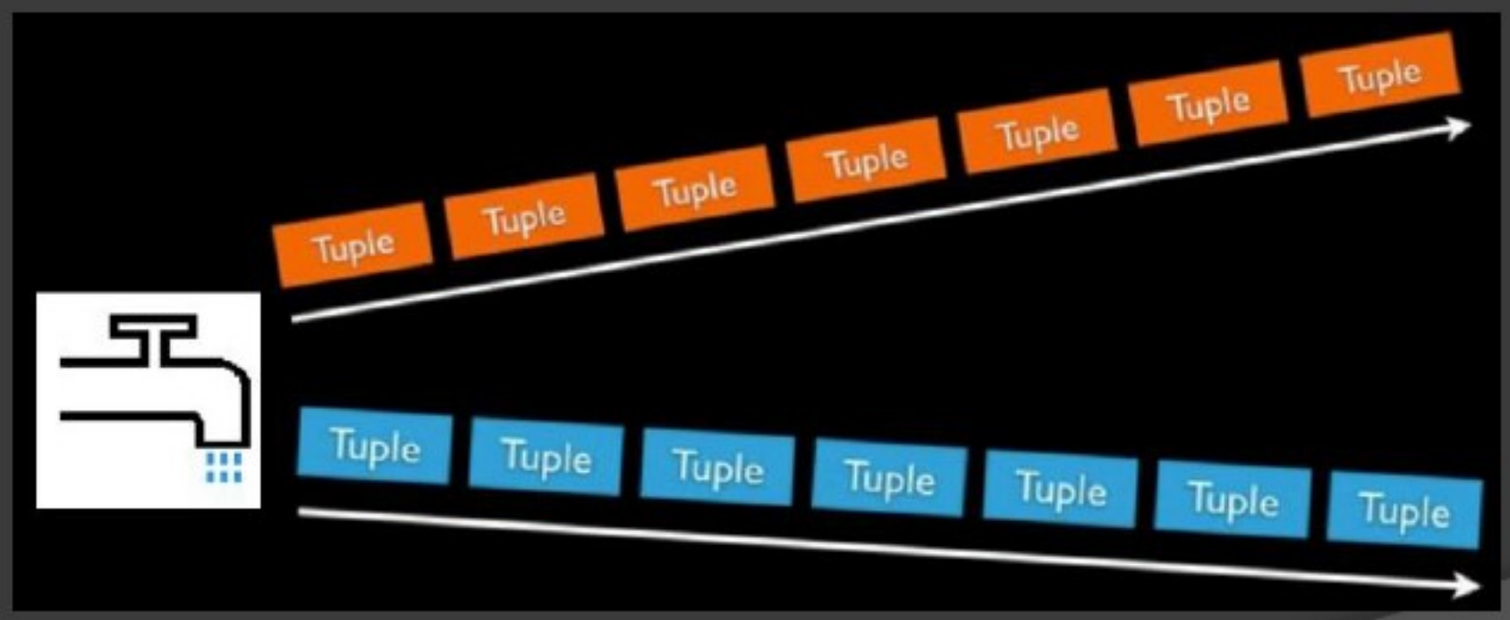
- **流式**：提供了一个简单的计算模型和API（类似MapReduce），可以方便的处理不断产生的数据
- **实时**：系统本身的效率很高，处理延迟在毫秒级
- **水平扩展**：通过简单加机器、提高并发数就可以提高整体处理能力
- **自动容错**：自动处理进程、机器挂掉的异常

Storm的位置


SaaS
分布式数据挖掘
如：Mahout
PaaS
分布式处理
如：MapReduce, JobKeeper
分布式数据库
如：HBase, 数据立方
IaaS
云存储
虚拟化
如：HDFS, cStor
如：VMware, OpenStack

几个重要概念：数据输入

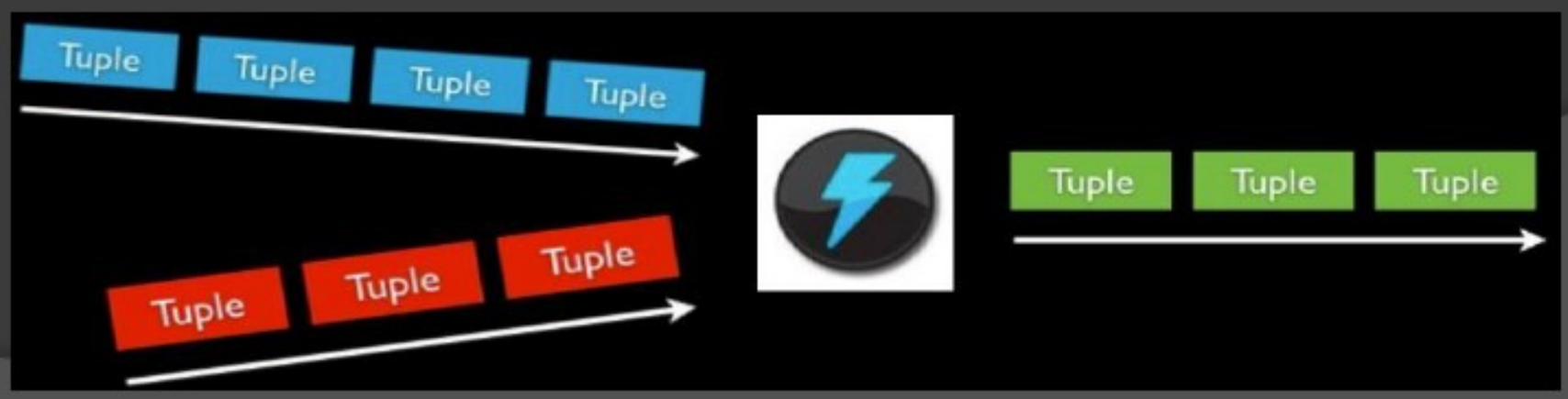
- Tuple：被处理的数据
- Stream:一群消息的集合
- Spout：产生数据源的组件



几个重要概念：数据处理

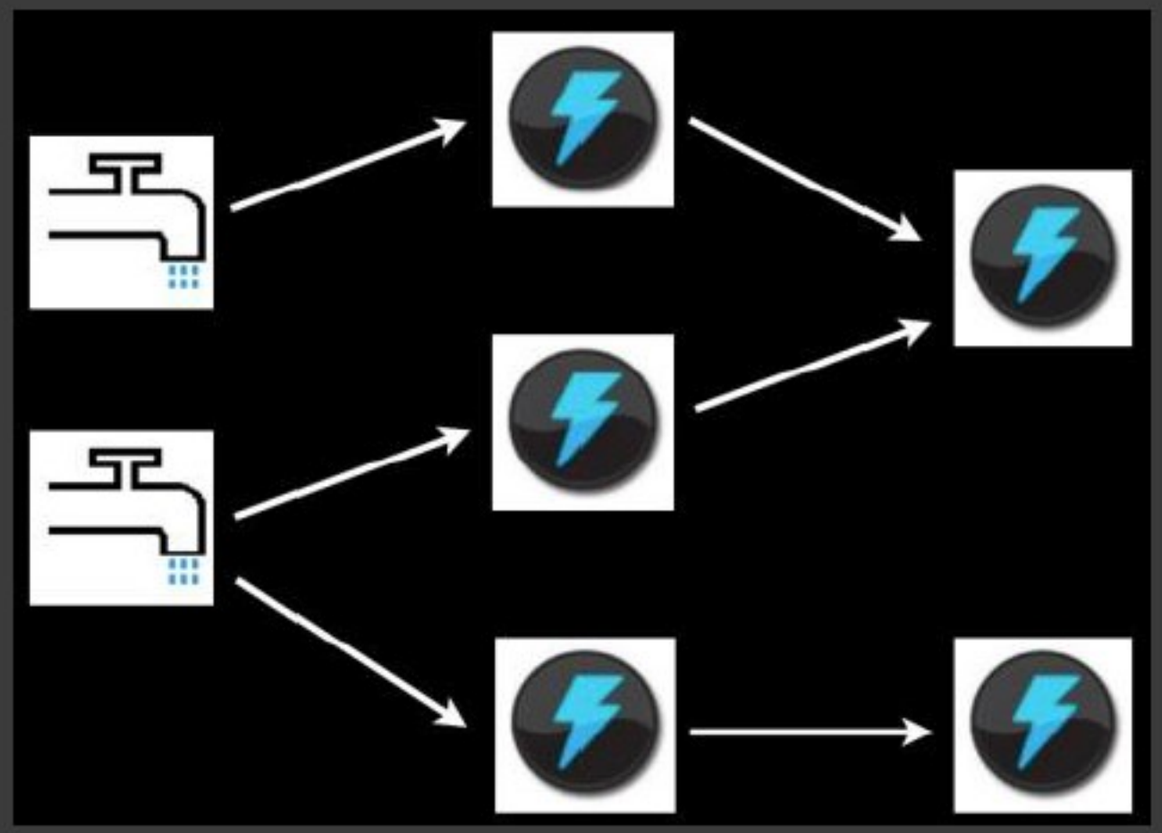
- Bolt:接受数据后处理数据的组件
- Worker：运行处理组件逻辑的进程
- Task：Work中每一个spout/bolt的线程

bolt的角色是处理数据，输入是上游（spout或bolt）的tuple，输出是发往下游（bolt）的tuple；bolt可以有多级，一般最后一级bolt会定期把结果写到外部存储



几个重要概念：组合

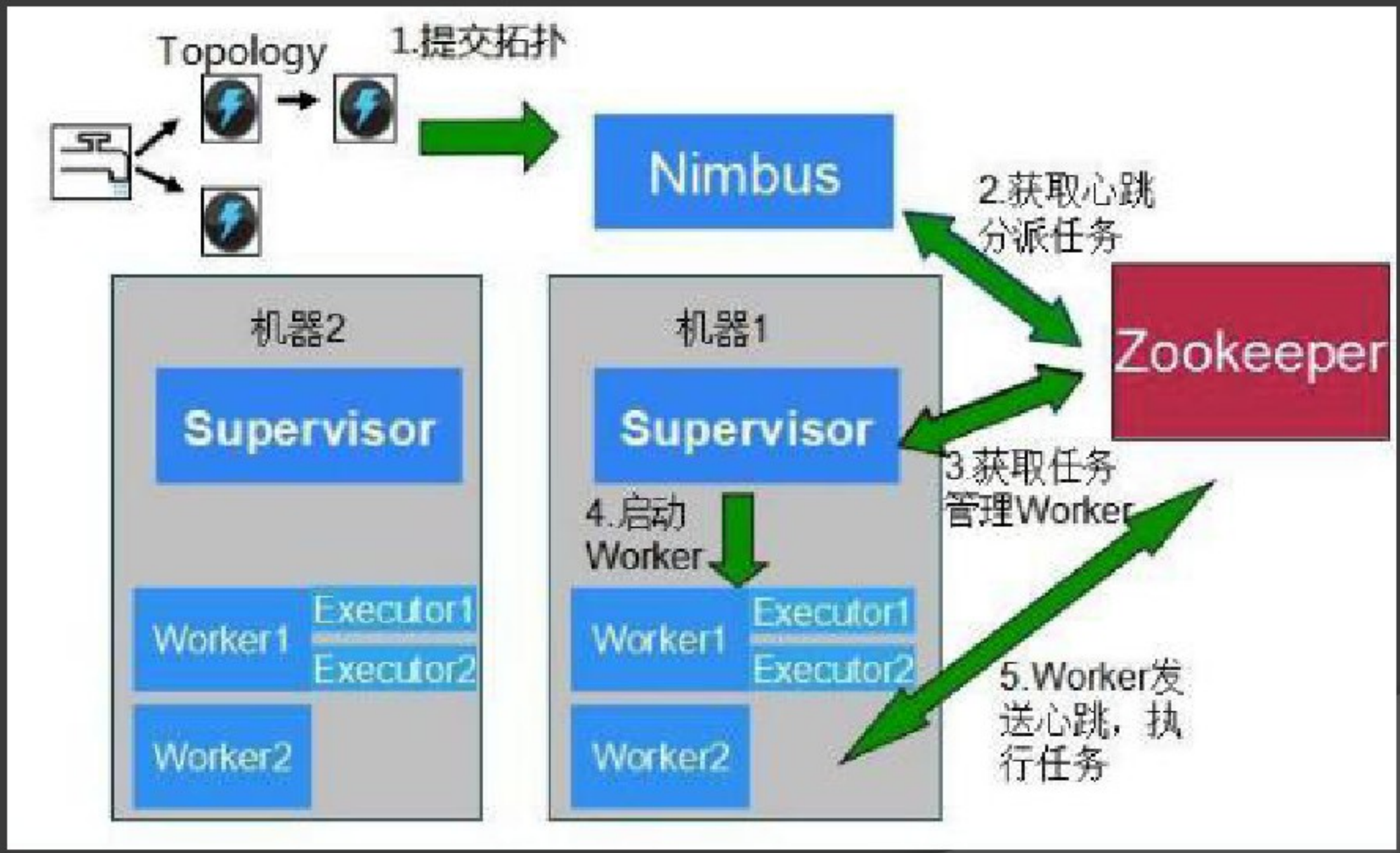
- Topology：对一个应用的spout、bolt类型、输入输出tuple/stream、关联关系的描述



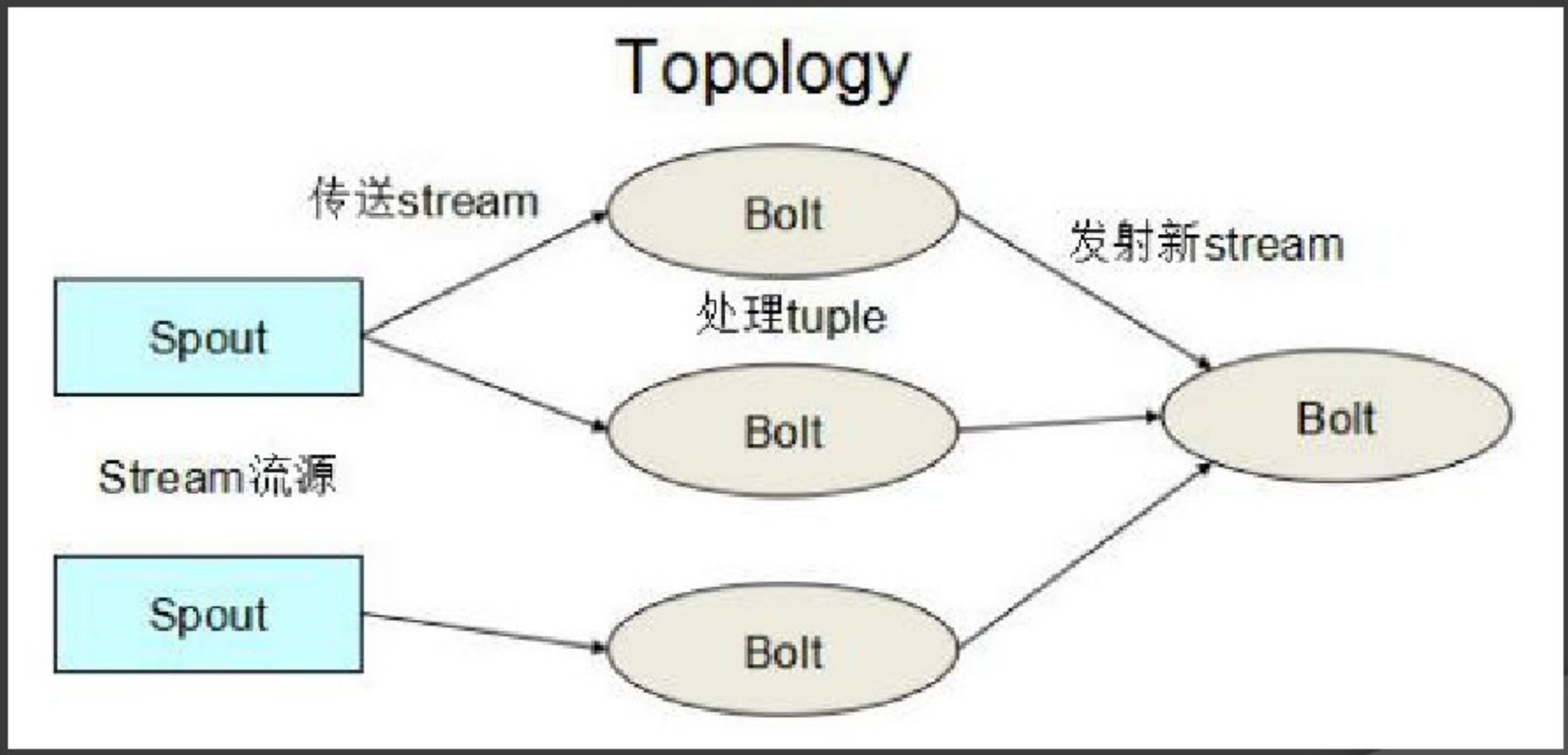
几个重要概念：集群控制

- Nimbus：主控节点。负责分配资源，任务调度，监控集群等
- Supervisor：接收任务，启动进程
- Zookeeper：协调Nimbus和Supervisor之间的工作，存放公共数据

几个重要概念：集群控制


Topology
1.提交拓扑
Nimbus
2.获取心跳
分派任务
Zookeeper
机器2
Supervisor
Worker1
Executor1
Executor2
Worker2
机器1
Supervisor
Worker1
Executor1
Executor2
Worker2
3.获取任务
管理Worker
4.启动
Worker
5.Worker发
送心跳，执
行任务

Storm处理流程


Topology
传送stream
Bolt
发射新stream
处理tuple
Spout
Bolt
Bolt
Stream流源
Spout
Bolt

演示：

- 开发机：220.181.150.112
- 展示页面：

http://w-m1.dfst.shgt.qihoo.net:8360/