集群通信系统概念设计

# 名词表

节点（Node）：对应到一台物理机器或虚拟机

集群（Cluster）：由节点组成的一个逻辑组

超级节点（Super Node）：节点的一种身份，为节点提供服务

普通节点(Ordinary Node)：相对于超级节点的节点

物理子网（Physical Sub-Net）：同一网段的物理节点集合

逻辑子网（Logical Sub-Net）：同一网段中属于一个Cluster的节点集合

超级节点组（SN Group）：由超级节点构成的一个逻辑组

# 工作原理

目标：以数据中心网络（Intranet）内的一组物理或虚拟主机构建一个逻辑上的集群，每台主机具有自己的逻辑名称，集群可作为一个管理单元，具有结构自组织，透明寻址，自动路由，高容错等特性。

接入集群通信系统的所有节点，在物理上是全联通的。集群通信系统使用**Cluster**来组织管理所有节点之间的通信。每个节点，至少隶属于一个Cluster。每个Cluster均有其特定不重复的标识符（Cluster Identifier）。

在少量节点的情况下，集群通信系统采用单Cluster模式；在跨数据中心，或者大量节点的环境下，集群通信系统采用层级式通讯管理，即多Cluster模式。（**本阶段暂未涉及**）



网络拓扑图

在一个Cluster中的节点，来自于多个物理子网（PSN）。节点分为两类，分别是“超级节点（SN）”和“普通节点（ON）”。普通节点是Cluster的基本元素和最小单位。在一个物理子网中属于同一个Cluster的节点，构成一个逻辑子网（LSN），在这个LSN的节点中通过一定的选举算法，产生一个SN。属于同一个Cluster的多个LSN选举出的SN构成一个超级节点组。同属一个超级节点组的SN互相交换彼此的所在本地LSN中的普通节点地址信息映射表，从而得知本Cluster的所有ON和SN的地址信息映射表，并将本Cluster的所有SN信息通知到各自LSN中的ON。ON可以藉由本地缓存的地址映射表和Cluster中的所有SN的地址信息表进行地址转换。

普通节点可以通过自身缓存的本地地址信息映射表直接访问逻辑子网中的其他节点，如果访问本逻辑子网外的节点地址信息，则需要通过访问超级节点获得。

超级节点持有所属Cluster的所有普通节点和超级节点的地址信息映射表。

超级节点组的功能：维护地址表的一致性，数据的完整性，通过多节点提供高可用性。

# 目标系统产生

整个系统的产生分为两个阶段：

第一阶段，在一个物理子网（PSN）中属于同一个Cluster的节点，构成一个逻辑子网（LSN），在这个逻辑子网（LSN）的节点中通过一定的选举算法，产生一个超级节点（SN），SN作为本LSN中的其他普通节点的Leader和Cluster中其他LSN中的SN通信；普通节点（ON）可以通过SN获得其他节点的最新地址信息。

第二阶段，属于同一个Cluster的多个LSN选举出的SN构成一个超级节点组，交换彼此的节点地址信息映射表。ON通过所属的SN可以获得其他LSN中属于同一个Cluster的节点的地址信息，至此整个网络自动构建完毕。

下面分别从单个节点和整个网络的层面分别描述系统的产生。

## 节点状态图



节点状态图

一个节点启动后，处于初始化状态（Initial），在本地物理子网（PSN）中进行广播，寻找本地PSN中的其他节点一起组成Cluster的本地逻辑子网（LSN），此过程中可以通过Neighbors Detecting机制交换彼此地址信息，从而构成本地LSN地址信息缓存表；如果本地LSN中没有超级节点（SN），则发起选举，直至产生SN；如果已有SN，则成为此SN管理下的普通节点（ON）。如果本地LSN中的SN发生异常引发故障侦测，本地LSN中的其他ON就会发起一轮新选举，提升一个ON作为新的SN，该SN继续加入Cluster的超级节点组同步其他LSN的信息。原SN如果重新启动，则作为ON加入。ON失败后也可重新加入Cluster。

**关于地址转换的过程**：

Cluster中的任意一个节点如果需要和其他节点通信，首先要进行地址转换，步骤如下：

1. 查询本地缓存表是否存在目标主机，不存在转3
2. 建立连接并验证cookie，通过结束，不通过继续
3. 查询SN节点，获得目标主机
4. 建立连接并验证cookie，并更新本地缓存表

由于节点加入Cluster时通过Neighbors Detecting已经可以得知所处LSN中的所有地址，加上本地LSN中的SN会广播其他LSN中节点的变动信息，可以大致认为寻址时间为O（1）

## 网络状态图



网络状态图

对于整个消息通信系统而言，在初始状态，没有Cluster，没有SN，所有节点都处于寻找其他节点的状态（Lookup）。

一个物理子网（PSN）中属于同一个Cluster的节点，构成一个逻辑子网（LSN），并选举出SN，此时已经处于Structured状态（可用），虽然只有一个LSN。

1. 当两个处于不同LSN的SN彼此发现对方，则需要构成SN Group（Structured），系统进入Recovery状态
2. 当新的SN加入已有的SN Group时，系统进入Recovery状态
3. 当SN Group中的某个SN由于系统崩溃等原因而不存在时（SN退出Group），系统进入Recovery状态

当系统进入Recovery状态时需要重新构建SN Group，多个SN重新交换彼此的地址信息映射表，此时暂不提供SN寻址服务，当信息交换完成，系统恢复提供SN寻址服务（Structured）。

当某个普通节点（ON）加入/离开时，不影响其他节点，整个网络的继续可用。

# 系统故障处理

## 普通节点故障

普通节点（ON）会定期的发起对所处逻辑子网（LSN）中的超级节点（SN）的心跳侦测，向SN告知自己的状态，SN通过为每一个ON设置一个Timeout来检查ON的存活，当超出Timeout时间后，SN会认为该ON已经故障，通知整个SN Group移除该ON的地址信息，并由相应的SN通知自己的LSN。

## 超级节点故障

当超级节点组（SN Group）中的某个超级节点由于网络故障、系统崩溃等原因而不存在时，整个集群通信系统进入Recovery状态，重组SN Group，此时暂不提SN的寻址服务，当重组完成，系统恢复提供SN的寻址服务（Structured）。

同时失败的SN所处的LSN中的某个ON通过心跳侦测发现该SN已不可用，该ON发起一次选举，按照选举规则重新选举出本地LSN中的SN，然后SN加入到SN Group中。需要注意的是，Cluster的SN寻址能力只在SN加入Group时受到影响（Recovery状态），也就是说即使本地LSN中不存在SN，由于其他的SN可用，在选举SN的同时（Structured），ON也可以通过SN Group获得最新的地址信息（可用性）。

## 灾难性故障

灾难性故障指的是所有的SN，即SN Group瞬间Crash，此时Cluster的可用性受到极大的影响，需要进行重新组网（选举SN，建组）。这个过程中ON可以继续使用本地的缓存地址信息，但无法透过SN获得最新的地址信息。

## 网络割裂的影响

由于系统拓扑结构的设计，假设割裂的发生是以物理子网（PSN）为单位，那么系统的自组网特性可以很好的支持多个割裂系统的正常运行，但割裂合并后的冲突如何解决，还是以后需要研究的问题。