(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107040407 B
(45)授权公告日 2020.02.18

(21)申请号 201710154402.0

(22)申请日 2017.03.15

(65)同一申请的已公布的文献号
申请公布号 CN 107040407 A

(43)申请公布日 2017.08.11

(73)专利权人 成都中讯创新科技股份有限公司
地址 610041 四川省成都市高新区天仁路387号2幢13层1306号

(72)发明人 谢滔

(74)专利代理机构 成都华风专利事务所(普通合伙) 51223
代理人 徐丰

(51)Int.Cl.
H04L 12/24(2006.01)

(56)对比文件
CN 102629941 A,2012.08.08,
CN 102495759 A,2012.06.13,
CN 104125165 A,2014.10.29,
CN 102929720 A,2013.02.13,
US 2016283335 A1,2016.09.29,

审查员 于晓溪

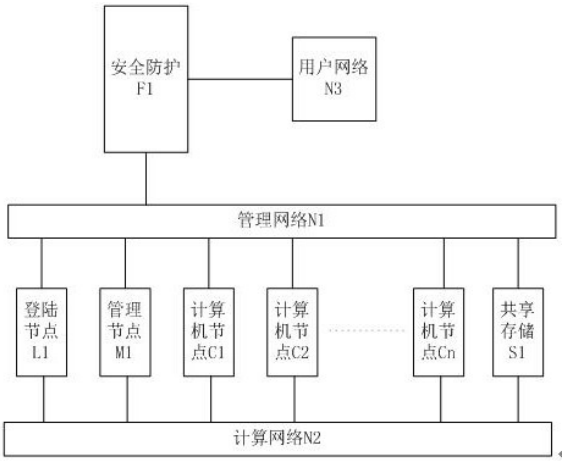
权利要求书3页 说明书7页 附图2页

(54)发明名称

一种高性能计算集群动态节点作业方法

(57)摘要

本发明提供了一种高性能计算集群动态节点作业方法,其基于Infiniband网络提供统一的文件系统空间和无盘启动,基于以太网提供集群作业调度、开关机控制、系统监控等功能。通过上述方式,本发明能够提高高性能计算集群部署效率,降低故障率,简化作业提交,提高能源和资金利用率,提升使用效率和性能。



CN 107040407 B

1.一种高性能计算集群动态节点作业方法,其特征在于,包括如下步骤:

步骤1:配置服务器,包括管理网络N1、计算网络N2和用户网络N3,以及用户网络N3和管理网络N1之间的安全防护网络F1、登陆节点L1、管理节点M1、若干计算节点CN;所述登陆节点、管理节点、计算节点分别与管理网络和计算网络建立通信连接;所述登陆节点和管理节点通过计算网络挂载共享存储S1;

步骤2:在管理节点M1上安装并配置启动管理服务(Flexboot)、动态主机配置协议服务(DHCP)、文件传输协议服务(TFTP)、域名系统服务(DNS)、共享服务、认证服务以及作业调度,配置完成后启动计算节点;

步骤3:用户网络N3中的用户,经过安全防护网络F1以安全外壳协议SSH登录到登陆节点L1上,通过SSH将需要计算的算例上传至登陆节点L1上的共享存储S1;

步骤4:用户在登陆节点L1上通过作业调度中间件,指定作业参数完成作业脚本的生成并启动提交给M1的作业调度进行资源分配和计算;

步骤5:管理节点M1收到登陆节点L1上用户提交的作业脚本后,首先将作业脚本中的函数调用请求交给管理节点中的Maui.d进行资源配额检查,如果配额不足则报错提示,如果配额足够,则将作业脚本转交给M1上作业调度pbs_server进行调度;

步骤6:管理节点M1的作业调度pbs_server收到作业脚本后,根据作业脚本要求的节点数、每节点核心数为作业分配进行计算的节点C1到Cn,如果节点数量不足,则将作业进行排队等待;如果节点数量足够,则根据作业脚本具体执行内容开始计算;

步骤7:当C1到Cn完成计算后,释放物理内存、处理器资源,并向M1的pbs_server反馈“所有核心均未占用,状态Free”的信号;M1的pbs_server对该信号进行记录,并允许后续作业在该节点上进行分配和计算;

步骤8:用户可以通过安全文件传输协议SFTP从L1上将S1中存储的计算结果信息取回到N3中,在本地电脑上打开进行处理和分析;

所述方法还包括计算机开机节点自动控制步骤:

步骤11:系统启动完成正常运行时,M1每60秒启动后台监控进程一次并检测启动时间,若未达到启动时间,则暂停60秒并重复步骤1;

步骤12:后台监控进程正常启动后,检查pbs_server是否有排队作业,若无排队作业并且处于工作状态的计算节点小于等于1时,则直接退出;若无排队作业并处于工作状态的节点大于1个,则通过管理网络N1和IPMI口关闭开机但位处于空闲状态的空闲节点数量-1个节点,只保留空闲节点中节点名排名最前的1个空闲计算节点处于开机状态以备用;

步骤13:若检查到有排队作业情况时,则分析处于排队作业的原因:若为用户超额,则直接退出系统;若用户未超额,则排队原因为资源不足,执行下一步骤;

步骤14:在步骤3之后,检查关机节点数量:若关机节点为0,则提示节点用尽并退出系统;若关机节点大于等于1,则根据处于等待状态各中作业各节点需求量的大小从小到大进行排序,设定N为等待状态作业最小作业节点需求数量,F为当前空闲状态节点数量,G为关机节点数量,比较当前需要开机节点数量(N-F)和G-1,若N-F≤G-1,表示关机节点数量比需求节点数量大,则在管理网络N1中通过IPMI接口开启N-F+1个节点,并重新进行步骤2;若N-F≥G-1,表示关机节点数量不足,则在管理网络N1中通过IPMI接口开启所有关机节点并提示节点数用尽,然后退出系统。

2.根据权利要求1所述的高性能计算集群动态节点作业方法，其特征在于：管理网络N1为以太网络，主要负责计算节点开关机控制、作业调度数据传输、系统监控功能，计算网络N1为Infiniband网络，提高设备的扩展性、数据的传输速率及通信延迟，主要负责数据及存储的IO、操作系统镜像分发、计算软件工作时各进程相互通信与数据同步功能，用户网络N3是高性能计算机用户所在的网络，安全防护F1为防火墙、UTM或路由器设备，提供用户网络到管理网络的端口映射、访问权限管理、异常流量监测、攻击防护功能，登陆节点L1、管理节点M1、若干计算节点CN统一为同一处理器架构，处理器具有完全相同的指令集，根据用户实际使用需求和高性能计算机总体计算性能要求，登陆节点L1、管理节点M1、若干计算节点CN可以通过集群(Cluster)方式进行横向扩展，共享存储S1为基于Infiniband的NFS服务器，所述共享存储S1包含底层的硬盘柜或磁盘阵列，或基于Infiniband的分布式存储系统，对外提供一个统一的文件系统空间，并且支持用户权限控制和容量配额。

3.根据权利要求1所述的高性能计算集群动态节点作业方法，其特征在于：系统第一次部署时将登陆节点L1的操作系统复制为镜像文件J1，并修改J1中包含的相关个性化参数配置文件为通用配置文件，其中包括修改网卡配置文件，去掉MAC地址、UUID唯一信息，修改为DHCP引导；将主机名修改为DHCP自动获取、将硬盘的挂载方式修改为设备名方式进行挂载、将系统环境变量存放目录修改为共享存储S1上特定目录。

4.根据权利要求1所述的高性能计算集群动态节点作业方法，其特征在于：所述计算节点的启动采用基于Infiniband网络的无盘启动方式，若干计算节点的启动方式一致，启动计算机节点C1具体步骤包括：

步骤1：将C1开机，设置为默认PXE引导，PXE默认设备为主机通道适配器(HCA)，采用FlexBoot模式；FlexBoot初始化HCA卡，检测端口协议及状态，以Infiniband方式启动端口，并以广播的方式发送DHCP客户端(Client)请求报文；

步骤2：管理节点M1的DHCP服务器从Infiniband网络收到请求报文后，将C1的IP地址、TFTP服务器、网络引导启动镜像目录发送给C1；C1接收到M1发出的报文后，根据收到的报文启动Infiniband网络，并从M1的TFTP服务器中下载启动镜像目录并加载到内存中，所请求的镜像目录内包含默认的启动镜像名称、默认启动镜像时间信息，通过C1上选择启动J1或者超过默认时间后自动选择启动J1，并向M1发送请求J1的报文；

步骤3：M1收到C1请求J1的报文后，将J1通过TFTP服务器发送给C1；C1的FlexBoot接收完J1后，将J1放入内存进行加载；

步骤4：C1依次加载J1的内核、根文件系统、网络、配置文件；C1在加载网络时广播DHCPClient的请求报文，M1的DHCPServer收到请求报文后再次将C1的IP地址发送给C1，C1收到IP地址报文后启动网络，并向M1请求主机名；M1的DNSServer收到C1请求主机名的报文后将C1的主机名发送给C1；C1网络启动完成后，首先加载挂载共享目录的配置文件，根据配置文件内容将S1通过Infiniband进行挂载；C1挂载完共享目录后，启动计算节点作业调度(pbs_mom)，并将作业调度状态反馈给管理节点M1的作业调度(pbs_server)，启动完成。

5.根据权利要求1所述的高性能计算集群动态节点作业方法，其特征在于：所述作业调度中间件中提前录入已知的调用函数，针对已知的计算软件进行作业提交流程的重构和标准化，用户在提交作业时通过调用函数输入作业类型、参与计算节点数量、每个计算节点参与计算核心数量、输入文件共计4个参数即可完成作业提交，避免了编写作业脚本的工作，

极大简化上级操作步骤。

一种高性能计算集群动态节点作业方法

技术领域

[0001] 本发明涉及高性能计算技术领域，特别是涉及一种高性能计算集群动态节点作业方法。

背景技术

[0002] 高性能计算(High performance computing，缩写HPC) 指通常使用很多处理器(作为单个机器的一部分)或者某一集群中组织的几台计算机(作为单个计算资源操作)的计算系统和环境。高性能计算在航空航天、材料、数学、生物、物理、化学、气象、环境、金融、媒体、电磁等多个行业具有较为广泛和重要的作用。当前高性能计算75%以上的系统都是通过X86服务器以Clustre架构进行构建，随着服务器节点的增多和对集群效率、实测计算峰值、能耗等各方面要求的提升，传统的高性能计算集群存在诸多瓶颈，需要广大科研人员和集群维护人员进行研究和探索。

[0003] 通过基于Infiniband网络的无盘部署、标准化的作业提交流程和步骤、自动控制和调整计算节点开机数量可以实现：

[0004] A、节能，基于无盘部署，计算节点不需要配置硬盘，降低了集群的功率开销和故障点，通过动态调整计算节点开关机数量，避免了大量计算节点开机空转的情况，提高了能源的使用率；

[0005] B、性能，基于Infiniband网络的无盘部署，将系统镜像通过低延时的高速网络加载到各计算节点内存中，提升了计算节点开机速度，并且充分利用了RAM DISK的IOPS性能优势，极大提升计算任务在计算节点单机内部的收敛速率。

[0006] C、标准化，由于高性能计算涉及行业较多、范围较广、海量的专业软件，导致了传统用户在使用高性能计算集群时需要去针对具体的计算软件进行了解学习后才能上机使用。通过对作业流程的重构和标准化，将海量的专业软件的作业提交流程通过中间件固定为同样的步骤和流程，极大简化了上机操作步骤，让传统用户能快速的入手并将集群充分使用起来。

[0007] D、节约，最大程度减少不必要的软硬件投入(如计算节点硬盘、计算节点操作系统)，提升资金使用率；

[0008] E、低故障率，传统高性能计算集群在每个计算节点上需要安装1块硬盘用于存放操作系统。机械硬盘价格便宜，使用年限久，但性能较差；固态硬盘性能较好，但成本太高，寿命太短。并且当集群意外断电时极易导致操作系统损坏。通过无盘部署，有效避免了由硬盘导致的故障，极大降低集群故障率。

[0009] F、高效率，传统高性能计算集群需要对所有的节点安装操作系统和配置环境变量才能工作，本申请所描述方式无需该环节，极大减少了集群部署时间，提升了集群部署的效率。

发明内容

[0010] 本发明主要解决的技术问题是提供一种高性能计算集群动态节点作业方法，能够提高高性能计算集群部署效率，降低故障率，简化作业提交，提高能源和资金利用率，提升使用效率和性能。

[0011] 为解决上述技术问题，本发明采用的一个技术方案是：提供一种高性能计算集群动态节点作业方法，其特征在于，包括如下步骤：

[0012] 步骤1：配置服务器，包括管理网络N1、计算网络N2和用户网络N3，以及用户网络N3和管理网络N1之间的安全防护网络F1、登陆节点L1、管理节点M1、若干计算节点CN；所述登陆节点、管理节点、计算节点分别与管理网络和计算网络建立通信连接；所述登陆节点和管理节点通过计算网络挂载共享存储S1；

[0013] 步骤2：在管理节点M1上安装并配置启动管理服务(Flexboot)、动态主机配置协议服务(DHCP)、文件传输协议服务(TFTP)、域名系统服务(DNS)、共享服务、认证服务以及作业调度，配置完成后启动计算节点；

[0014] 步骤3：用户网络N3中的用户，经过安全防护网络F1以安全外壳协议SSH登录到登陆节点L1上，通过SSH将需要计算的算例上传至登陆节点L1上的共享存储S1；

[0015] 步骤4：用户在登陆节点L1上通过作业调度中间件，指定作业参数完成作业脚本的生成并启动提交给M1的作业调度进行资源分配和计算；

[0016] 步骤5：管理节点M1收到登陆节点L1上用户提交的作业脚本后，首先将作业脚本中的函数调用请求交给管理节点中的Maui.d进行资源配额检查，如果配额不足则报错提示，如果配额足够，则将作业脚本转交给M1上作业调度pbs_server进行调度；

[0017] 步骤6：管理节点M1的作业调度pbs_server收到作业脚本后，根据作业脚本要求的节点数、每节点核心数为作业分配进行计算的节点C1到Cn，如果节点数量不足，则将作业进行排队等待；如果节点数量足够，则根据作业脚本具体执行内容开始计算；

[0018] 步骤7：当C1到Cn完成计算后，释放物理内存、处理器等资源，并向M1的pbs_server反馈“所有核心均未占用，状态Free”的信号；M1的pbs_server对该信号进行记录，并允许后续作业在该节点上进行分配和计算；

[0019] 步骤8：用户可以通过安全文件传输协议SFTP从L1上将S1中存储的计算结果等信息取回到N3中，在本地电脑上打开进行处理和分析。

[0020] 优选地，管理网络N1为以太网络，主要负责计算节点开关机控制、作业调度数据传输、系统监控等功能，计算网络N1为Infiniband网络，提高设备的扩展性、数据的传输速率及通信延迟，主要负责数据及存储的IO、操作系统镜像分发、计算软件工作时各进程相互通信与数据同步等功能，用户网络N3是高性能计算机用户所在的网络，安全防护F1为防火墙、UTM或路由器设备，提供用户网络到管理网络的端口映射、访问权限管理、异常流量监测、攻击防护等功能，登录节点L1、管理节点M1、若干计算节点CN统一为同一处理器架构(如X86架构、MIPS架构、ARM架构、Power架构、Spark架构等)，处理器具有完全相同的指令集，根据用户实际使用需求和高性能计算机总体计算性能要求，登录节点L1、管理节点M1、若干计算节点CN可以通过集群(Cluster)方式进行横向扩展，共享存储S1一般为基于Infiniband的NFS服务器其包含底层的硬盘柜或磁盘阵列，或基于Infiniband的分布式存储系统，对外提供一个统一的文件系统空间，并且支持用户权限控制和容量配额。

[0021] 优选地，系统第一次部署时将登陆节点L1的操作系统复制为镜像文件J1，并修改J1中包含的相关个性化参数配置文件为通用配置文件，其中包括修改网卡配置文件，去掉MAC地址、UUID等唯一信息，修改为DHCP引导；将主机名修改为DHCP自动获取、将硬盘的挂载方式修改为设备名方式进行挂载、将系统环境变量存放目录修改为共享存储S1上特定目录等。

[0022] 所述计算节点的启动采用无盘启动方式，具体步骤包括：

[0023] 步骤1：将C1(或Cn)开机，设置为默认PXE引导，PXE默认设备为主机通道适配器(HCA)，采用FlexBoot模式；FlexBoot初始化HCA卡，检测端口协议及状态，以Infiniband方式启动端口，并以广播的方式发送DHCP客户端(Client)请求报文；

[0024] 步骤2：管理节点M1的DHCP 服务器从Infiniband网络收到请求报文后，将C1的IP地址、TFTP服务器、网络引导启动镜像目录发送给C1；C1接收到M1发出的报文后，根据收到的报文启动Infiniband网络，并从M1的TFTP服务器中下载启动镜像目录并加载到内存中，所请求的镜像目录内包含默认的启动镜像名称J1、默认启动镜像镜像时间等信息，通过C1上选择启动J1或者超过默认时间后自动选择启动J1，并向M1发送请求J1的报文；

[0025] 步骤3：M1收到C1请求J1的报文后，将J1通过TFTP服务器发送给C1；C1的FlexBoot接收完J1后，将J1放入内存进行加载；

[0026] 步骤4：C1依次加载J1的内核、根文件系统、网络、配置文件等；C1在加载网络时广播DHCP Client的请求报文，M1的DHCP Server收到请求报文后再次将C1的IP地址发送给C1，C1收到IP地址报文后启动网络，并向M1请求主机名；M1的DNS Server收到C1请求主机名的报文后将C1的主机名发送给C1；C1网络启动完成后，首先加载挂载共享目录的配置文件，根据配置文件内容将S1通过Infiniband方式进行挂载；C1挂载完共享目录后，启动计算节点作业调度(pbs_mom)，并将作业调度状态反馈给管理节点M1的作业调度(pbs_server)，启动完成。

[0027] 进一步的，所述作业中间件中提前录入已知的调用函数，针对已知的计算软件进行作业提交流程的重构和标准化，用户在提交作业时通过调用函数输入作业类型、参与计算节点数量、每个计算节点参与计算核心数量、输入文件(如果有)共计4个参数即可完成作业提交，避免了编写作业脚本的工作，极大简化上级操作步骤；

[0028] 进一步的，计算机开机节点自动控制包括如下步骤：

[0029] 步骤1：系统启动完成正常运行时，M1每60秒启动后台监控进程一次并检测启动时间，若未达到启动时间，则暂停60秒并重复步骤1；

[0030] 步骤2：后台监控进程正常启动后，检查pbs_server是否有排队作业，若无排队作业并且处于工作状态的计算节点小于等于1时，则直接退出；若无排队作业并处于工作状态的节点大于1个，则通过管理网络N1和IPMI接口关闭开机但位处于空闲状态的空闲节点数量-1个节点，只保留空闲节点中节点名排名最前的1个空闲计算节点处于开机状态以备用；

[0031] 步骤3：若检查到有排队作业情况时，则分析处于排队作业的原因：若为用户超额，则直接退出系统；若用户未超额，则排队原因为资源不足，执行下一步骤；

[0032] 步骤4：在步骤3之后，检查关机节点数量。若关机节点为0，则提示节点用尽并退出系统；若关机节点大于等于1，则根据处于等待状态各中作业各节点需求量的大小从小到大进行排序，设定N为等待状态作业最小作业节点需求数量，F为当前空闲状态节点数量，G为

关机节点数量。比较当前需要开机节点数量(N-F)和G-1。若N-F≤G-1,表示关机节点数量比需求节点数量大,则在管理网络N1中通过IPMI接口开启N-F+1个节点,并重新进行步骤2;若N-F≥G-1,表示关机节点数量不足,则在管理网络N1中通过IPMI接口开启所有关机节点并提示节点数用尽,然后退出系统。

[0033] 区别于现有技术的情况,本发明的有益效果是:

[0034] 1、节能:基于无盘部署,计算节点不在需要硬盘,降低了集群的功率开销和故障点。通过动态 调整计算节点开关机数量,避免了大量计算节点开机空转的情况,提高了能源的使用率。

[0035] 2、性能:基于Infiniband网络的无盘部署,将系统镜像加载到各计算节点内存中,提升了计算节点开机速度,并且充分利用了RAM DISK的IOPS性能优势,极大提升计算任务在计算节点单机内部的收敛速率。

[0036] 3、标准化:由于高性能计算涉及行业较多、范围较广、海量的专业软件,导致了传统用户在使用高性能计算集群时需要去针对具体的计算软件进行了解学习后才能上机使用。通过对作业流程的重构和标准化,将海量的专业软件的作业提交流程通过中间件固定为同样的步骤和流程,极大简化了上机操作步骤,让传统用户能快速的入手并将集群充分使用起来。

[0037] 4、节约:最大程度减少不必要的硬件投入(如计算节点硬盘),提升资金使用率。

[0038] 5、低故障率:传统高性能计算集群在每个计算节点上需要安装1块硬盘用于存放操作系统。机械硬盘价格便宜,使用年限久,但性能较差;固态硬盘性能较好,但成本太高,寿命太短。并且当集群意外断电时极易导致操作系统损坏。通过无盘部署,有效避免了由硬盘导致的故障,极大降低集群故障率。

[0039] 6、高效率:传统高性能计算集群需要对所有的节点安装操作系统和配置环境变量才能工作,本申请所描述方式无需该环节,极大减少了集群部署时间,提升了集群部署的效率。

附图说明

[0040] 图1是本发明实施例高性能计算集群系统拓扑图。

[0041] 图2是本发明实施例动态节点控制流程图。

具体实施方式

[0042] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅是本发明的一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0043] 参见图1提供的一种高性能计算集群动态节点作业方法,其特征在于,包括如下步骤:

[0044] 步骤1:配置服务器,包括管理网络N1、计算网络N2和用户网络N3,以及用户网络N3和管理网络N1之间的安全防护网络F1、登陆节点L1、管理节点M1、若干计算节点CN;所述登陆节点、管理节点、计算节点分别与管理网络和计算网络建立通信连接;所述登陆节点和管

理节点通过计算网络挂载共享存储S1；

[0045] 步骤2:在管理节点M1上安装并配置启动管理服务(Flexboot)、动态主机配置协议服务(DHCP)、文件传输协议服务(TFTP)、域名系统服务(DNS)、共享服务、认证服务以及作业调度,配置完成后启动计算节点；

[0046] 步骤3:用户网络N3中的用户,经过安全防护网络F1以安全外壳协议SSH登录到登陆节点L1上,通过SSH将需要计算的算例上传至登陆节点L1上的共享存储S1；

[0047] 步骤4:用户在登陆节点L1上通过作业调度中间件,指定作业参数完成作业脚本的生成并启动提交给M1的作业调度进行资源分配和计算；

[0048] 步骤5:管理节点M1收到登陆节点L1上用户提交的作业脚本后,首先将作业脚本中的函数调用请求交给管理节点中的Maui.d进行资源配额检查,如果配额不足则报错提示,如果配额足够,则将作业脚本转交给M1上作业调度pbs_server进行调度；

[0049] 步骤6:管理节点M1的作业调度pbs_server收到作业脚本后,根据作业脚本要求的节点数、每节点核心数为作业分配进行计算的节点C1到Cn,如果节点数量不足,则将作业进行排队等待;如果节点数量足够,则根据作业脚本具体执行内容开始计算；

[0050] 步骤7:当C1到Cn完成计算后,释放物理内存、处理器等资源,并向M1的pbs_server反馈“所有核心均未占用,状态Free”的信号;M1的pbs_server对该信号进行记录,并允许后续作业在该节点上进行分配和计算；

[0051] 步骤8:用户可以通过安全文件传输协议SFTP从L1上将S1中存储的计算结果等信息取回到N3中,在本地电脑上打开进行处理和分析。

[0052] 优选地,管理网络N1为千兆以太网络,主要负责计算节点开关机控制、作业调度数据传输、系统监控等功能,计算网络N1为Infiniband网络,提高设备的扩展性、数据的传输速率及通信延迟,主要负责数据及存储的IO、操作系统镜像分发、计算软件工作时各进程相互通信与数据同步等功能,用户网络N3是高性能计算机用户所在的网络,安全防护F1为防火墙、UTM或路由器设备,提供用户网络到管理网络的端口映射、访问权限管理、异常流量监测、攻击防护等功能,登录节点L1、管理节点M1、若干计算节点CN统一为同一处理器架构(如X86架构、MIPS架构、ARM架构、Power架构、Spark架构等),处理器具有完全相同的指令集,根据用户实际使用需求和高性能计算机总体计算性能要求,登录节点L1、管理节点M1、若干计算节点CN可以通过集群(Cluster)方式进行横向扩展,共享存储S1一般为基于Infiniband的NFS服务器其包含底层的硬盘柜或磁盘阵列,或基于Infiniband的分布式存储系统,对外提供一个统一的文件系统空间,并且支持用户权限控制和容量配额。

[0053] 其中、系统第一次部署时将登陆节点L1的操作系统复制为镜像文件J1,并修改J1中包含的相关个性化参数配置文件为通用配置文件,其中包括修改网卡配置文件,去掉MAC地址、UUID等唯一信息,修改为DHCP引导;将主机名修改为DHCP自动获取、将硬盘的挂载方式修改为设备名方式进行挂载、将系统环境变量存放目录修改为共享存储S1上特定目录等。

[0054] 具体的,所述计算节点的启动采用无盘启动方式,具体步骤包括：

[0055] 步骤1:将C1(或Cn)开机,设置为默认PXE引导,PXE默认设备为主机通道适配器(HCA),采用FlexBoot模式;FlexBoot初始化HCA卡,检测端口协议及状态,以Infiniband方式启动端口,并以广播的方式发送DHCP客户端(Client)请求报文；

[0056] 步骤2:管理节点M1的DHCP 服务器从Infiniband网络收到请求报文后,将C1的IP地址、TFTP服务器、网络引导启动镜像目录发送给C1;C1接收到M1发出的报文后,根据收到的报文启动Infiniband网络,并从M1的TFTP服务器中下载启动镜像目录并加载到内存中,所请求的镜像目录内包含默认的启动镜像名称J1、默认启动镜像镜像时间等信息,通过C1上选择启动J1或者超过默认时间后自动选择启动J1,并向M1发送请求J1的报文;

[0057] 步骤3:M1收到C1请求J1的报文后,将J1通过TFTP服务器发送给C1;C1的FlexBoot接收完J1后,将J1放入内存进行加载;

[0058] 步骤4:C1依次加载J1的内核、根文件系统、网络、配置文件等;C1在加载网络时广播DHCP Client的请求报文,M1的DHCP Server收到请求报文后再次将C1的IP地址发送给C1,C1收到IP地址报文后启动网络,并向M1请求主机名;M1的DNS Server收到C1请求主机名的报文后将C1的主机名发送给C1;C1网络启动完成后,首先加载挂载共享目录的配置文件,根据配置文件内容将S1通过Infiniband方式进行挂载;C1挂载完共享目录后,启动计算节点作业调度(pbs_mom),并将作业调度状态反馈给管理节点M1的作业调度(pbs_server),启动完成。

[0059] 进一步的,所述作业中间件中提前录入已知的调用函数,针对已知的计算软件进行作业提交流程的重构和标准化,用户在提交作业时通过调用函数输入作业类型、参与计算节点数量、每个计算节点参与计算核心数量、输入文件(如果有)共计4个参数即可完成作业提交,避免了编写作业脚本的工作,极大简化上级操作步骤;

[0060] 如图2所示,计算机开机节点自动控制包括如下步骤:

[0061] 步骤1:系统启动完成正常运行时,M1每60秒启动后台监控进程一次并检测启动时间,若未达到启动时间,则暂停60秒并重复步骤1;

[0062] 步骤2:后台监控进程正常启动后,检查pbs_server是否有排队作业,若无排队作业并且处于工作状态的计算节点小于等于1时,则直接退出;若无排队作业并处于工作状态的节点大于1个,则通过管理网络N1和IPMI接口关闭开机但处于空闲状态的空闲节点数量-1个节点,只保留空闲节点中节点名排名最前的1个空闲计算节点处于开机状态以备用;

[0063] 步骤3:若检查到有排队作业情况时,则分析处于排队作业的原因:若为用户超额,则直接退出系统;若用户未超额,则排队原因为资源不足,执行下一步骤;

[0064] 步骤4:在步骤3之后,检查关机节点数量。若关机节点为0,则提示节点用尽并退出系统;若关机节点大于等于1,则根据处于等待状态各中作业各节点需求量的大小从小到大进行排序,设定N为等待状态作业最小作业节点需求数量,F为当前空闲状态节点数量,G为关机节点数量。比较当前需要开机节点数量(N-F)和G-1。若N-F≤G-1,表示关机节点数量比需求节点数量大,则在管理网络N1中通过IPMI接口开启N-F+1个节点,并重新进行步骤2;若N-F≥G-1,表示关机节点数量不足,则在管理网络N1中通过IPMI接口开启所有关机节点并提示节点数用尽,然后退出系统。

[0065] 通过上述方式,本发明实施例的高性能计算集群动态节点作业方法,提高高性能计算集群部署效率,降低故障率,简化作业提交,提高能源和资金利用率,提升使用效率和性能。

[0066] 以上所述仅为本发明的实施例,并非因此限制本发明的专利范围,凡是利用本发明说明书及附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技

术领域,均同理包括在本发明的专利保护范围内。

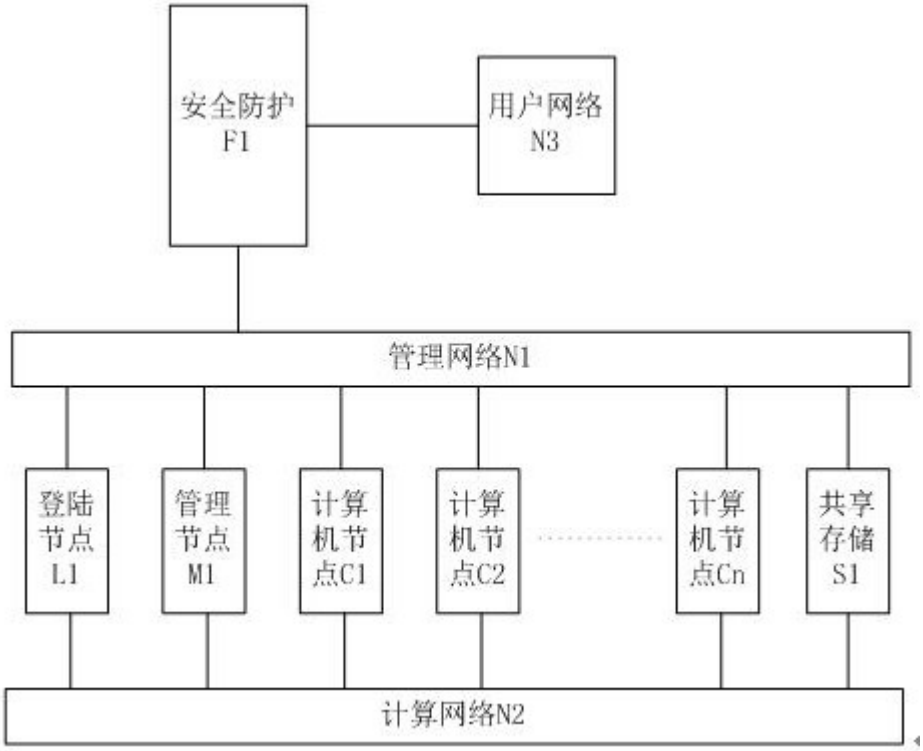

安全防护
F1
用户网络
N3
管理网络N1
登陆
节点
L1
管理
节点
M1
计算
机节
点C1
计算
机节
点C2
计算
机节
点Cn
共享
存储
S1
计算网络N2

图1

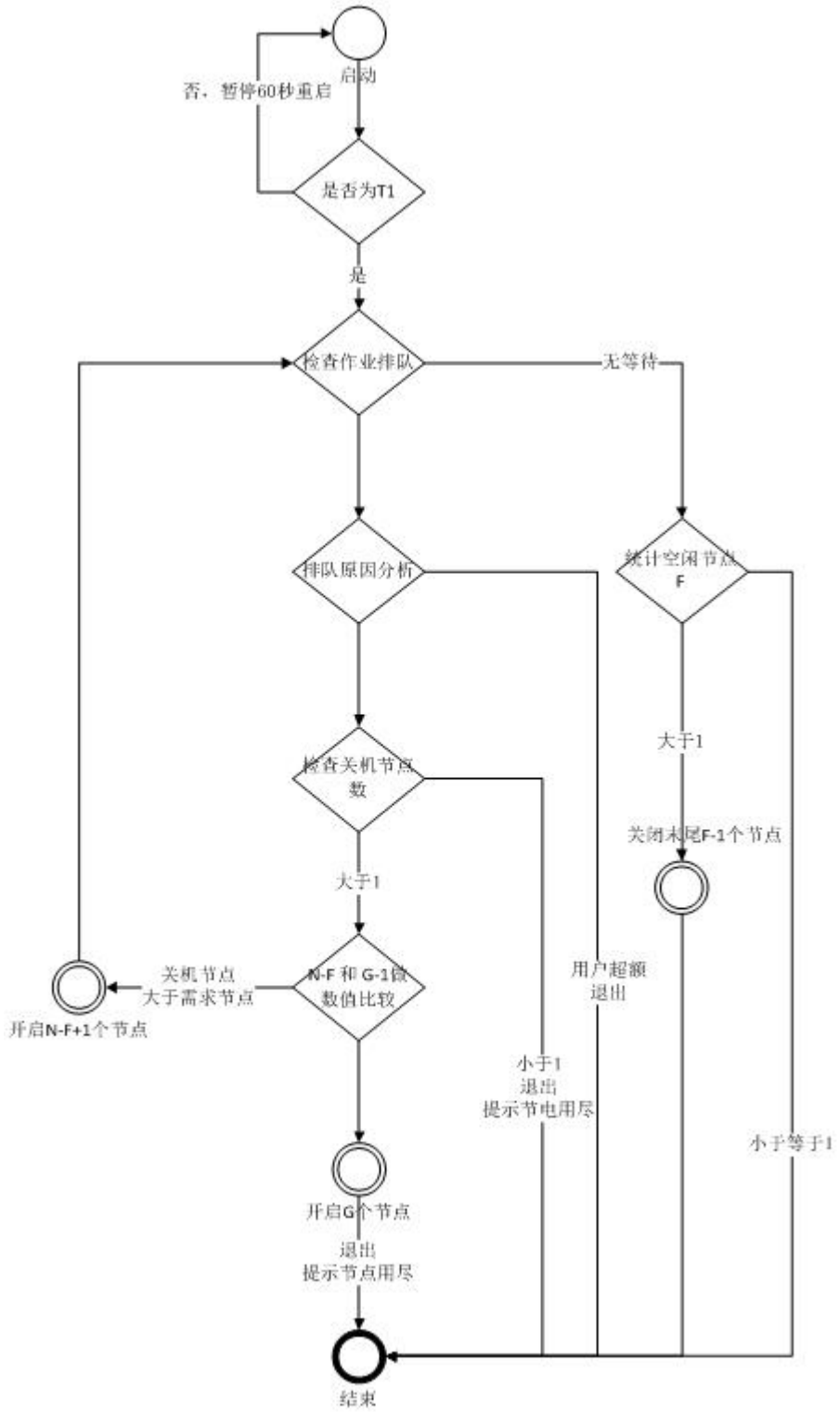

启动
否，暂停60秒重启
是否为T1
是
检查作业排队
无等待
排队原因分析
统计空闲节点F
检查关机节点数
大于1
关闭末尾F-1个节点
大于1
关机节点大于需求节点
N-F 和 G-1做数值比较
开启N-F+1个节点
用户超额退出
小于1
退出
提示节电用尽
小于等于1
开启G个节点
退出
提示节点用尽
结束

图2