北京计算科学研究中心

天河 2 集群用户手册

北京计算科学研究中心二〇一六年一月

目　录

JK 天河 2 号集群用户手册

1 集群系统的整体概括

1.1 集群硬件介绍

序号	设备用途	分类	要求配置	数量
1.	计算处理系统	瘦计算节点	天河刀片，双路 Intel Xeon 新一代 Haswell 系列 E5-2600 系列 v3 高性能 10 核 CPU，主频 2.6GHz；内存 192GB，1 个自主高速互连接口，2 个以太网接口，1 块 2.5 寸 500GB SATA 硬盘	546
			高性能加速部件：Intel Xeon Phi（双精度峰值性能 1.003Tflops）	256
		异构计算节点	天河刀片，双路 Intel Xeon 新一代 Haswell 系列 E5-2600 系列 v3 高性能 10 核 CPU，主频 2.6GHz；配置 1 个高性能加速部件 GPU：1 块 NVIDIA Tesla K40；内存 192GB，1 个自主高速互连接口，2 个以太网接口，1 块 2.5 寸 500GB SATA 硬盘	30
		胖计算节点	八路结点，8 个 Intel Xeon 高性能 E7-8880 v2 15 核 CPU，主频 2.5GHz；内存 1TB DDR3 ，1 个 IB FDR 接口，1 个万兆以太网接口，2 个以太网接口，2 块企业级 SAS 硬盘	20
2.	服务处理系统	管理和登录节点	2U 机架式含上架套件；双路 Intel Xeon 新一代 Ivy Bridge 系列 E5-2692 v2 高性能 12 核 CPU，主频 2.2GHz；内存：64GB DDR3 ；2 块 300G 10K SAS 硬盘；2 个千兆网口，1 个自主高速互连接口，1 块 IB HCA 卡，电源：冗余电源；集成 IPMI2.0 远程控制模块，超薄 DVD-RW 光驱	8
3.	互连通信系统	自主互连网络	192 口光交换机，配套 NRM 板、SWM 板、DSWM 板及光模块；点点通信带宽 112Gb/s	3
			QSFP FDR 光纤	816
		IB FDR 网络	108 口 IB FDR 交换机，配套光模块；点点通信带宽 56Gb/s	1
			QSFP FDR 光纤	108

序号	设备用途	分类	要求配置	数量
		以太网络	以太网交换模块	18
			万兆以太网交换机，48 个万兆端口，配套光模块和光纤	1
			以太网交换机，48 个千兆端口，2 个万兆光口；配套网线	1
4.	全局存储系统	I/O 存储节点(MDS+OSS)	双路 Intel Xeon 新一代 Ivy Bridge 系列 E5-2692 v2 高性能 12 核 CPU，主频 2.2GHz；内存：64GB DDR3；2 个千兆网口，1 块 IB HCA 卡，2 块双口 FC HBA 卡	22
		磁盘阵列(OST）	双 RAID 控制器，8 个 8Gb/s 主机通道；一个主控框+3 个扩展框;共 72 块磁盘,采用 3.5 寸 2TB SATA3 磁盘，总容量 144TB	10
		磁盘阵列（MDT)	双 RAID 控制器，8 个 8Gb/s 主机通道；一个主控框；配置 10 块 600GB 15K SAS 磁盘；用于元数据存储	1
5.	后备存储系统	I/O 存储节点	双路 Intel Xeon 新一代 Ivy Bridge 系列 E5-2692 v2 高性能 12 核 CPU，主频 2.2GHz；内存：64GB DDR3 ；2 个千兆网口，1 块 IB HCA 卡，2 块 RAID 卡；配置 36 块 4TB SATA3 磁盘,，总容量 144TB	16
6.	监控诊断系统	监控管理网络	监控管理模块	18
			48 口千兆以太交换机，配套网线	2
		一体化 KVM	机架式，显示平台，支持 KVM over IP	1
7.	基础架构	TH2 机柜	简化右托盘，配 4 个 PDU，24 个 3KW 电源模块	5
		TH2 列头柜	两组，共 4 个液晶屏	2
		I/O 机柜	1.2 米深 19"机柜，前后封闭玻璃门，配 2 个 PDU	14
		列式空调 1	LCP30	7
		列式空调 2	LCP80	12

1.2 集群软件

(1) 64 位 Kyliin Linux 操作系统，符合国际相关标准和规范。

(2) 配置 INTEL 和 GNU 编译器，支持 C/C++、Fortran、OpenMP、CUDA、OpenCL

编程语言；支持 OpenMP 与 MPI 并行编程接口，遵循 MPI3.0 标准和 OpenMP3.1 标准，支持 MPI 中间件。

(3) 全系统支持 MPI、高性能计算结点内部支持 OpenMP。

(4) 提供并行程序调试器 Allinea DDT，性能分析工具 YHProf；系统提供 MKL、BLAS、LAPACK、ScaLAPACK、FFTW 等数学库。

(5) 具有高效、易用的作业管理系统及资源统计管理软件，用户记账系统完备，系统资源可配额分配。

(6) 系统具有监控管理软件，能实现对网络性能和设备运行状况的监控管理，完成结点状态的监控、调试与诊断、管理与维护。

集群登陆节点

登陆节点 4 个节点，管理节点 4 个

登陆节点 ln0-ln3 IP 设置为： 10.0.0.101-10.0.0.104

管理节点不设置登陆 IP。

胖节点：10.0.0.102 也可以登陆 瘦节点之后 ssh ln1

2 作业队列说明

2.1 天河 2-JK 作业队列具体设置权限规定

天河 2 号集群队列 576 除掉 32 个 IO 节点包括 544 个 smalltask, middletask, bigtask, test, gpu, mic, engeering 队列, 我们规定 smalltask, middletask, bigtask 三个队列涵盖了除 test, gpu, engeering 之外的所有计算节点

smalltask [1-25 nodes: 20-500 cores] 默认分区，提交作业最多允许使用 25 个计算节点，每个账号限制 2 个作业运行，排队总数 10 个；最长运行时间为 2 天；计算节点使用范围 cn[18-230,233-307,310-383,416-441,444-543]。

middletask [26-100 nodes: 520-2000 cores] 最长运行时间为 1 天；计算节点使用范围 cn[18-230,233-307,310-383,416-441,444-543]。

bigtask [101-400 nodes: 2020-8000 cores]最长运行时间为 1 天；计算节点使用范围 cn[18-230,233-307,310-383,416-441,444-543]。

test [1-20 nodes: 20-200cores]分区共 20 个节点，每个任务最多可使用 3 个节点；最长运行时间为 1 小时；计算节点使用范围 cn[8-17]。

engineer 最长运行时间为 5 天；计算节点使用范围 cn[0-7,231-232,308-309,442-443,558-559]。

mic 最长运行时间为 5 天；计算节点使用范围 cn[0-127]。

gpu 最长运行时间为 5 天；计算节点使用范围 cn[544-557,560-575]
(only available to codes shown to be suitable for GPU implementation)

提交作业命令如下：

yhrun –p smalltask –N 10 -n 200 -t 2-00 –J JobName ~/work/example.sh

其中各参数意义如下：

-p 提交作业在指定分区运行

-N 请求为作业分配指定数量节点。

-n 指定要运行的任务数。请求为指定数量个任务分配资源。默认为每个任务一个 core。

-t 作业运行的总时间限制。如果作业没有在指定时间内完成，将自动退出；该时间不能设置大于分区时间限制。2-00 表示作业限制为 2 天。13:30:30 表示为限制 13 小时 30 分钟 30 秒。

-J 指定该作业名称，通过 yhq 查询作业队列时显示该名称；若未设置，默认显示可执行程序。

另外胖节点 fn [0-19]

胖节点运行： yhrun –n 120 ./exe > out.log &

登陆胖节点： ssh ln1 即可。

2.2 每个 job 的优先权是如何计算

TH2-JK 使用基于综合优先级的作业排队机制。作业的优先级为 32 位无符号整数，由作业属性与管理员配置的权重综合计算得出：

作业优先级=PriorityWeightAgeX 作业年龄因子
+PriorityWeightFairshareX 公平份额因子
+PriorityWeightJobSizeX 作业大小因子
+PriorityWeightPartitionX 分区因子
+PriorityWeightQOSXQOS 因子
+作业的优先级偏移

上式中，各优先级权重为管理员可在系统配置文件中设置的权重值，为 32 位无符号整数；各因子根据作业属性计算得到，值介于 0.0 和 1.0 之间；优先级偏移在提交作业或修改作业时指定，值为介于-10000 和 10000 之间的整数。如果用户直接设置了作业的优先级，则系统使用用户指定的值，而不再为其进行计算。 a) 作业的年龄指从可以运行算起，

作业在队列中等待的时间。作业因依赖关系、启动时间限制等不能运行时在队列中等待的时间不计算在内。年龄因子为作业的年龄与系统配置的最大作业年龄（系统配置文件中 PriorityMaxAge 选项）的比值，并被舍入到 0.0 和 1.0 之间。

b) 公平份额因子根据作业的关联的公平份额，及其实际资源使用情况计算得来。

c) 作业大小因子是作业请求的最少节点数目与系统中的总节点数之间的比值；但是如果系统配置文件中设置了选项 PriorityFavorSmall，则为 1 减去该比值。 d) 分区因子为作业所在分区的归一化优先级，即分区优先级与系统中优先级最高的分区的优先级之间的比值. 分区的优先级由管理员在系统配置文件中设置。

e) QOS 因子为作业所使用的 QOS 的归一化优先级，即 QOS 的优先级与系统中优先级最高的 QOS 的优先级之间的比值。QOS 的优先级由管理员在定义 QOS 时设置。通过定义不同的优先级权重，管理员可以定义出不同的调度策略。先来先服务（FCFS），即排队时间越长，优先级越高：

PriorityWeightAge=1000

小作业优先：

PriorityWeightJobSize=1000

PriorityFavorSmall=1

按 QOS 确定优先级：

PriorityWeightQOS=1000

2.3 如何查询节点的状态（是否正常，繁忙等）

在 mn[0-3]、ln[0-3]上使用命令"yhi"查看节点状态，节点状态可能如下：

a) UNKNOWN：未知状态

仅见于系统初启阶段，在控制进程还未能与节点监控进程通信，获得节点的状态之前存在。此状态在 yhi 的输出中为“unk”。

b) DOWN：宕机状态

处于此状态的节点不能分配给作业使用。节点进入 DOWN 状态的可能原因包括：

–节点上资源数量少于系统配置的数量

–通信故障导致的节点持续不响应

–节点运行作业的 Prolog/Epilog 程序出错

–管理员直接设置

节点进入 DOWN 状态时，在节点上运行的作业将被终止，作业状态

标记为 NODE_FAIL。此状态在 yhi 的输出中为“down”。

c) IDLE：空闲状态

节点状态正常，且没有分配给任何作业。此状态在 yhi 的输出中为“idle”

d) ALLOCATED：分配状态

节点状态正常，且已经分配到一个或多个作业。此状态在 yhi 的输出中为“alloc”。除状态之外，与节点相关的还有一些标志，这些标志与节点的状态一起，完整地表示节点的信息。

标志与状态相对独立，即节点可以在多种状态下都具有某个标志。可能的标志有：

e) DRAIN：排空标志

具有此标志的节点不再被分配到作业，但已经在节点上运行的作业不受影响。

在 yhi 的输出中，具有此标志的节点上如果有作业运行，则显示为“drng”，否则显示为“drain”。

f) COMPLETING：正在退出标志

表示节点上有作业处于退出过程中，正在释放资源。在作业运行结束时（包括作业运行成功、失败、节点故障、超时、被取消等各种情形）其分配的节点被释放，同时这些节点被设置 COMPLETING 标志。具有 COMPLETING 标志的节点不能被分配到需要独占节点的作业。在控制进程确认作业的所有进程都已经在节点上退出，即节点上的资源已被完全释放后，节点的 COMPLETING 标志被清除。在 yhi 的输出中，如果一个节点上有多个作业，且有的处于运行状态，有的处于退出过程中，则节点状态显示为“alloc+”；如果节点上所有的作业都处于退出过程中，则节点状态显示为“comp”。如果 yhi 显示有节点具有 COMPLETING 标志，则 yhq 命令可以看到对应的具有 COMPLETING 标志的作业。

g) NO_RESPOND：无响应标志

表示节点监控进程与控制进程的通信有故障。

为了维护节点的状态，控制进程周期性地 ping 所有的非 DOWN 状态的计算节点，以测试节点是否响应。

系统的一些其他动作也涉及到控制进程与节点上监控进程之间的通信。若这些通信失败，节点将被设置 NO_RESPOND 标志。

如果在系统配置的周期中，与节点的通信连续失败，则节点将被置为 DOWN 状态。

在 yhi 的输出中，具有此标志的节点的状态后带有“*”符号。

h) POWER_SAVE：节能标志

表示此节点目前被控制进程设置为节能状态。在 yhinfo 的输出中，具有此标志的节点的状态后带有“?”符号。

i) FAIL：失效标志

仅由管理员设置。具有此标志的节点不会被分配给作业。

在 yhi 的输出中，具有此标志的节点上如果有作业运行，则显示为“failg”，否则显示为“fail”。

j) POWER_UP：启动标志

表示节点正在从节能状态恢复。具有此标志的节点在 yhinfo 的输出中显示为“power_up”。

k) MAINT：系统维护标志

表示节点正在进行系统维护，即节点处于系统维护性预约中，且当前时刻在该预约的起止时间内。

具有此标志的节点在 yhinfo 的输出中显示为“maint”。

3 集群账号创建及登陆方法

3.1 集群账号创建

账号使用者需经过本单位有直接联系的研究人员许可后，再经过研究部主任批准，方可拥有集群账号的权限。然后去 B320 办公室填写表格创建自己的集群账号。

3.2 如何登录机群（中心内，中心外，国外）

登录节点：
ln0 10.0.0.101 用于瘦节点任务
ln1 10.0.0.102 用于胖节点任务
ln2 10.0.0.103 用于瘦节点任务

中心内登录：
通过 ssh 方式进行登录，例： ssh jk2@10.0.0.101
Account name: jk2
Code: ********

中心外登录： 需要联系科研管理办公室 A (108)焦明辉（010-56981709）处 ，获得 vpn 权限，登录到中心内网之后，按照中心内网的登录方式进行登录即可。

4. 机子时间统计与查询

4.1、机时统计与查询

使用如下命令查询用户个人作业已完成记录，包括作业使用的核与机时：

yhacct –S 2015-10-01T00:00 –E 2015-10-05T00:00 --format=JobName,Jobid,AllocCPUs…

-S 指定时间段的起始时间

-E 指定时间段的结束时间

--format 指定输出内容列；可通过使用命令"yhacct -e"获得可输出列字段。

所有用户可在 ln0 服务器/vol7/home/目录下使用脚本查询统计已使用机时：

用户统计指定时间段内已使用机时

/vol7/home/checksum.sh –s 2015-10-01T00:00 –e 2015-10-05T00:00

管理员统计指定时间段内指定用户组或指定用户已使用机时

/vol7/home/checksum.sh [-g nanowire | -u jk15] –s 2015-10-01T00:00 –e 2015-10-05T00:00

5 其他常用事项。

yhi:查看结点的作业执行的状态信息。

yhq:查看作业队列中作业执行的状态信息。

bjobs:查看用户作业执行状态信息

bhosts:查看用户可见分区状态信息

yhcontrol showpartitions Partition Name:显示分区详细信息

yhcontrol showjobs job_id: 查看作业详细信息

yhcontrol showsteps job_step_id:查看作业步信息

yhattach job_step_id:查看作业步任务输出

yhattach–layout job_step_id:查看作业步的任务布局

yhcancel:取消作业-u<user_name>|-p<partition_name>|-t<job_state>

当用户提交作业处于排队状态，使用"yhq"命令查询到如下排队原因

a) (Resources)
   计算节点资源不足，等待其他用户提交作业完成后释放可用节点。
b) (GrpsJobLimit)

最多同时运行作业数量限制，用户已提交最大上限作业数量，需等待用户其他作业完成后才可运行。

6 环境变量设置

```
# User specific aliases and functions
ulimit -s unlimited
ulimit -c unlimited
source /opt/intel/composer_xe_2015.0.090/bin/compilervars.sh intel64
export PATH=/usr/local/mvapich2/bin/:$PATH
export MV2_USE_SHARED_MEM=0
export MPICH_COLL_ALIAS_CHECK=0
```