

## C3 Cluster Power Tools: *increasing performance & expanding scalability*

Stephen L. Scott\* ; Oak Ridge National Laboratory



### Summary

*C3 Cluster Power Tools (cluster command & control) version 4.0 release increases its performance on large clusters by over a factor of 3x while expanding scalability into 1000's of compute nodes. C3 is a suite of cluster tools, available to administrator and user, which provide a single system illusion (SSi) top down view of a high-performance computing cluster such that the cluster may be viewed as a single machine.*

**Introduction:** In 1999 the C3 project was initiated at Oak Ridge National Laboratory (ORNL) with its goal to develop a set of tools to facilitate the use and administration of clusters in a *Single System illusion (SSi)* style such that a single command may run as one command across numerous machines. For example, a single password change will effect change on all 1000-nodes of a cluster – immediately, as if typed on each machine. This effort resulted in a very useful suite of tools that went further than just one cluster SSi to extend the paradigm of SSi across the entire computing range of: single machines - to- clusters -to- multiple clusters across the Internet (Grid) – and any subset of these via both the command line and scripting languages. Thus the same tool that is used to manage or access a group of office workstations may be used across a high-performance computing cluster or even a group of such clusters spanning the Grid.

**Command Overview:** C3 consists of ten general use tools: *cpushimage*, *cshutdown*, *cpush*, *crm*, *cget*, *cexec(s)*, *ckill*, *clist*, *cname*, and *cnum*. The *cpushimage* and *cshutdown* are both system administrator tools that may only be used by the root user. The other eight tools may be employed by any cluster user for both system and

application level use. A brief description of each core tool follows: The **cexec** command is the C3 general utility tool as it may execute any command on each cluster node. The **cget** command is effectively the *gather* operation – it retrieves files from each cluster node and deposits them on the local machine. The **cpush** command is the *scatter* operation – it pushes a file to specified nodes. The **ckill** tool runs the standard Linux kill command on each machine for a specified process name – stopping and removing that process from each. **cpushimage** enables a system administrator to push a machine disk image across the cluster with the option to reboot – allowing one to dynamically alter the operating environment. The **crm** is a cluster version of the standard Linux rm – for cluster wide file/directory removal. **cname** translates node name into a C3 position index value. **cnum** takes a range argument and returns the node names of those positions – no range returns all nodes. **clist** returns a list of clusters and their type. These last three commands are very important for dynamic and multi-cluster or Grid operations.

**Basic C3 Operation:** C3 recognizes three modes of cluster use. The first mode is *direct local (Figure 1)*. Direct refers to the

\* 865-574-3144, scottsl@ornl.gov

command being invoked from the cluster head-node. Local means that the invoking machine (head-node in this case) has knowledge of all nodes that exist in the cluster. Thus, direct local means that the command is invoked from the cluster head-node and the head-node has knowledge of all nodes that exist in the cluster via a local C3 configuration file.

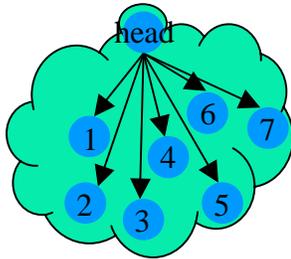


Figure 1: Direct local scope.

The second mode is *direct remote* (Figure 2). Remote indicates that the machine from which the command is invoked is not the head-node. Thus, direct remote means the command is invoked from a machine other than the head-node and because it is direct and the invoking machine has knowledge of the targeted compute nodes.

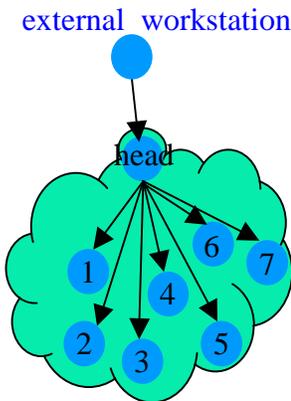


Figure 2: Direct & indirect remote scope.

The third mode is *indirect remote* (Figure 2). Indirect refers to the command being invoked from a machine that does not have knowledge of all nodes within the cluster. In this case, the invoking machine is not the

cluster head-node and only knows of the cluster's existence and not anything of the individual nodes within the cluster. This is the typical case for off-cluster access where the user's desktop machine only has a reference to the cluster (no individual node information) and the physical cluster configuration is maintained on the cluster's head-node.

**Scalability & Performance:** Early versions of C3 were serial in nature, executing the same command one-at-a-time across all nodes in the cluster as quickly as the head node could fire and retrieve results. Today, commands execute in parallel using a multi-process technique combined with a fan-out style (Figure 3) that enables the version 4.0 release (*C3 scalability release*) to talk to over 4096-nodes in approximately the same amount of time that the previous version took to reach 128-nodes.

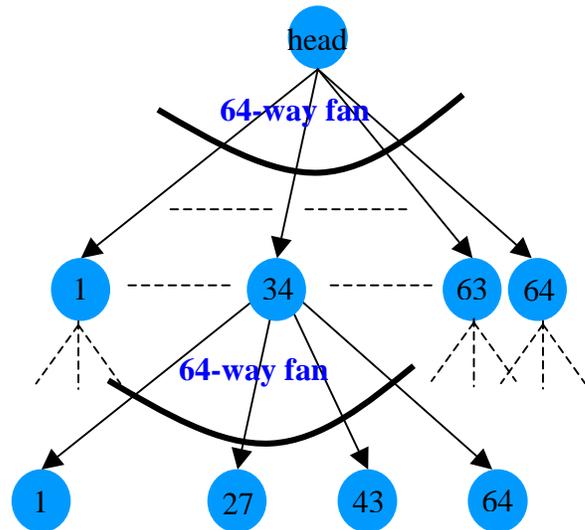


Figure 3: 64x64-way fan out reaches 4160 nodes.

**For further information on this subject contact:**  
 Dr. Fred Johnson, Program Manager  
 Mathematical, Information, and Computational  
 Sciences Division  
 Office of Advanced Scientific Computing Research  
 Phone: 301-903-3601  
 fjohnson@er.doe.gov