CLUSTER MANAGEMENT

Clusters have open-source roots

When setting up and configuring an HPC cluster, there is a wide variety of choices when it comes to tools: proprietary software from hardware or software vendors, many open source tools, and open source tools with commercial support. **Paul Schreier** provides a quick overview of the major players and their offerings

Scientific Computing World**: April/May 2009**

Setting up a computing cluster is much more than simply hooking up additional machines to a network. Somewhere, something has to set up the machines to recognise and work with each other; something has to make sure that compute tasks are assigned to the various nodes in the best fashion. A special class of middleware, which sits between the operating systems and the applications, has evolved specifically to handle these tasks, and it goes by the name of cluster management software.

Here, terminology is not completely unified, but most agree that cluster management involves defining a software stack and installing it first on a head node that holds user data plus tools for scheduling and monitoring. Then, in what is known as provisioning or imaging, it also sends an operating system plus other required software to compute nodes. In the definition of cluster management, however, some people also include workload management tools, the software that distributes the jobs among the nodes and creates detailed reports to show how efficiently each is being used.

‘Without resource management, you achieve 20 to 30 per cent system utilisation,’ says Michael Jackson, president of Cluster Resources. ‘By adding basic resource management, utilisation can reach 70 per cent, and by adding workload management and intelligent scheduling, cluster utilisation can reach 90 to 99 per cent.’

Open-source roots hold strong

Computer clusters first emerged in universities and research centres where this extra power was especially needed. Some of the things that characterise these organisations include tight budgets and people with computer expertise. Thus, it’s not surprising that cluster management software grew up primarily from open source Linux projects that cost almost nothing.

Meanwhile, the market for cluster management software resembles that for Linux itself. A number of open source distributions are available, but many of today’s HPC users don’t want to get into the messy details. Instead, they turn to companies that have packaged these distributions in an easy-to-use format. ‘Early on, users selected among different tools for provisioning/imaging and cluster management,’ explains Alanna Dwyer, HPC cluster marketing manager at Hewlett Packard. ‘Some users can handle this with a custom stack, but more and more people don’t want to go to 10 different projects and get the software working together. Many users today have much less familiarity with Linux and open source software, and they need a simpler, more comprehensive solution. This is even more important with today’s challenge: how do you deal with scale? How do you upgrade and do imaging quickly?’

The ecosystem chart from Cluster Resources (below) shows that the firm takes an expanded view of cluster management by also including workload management tasks.

Beowulf, cluster-building kits and batch systems

Many people give credit for clusters as we know them today to Donald Becker and Thomas Sterling who in late 1993 began to outline a commodity-based cluster system as a cost-effective alternative to large supercomputers and they called it the Beowulf Project. In 1999, Becker founded Scyld Computing and led the development of the next-generation Beowulf cluster operating system; he then joined Penguin Computing as CTO through its acquisition of Scyld Computing. Scyld ClusterWare, the commercial version of Warewulf, has an architecture based on three principles. First, nodes are provisioned through a network boot process from the cluster's master node into local RAM; an OS installation to local disk is not required. Libraries and drivers are provided to compute nodes by the head node on demand. Second, compute nodes run a lightweight OS; only one system service, used for communication with the master node, is launched on each one. Third, ClusterWare provides a unified process space across all nodes.

Open source technology also finds early roots in OpenPBS, a portable batch system developed by NASA Ames Research Center, Lawrence Livermore National Laboratory and Veridian Information Solutions. It monitors resources, queues up workload requests and then executes tasks on compute nodes when requested to do so. This open-source project was championed by Altair Engineering, which then produced a commercial offering known as PBS Professional and also by Cluster Resources, which added scalability, usability and other enhancements to the open-source product and renamed the enhanced version Torque. That code is now downloaded more than 100,000 times a year, making it one of the most broadly used resource managers today.

Another long-established and widely used solution for cluster distribution is Rocks, and more than 10,000 clusters have been deployed says Tim McIntire, president of ClusterCorp, a firm that sells a commercial distribution called Rocks+. That end-to-end HPC software stack includes the OS, cluster-management middleware, libraries and compilers. Users can add support for other hardware or software through modules called rolls. For instance, Clustercorp recently announced a roll for Platform LSF, a workload-management suite from Platform Computing. With it, users can add it to any Rocks+ cluster simply by clicking on a check box at install time or running a few commands post-install. The software is tightly integrated into the Rocks framework, which automates deployment and configuration across an entire compute cluster without hands-on intervention from a system administrator. Rolls are also available for other management software tools including Moab from Cluster Resources (which can be used to support multiple operating systems, including Windows), the open source Sun Grid Engine and Torque as well as Intel Cluster Ready, and it also supports hardware based on Nvidia and AMD chips.

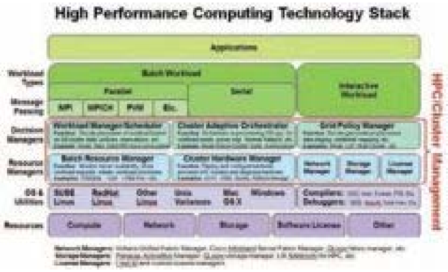

Image courtesy of Cluster Resources

Yet another cluster management toolkit to keep an eye on is OSCAR – Open Source Cluster Application Resources. It was created in 2000 by a group of organisations when it became apparent that assembling cluster software from many components was challenging and tedious. This group decided to choose 'best practices', selected among the many open-source solutions and include them in a software stack so as to make the installation, configuration and management of a modest-sized cluster easier. OSCAR is a bit different in that it installs on top of a standard installation of a supported Linux distribution. It then creates customised disk images used to provision the nodes.

Also worth mentioning is Warewulf, developed by the Scientific Cluster Support Program at the Lawrence Berkeley National Laboratory. This open-source toolkit is similar in many ways to Rocks and OSCAR. It works by allowing compute nodes to boot a shared image from a master node so that a systems administrator only needs to support the master node and the shared image for the rest of the system. In 2005, more than 20,000 downloads were recorded. Note that starting with Warewulf 3, cluster-provisioning features have been replaced by a project called Perceus.

Open source vs commercial

The open source/commercial route is evident in products from Platform Computing, whose Open Cluster Stack 5 was designed on the base of the open-source software in Project Kusu. The company now offers it as Platform OCS and charges for maintenance and support. If you just need tools to set up a cluster, this product is fine, comments Christoph Reichert, VP of HPC sales EMEA. The next step up is a proprietary package, Platform Manage, which has no open-source components, but adds a rich set of monitoring and reporting features. This software was previously known as Scali Manage; roughly 18 months ago Platform Computing acquired that software and about six months ago acquired the remaining parts of the company.

Platform Computing has entered into some interesting business relationships. Red Hat has teamed with it to offer the Red Hat HPC Solution, which integrates the OS, Platform OCS, drivers, installer, management, monitoring and other tools. This software is designed for smaller departmental clusters (<100 nodes) running on x86 64-bit hardware. Similar in scope is Platform Open Cluster Stack – Dell Edition, which besides a variety of tools and utilities also includes the Platform Lava job scheduler.

Several Platform Computing employees left to work at Univa UD, and they took KUSU and rewrote significant portions of it to create UniCluster. This company straddles the line between proprietary and open source code; it assembled the UniCluster system out of components from multiple sources: internally developed programs, commercial partnerships and open-source components. The company believes that its 'value add' is in the evaluation, selection and integration of these components.

And while UniCluster can be downloaded for free (last year there were nearly 8,000 downloads), the company sells a more-capable commercial version as well as a number of optional kits called Pro Packs that ease cluster setup and above all serve for reporting and analytics. Says president Gary Tyreman: 'It's ironic that various cluster-management packages rely on a narrow set of tools. Sure, several packages provide provisioning and monitoring, but what's missing are tools for operators to actually "manage" the cluster. Systems change over time, and new nodes are not likely a copy-exact replacement even if from the same manufacturer. Often there are multiple admins and even contractors that care for the cluster. Vast amounts of time can be spent figuring out what happened, who did what and when it was done.' Univia plans to release substantial improvements to UniCluster's systems management capabilities based on its experience with the Texas Advanced Computing Center, whose largest cluster has nearly 4,000 nodes.

Dual Linux/Windows support

Cluster Resources' Moab Cluster Builder, a proprietary package that includes the Moab Cluster Suite, leverages SUSE Linux Enterprise Server's pattern deployment capability to apply the needed software to the head node and compute nodes. Upon restart, a Moab cluster-building wizard queries the user about the number of racks, nodes per rack and network hardware. Moab Cluster Builder then automates the software configuration and the installer need only power up the nodes in a logical order. The software runs diagnostics and further validates proper installation by submitting serial and parallel jobs that serves as an acceptance test.

In addition, Moab Cluster Suite is a heterogeneous workload manager that increases utilisation rates to 90-99 per cent, provides a job-submission portal, helpdesk interface and management reports and can work with any of the popular resource managers including Torque, OpenPBS, Platform LSF, PBS Pro, Sun Grid Engine, Microsoft's Resource Manager and others.

Concerning hardware management, it is becoming important to provision and deploy software onto computers and manage nodes from a power perspective. Moab software thus connects the intelligence of the workload to energy savings such as to allow jobs to go to the most energy-efficient nodes; it can also perform thermal balancing and send jobs to the coolest computers.

A big advantage of software from Cluster Resources is that traditionally most cluster applications have been Linux based, but there is a gradually increasing demand to run Windows apps. And while many cluster software vendors talk about their support for heterogeneous systems, not that many actually support a mixed Linux/Windows setup. And some of those who can actually do this rely on software from Cluster Resources. Specifically, its Moab Hybrid Cluster allows users to provision nodes with a dual-boot manager so the scheduler can dynamically adapt the underlying OS between Linux or Windows depending on what environment the application or workload requires.

As noted, the vast majority of HPC clusters are Linux-based. However, Microsoft is not letting this market pass by, and in response it offers the Windows HPC Server 2008. It includes a template-based provisioning based on Windows Deployment Services technology. Hewlett Packard is one company happy to see this development. Says HP's Dwyer: 'We needed something for Windows, and it's great that Microsoft has come out with this, because it complements our Linux solutions.'

Vendor solutions

Hewlett Packard, like most other hardware vendors, offers several choices depending on user requirements. Its XC Systems Software is a dedicated environment that includes everything users need in one package, and it's for users who wish to reduce IT administrative tasks; they don't want to create a stack, and they have a fairly static environment. Next is CMU (Cluster Management Utility), which is used more by customers who have a dynamic system where they need to change the OS frequently or want variations on the stack, and it appeals to HPC laboratories, university research sites with in-house Linux expertise. For enterprise data centres, HP offers Platform HPC, based on software from Platform Computing.

For its part, IBM in the early 1990s set up the SP (Scalable POWERparallel) programme to build a cluster on AIX and later Unix with the aim of making the distributed hardware appear to function as a single computer. For this, the company wrote a proprietary package called CSM (cluster systems management). This product has continued to develop such as to add support for Linux and features targeted at the commercial world. In the late 1990s when Linux clusters started to pervade in HPC, some people felt that CSM was too broad to meet HPC requirements, so Dr Egan Ford wrote xCAT (Extreme Cluster Administration Toolkit), which today is IBM's strategic product for large HPC clusters running parallel jobs. With Version 2, xCAT has picked up many (but not all) functions from CSM, and in the meantime IBM has turned xCAT over to the open source community.

Finally, for diskless systems, IBM developed DIM (Distributed Image Management for Linux Clusters), which again is distributed at no charge. It was first developed for use in the MareNostrum supercomputer, which is managed by the National Supercomputing Center in Barcelona. Today DIM supports PowerPC, x86 and Cell B.E. based nodes. The MareNostrum supercomputer is composed of 2560 JS21 blades in 821 BladeCenters.

References

G. Pfister, In Search of Clusters, 2nd edition, Prentice Hall PTR, 1997. It's not easy to teach someone about cluster hardware and software in an entertaining manner, but while working at IBM, Pfister wrote this book, which does a very good job.