HRG Assessment:

Clusters - An HRG Assessment written in 1996

Clustering is one good way to meet the High Availability Challenge. The High Availability Challenge is the challenge faced by businesses with mission critical or business critical applications which require enhanced or increased system uptime and or availability. Clustering has become a pervasive technology in the arsenal of computer vendors and users because it can provide cost effective solutions to a wide variety of generic computer issues. We will briefly explore the primary issues which clustering addresses from an historical perspective. We will explore system availability in somewhat greater detail since this is a term which is frequently used in different contexts. The earliest cluster implementations will be discussed with regard to the generic system issues they addressed. Finally, criteria for evaluating current cluster implementations will be proposed

Brief History of Computing

Since the advent of the first computers, there has always been a desire and a need for more – more computes; more throughput – data in and data out; more data that is accessible; more memory to store programs and data; more up-time. The original thrust in providing these "mores" was to build bigger memories; faster processors with wider I/O busses and channels. The problem with this approach was that "more" was very expensive. In particular, it was very expensive even for the users who did not necessarily need more. Building a computer system that could accommodate larger memory sizes, faster processors, and more disks required a greater cost in the computing infrastructure that is rarely considered. Large systems require more robust cabinetry that can handle more and heavier components; larger power supplies that can generate the power necessary to support the largest configurations; longer cables with better noise immunity so all the necessary disks can be attached to the system; more powerful fans to provide the necessary cooling; and wider and faster internal paths that can bring data from memory to the processor.

The users with the greatest system requirements would go out and buy the largest system available and hope that it would meet their needs for years to come – or at least long enough that they could depreciate the investment. And when the business grew, or new applications were needed and the current system no longer could support all the users who needed it or the applications that needed to be run, the system was rolled out and the latest, newest model with increased capacity was rolled in. This was clearly a route that only the largest, richest companies could afford.

And, all the latest equipment did not necessarily satisfy the need for increased up-time. Up-time requirements were modest by today's standards: 16 X 6 (16 hours a day, six days a week). There were 2 shifts on Sunday and 1 shift each day for the remainder of the week for housekeeping and maintenance.

Computers were expensive and tended by a cadre of professionals with access to the machine strictly reserved for the operations staff. Preventive maintenance was the holy grail – it was performed on a rigid schedule dictated by the manufacturer. Programmers rarely had direct contact with the actual hardware let alone mere users.

Harvard Research Group, Inc.
P.O. Box 297
Harvard, MA 01451 - 0297 USA
Copyright © 2003 Harvard Research Group, Inc.
Tel. (978) 263-3399
Fax. (978) 263-0033
hrg@hrgresearch.com

Up-time Requirements Increase

Yet, as more and more business processes were computerized and manual backup procedures eliminated, the need for the system to be up when required became greater.

The earliest approaches to keeping the computer operational longer were to use more reliable components—components that would fail less frequently. For the most critical applications (usually military or life-threatening) specialized computers with redundant components and voting logic would be built. These were VERY expensive, one of a kind systems funded by governments (the only enterprise with big enough pockets to afford the large one-time engineering costs).

Other approaches were also employed to increase system up-time. The types of errors that caused system crashes were investigated. Certain classes of errors were recognized as being impossible to totally eliminate – memories having bits changed due to random Alpha particle hits; signals in wires interfering and changing signals in wires close by. Various check codes began to be appended to the data (and instructions) to detect such errors and eventually the codes became sufficiently sophisticated to enable correction of some of the simpler more common types of errors (ECC memory correcting single-bit memory errors and detecting double-bit errors). Mainframe manufacturers, led by IBM were the first to deploy these techniques since it was their systems that ran the business.

Programming errors were a frequent cause of crashes. Over time, increasingly sophisticated tools were developed to capture and represent the state of the system when a fatal error condition was encountered. These tools speeded the process of finding and correcting programming flaws.

Recognizing that systems did crash, steps were taken to ensure that work already completed would be saved and that long-running programs would not have to be totally rerun (there were programs that took longer to execute than the mean time between failures). Checkpoint/restart was born – one of the earliest high availability software techniques.

High Availability or Increasing Up-Time

For each user of a system, the system being up or available means that every distinct piece in the computer(s) and communications systems that is needed to get work done is operational. In a client/server environment, a user needs to write and print a report where the printer is on a departmental print server and the data that the report will be based upon is on a corporate server. In order to complete their work the user needs all of the following elements to be operational:

- Their PC or workstation
- The LAN that connects to the print server
- The print server
- The printer
- The LAN or WAN that connects to the corporate server
- The corporate server
- The disks on the corporate server where the data resides

Any failure in this chain will prevent this user from doing his work. The specific piece of equipment that is not operational will determine how many other users perceive the system as down. If the printer is not operational, it can potentially make the system seem down for many users – namely all users in the department. Similarly, failures in the network or the corporate server have the potential of creating outages for even greater numbers of users.

Copyright © 2003 Harvard Research Group, Inc.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

The design of a high availability system focuses on reducing the frequency of unplanned outages and their duration. Recognizing that mechanical and electrical equipment fails; that human beings make mistakes or more charitably that they have oversights and don't envision all possibilities; high availability system designers attempt to mask the effects of such failures. This is done primarily through ensuring that there is a backup for each component and that the backup can take over the function of the failed component rapidly. A good example of this approach is the use of disk mirroring and certain RAID levels to mask the effect of a disk failure. From a user perspective, the data is still there for him to use but the device(s) the computer retrieves the data from may be different.

The Earliest Clusters

In the late 1970s Tandem introduced the first "non-stop" commercial system running the Guardian operating system. This system was an integrated hardware/operating system/application engine designed to support on-line transaction processing. The system was built with redundant I/O paths; duplicated disk drives and controllers and processes that could be shadowed on another processor. Of course, the application program had to explicitly determine when data and state information would be updated on the shadow process but if the primary processor failed, there was another processor which had access to the data, and sufficient saved state that it could be restarted with no discernible interruption to users accessing the system.

Furthermore, the nature of transaction processing is such that "more" translates into more users doing the same things and transactions being processed are, for the most part, independent of each other. (Airline reservation systems were one of the earliest transaction processing applications – rarely will two telephone calls be simultaneously in progress where both customers want to book a seat on the same flight at the same time, let alone want the same seating location.) Due to the characteristics of transaction processing, it was easy to "grow" the Guardian system by adding additional processors in pairs and disk drives in pairs.

During the early 1980s, Digital introduced the first VAXcluster systems (the origination of the term clusters comes from these systems). These systems were intended to compete with mainframe systems in non monolithic applications – multiple processes (users) doing their work simultaneously. Users could log into the cluster and would automatically be logged into the system with the fewest users. Regardless of what physical system the user was actually connected to, the set of resources that was available was identical and could be accessed identically. To all intents and purposes, the multiple computers that comprised the cluster appeared to the user as a single system.

Interestingly, the underpinnings of the VAXcluster were not dramatically different than the underpinnings of the Guardian systems. Both systems had a high speed, dual rail bus which connected the processors with each other. In the Tandem system, disks were owned by a single processor but dual ported to a backup processor for availability reasons. In the VAXcluster, disks were connected to a separate processor, called a hierarchical storage controller that was equally accessible to all processors in the cluster. Optionally, disks could be dual ported between two separate hierarchical storage controllers. This ensured that if a controller failed that the drives and data could be accessed through the other storage controller. In addition, the storage controller could also mirror a disk drive.

Cluster Advantages

Clusters became, and still are, a very effective computing paradigm since they address many computing issues in a very cost-effective manner.

More processing – just increase the number of processors in a cluster or alternatively, make each processor more powerful. Currently available clusters permit nodes (where a node is a executing copy of the operating system) to be a mix of uniprocessor and multiprocessor systems of different capacities of the same architecture running the same operating system. (There are a few cases where the same architecture restriction is waived.)

Copyright © 2003 Harvard Research Group, Inc.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

Because clusters generally allow for a mix of similar processors of different capabilities, if more capacity is needed, a new system or additional storage can be added to what already exists. It is no longer necessary to throw out what is already there and totally replace it.

Tandem's Guardian system addressed the need for increased up-time explicitly. All elements of the system required to do the work were duplicated, connected by high-speed, redundant paths so that any failed component (including the operating system and application) could be rapidly (although not necessarily totally transparently) replaced by its duplicate.

The VAXcluster approach to increased up-time was more implicit. If the processor you were logged into went down, you lost the session in progress and had to reestablish a connection with another member of the cluster. However, this took far less time than would be taken if, in a single computer system, the failure could be corrected by simply rebooting the operating system which is a best case scenario. In a single computer system (including multiprocessors), if the action required to bring up the system includes a service call with parts replacement, the outage could easily extend to several hours if not days.

With a VAXcluster, once you were logged into another node in the cluster, all your saved work (the disks you were using) was accessible. The VAXcluster system could be configured so that disk drives were connected to a backup controller and the drives could be mirrored so that disk crashes or controller failures did not necessarily translate into users not being able to do their work.

The Rise of UNIX

The late 1980s and early 1990s saw the increasing market penetration of UNIX-based systems. The proprietary systems of the 70s and early 80s which included both VAXclusters and Tandem's Guardian systems saw their market share diminish as UNIX-based systems became more prevalent. UNIX-based systems bestowed many advantages on both users and applications developers. For businesses, using a proprietary system could easily mean being forced to buy systems from the same vendor since the cost of conversion could be enormous – converting applications, rewriting procedures, and retraining of operations staff, programming staff, and users. Since conversion costs were high, vendors of these proprietary systems could and did charge inflated prices and frequently delivered systems that lacked leading edge features and functions.

For application software vendors, the proliferation of unique operating environments implied massive investments of both time and money in rewriting an application to operate in other environments.

These two factors encouraged businesses to acquire UNIX based systems for their new applications. First, they could buy the best hardware available at a competitive price and if they later decided to change vendors the conversion was easily managed. Second, it was now possible to buy many more applications rather than developing them in-house.

The early UNIX systems came out of telephony and the universities and did not necessarily have the robust, bet your business characteristics of mainframe computers or the best in breed of the mid-range proprietary systems. Over time, the UNIX offerings became more robust and functional and system vendors realized that clustering was a cost effective methodology for incorporating many of the features businesses were demanding – more processing power, large storage capacity, more up-time, longer productive product life and simplified management.

UNIX Clusters

Today, almost every major vendor has a UNIX offering and a UNIX cluster offering. UNIX clusters are available from IBM, Sun, Digital, HP, NCR, Data General, Sequent, and Tandem.

Copyright © 2003 Harvard Research Group, Inc.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

Despite the proliferation of UNIX clusters, VMSclusters (the new name for VAXclusters since they now include both VAX and Alpha-based processors) is still the standard for comparison.

What follows is a recommended list of issues and features which perspective purchasers of clustered systems should take into consideration.

Capacity and Scalability

- How many nodes can a cluster accommodate?
- Can the nodes be a mix of uni and multiprocessors?
- Can the cluster accommodate mixed processor types?
- What is the smallest node that can be part of a cluster?
- What is the largest node that can be part of a cluster?
- How much disk storage can be put on the cluster?
- How many communications adapters? LANs, WANs
- Can tape drives be shared?
- How is disk storage allocated? Is it owned by one node and served to other nodes or do all nodes have equal access?
- Can the cluster, at job initiation, distribute the workload to maximize throughput?
- Can the cluster automatically redistribute the workload to maximize resource utilization?

Investment protection

- Does the cluster allow for different processor types?
- Does the cluster allow for different disk types with different capacities?

Availability

- Can additions (nodes, storage) be made to a cluster while the cluster is still operational?
- Can disks be mirrored? Can they be mirrored across controllers? Can they be mirrored across processors?
- Can the ownership of disks be switched between processors?
- Can any node take over the IP address of a failed node?
- In the case of a node failure, for how long is the rest of the cluster inaccessible?
- Can the cluster run two different versions of the operating system? (Otherwise, an operating system upgrade may require significant downtime for the entire cluster)
- Can a hung process or node be detected? Can it be restarted?
- Do the processors support RAID controllers?
- What provisions can be made for disaster recovery? Can the nodes that comprise a cluster be split across an arbitrarily large geographic distance?

Copyright © 2003 Harvard Research Group, Inc.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

System Administration

- Can applications be installed once and be accessible to every node in the cluster?
- Does the cluster automatically distribute the workload or is this a job for the operator?
- Can the entire cluster be viewed as a single system for the majority of operations or does the operation have to be repeated for every individual node in the cluster?
- Does the operator have a single topological view of the cluster and its connections?
- Can the cluster inform the operator of failure or unsafe conditions (close to capacity usage) in the cluster?

Wolfpack – what is it? When will it be here?

Wolfpack is a WNT clustering initiative that will provide the traditional cluster benefits of performance scaling, system availability, investment protection and simplified systems management. Unlike proprietary or UNIX cluster offerings, Wolfpack will allow systems from multiple vendors to be combined into a single cluster system. Wolfpack is a set of application programming interfaces that will allow WNT systems running on different vendor hardware platforms to provide cluster benefits to their client systems.

The initial partnership to provide WNT clustering was between Microsoft, Compaq and Tandem and occurred in October, 1995. Although a set of APIs has been published, actual products have not yet appeared. Initial Phase 1 release is planned for Q1 1997 (Compaq talks about 1st half of 1997). Microsoft states that Phase 1 will provide a two-node cluster capability with failover support and capable of executing on a "limited number of platforms" and released "in a series of carefully controlled, well tested phases". Phase 1 will additionally provide for remote mirrored storage using Servernet (developed by Tandem, and licensed to Compaq, Dell and others) and will provide support for Oracle and SQL. A subsequent release scheduled for 1997, will support 2 node parallel execution (as opposed to 2 node failover) with support for Oracle Parallel Server.

Phase 2 BETA is planned for 1998, at which time multiple-node clusters will be supported.

Scalability appears to be a weakness of Wolfpack. The Microsoft Windows NT Server Cluster Strategy White Paper (dated 11/8/95) observed that, "applications which require maximum scalability should use the cluster's shared nothing support".

While the Microsoft Windows NT Server Cluster Strategy White Paper talks about open standards and industry collaboration, the Wolfpack vision emphasizes the use of Microsoft's BackOffice components, namely, SQL Server, and Exchange Server. The final comment under the Server Application Support section reads, "Of course, Microsoft will encourage other vendors to leverage Windows NT Server Clusters."

Other vendors, including Digital, NCR and IBM have Wolfpack phase 1 functionality today—namely 2 node failover support. These vendors bring to WNT their extensive knowledge and experience with the issues of performance scaling, availability, and system management from the viewpoints of the enterprise, the department, and the desktop. Microsoft is coming to clustering, and enterprise level issues, from their desktop experience.

Today, Microsoft is experiencing major problems as it attempts to deploy Exchange as an enterprise level mail system. Are we sure that Wolfpack won't suffer from the same fate!

Conclusion

Copyright © 2003 Harvard Research Group, Inc.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

Any Highly Availble Cluster offering, improperly, inappropriately, and inconsistently used could easily cause worse problems than it is intended to cure. In other words, thinking ahead, planning, and communicating through complete specifications invariably reduces the risks (and costs) associated with implementation.

Copyright © 2003 Harvard Research Group, Inc.

Tel. (978) 263-3399
(978) 263-0033
hrg@hrgresearch.com

Harvard Research Group is an information technology market research and consulting company. The company provides highly focused market research, consulting services, and business modeling tools to vendors and users of computer hardware, software, and services. For more information contact Harvard Research Group as follows:

Harvard Research Group™
740 Massachusetts Avenue
Harvard, MA 01451 USA
Tel. (978) 263-3399
Fax (978) 263-0033

E-mail: hrg@hrgresearch.com
http://www.hrgresearch.com

Copyright © 2003 Harvard Research Group, Inc.