DA-301 Experiences

Terence L. Jones, Grumman Data Systems, Inc.

ABSTRACT:
In early 1995, Grumman Data Systems installed a C916 system at a customer's site to upgrade the existing computational platform. A key component in the design of the C916 system is the utilization of disk arrays. The CRAY DA-301 disk array was selected over conventional high-performance disk to provide cost-effective, high-performance, fault tolerant disk to support I/O intensive applications and critical files. This paper highlights the reasons for selection of the DA-301 as the high-performance component of the disk subsystem and describes the real-world experience gained with these devices.

Background

In late 1993, Grumman Data Systems was tasked with upgrading a customer's existing Y-MP8 system. The customer specified that the upgrade system must be capable of providing premium service to the largest of jobs, termed "grand challenge applications", while still affording adequate measures of throughput to the smaller jobs. With this directive as the desired objective for the upgrade system, design focused on the high-end C90 systems with particular attention paid to the design of the I/O subsystem. Numerous configuration approaches were considered, each with their own merits; however, the requirement for premium service to the largest of processes possible eliminated all but the most expansive designs. The goal became that of providing a system with the largest capacity possible, optimizing the tradeoffs between performance and fiscal considerations to produce a robust, feature-rich hardware environment capable of supporting leading-edge research and development for the next five years. In short, the system had to deliver the highest attainable performance currently available on the market. The capacity of the system would be limited only by budgetary constraints.

Particular attention was focused on the peripheral subsystems. Requirements existed to support high-speed network connectivity, conventional 36-track tape, and high-performance tape storage media. Beyond those requirements, however, came the task of configuring an optimal disk subsystem to support the large scale activities anticipated. A two-tiered disk subsystem was devised, providing high-performance disk for temporary working space and swap support, and a lower-speed, high volume disk environment for on-line storage of user's files. After several iterations, a configuration based on DD-60 disk for the high speed complement and DD-301 disk for the nominal performance element was devised.

However, the DD-60 was less than desirable as it was older technology and very expensive compared to the DD-301. The high-cost of the DD-60 constrained the amount of nominal storage that could be attained within the budget limitations. The performance of data migration on the existing system, when combined with the projected growth of the existing workload and anticipated new workload requirements necessitated as large of a user storage area as possible. Further, large-scale applications typically have long run times, and a failure in the high-performance disk subsystem would mean the loss of the CPU time investment for such jobs. A more robust, cost-effective approach was desired.

Disk arrays were discussed initially; however, the high cost of the controller made them appear less than attractive. DD-60 technology utilized the DCA-2 adapter with a then list-price of approximately $7,500 where as disk arrays required the DCA-3 adapter with a then list price of $50.000. Work performed by Grumman Data Systems at another customer site initiated a reinvestigation of disk arrays, ultimately leading to the replacement of the DD-60 technology with DA-301 technology as the high-performance disk component. The results of this analysis can be found in [1]. The conclusion of this analysis was that the DA-301 was actually more cost-effective, providing higher storage capacity and greater bandwidth than the DD-60 for less cost. The DA-301 also provided fault-tolerance, something not available using conventional, non-array disk. Maintenance costs were far lower, and the maintenance procedures were simpler with the DD/DA-301 equipment than the DD-60 devices. Finally, DD/DA-301 disks represented current generation technology and provided a growth path for future expansion. DD-60 technology was older technology and had reached the end of its growth path. Therefore, while the "infrastructure" necessary to support DA-301 disk arrays was initially higher in cost than the DD-60 technology, these higher costs were rapidly amortized through the higher performance, storage capacity, maintainability, capacity increments, and future potential, and their considerably lower maintenance costs over the lifetime of the hardware.

System Configuration

The installed I/O Subsystem configuration was ultimately limited by budgetary constraints. Considerable effort was expended to achieve a minimum of 300 GB of disk storage in total. Through the use of various options, and the subsequent announcement of the DD-302 technology, a disk subsystem design exceeding this capacity was attained. Additionally, it was not sufficient to simply increase the storage capacity to the highest possible limits without a corresponding increase in I/O subsystem bandwidth. Large quantities of data, accessed by a large number of users and I/O intensive "Grand Challenge" applications necessitated the highest bandwidth possible. Efforts to design an I/O subsystem configuration which maximized both storage capacity, data accessibility, and I/O bandwidth required detailed analysis and numerous tradeoffs. The final configuration installed is as follows:

- CRAY C916/16
- 1024 MW Central Memory
- 4096 MW Solid State Storage Device
- 6 I/O Cluster Model E I/O Subsystem
- 178 DD-301 disks initially configured as follows:
- 8 DA-301 disks/8 DCA-3 controllers
- 138 DD-301 disks/48 DCA-2 controllers

Constructing DA-301 devices requires five DD-301 disks. Four are used to store data and the fifth is used to store parity data. This fifth parity device is responsible for providing the fault tolerance feature which significantly enhances the desirability of disk arrays. The initially installed configuration utilized a single DA-301 device per controller to provide the highest accessibility possible. Daisy chaining DA-301 devices allows for creations of strings 8 devices deep; however, as only one device can be transferring data at a time on a chain, this affects the accessibility of the data on the devices. The design criteria for the high-performance complement of the disk subsystem limited the length of daisy chains to at most two. A second consideration to configuration was the maximum recommended DCA-3 controllers in a I/O Cluster (IOC). Cray Research recommends at most three DCA-3 controllers per cluster and preferably two.

The installed configuration uses four of the six I/O clusters exclusively for disks. The remaining clusters are used exclusively for communications and tape channels. This allowed the configuration to operate at peak efficiency with only two SSD VHISP Channels, each of which provides for HISP paths between two IOCs and the SSD. Thus, four I/O clusters are configured with "SSD Backdoor" channels, and these four clusters are those that exclusively service disks. All four of the disk IOCs are configured identically to enhance maintenance procedures. IOP0 holds two DCA-3 adapters and IOP1, IOP2, and IOP3 each hold four DCA-2 adapters.

Experiences

Since installation, the DA-301 (and underlying DD-301) technology has performed extremely well, considering the quantity of spindles installed. In total, five spindles have been replaced, three of which were members of a DA-301 disk array. None of the disk array element failures resulted in a system outage and all were replaced with the system fully operational. A command sequence is available using the *pddstat* and *ddms* utilities to perform the following steps in a maintenance procedure:

1. Disable the failing spindle, placing the DA-301 device into 4-device mode
2. Spindown and power off the disabled spindle. Spindle can now be disconnected from the other four still-functioning spindles, removed from the cabinet, and replaced with a new spindle.
3. Following certification of the new spindle, the DA-301 is restored to five-spindle mode
4. A final command is issued to reconstruct the DA-301 data across all five spindles, restoring the parity and fault tolerance.

The reconstruction of the data on an active array requires approximately 15 to 20 minutes and diminishes the device bandwidth by one-third. However, the system remains operational during the entire service interval. The performance of the disk array operating in four-spindle mode is no different than the performance of the array operating in five-spindle mode (peak of 32.8 MB/second). The site engineers report that maintaining DD-301 technology is far simpler and faster than the DD-60, DD-62, and older DD-41 and DD-49 technology in use at the customer's site. Cabling requirements are vastly simplified in the DD-301 disk family, requiring about 45 minutes less to prepare the disk for service than the other disk hardware. DA-301 devices are an engineer's dream because unless they must offload the filesystem, they do not require dedicated time to fix a failing or failed drive. These repairs can be typically effected during prime shift with no impact to the users. Even DD-301 disks themselves are far easier to work on due to the ability to rapidly access the hardware.

Further, a dump/restore of a DD-301 takes only about 5 minites to accomplish.

Performance

The four data disks of the DA-301 disk operate in parallel, effectively quadrupling the transfer rate of the single DD-301 elements. Thus, the maximum bandwidth of the DA-301 should be approximately 32.8 MB/seconds. Measurements conducted by Grumman Data Systems show that transfer rates approximating this maximum can be readily obtained. The most limiting factor appears to be the I/O method used and the length of the transfer. An application that needs to transfer a large amount of unformatted data should be able to easily attain the maximum rated speed of the DA-301 disks. In reality, a stripe group of four DD-301 disks should be able to attain the same bandwidth on transfers. However the DA-301, being "hardware striped" synchronizes the rotation of all five spindles in the array. This provides for a increase in performance for smaller sized transfers. A special wire pair runs along the back of the devices to provide the spindle sync signal.

Usage

The DA-301 disk arrays were originally planned to support the swap device (SWAPDEV) as an augmentation to a 2048 MW SSD. A later design phase introduced a 4096 MW SSD which could contain all of SWAPDEV and still support ample quantities of user Secondary Data Segments (SDS). The DA-301 disks were then re-deployed in the design to support the /tmp file system. This provided a high-performance scratch disk area with fault tolerance to protect the CPU time investment of long-running jobs. A further review of system design highlighted the use of the DA-301 arrays to support the primary segments of the user HOME file system. The primary segments were allocated so as to contain inodes only, and to reside entirely on a DA-301 device. By necessity, the HOME file system was allocated as a single file system of over 5 million inodes. This file system takes hours to load and unload, making the protection of the inodes of paramount importance. Additionally, with this large number of inodes allocated, performance issues became a concern. The high bandwidth and fault tolerance of the disk array was a logical solution that has worked very well to date.

The Future

Shortly after installation of the upgrade system, Cray Research offered the DD-302 device as an upgrade to the DD-301. The DD-302 provides for 1.8 GB/spindle and 9.8 MB/second bandwidth. At the same time, the DD-302 was introduced at the same price as the DD-301 and the list price of DD-301s were lowered. The DD-302 is the same disk spindle as the DD-301; however, it employs upgrade electronics to achieve the higher capacity and bandwidth. The DD-302 uses the same DCA-2 and DCA-3 controllers as the DD-301, and fits into the same DE100 enclosures. First availability of the DD-302 spindles is targeted for end of 2Q95 and plans are in work to double the disk array complement on the system. When completed, the planned upgrades will yield up to 107 GB of Array Disk Storage and nearly 220 GB of DD-301/DD-302 disk storage, bringing the total storage capacity to over the desired storage capacity of 300 GB.

Conclusion

The DA-301 disk array was successfully used to base the design of the disk subsystem for a leading edge High Performance Computing system. The advantages afforded by the array, including the increased performance resulted in its selection over the mature DD-60 disk. Actual experiences in the field at a customer's site have proven that the choice of Cray disk arrays was a good one. Disk maintainability and performance has been proven through actual real world experience. Upgrades are now in work to replace the DA-301 devices with DA-302 devices to provide increased performance and capacity.

References

1. Jones, T. L., and Welford, A. A., *An Analysis of DA-301 Disk Arrays; Proceedings of the Fall 1994 Cray User Group Meeting*, October, 1994