FNALdCachesetup

for the Tier II dCache workshop, June 2006
by Jon Bakken, FNAL

1. Head node setup

The head node functionality is split among 10 Nodes. 7 of them are dCache specific and 3 perform monitoring, controlling and sanity scans.

1. Core Head Nodes
    - Location Manager
    - dCache PoolManager
    - hsm Controller
2. dCap Doors
    - dCap Door 0 - 2
3. Management
    - Copy 0 – 5
    - dCap Door 3 – 6
    - gsiFtp Door
4. SRM – utils
    - SRM
    - SRM-V2
    - PinManager
    - Space Manager
    - Tomcat
5. dCap Doors – 2
    - dCap Door 7 – 10
    - gridFtp Door
6. Replica
    - Replica Manager
    - gridFtp Door
7. Information System
    - Billing
    - Spy
    - Httpd
    - infoProvider

2. Stage Pool Node Area

In order to get optimized throughput from tape to disk (restore), a set of pools is dedicated to restoring the tape based data to disk. User access to those pools is disabled. If a tape-only file is requested, the file is restored onto one of the 'stage pool nodes', subsequently copied to the actually read pools and from there delivered to the client processes.
The Stage Pool Node area consists of 8 nodes with a total capacity of about 9 TB of space.

3. Read/Write Pool Node Area

Currently there are about 110 TB of disk on 23 pool nodes dedicated to the regular read and write space. In summer another 100 nodes will be added with 600 TB disk space. All nodes have 2 bonded GE interfaces. Disk storage is either Infortrend with one node per array holding 2 partitions or Nexsan with 2 nodes per array holding 4 partitions. Each partition has 2 pools. One large production pool and one small volatile pool (about 10 GB). Each node runs one gridFtp server, typically only known by the SRM server. Each pool has 2 queues. One, so called LAN queue allowing up to 600 active movers (mostly dCap random I/O) and another one, so called WAN queue, allowing up to 3 active movers for file streaming.

4. Resilient dCache on worker nodes

About 500 worker nodes with a total space of about 55 Tbytes build the resilient dCache area. The resilient manager takes care that each file has at least 3 copies on different physical pool nodes.

5. Hardware

Nodes mostly dual processor intel. New nodes for the Nexsan storage will be dual-core dual-cpu opterons. Nodes have 2 GE bonded. We get 400 MB/sec per nexsan storage array.

6. Various

The WAN vers LAN (multiple I/O queues) dramatically improved reliability of the system, because we can better differentiate between 20-stream file transfers and posix I/O transfers and set appropriate max limits for each of them.

We use postgres 8.1.3 (resp 8.1.4) Journals are copied to different disks and then to different nodes. Complete backups twice per day, to tape.

Daily VOMS updates to dcache.kpwd, final testing of dynamic gPlazma GUMS authorization.

7. Problems

Tape is the hardest problem. The dCache-alone portion is much easier.