T10KC Technology in Production

Nick Balthaser
Wayne Hurlbert
LBNL/NERSC Storage Systems Group
May 9, 2013
Agenda

• **Environment**
  – Number T10KC drives
  – Length of time in production
  – Drive features in use

• **Data Volume**
  – Carts, TB, files

• **Workload**
  – Exchanges/unit time
  – IO Rates: Raw vs. HPSS

• **Error Rates**

• **Data Loss**

• **Conclusion**
Environment

• **Currently 34 T10KC in production**
  – Total population of 162 Oracle/STK tape drives in 4 SL8500s
  – First set of 18 C drives put into production on 01/25/2012
  – Second set of 16 drives put into production in 07 - 08/2012
  – Adding third set of 10 drives ASAP 2013 (drives on site)
  – Intending to purchase another set ASAP

• **No optional drive features in use**
  – We do not use encryption
  – We do not use the tape length extension technique

• **HPSS archival storage application**
Data Volume

- **Quantity of data on T10KC since 01/2012:**
  - 3,242 T10KC Cartridges
  - 20,783 TB
  - 25,767,059 Files
Workload: Exchange Rates

- Exchange Rates
  - total 458,234 exchanges in 2012 (01/25 – 12/31 2012)
  - Average: 1,340 T10KC total exchanges per day
    - Roughly 40 exchanges per drive per day
  - Highest:
    - 5,117 exchanges for drive 1,13,1,9 in 06/2012
    - 1,674 exchanges for cartridge EP0652
Workload: Raw IO Rates

- Raw IO to drive
  - 4GB uncompressable file using Unix "dd" utility in loop:
    - Max read: **248MB/sec** (sequential read from drive to /dev/null)
    - Max write: **201MB/sec** (sequential 4GB writes from local disk)

<table>
<thead>
<tr>
<th>Host</th>
<th>Action</th>
<th>Block</th>
<th>File</th>
<th>Blocks</th>
<th>File rate</th>
<th>Wall clock</th>
</tr>
</thead>
<tbody>
<tr>
<td>phish</td>
<td>write</td>
<td>256 KiB</td>
<td>671088640</td>
<td>2560</td>
<td>103 MB/s</td>
<td>102 MB/s</td>
</tr>
<tr>
<td>shins</td>
<td>write</td>
<td>256 KiB</td>
<td>671088640</td>
<td>2560</td>
<td>102 MB/s</td>
<td>102 MB/s</td>
</tr>
<tr>
<td>phish</td>
<td>read</td>
<td>256 KiB</td>
<td>671088640</td>
<td>2560</td>
<td>251 MB/s</td>
<td>246 MB/s</td>
</tr>
<tr>
<td>shins</td>
<td>read</td>
<td>256 KiB</td>
<td>671088640</td>
<td>2560</td>
<td>251 MB/s</td>
<td>249 MB/s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-----</td>
</tr>
</tbody>
</table>

| phish| write  | 256 KiB| 4 GB      | 15258  | 201 MB/s   | 200 MB/s   |
| shins| write  | 256 KiB| 4 GB      | 15258  | 200 MB/s   | 200 MB/s   |
| phish| read   | 256 KiB| 4 GB      | 15258  | 248 MB/s   | 248 MB/s   |
| shins| read   | 256 KiB| 4 GB      | 15258  | 248 MB/s   | 248 MB/s   |

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-----</td>
</tr>
</tbody>
</table>

| phish| write  | 512 KiB| 1342177280| 2560   | 145 MB/s   | 144 MB/s   |
| shins| write  | 512 KiB| 1342177280| 2560   | 144 MB/s   | 144 MB/s   |
| phish| read   | 512 KiB| 1342177280| 2560   | 246 MB/s   | 246 MB/s   |
| shins| read   | 512 KiB| 1342177280| 2560   | 246 MiB/s  | 246 MiB/s  |
Workload: HPSS IO Rates

- **HPSS Migration (writes from application):**
  - SC 8 migration: **145 MB/sec** per drive. Includes overhead and drive start/stop as each file migrates
  - Regent direct-to-tape (fwfs): **151 MB/sec** per file across the network (HPSS direct to tape drive).

- **HPSS transfers from GPFS**
  - 1TB uncompressable file to Direct-to-Tape HPSS COS:
    - **Read:** **224 MB/sec** (T10KC via HPSS read to /dev/null)
    - ```
      time hsi -q -s hpss 'set cos=14; get /dev/null : /home/n/nickb/testfiles/1TB'
      get '/dev/null' : '/home/n/nickb/testfiles/1TB' (2013/04/21 14:29:50 1099511627776 bytes, 224056.2 KBS )
    ```
    - real 83m31.609s
    - user 0m32.688s
    - sys 12m15.633s
    - **Write:** **156 MB/sec** (NGF to HPSS T10KC direct-to-tape)
    - ```
      time hsi -q -s hpss 'set cos=14; put 1TB : /home/n/nickb/testfiles/1TB'
      put '1TB' : '/home/n/nickb/testfiles/1TB' ( 1099511627776 bytes, 156125.7 KBS (cos=14))
    ```
    - real 118m0.818s
    - user 0m7.273s
    - sys 12m1.197s
  - **Native read from NGF (DTN node):** 2.5 – 3 GB/sec
    - ```
      time cat 1TB > /dev/null
    ```
    - real 7m2.365s
    - user 0m0.172s
    - sys 3m39.517s
Error Rates

• AIX data movers report device errors via error report (errpt) facility

• Total 1,847 device errors on 335 distinct cartridges:
  – 1,741 FSF error due to early FW bug
  – 28 EOM – known AIX driver issue
  – 12 errors on test tapes
  – 4 errors on damaged cartridge EP1041
  – Total: 62 unexplained read or write errors
    • We were subsequently able to retrieve the data after retries
Data Loss

- **Partial loss of one cartridge Sept 2012**
  - Cartridge would not mount although we could manually wind it
  - Sent to Oracle for recovery:
    - Large section of tape somewhere near the end was crumpled
    - Would not mount due to increased spool diameter
    - They had to cut out the crumpled section and splice in new tape to maintain spool diameter
    - They were able to recover undamaged section using a labor-intensive manual process
  - **Data Lost:**
    - 1,660 files out of 12,951 on cartridge
    - Approximately 1TB out of 6.5TB on cartridge
Conclusions

• Data loss and error rates have been low so far
  – < 0.005% data loss by volume
• HPSS appears to incur write overhead
  – ~55MB/sec decrease vs. raw IO
• We’d be interested in how our statistics compare to T10KC stats from other sites
  – We expect we have a relatively high volume of exchanges relative to other archival sites, largely due to read activity
• Comparison with T10KB stats would be interesting
• Thanks for listening!
National Energy Research Scientific Computing Center
Section Title