PersonalRAID

Jason Wilson

wacko@wpi.edu
Introduction

- Disconnected devices
- Data redundancy
- Log-structured design
- Personal usage
Desirable Features (Goals)

• **Availability of a single coherent name space**
  - No manual hoarding or propagating
  - Same space, regardless of location (as if all data was on mobile storage device)
  - No modification of existing applications

• **Reliability**
  - Mobile devices are not second-class citizens

• **Acceptable Performance**
  - No significant overhead
  - Performance close to local storage
Virtual - A

- Removable Storage
- Capacity greater than actual capacity
- 1 GB Microdrive on PCMCIA
Operation

(a) Record
(b) Disconnect
(c) Connect
(d) Read
(e) Replay
Log Structure

- Data is buffered into large memory segments
- Buffers prevent overwritten data from ever reaching the disks and large segment-sized writes are efficient
- Large segment-sized writes are more efficient
- Granularity allows full bandwidth of host disk resulting in fast replay
- Segment cleaning on replay
# Data Structures

<table>
<thead>
<tr>
<th>Segment Summary</th>
<th>Local Disk</th>
<th>Virtual-A</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Checkpoint</td>
<td>PA → LA → t</td>
<td>PA → LA → t</td>
<td>Disk</td>
</tr>
<tr>
<td>Map</td>
<td>LA → PA</td>
<td>LA → PA</td>
<td>Memory</td>
</tr>
</tbody>
</table>

- **PA**: physical device address.
- **LA**: logical (virtual) device address.
- **t**: time stamp (global counter).
- **bitmap**: one bit ($b_i$) per host, $b_i=1$ if host $i$ needs propagation.
- **state**: 4 bits of state information for a block.
  - $s_0$: 1 iff the block needed to be propagated to the current host at the beginning of the current session.
  - $s_1$: 1 iff the block needed to be propagated to any of the other hosts at the beginning of the current session.
  - $s_2$: 1 iff the block has been propagated from the VA to the current host in the current session.
  - $s_3$: 1 iff the block has been overwritten in the current session.
Recording

• Append new data to both logs
• In-memory maps updated with latest locations
• Set $s_3 = 1$ to mark creation during session
Disconnect

- Unmount and flush dirty file-system buffers to logical disk
- Write in-memory map to checkpoint region
- On VA, compute new bitmap fields and write
  - i.e. If $s3 = 1$, set all bits for this block
- After crash, local disk and VA must be mutually consistent
  - Unwritten data blocks
  - Restore bitmaps in VA checkpoint
Connection and Reading

• Initiate in-memory maps from checkpoints
• Compute state field on VA
  – i.e. If \( s_0 = 1 \) means local disk copy is obsolete and need to nullify L-to-P mapping
• Reading uses in-memory maps to determine location of most recent copy of data
Replay

- If \( s_0 = 1 \), need to write block to log for local disk

- Segment cleaning
  - If block copied to all hosts \( (s_1 = 0) \), it can be removed from VA
  - Else leave on disk or append to VA log
Recovery

? Host disk loss
  – Sync VA with surviving host
  – Mirror local disk and use to replace lost disk

? VA device loss
  – Visit all hosts twice to determine what data was lost
    ? Recreate bitmaps
    ? Retrieve lost data
Implementation

- PR Driver: device driver to virtual disk
- PR Server: implements LLD
- Main-memory segments for each device
- Segment cleaning when below a threshold
### Performance - Recording

<table>
<thead>
<tr>
<th>Method</th>
<th>mkdir (s)</th>
<th>lwrite (s)</th>
<th>swrite (s)</th>
<th>touch (s)</th>
<th>du (s)</th>
<th>grep (s)</th>
<th>wc (s)</th>
<th>total (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UFS-Local</td>
<td>156</td>
<td>20</td>
<td>398</td>
<td>806</td>
<td>6</td>
<td>142</td>
<td>368</td>
<td>1896</td>
</tr>
<tr>
<td>UFS-Upcalls</td>
<td>397</td>
<td>33</td>
<td>675</td>
<td>951</td>
<td>15</td>
<td>220</td>
<td>388</td>
<td>2679</td>
</tr>
<tr>
<td>LLD-Local</td>
<td>184</td>
<td>28</td>
<td>229</td>
<td>330</td>
<td>5</td>
<td>66</td>
<td>94</td>
<td>936</td>
</tr>
<tr>
<td>UFS-MD</td>
<td>398</td>
<td>320</td>
<td>725</td>
<td>1440</td>
<td>1200</td>
<td>535</td>
<td>880</td>
<td>5498</td>
</tr>
<tr>
<td>PR-VA</td>
<td>325</td>
<td>239</td>
<td>484</td>
<td>350</td>
<td>167</td>
<td>85</td>
<td>127</td>
<td>1777</td>
</tr>
<tr>
<td>PR-VVA</td>
<td>192</td>
<td>55</td>
<td>263</td>
<td>336</td>
<td>151</td>
<td>75</td>
<td>100</td>
<td>1172</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>MMAB total (s)</th>
<th>MOZ1 (s)</th>
<th>MOZ2 (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>post-replay</td>
<td>pre-replay</td>
<td></td>
</tr>
<tr>
<td>UFS-Local</td>
<td>1896</td>
<td>75</td>
<td>476</td>
</tr>
<tr>
<td>UFS-Upcalls</td>
<td>2679</td>
<td>126</td>
<td>481</td>
</tr>
<tr>
<td>LLD-Local</td>
<td>936</td>
<td>66</td>
<td>464</td>
</tr>
<tr>
<td>UFS-MD</td>
<td>5498</td>
<td>404</td>
<td>640</td>
</tr>
<tr>
<td>PR-VA</td>
<td>1777</td>
<td>395</td>
<td>548</td>
</tr>
<tr>
<td>PR-VVA</td>
<td>1172</td>
<td>102</td>
<td>482</td>
</tr>
</tbody>
</table>
Performance - Replaying

- VA --> LFS limited by read performance of Microdisk

<table>
<thead>
<tr>
<th></th>
<th>MMAB (s)</th>
<th>MOZ1 (s)</th>
<th>MOZ2 (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VA→LFS</td>
<td>553</td>
<td>340</td>
<td>56</td>
</tr>
<tr>
<td>VA→UFS</td>
<td>988</td>
<td>547</td>
<td>100</td>
</tr>
<tr>
<td>VVA→LFS</td>
<td>159</td>
<td>90</td>
<td>17</td>
</tr>
</tbody>
</table>
Final Thoughts

- VA to use flash memory device (e.g. SD or CF)
- Weakly connected hosts
- Nonpersonal usage and conflicts
- No access without VA