6 Gb/s SAS Disk Drive Considerations

Richard Uber
Maxtor
7-NOV-2005
• Other interfaces are already running faster than our 3 Gb/s SAS and SATA interfaces (including 4.25, 5.0, 6.25, 8, 10 Gb/s)

• This differs from parallel SCSI where we were the only industry group trying to build multi-drop, parallel, backwards compatible buses at U320 and U640 rates

• Challenges are:
  – Choosing an optimal set of feature additions for 6 Gb/s SAS to serve the “set of SAS applications”
  – Figure out what “set of SAS applications” is now, and what it will grow into
Publicly reported Serial interface test chips, Sept-05

### 6Gb/s + Backplane Transceivers

(most developed under OIF CEI-02.0)

<table>
<thead>
<tr>
<th>Company</th>
<th>Data Rate Gb/s</th>
<th>Tx</th>
<th>Rx</th>
<th>Channel Loss at Nyquist Frequency</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agere</td>
<td>3.6.25</td>
<td>PAM-4/NRZ</td>
<td>PAM-4/NRZ</td>
<td>Legacy backpl</td>
<td>Macrocell</td>
</tr>
<tr>
<td>IBM</td>
<td>4.9-6.4</td>
<td>4-tap FIR</td>
<td>Peak+5-tap DFE</td>
<td>32 dB 25 dB (DFE)</td>
<td>Core IP</td>
</tr>
<tr>
<td>LSI Logic</td>
<td>4.8-6.4</td>
<td>2-tap FIR</td>
<td>4-tap DFE</td>
<td>40” FR4</td>
<td>Core IP</td>
</tr>
<tr>
<td>PMC-Sierra</td>
<td>4.9-6.4</td>
<td>2-tap FIR</td>
<td>10-tap DFE</td>
<td>25dB</td>
<td>Device</td>
</tr>
<tr>
<td>Synopsis</td>
<td>0.6-9.6</td>
<td>3-tap FIR</td>
<td>1-tap DFE</td>
<td>36” BER=10^{-12}</td>
<td>Core IP</td>
</tr>
<tr>
<td>Texas Instruments</td>
<td>3.5-6.25</td>
<td>4-tap FIR</td>
<td>4-tap DFE</td>
<td>30” FR4</td>
<td>Device</td>
</tr>
<tr>
<td>Vitesse</td>
<td>5-Gb/s</td>
<td>3-tap FIR</td>
<td>FFE +3-tap DFE</td>
<td>9.9dB</td>
<td>Device</td>
</tr>
</tbody>
</table>

- Performance demonstrated by IC industry already stretches beyond SAS-1.1 backplane rich of 0.5m @ FR-4-2. IC industry is ready for SAS-2
- SAS-2 means fitting existing interconnect technology (not only OIF-based) to the SAS-2 specifics
• Keep points of failure isolated to Field Replaceable Units, namely HBAs and HDDs
• Backplane, cables all passive
• “Multiplex function” contained in remote HBA card
• Each HDD runs at full transfer speed of storage system
• Substantial active silicon devices on or near backplane
• Insufficient energy expended to date on electronics cost tradeoffs between HDD ASICs and Expander ASICs
Multiplexing of SAS Channels

• Considered early, but not yet implemented or documented in specifications
• No means for combining 1.5 or 3.0 Gb/s channels into a single 6 Gb/s box-to-box channel
• Use of 6 Gb/s anywhere forces all components to also operate at 6 Gb/s
• If hardware allowed SAS channels to be multiplexed into higher speed channels:
  – Box-to-box speeds could quickly go to 6, 10, or 12 Gb/s
  – Drive interface speed could track disk data rates:
    • Lowest cost brick
    • Add cost for higher speeds only when necessary
Seagate view of Max HDD Transfer Rate, Sept-05, from STA web site

- Outer track
- 3.5” drive
- No seeks
- Peak rate, not the sustained rate
- Need 6 Gb/s after 2011

6 Gbps SAS – Projected Disc Data Rate

- Data rate into the drive buffer from an Outer Diameter Track
  - Single Track. No allowance for Seeks or Rotational Latency
  - Format efficiency taken into account
• DRAM required bandwidth is dominated by the Serial I/O ports
  • Max I/O @ 3 Gb/s is 4 x 300 MB/s = 1200 MB/sec
  • Max I/O @ 6 Gb/s is 4 x 600 MB/s = 2400 MB/sec
• Sustainable disk rate ~ 100 MB/sec
### SAS HDD Bandwidth Possibilities

<table>
<thead>
<tr>
<th>Case</th>
<th>Class</th>
<th>Port 0</th>
<th>Port 0</th>
<th>Port 1</th>
<th>Port 1</th>
<th>Disk</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3G</td>
<td>300 MB/s</td>
<td>misc</td>
<td>misc</td>
<td>misc</td>
<td>100 MB/s</td>
<td>400+</td>
</tr>
<tr>
<td>2</td>
<td>3G</td>
<td>300 MB/s</td>
<td>300 MB/s</td>
<td>misc</td>
<td>misc</td>
<td>100 MB/s</td>
<td>700+</td>
</tr>
<tr>
<td>3</td>
<td>3G</td>
<td>300 MB/s</td>
<td>300 MB/s</td>
<td>300 MB/s</td>
<td>misc</td>
<td>100 MB/s</td>
<td>1000+</td>
</tr>
<tr>
<td>4</td>
<td>3G</td>
<td>300 MB/s</td>
<td>300 MB/s</td>
<td>300 MB/s</td>
<td>300 MB/s</td>
<td>100 MB/s</td>
<td>1300+</td>
</tr>
<tr>
<td>5</td>
<td>6G</td>
<td>600 MB/s</td>
<td>misc</td>
<td>misc</td>
<td>misc</td>
<td>100 MB/s</td>
<td>700+</td>
</tr>
<tr>
<td>6</td>
<td>6G</td>
<td>600 MB/s</td>
<td>600 MB/s</td>
<td>misc</td>
<td>misc</td>
<td>100 MB/s</td>
<td>1300+</td>
</tr>
<tr>
<td>7</td>
<td>6G</td>
<td>600 MB/s</td>
<td>600 MB/s</td>
<td>600 MB/s</td>
<td>misc</td>
<td>100 MB/s</td>
<td>1900+</td>
</tr>
<tr>
<td>8</td>
<td>6G</td>
<td>600 MB/s</td>
<td>600 MB/s</td>
<td>600 MB/s</td>
<td>600 MB/s</td>
<td>100 MB/s</td>
<td>2500+</td>
</tr>
</tbody>
</table>

- Case 2 is most common implementation is first-year SAS products
- Customer inputs are diverse:
  - Some will use only port 0
    - Port 1 for fail over
    - Port 1 unused, to fit in same socket as a SATA derive
  - Some are interested in cases 3 and 4 for dual access paths to data
  - Some want case 6 to avoid changing expander architecture for 6 Gb/s
- Note that case 4 and case 6 have identical DRAM bandwidth requirements
Cost within HDD

- ASIC die area → Part cost
  - Equalizer
  - DFE?
  - Higher speed RX, TX
  - State machines for training, tuning
  - Larger data FIFOs
- ASIC Package → Part cost
  - More DRAM interface pins
- Printed Circuit Board → PCB cost
  - Better connector?
  - Better PCB signal integrity
- Power → Trade off vs reduced reliability
  - Double DRAM bandwidth
    - Double DDR I/O pins?
    - Double interface speed?
    - Double number of DRAM chips?
  - More DRAM power
  - More FIFO power
  - Additional analog PHY power
HDD Cost Estimates for 2x PHY Rate + 2x DRAM

- Assume that PHY with Equalizer and modest DFE will grow by 50% of die area
- Define PHY cost adder as one cost unit, then:
  - Add 2 cost units for added DRAM bandwidth within the ASIC
  - Add 4-10 cost units to double DRAM bandwidth
    - Double data bus width, single DRAM
    - Double data bus width with second DRAM
    - Double speed of ASIC / DRAM interface
  - Add ??? If DRAM buffer capacity doubles with
  - Add at least 1 cost unit for ASIC package and PCB
  - Total cost add = 7x – 13x cost of faster PHY
- Conclusion is that the cost of migrating SAS HDDs to 6 Gb/s is dominated by DRAM bandwidth and size, and not by the cost of doubling PHY speed
Non-HDD SAS Applications

• For box to box, some interest in very long cables, requiring very complex electronics for recovering small, noisy signals
• Tape drives have roadmaps for speeds exceeding 3 Gb/s SAS
• Reduction of the number of 3 Gb/s physical links between nearby boxes
Only of use on the subset of drives used in SAS/SATA arrays

Cost of 6 Gb/s upgrade would be less than SAS
  – One port instead of two
  – One full data rate stream instead of 2-4 streams
  – Lower bandwidth than SAS, so smaller DRAM

Bandwidth cost

SATA cost sensitivity many times greater than SAS, overwhelming even the lesser cost for 6 Gb/s

The viability of mixed SATA/SAS arrays hinges on being able to use mainstream SATA drives
Partitioned Specs for 6 Gb/s SAS?

- PHYs are available with the capability of recovering signals from long cables after 30 db of attenuation
- HDDs in backplanes face different issues
  - Reflections
  - Cross talk
  - Attenuation << 30 db
- Creating separate specifications for backplanes and long cables seems desirable
  - Cable 6 Gb/s for box-to-box soon
  - HDD / backplane 6 Gb/s much later
Summary

• The transition from 3 Gb/s to 6 Gb/s presents an opportunity to alter the distribution of complexity between the HDD and the Expanders.
• SAS data streams dominate DRAM bandwidth usage:
  – SAS data ~ 6x disk data rate at 3 Gb/s
  – SAS data ~12x disk data rate at 6 Gb/s
• At 6 Gb/s, the cost of added memory bandwidth is much larger than the cost of a faster PHY block.
• Consideration of aggregating bandwidth in what is currently the expander block is recommended.
• Allowing different specifications (different levels of PHY complexity / cost) between HDD and box-to-box PHYs should be considered.