Low Level Simulation of Storage Systems

Tim Courtney

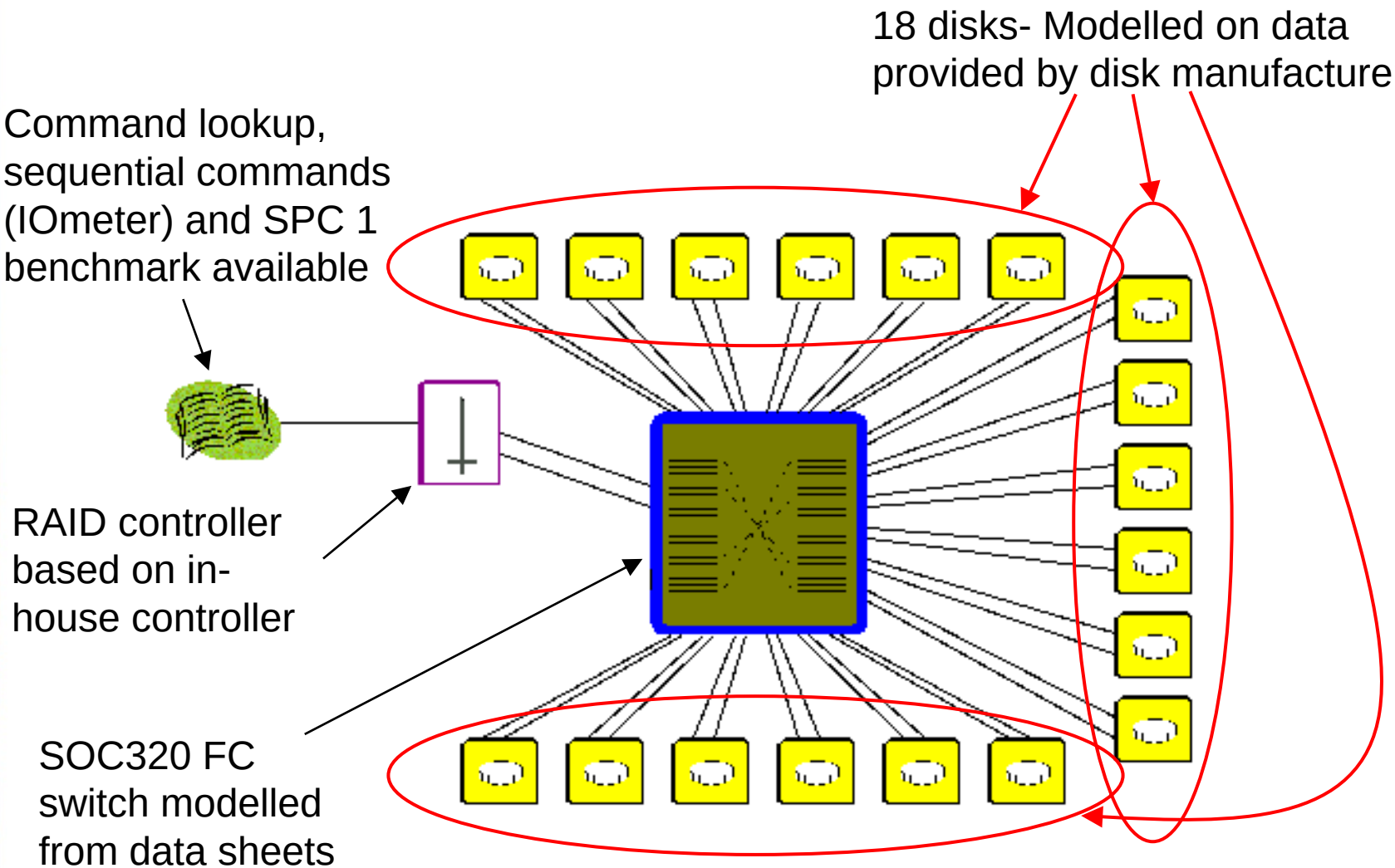
tim_courtney@xyratex.com

NASA/IEEE MSST 2004
12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies
The Inn and Conference Center
University of Maryland University College
Adelphi MD USA
April 13-16, 2004

Simulation Environment

- HASE
  - Discrete event simulator
    - Written by University of Edinburgh for teaching of processor architectures
  - Uses C++ for model components
    - Thread control and message passing dealt with by HASE environment
    - Allows control of what a message is and when they are passed
  - Provides GUI
    - See model connectivity
    - See animation of model with message passing
    - To be extended to allow “drag and drop” model design

The model


18 disks- Modelled on data provided by disk manufacturer
Command lookup, sequential commands (IOmeter) and SPC 1 benchmark available
RAID controller based on in-house controller
SOC320 FC switch modelled from data sheets

Communication abstraction

- Modelling every fibre word too time consuming
- Statistical modelling insufficiently accurate
- Active word only abstraction
  - IDLEs not transmitted
  - Only single ARB needed
  - Assume model initialised & running
    - Undergoing change to model initialisation also
  - Aim to accurately represent all the steps in FCP
- Allows many fewer events to be modelled whilst maintaining low-level accuracy

Initial Findings

- Uses 10 disks configured as two 4+1 RAID 5 sets
- Based on IOmeter style sequential I/O requests
  - Allows comparison with test data from RAID controller
    - Test of early silicon of controller
  - Test data assumes cache miss and so cache not modelled
    - Will be added and active in future models
  - Gives good comparison to test results (see next slide)
- Easily extended to SPC 1.0 benchmark I/O profile once comparison data is available
  - No comparison data available
  - Useful data flow in order of 110Mbyte/s bottlenecked by physical disk platter I/O appears reasonable

IOmeter style results

- Operation performed is full strip read

Strip Size (Kbytes)	I/O per second from model	I/O per second from test data	% Error
8	13256	12500	4·4
16	7636	7500	1·8
64	2139	2200	-2·7

- Further refinement needed
  - Obviously too much dependence on the size of the operation
    - Possibly due to incorrect fibre channel L_PORT characteristics being used
  - Undergoing correction to match model to data that is available

On-going Work

- Refinement of model
  - Include cache
  - Extend to multiple controllers
  - Generate SPC 1·0 benchmark figures from model
    - Compare with test data once available
  - Addition of errors to the data links
- Experimentation
  - Performance implications of variation in cache size
  - Performance implications of error rates with move to optical connectivity (10Gbit/s and beyond)
  - Possible new topologies to allow reduction in cost of optically connected solution

Any Questions?

E-mail: tim_courtney@xyratex.com
fcheval1@inf.ed.ac.uk