ICF-connected Disks

Jerry Leitherer
Scalable Systems
I/O Architecture

Advice & Contribution
Prof Alok Choudhary, Syracuse University
Chris Dodd, Intel
Carl Pick, Genroco
Simon Shapiro, Intel

Problem Statement

- Type 2 Clustered Server Issues
  - Distributed client-server limitations
    - Transaction Rate
    - I/O Op Rate
    - Bandwidth
    - Connectivity
  - Load balancing
    - Process Migration
    - Data Migration
  - Availability
    - Alternate Peripheral Paths
Disk Access Distribution

- Client-Server Model
  - host node entirely involved in all accesses

- Direct Access Model
  - split IOPH driver
  - host node only involved in access management

Stream Flow - Node

- Distributed client-server limitations
  - Bandwidth assuming 4 data transfer paths
    - PCI bus limit ~66MB/s
    - 66MHz yields 266MB/s peak theoretical, ~170MB/s
    - widen bus to double transfer rate, not to op rate
    - OPB limit ~90 MB/s
    - new arbitration scheme and buffer plan

353/588=60%
opb=25%
Database Architectures

Shared Nothing

Shared Disk

Shared Everything

Disk Sharing

Multi-initiator Sharing

Multi-port Sharing

Psuedo Sharing

All three models may be referred to as Multi-host
Problem Statement - Connect

- Distributed server limitations
  - Connectivity for CBP-2
    - 3 PCI slots, 2 PCI buses
      - soil: 1 on board, 3 plug-ins
      - 30m, 1x15=45 devices
      - tcal: 3 plug-ins
      - .5km, 3x120=360 devices
  - Connectivity for PCI
    - 3 x 40 MB/s controllers
      - Aggregate bus performance shared
      - Distance limit
    - 1 PCI per OPB
      - 8 loads per Orion bus

Access Models

Uniform Access Model
- given n compute nodes, the remote workload seen by an i/o system is (n-1) λ / n, where
  - λ is per node i/o access rate
- each node's resources heavily used to serve other nodes
- access model likely for balancing workload, eg - video/medial server or query engine

Completely Local Model
- no ICF i/o traffic
- complete knowledge of access patterns must be known at design
  - time so data is partitioned
- migrate task or function to data
- load balancing impractical
Solution Philosophy

Apply core competencies
- interconnect
- scaling
- analysis

Focus on scalable building blocks
- both HW and SW

Maintain the technology treadmill
- processors
- ICF and NICs
- peripherals

Keep time-to-money near desktops
- silicon to system in 2-3 quarters

Solution Goals

A fully scalable, balanced architecture
- processors and memory (capacity and bandwidth)
- I/O (using the ICF in concert with PCI)
  - peripheral interface balance
  - bypass bottlenecks
  - dynamic path load balancing
  - RAS
    - modularity

Minimum number of mechanisms
- common infrastructure for all board types
- simplified system management to improve ease of use
- commodity parts wherever possible
Key Insight:
Decoupled 3rd-party I/O

Traditional client-server among cluster
- heavyweight OS-to-OS communication from nodes needing data to nodes with I/O devices containing data
- 2-party (host-peripheral) data transfer for disk access followed by second 2-party transfer for cluster ICF access

Traditional host-controlled I/O with DMA
- processor and memory bus tied up for I/O protocol setup and long path I/O request and controller memory operations

Approach overlooked 3rd-party device access with decoupled I/O protocol processing
- opportunity to integrate I/O devices into ICF
  - no need for full OS on node
  - easier to exploit light-weight protocols

I/O Access Packaging

Access to peripherals via the ICF

I/O control logic should be just what’s necessary
- access protocol handling
- device command issue
- speed-match buffering
- no full OS support

Leverage PCI I/O interfaces
- low cost
- standard from factor
- standard platform administration
ICF-disk Description

- use cluster/server inter-connection fabric (ICF)
  - CBP-2 ICF
  - Type 2 ICF (ATM or SONET)
  - Type 1 LAN as ICF

- attach disk subchannels to a processor cluster
  - switched virtual channels operating over ICF network
  - includes fibre channel or scsi-like bus
  - connects paths between
    - disk storage devices
    - processor memory via ICF

TURBOstor Architecture- proposed*

*Based upon "U. S. Pat. No. 5,420,984"
ICF-disk Node Stream Flow

Benefits

Better performance
- analytical model says "better", but need simulation

Fully integrated RAS architecture
- redundant paths for peripheral access
- system configuration can be determined during boot
- devices could be DMI-ready
- system management simplified

Elegant solution for shared-disk DBMS
- opportunity to support locking protocol at the device
**Additional Benefits**

Higher degree of scalability
- i/o services
- distributed resources
- load balancing strategies
- fault tolerance
- recovery schemes

**Stream Flow - base case**
Stream Flow - ICF-disk

SPP ICF-disk Project Summary

Project Definition
- Demonstrate and productize Disk I/O using SSP interconnect

Key Deliverables
- PCI Card with ICF Interface (Genroco)
- Driver and Firmware for PCI Card Support
- Disk and Path Management Software for 3rd Party Subchannels

Target Market/Expected Benefits
- All SPP OEMs
  ✓ Dramatically reduce total system cost (memory and disk)
  ✓ Increase use of SSP interconnect for Disk I/O
  ✓ Increase system performance by improving I/O balance

Investment Estimate

<table>
<thead>
<tr>
<th></th>
<th>1995</th>
<th></th>
<th>1996</th>
<th></th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H/C</td>
<td>S</td>
<td>H/C</td>
<td>S</td>
<td>$</td>
</tr>
<tr>
<td>ICF-disk</td>
<td>4</td>
<td>0.41</td>
<td>4</td>
<td>0.72</td>
<td>1.13</td>
</tr>
</tbody>
</table>

Page 10
Project Outline

Objectives
• Stage 1 - bandwidth
• Stage 2 - connectivity (3rd party)
• Stage 3 - distributed device subsystem
  - shared disks
  - hierarchical storage

Milestones
• Architecture Definition Paper - 7/31
• Simulation Evaluation - 10/30
• Embodiment - 1q96

Embodiment Stages

Hardware - Genroco
• HIPPI & FC on PCI - today
• FC to NIC, PCI - 1q96
• 2FC to NIC, PCI - 3q96

Software - Intel, ?
• Local Disk via ICF - 1q96
• 3rd Party Disk via ICF - 3q96
• Shared Disk - ?
Intel Scalable, Multi-Server Architecture - CBP-II

Interconnect: 3-D mesh - 400/800 MB/s via mem bus
Nodes: 2xP6, 530MB/s mem bus
Storage: PCI-based FCAL, SCSI-fast20 (disks or arrays)
Intrax/Divr: PCI-based ATM, FC, HIPPI, FDDI, E-net100

Z-Ring (Shared MRC) Topology

Peak (sustainable) Bi-Directional Bandwidth
800 MB/sec (700)
600 MB/sec (760)
800 MB/sec (500)

Full duplex 200 MB/sec (390)
Memory 533 MB/sec (590)

Two Nodes on each Keastrel Board

Intel Confidential
Page 12
TURBOstor Architecture - today*

* "U. S. Pat. No. 5,420,984"
The Issues

Shared peripherals architecture
Device access protocols over the ICF
RAS implications
Performance
Packaging
Cost and cost/performance

Memory Board (Ram Disk)

DRAM or Flash
- even better, Flash
  backing up DRAM
- could aide warm restart

Potential uses
- paging store
- transaction logs