FAULT TOLERANT SYSTEMS

http://www.ecs.umass.edu/ece/koren/FaultTolerantSystems

Part 18

Chapter 7 – Case Studies

Introduction

- Illustrate practical use of methods described previously
- Highlight fault-tolerance aspects of six different computer systems
  - NonStop – Tandem/HP
  - Stratus
  - IBM G5
  - IBM Sysplex
  - Intel's Itanium
  - Intel's Xeon
  - Oracle's and Fujitsu's SPARC
- All have various fault-tolerance techniques implemented in their design
- High-level description - no comprehensive low-level details

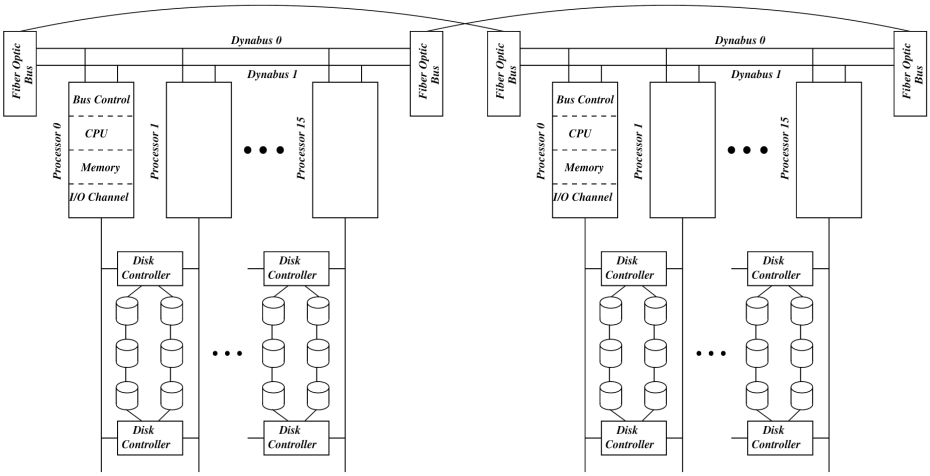
NonStop Systems - Principles

- Goal: online transaction processing
  - reliable real-time operation
- Design principles
  - Modularity: HW & SW modules of failure, diagnosis, service and repair
  - Fail-fast operation: modules works properly or stops
    - » Self-checking HW
  - Single-failure tolerance
  - Online Maintenance: disconnect/connect with no disruption

NonStop Systems – Architecture (1)

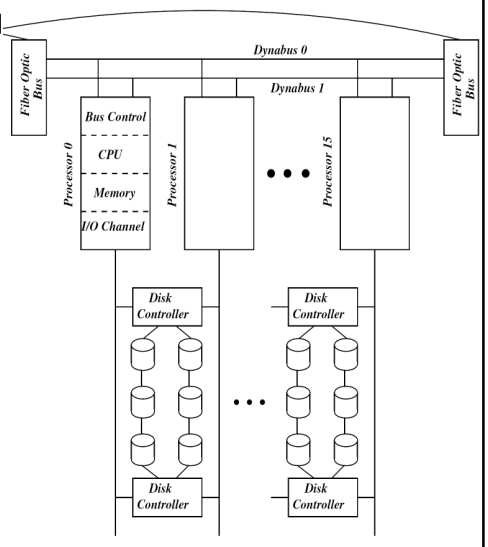
- Cluster of computers – up to 16 per cluster
- Computer: CPU, memory (own OS copy), bus, IO
  - Extensive error checking CPU for fail-safe
    - » parity check & prediction
    - » Serial-scan registers for testing
  - Hamming code in memory – SEC/DED
    - » Parity on address
    - » Cache retry for transient errors
    - » Spare module for permanent errors
  - Parity checking in datapath
    - » Parity prediction for simple arithmetic operations, e.g, add
    - » Multiply – redo with operands exchanged and one shifted
      - Recomputing with shifted operands – also detects permanent faults

NonStop Systems – Original Architecture



NonStop Systems – Architecture (2)

- CPUs communicate through dual Dynabuses
- Disks – dual ported controller connected to 2 CPUs
  - Dual ported IO devices & dual ported controllers – 4 paths
  - Data parity checked + watchdog timer detects controller stops
- Two power supplies + battery backups
- Disk mirroring – 8 paths for read/write
  - Data checksum for error detection – mirroring for recovery



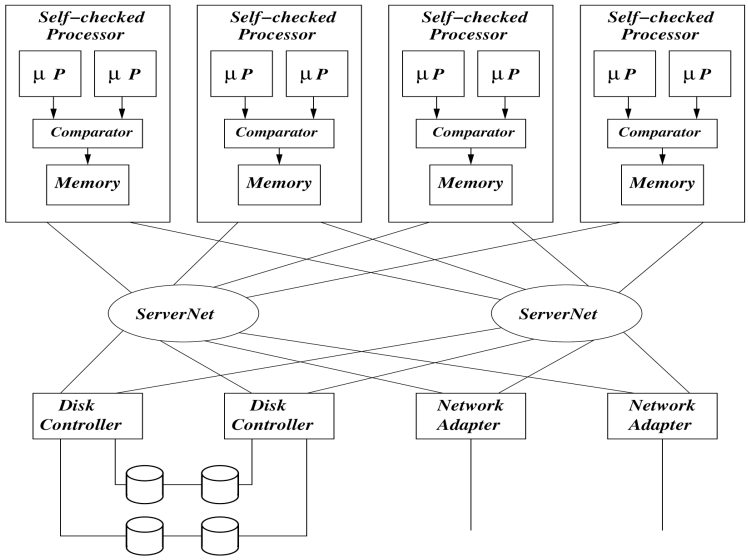
NonStop - Maintenance and Repair Aids

- Automatically detect errors, analyze and report to remote support centers
- Maintenance processor
  - Collects failure info and report to remote center
  - Reconfigure system in response to failures
  - Capable of fault diagnosis using a knowledge database
  - Monitors sensors for voltage, temperature, fans etc
- Diagnostic unit in each CPU
  - Monitors & reports to Maintenance processor
  - Can force CPU to single-step and can access scan paths
  - Can generate pseudo-random tests and run them

NonStop Systems - Software

- Process pairs – primary fault-tolerance scheme
  - OS generates a backup process for each new primary process
  - Checkpoints taken during execution sent to backup process
  - If primary fails, OS orders backup to start
  - Disk access also through primary/backup process pair
- CPUs check on each other
  - "I am Alive" messages/second to all including itself
  - If missing – CPU declared faulty & communication stopped
- Transaction monitoring module to guarantee ACID
  - Atomic – all or none database (DB) updates executed
  - Consistent – successful transaction preserves DB consistency
  - Isolated – events within a transaction isolated from other transactions
  - Durable – once transaction commits, its result survives failures
- Software failures – consistency tests in each module, upon a failure detection processor halted and the backup started

NonStop Systems – Modified Architecture

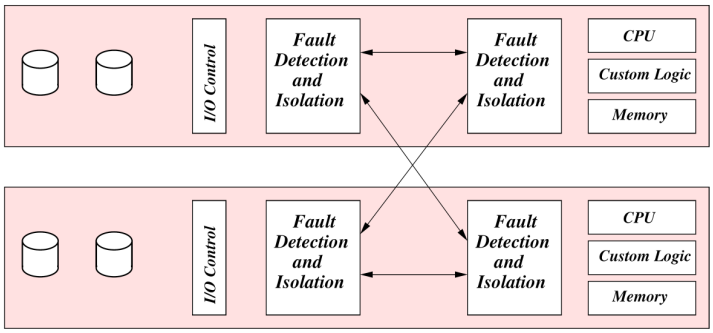


NonStop Architecture Modifications

- COTS μProc instead of custom-designed
  - No self-checking circuits to support fast-fail
  - Lockstep operation of pairs – memory op executed only if both requests identical
- Packet-switched *ServerNet* – 2 independent fabrics
  - High bandwidth and low latency
  - Better support for detection & isolation of errors
    - » CRC per packet – checked by each router – flagged if error
- Lockstep no longer practical
  - Multiple clocks on chip & asynchronous interfaces
  - Variable voltage/frequency for power management
  - Soft error handling
  - Multiple cores – failure in one will disrupt all
- Loose lockstep – only compare outputs of IO operations
- Allow TMR configurations

Stratus Systems

- Similarities to NonStop
  - Units replication – avoid single points of failure
- Mainly relies on hardware redundancy
  - Use pair-and-spare (2 CPUs in lockstep)
  - Upon pair mismatch – it declares itself faulty
  - Only IO outputs compared



Stratus Systems - Details

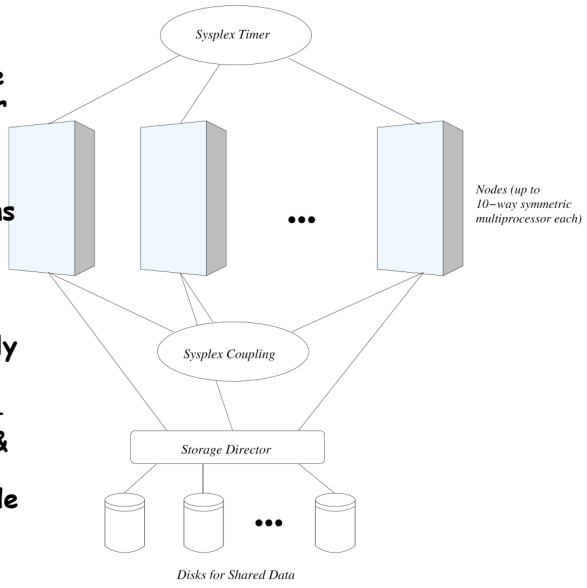
- TMR configurations allowed
- Memories also duplicated (unlike NonStop)
- Device drivers hardened
- Sanity checks on inputs to IO devices
- Upon system crash – automatic reboot
  - Dump memory to disk for analysis
- Report faults to remote support center
- If permanent fault detected – ship hot swappable parts.

IBM G5 Processor

- Fault tolerance in CPU, memory and IO to recover from transient faults
  - CPU and IO – replication
  - Hardware support for rollback recovery
  - Memory – error detection and correction codes (ECCs)
  - CPU: I and E units duplicated & lockstep; R unit stores checkpointed state to allow rollback – registers use ECC
  - Write to L1 cache also written to L2 – serves as backup
  - Memory and L2 use (72,64) SEC/DED Hamming code
  - Address bus uses one parity bit per 24 bits
  - Memory scrubbing to prevent accumulation of transients
  - Spare DRAM to replace mulfunctioning memory chip
  - L1 uses simple parity
- Responses to errors
  - Local errors in registers and L2 – corrected using ECC
  - Errors in L1 – detected (parity) and line replaced from L2
  - Errors in processor – instruction retry
  - If recovery fails – checkpoint data transferred to spare processor

IBM Sysplex

- Up to 32 nodes
  - Each node – single or multi-processor
- Shared storage – multiple disk systems
- Redundant connections to disks
- Storage uses coding or replication
- Node send periodically "I am alive"
- Upon a node failure – try to restart node & restart applications executed on that node



IBM Sysplex – Automatic Restart Manager (ARM)

- When a node fails – ARM takes charge
  - Balance the load when migrating processes
  - Check whether failed nodes are down – avoid duplicates
  - Not allow node that lost access to global state restart – disallow duplicates
- ARM support hot standby mode
  - Primary and secondary for given application
  - When primary fails, secondary takes over immediately

Intel's Itanium

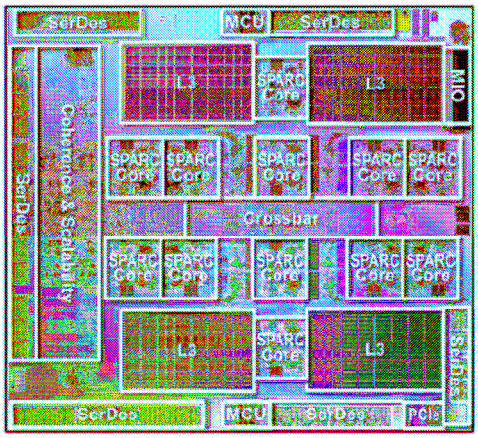
- 64-bit Explicitly Parallel Instruction Computer (EPIC) – VLIW
- Extensive use of ECCs
  - L1 (I & D): byte-parity. Upon error – cache invalidated
  - L2: (72,64) SEC/DED Hamming for data, parity for tag
  - Same for L3
- If an error is not hardware-correctable
  - If error containment required – bus reset
  - If not – Machine Check Abort
- Error handling done layer by layer
  - Hardware layer
  - Processor abstraction layer
  - System abstraction layer
- Uncorrectable erroneous data – marked as such (*data poisoning*)
  - At the L2 level
- Itanium used in recent designs of NonStop and other fault-tolerant systems

Intel's Xeon E7 (servers)

- Up to 10 cores and 20 threads
- Support for self-monitoring and self-healing
- The HW FT features communicate with the SW/OS
  - Inform OS after error correction for logging and analysis (e.g., identify failing memory chips that should be replaced)
  - OS may abort a task upon an uncorrectable error
- Main emphasis on memory - claimed to be responsible for most errors
- ECC in memory, caches and registers
  - Memory scrubbing – correct or tag as "poisoned" to prevent spreading of error
  - Memory thermal throttling – reduce access rate when overheating
  - Selective memory mirroring
  - Spare memory unit – failing unit's contents copied to spare
- System interconnect (cores and I/O): CRC detecting burst errors of up to 8 bits – retry if error detected
  - Can reduce the width to half if error persists
- Migrating workload to a spare core if a CPU or its memory fail

Oracle's SPARC M6 (2013)

- Parity – Registers & Busses
- CRC – external links (retry)
- Cache units
  - ECC for data
  - Parity for address
- Memory
  - ECC
  - "Scrubbing" to prevent accumulation of soft errors
  - Word line failure – OS retires page
  - Bit line failure – use spare column



Fujitsu's SPARC X+ (2013)

- INT/FP registers – ECC
- Other registers - parity
- Cache units
  - ECC for data
  - Parity for address
  - Dynamic degradation
- ALU – Residue
- Instruction retry
- Core dynamic degradation