zSeries hardware reliability

The RAS features that put zSeries hardware
in a league of its own and their benefit for Linux

Malcolm Beattie, IBM UK

beattiem@uk.ibm.com
Linux Technical Consultant
IBM EMEA Enterprise Server Group

Motivation

We know zSeries hardware is reliable.
We know zSeries RAS leads the world.

Other hardware platforms are improving too.
What makes zSeries hardware so much better?
Why is zSeries hardware *so* reliable?
How does this benefit Linux on zSeries?

This presentation fills in some of the answers.

Introduction

- We mostly discuss hardware, not software
- The RAS features that apply to all O/Ses
- We focus on zSeries hardware
  - latest and best of a rich reliability history
  - fine detail is for z900, other zSeries may differ
- We cover three different levels
  - microarchitecture: MCM and chips
  - higher level: power, cooling, call-home
  - I/O features: STI, channels and cards

Acknowledgment, Reference and Disclaimer

- Almost all the content of the first part of this presentation is taken from the paper:

  RAS design for the IBM eServer z900,
  L.C. Alves et al,
  IBM J Res & Dev, Vol 46, No 4/5, July/Sept 2002
- I have produced “presentation” format from the original “research paper” format
- Any errors that have crept in are my fault
  - I am not a hardware RAS expert
  - I am not a microarchitecture expert
  - but I hope to communicate what I have learned

Microarchitecture reliability features

- MCM and on-chip design features
- Goal: Continuous Reliable Operation (CRO)
  - *Continuous*

    run the customer's operation without interruption caused by errors, maintenance, or change in server hardware or Licensed Internal Code (LIC), ...
  - *Reliable*

    ...while ensuring error-free execution and data integrity.

The seven building blocks of CRO strategy

- error prevention
- error detection
- recovery
- problem determination
- service structure
- change management
- measurement and analysis

Processor (PU) reliability

- Each PU has dual instruction/execution engines
- The two engines execute each instruction in lockstep
- Output is compared at each checkpoint
  - On mismatch, PU retries from checkpoint
  - On continued failure, dynamic sparing occurs
- Following PU sparing detail is taken from:

  RAS strategy for IBM S/390 G5 and G6,
  M. Mueller et al,
  IBM J Res & Dev, Vol 43, No 5/6, Sept/Nov 1999
- Applies to zSeries too

Dynamic Sparing of a PU

- At least one spare PU is available
  - on any G5, G6, z900, z800, z990, z890...
  - ...except for fully populated z800 or z890 (4-way)
- Failing PU is fenced
- Clock chip, other PUs and SE are signalled
- SE and SAP combine to locate spare PU
  - If failing PU is a “master” SAP and a spare PU is not available, an active CP is reassigned as master SAP
- Target spare begins “self-initiated brain transplant”
- Target spare takes identity of fenced PU
- Begins executing at same instruction
- Transparent to O/S

L1 cache protection

- parity-checked
- store-through to L2
- refresh capability (get valid copy from L2)
- supports cache-line delete (via registers)
  - compartment delete
  - one-line delete
- supports cache-line sparing
  - “fuse” relocation technology
  - replaces failing word lines with spare word lines
  - SE replaces defective word line at next FPOR

L2 cache and directory protection

- Defective cache-line relocation capability
- Error-correction code (ECC) protection
  - L2 cache: (72, 64) ECC
    - i.e. each 64-bit data word has 8 check bits
  - Directory address field: (25, 19) ECC
  - Directory ownership field: (11, 5) ECC
- Single-error correction; double-error detection
- L2 cache can delete any combination of lines

Action taken by L2 to handle errors

- PU fetches a double line (256 bytes) from L2
- ECC logic
  - corrects correctable errors (CEs)
  - detects uncorrectable errors (UEs)
- In either case
  - failing location is saved and logged later
  - hardware purges the cache line

Handling a correctable error in an L2 line

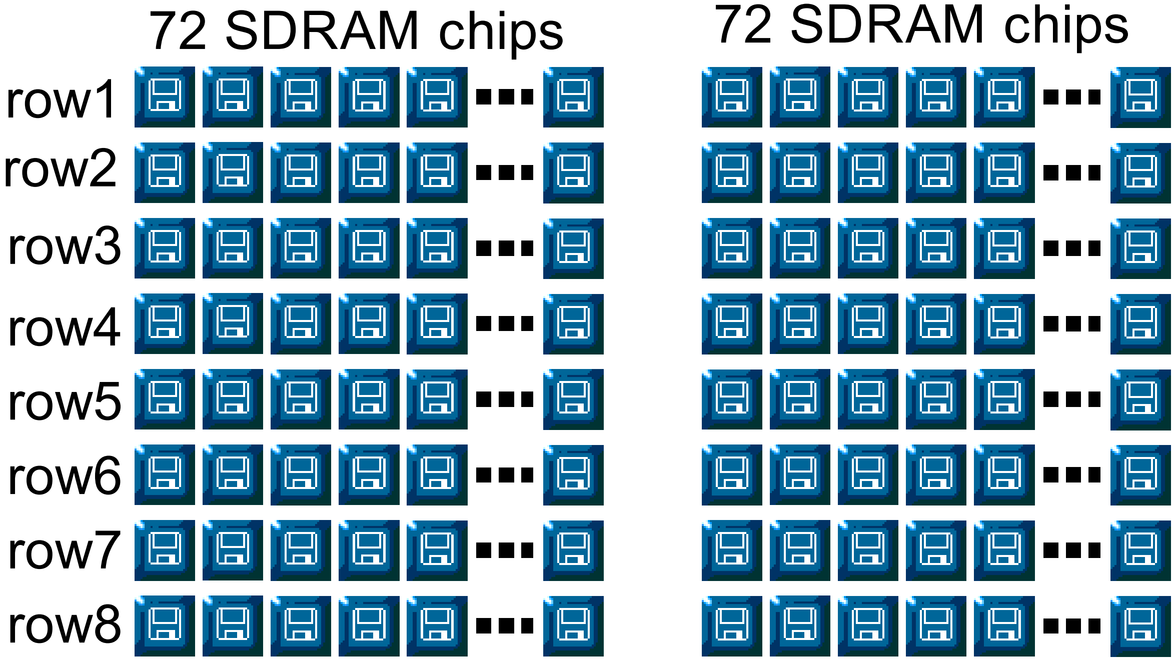
- Data is corrected and sent back to main storage
- Compare with info about previous failure
  - each cache area traps such information
- Is the new failure in the exact same bit as before?
  - if so, line is purged and deleted
  - it will not be used again

Handling an uncorrectable error in an L2 line

- PU is notified that data is invalid
- If readonly or unchanged,
  - just invalidate it
  - correct data can be retrieved from main storage
- If changed,
  - send it to main storage and mark permanently bad
- OS is notified of the failed storage address
- the cache line is purged and deleted
- this schedules a “fuse” repair for next POR
- At next POR
  - fuse reallocates cache line
  - ABIST (Array Built-In Self Test) verifies repair

Error correction/detection in main memory

- z900 has 4 memory cards (z990 has 2 per book)
- Each has 8 rows of 144 SDRAM chips
- Each store goes to one row, two bits per chip
- This splits as two 144-bit data words



ECC in main memory

- To protect the data, a (140,128) ECC code is used
  - i.e. 128 data bits and 12 check bits
- This code corrects any of the following errors:
  - any single-bit failure
  - any single symbol failure (2-bit failure in one chip)
  - i.e. if a DRAM chip is completely broken and gives unpredictable results, the hardware can correct it
- The code can also detect when 2 of the 72 DRAM chips in the same data word are broken
- Only 140 of the 144 bits are needed
  - 2 chips of each 72 are used for sparing (32 per card)

Address protection and failure isolation

- A (144,132) code was specially designed
  - Code accommodates 132 data bits
  - But only 128 bits needed for storage
- Remaining 4 bits available for use
  - 2 used for memory address protection
    - protects against fetch from erroneous location
  - 2 used for failure isolation: special bit patterns for
    - cache or non-memory failure: “cache special UE”
    - memory error: “memory special UE”
    - memory store interface error: “interface special UE”

Error avoidance

- DRAMs with many defects are spared
  - even though ECC can correct 2 bits per DRAM
  - prevents defects in data word lining up with other defects
- DRAM sparing can be done at POR or dynamically
  - At FPOR, memory controller runs self-test
  - stores and fetches fixed and random patterns
  - accumulates error counts
  - special patterns used across all 144 bits
  - these detect simultaneous errors and exercise ECC logic
- DRAMS with high failure counts can be spared
  - DRAMs with high failure counts are not used for spares
  - spares can be spared too!

Memory scrubbing and key protection

- Constant memory scrubbing runs in background
  - fetch and store from all addresses in memory
  - correct errors and store back into memory
  - corrects “soft” errors from alpha particles etc.
- Error counts are accumulated
  - DRAMs with high counts are spared
  - data written to both original and spare DRAM
  - after scrubbing pass, spare DRAM used instead of old
- Key protection
  - three copies kept of every key (plus parity)
  - All three copies read and compared on each access

Protecting other parts of memory subsystem

- Configuration array
  - translates physical address to absolute array address
  - seldom written during normal server operation
  - (16,10) ECC (corrects 1-bit errors; detects 2-bit errors)
  - two extra “valid” bits (for redundancy) on each entry
  - each “valid” bit has an associated parity bit
  - entire array scrubbed every 8000-16000 memory cycles
- L2 cache-memory controller interface
  - parity protection of commands/addresses
  - nonzero null pattern on command bus
  - 1 bit error while idle cannot look like a valid command

I/O Subsystem RAS features

- All cards can be replaced concurrently
  - except for FIB and CHA in “compatibility cage” on z900
  - “Repair & Verify” (R&V) function on SE guides the ~~CE~~ Service Representative through the process step by step
- ESCON-16 port sparing
  - Each ESCON-16 card has 16 ESCON ports
  - If a port fails, Service Representative uses R&V on SE
  - SE blinks LED of defective port; I/O cable unplugged
  - SE blinks LED of spare port; I/O cable replugged
  - Original CHPID now refers to this port
  - SE CHPID mapping table updated automatically
  - 1 port always left “spare” but any reserved port usable

Power and cooling

- Power/cooling subsystem comprises
  - Central Bulk Power Assembly (BPA)
  - Point of load direct current assembly (DCA)
  - Air-moving devices (AMDs)
  - Modular Cooling Units (MCUs)
- Each critically important in achieving high availability

Central Bulk Power Assembly (BPA)

- BPA is duplicated, each powered from own A/C cord
- BPA converts 3-phase line into 350V distribution bus
- Two such buses produced by two BPAs
- Each of these is capable of powering entire server
- Active/Active mode: power load shared when both work
- Additional robustness from active phase switching
  - If one phase lost; BPA switches to single phase
  - Will run indefinitely thus for most server configurations
  - Most power losses are single phase transient dips
  - Hence server is impervious to almost all power line disturbances, even without a battery or UPS
- BPAs provide power with “cross-coupled redundancy”

DCAs and AMDs

- Point of load direct current assembly (DCA)
  - DCAs provide precise voltages for logic functions
  - DCAs (fed by BPAs) are (N+1)-redundant
  - Redundancy is “active”: share load when working
- Air-moving devices (AMDs)
  - Blowers with intelligent speed control devices
  - Air cooling and heat removal for everything except MCM
  - CPC cage and I/O cage have two blowers in parallel
  - Power cage has two blowers in series
  - Active redundancy: one fails; the other speeds up
  - But “sped-up blower increases acoustic noise beyond the specification for a fault-free server” (i.e. it's loud)

Modular Cooling Units (MCUs)

- MCU is the refrigeration unit which cools the MCM
  - the MCM (Multi-Chip Module) is the heart of the system
- MCU comprises modular refrigeration unit, condenser, intelligent controller, cooling fan, evaporator, cavity humidity sensor, board heater block, ………
- MCUs are duplicated
  - redundancy mode is “standby”
  - operate separately to lengthen compressor service life
  - scheduled switchover each 160 hours
  - failure of the running MCU switches to the standby

Concurrent replacement/repair

- All parts of BPAs are concurrently replaceable
  - Nine Field Replaceable Units (FRUs) in each
- All parts of DCAs are concurrently replaceable
- All parts of MCUs can be concurrently replaced or repaired except for the evaporator
  - evaporator has two independent loops, one to each MCU
- DCA and parts of the MCU do
  - fault isolation
  - error logging
- Microcode ensures repair/replacement is correct
  - microcode level of each FRU auto-checked/updated

Alternate SE auto switchover

- Two SEs: one primary; one alternate
- Both can communicate at the same time with
  - each other
  - the server
- communication detects problems
  - failing SE is fenced
  - alternate SE takes over automatically
  - “soft” switch on HMC can initiate SE switchover
  - no physical presence at server necessary (cf. G5/G6)

Remote support subsystem

- First error data capture (FEDC)
  - Gather pertinent data as soon after a failure as possible
- Error logs stored in HMC or SE and queued to RETAIN
  - IBM REmote Technical Assistance Information Network
- FEDC filter
  - auto-downloaded list of files to send to RETAIN
- Each HMC now acts as a “phone home” server
  - SE can contact multiple HMCs to phone home
  - If all fail, it queues up files for later transmission
- Transmit System Availability Data (TSAD)
  - Call-home done weekly for transmission of data
  - power status, SE status, microcode status, recovery info

Channel SubSystem (CSS) RAS features

- CSS hardware reliability
  - Generic: STI and channel independence
  - Common I/O card platform (CIOP)
    - details follow
  - ESCON-16 card
    - details earlier
- Multipathing
- Measurement (performance/activity)
- Administration

Common I/O card platform (CIOP)

- For details, see

  IBM eServer z900 I/O subsystem,
  D.J. Stigliani, Jr. et al,
  IBM J Res & Dev, Vol 46, No 4/5, July/Sept 2002
- zSeries I/O programming model
  - SSCH, subchannels, CCWs, IDAWs
  - initiate, terminate, report status, present interrupt
- PCI I/O model is memory mapped
- CIOP bridges these and is currently used for
  - FICON-Express
  - OSA-Express
  - Crypto: PCICC, PCICA, PCIXCC

Common I/O card platform (CIOP)

- PCI-to-STI adaptation bridge with extensive RAS features
- Dual cross-checked PowerPC microprocessors
  - 333MHz processors, 66Mhz memory controller interface
  - Run in lockstep, checked cycle by cycle
- Dual cross-checked PowerPC bridges
  - combined L2 cache, L3 memory controller
  - up to 128MB ECC memory
  - 2MB on-card flash memory
  - master and checker controllers: checker compares memory and PCI operations done by master with what the checker “would do”

Common I/O card platform (CIOP)

- Uniquely designed STI-PCI bridge ASIC
  - supports 32-bit/64-bit PCI cards at 33MHz or 66 MHz
  - data buffer controller with up to 128 MB ECC memory
  - two independent data-mover queue (DMQ) engines
    - move addressable zSeries storage to SDRAM or PCI
  - High-priority storage requester (HSPR) engines
  - PCI 2.2 interface with 2 x 256-byte buffers
- MMU includes many features
  - bus arbitration, multi-level priority, self-refresh, ECC, bus-locking, full-page burst, memory test engine
- two PCI adapters for two standard PCI cards

CSS multipathing

- up to 8 channel paths per device
- SSCH (Start SubChannel) instruction initiates I/O
  - just specifies device (subchannel) number (sorta...)
  - Service Assist Processor (SAP) finds a path which is
    - installed
    - administratively enabled
    - operational
  - SAP initiates the I/O
    - may reconnect over different channels if necessary

CSS I/O measurements

- CSS has built-in support for I/O measurement
  - performance and activity
- Each subchannel (device) can be measured
  - can have a Subchannel Measurement Block
  - records counts, connect time, disconnect time, control unit queuing time, busy time, ...
- Low overhead
- Good support by O/S and additional software
  - z/OS RMF
  - z/VM Performance Toolkit
  - ISV software available too

- Questions?

Thank You

Malcolm Beattie
beattiem@uk.ibm.com