

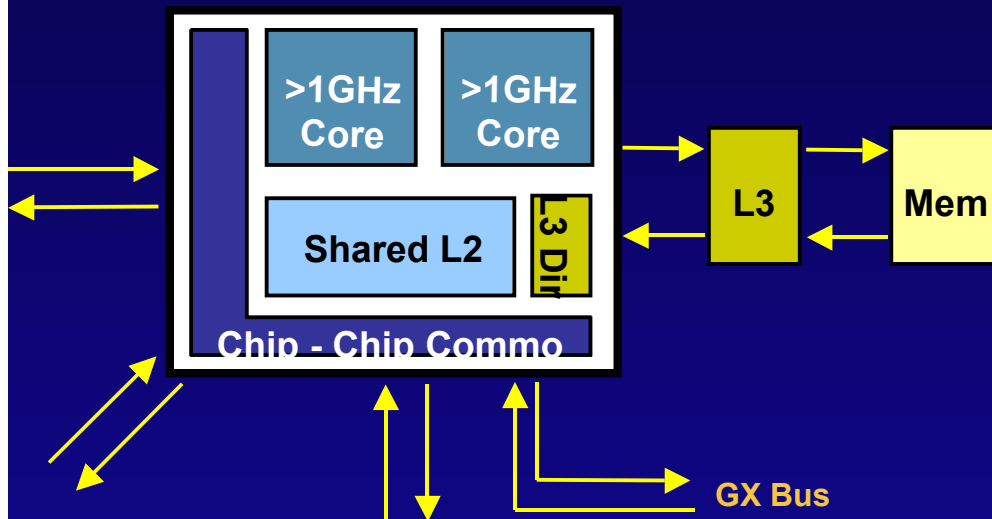
# POWER4 Systems: Design for Reliability

Douglas Bossen, Joel Tendler, Kevin Reick

IBM Server Group, Austin, TX

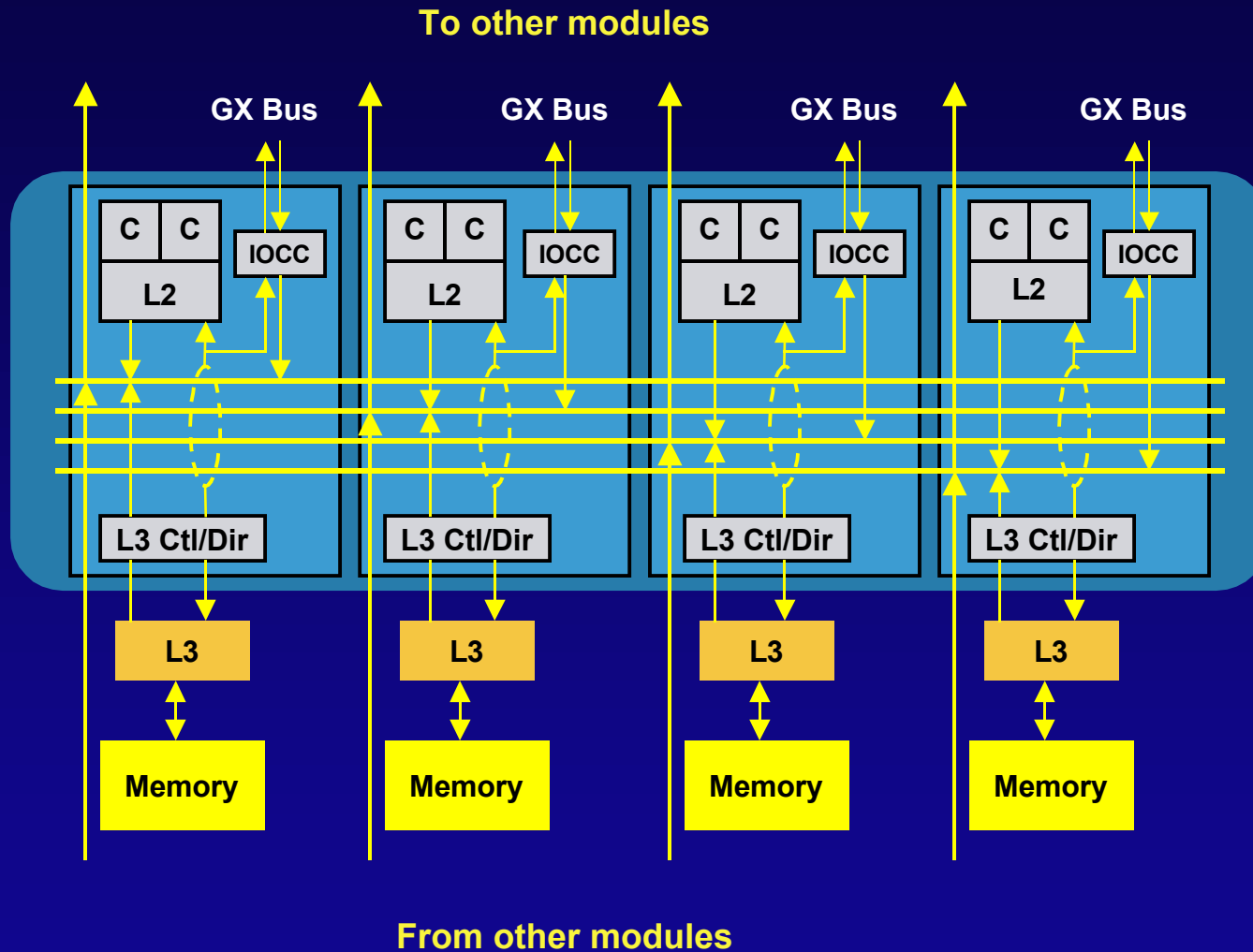


# POWER4 Microprocessor



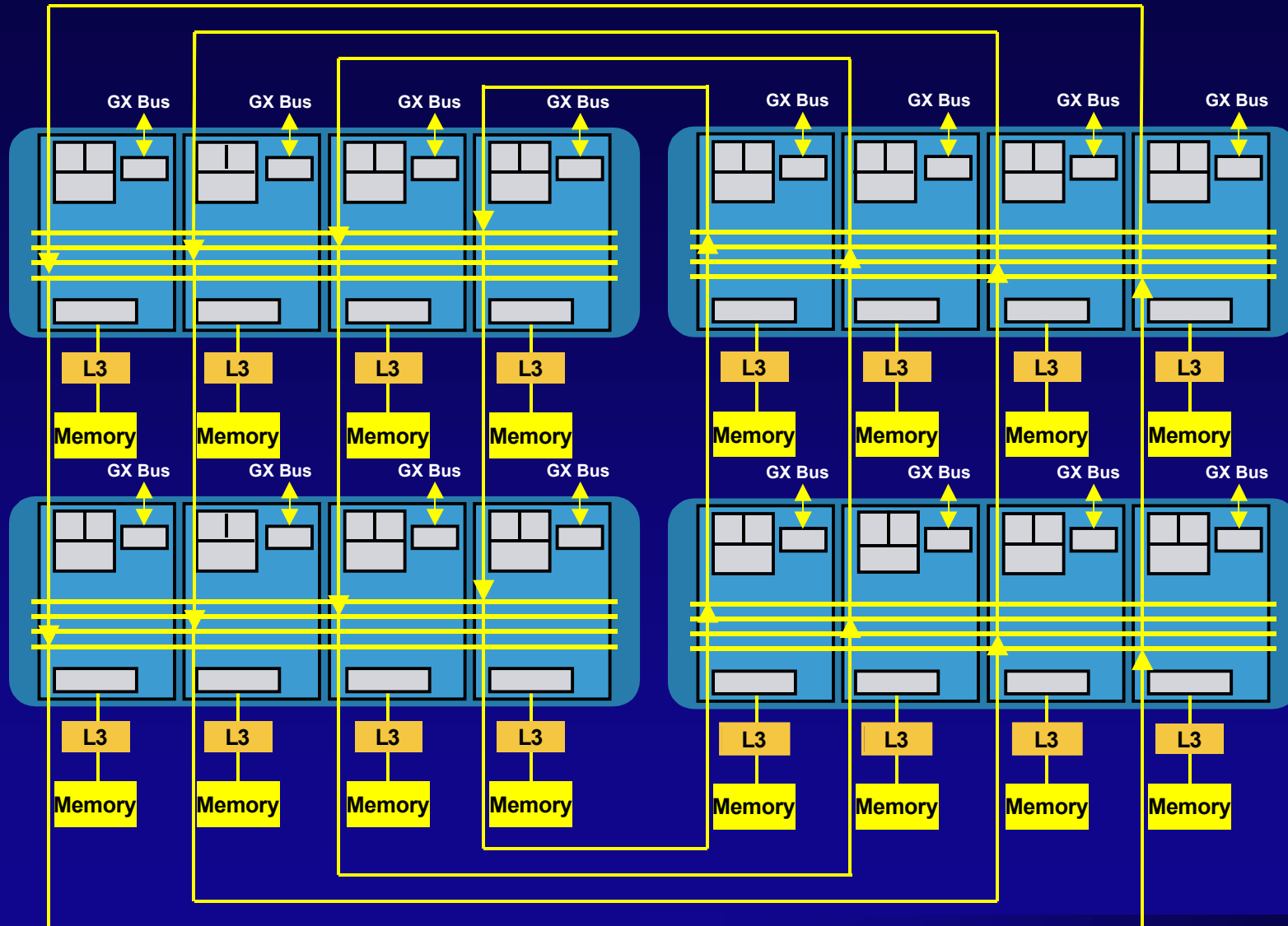
- 2-way SMP system on a chip
  - > 1 GHz processor frequency
- Storage Hierarchy
  - L1: 32 KB Data, 64 KB Instruction per processor
  - L2: ~1.5 MB per chip
  - L3: 32 MB per chip
- Chip interconnect:
  - New *Distributed Switch* design
  - Buses operate at \_processor speed
- Technology:
  - 0.18  $\mu\text{m}$  lithography
  - Copper, SOI
  - 174 million transistors

# System Building Block



- 4-chip, 8-way SMP module
- New *Distributed Switch* design
- Hybrid switch and bus configuration
- Enables aggressive cache-to-cache transfers
- Buses extended in multi-module configurations

# 32-way System Logical Structure



POWER4

# 32-way Components

- Hardware

- 4 multi-chip modules with 16 POWER4 chips
- 16 L3 dual chip modules with 32 16MB EDRAM chips
- Up to 16 memory controllers
- Thousands of DRAM chips to support up to 256 GB RAM
- GX to Remote I/O Bridge Chips
- PCI Host Bridge Chips
- PCI-PCI Bridge Chips
- Redundant Power Supplies and Air Moving Facilities
- I/O devices

- Code

- Service Processor Code
- System Firmware Code
- Operating System Code

System comprised of hardware & software components interacting and affecting system RAS

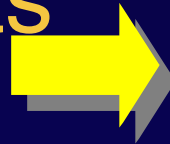


POWER4

# Chip Design Driven by System

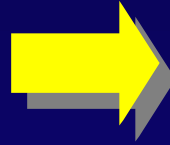
## Requirements

Fault Avoidance



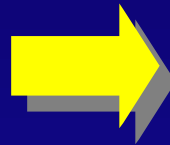
- Component minimization
- Intrinsic SER Mitigation

Recovery



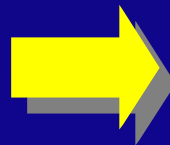
- Local chip level: ECC, refresh
- Chip interactions: bus/command retry
- System error handling: UE, PCI bus retry
- Run time / IPL diagnostics

Diagnosis and Reconfiguration



- Local chip level: spare bits, lines elements

Repair Policy



- System level: CPU, cache, memory sparing + degraded modes of operation
- Minimize system down time
- Concurrent and deferred maintenance

# Fault Avoidance: SER Mitigation With ECC

Failure Category	Failure Rate (Relative)	Source	Chip Example	
			FITS	MTBF
Intrinsic	1	Vendor DB	10-100	> 1140 yrs
Intermittent Marginal Pattern & Stress Sensitive	2-3	Empirical Field Avg	20-300	> 380 yrs
Aggregate Chip SER	150-2000	Vendor Appl Notes	1500-140000	0.8-76 yrs

## POWER4 Chip Set Design Rule:

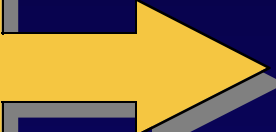
Implement ECC or equivalent recovery on all arrays sufficient to keep residual SER at or below each chip's IFR



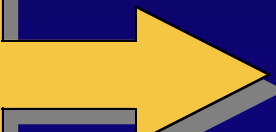
POWER4

# Recovery: ECC, Retry, UE Handling

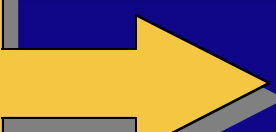
*ECC or hardware refresh:*  
Memory, Caches,  
ERAT, TLB



*Retry:*  
Module, GX &  
PCI buses, and  
I/O link



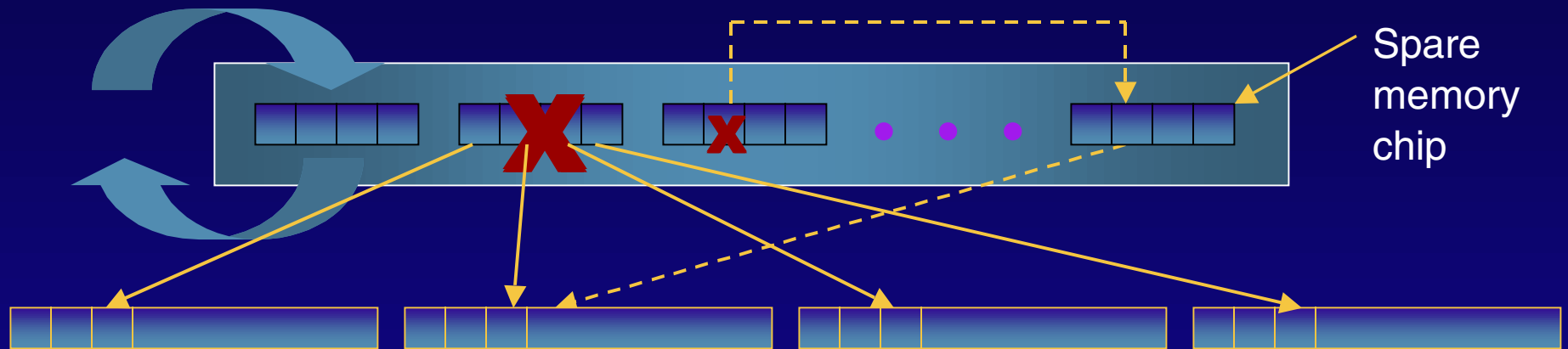
*UE handling:*  
ECC protected  
arrays, buses  
using retry



- Hamming SEC-DED augmented for special uncorrectable error (UE) handling
- Mainstore ECC designed for chip kill + redundant bit steering
- Masks intermittent errors
- Supports UE handling
- Increases unmasked MTBF from 4 months to > 20 years
- Marking, moving UE data
- AIX to support process terminate and software partition reboot

# Recovery: Main Store ECC and Extensions

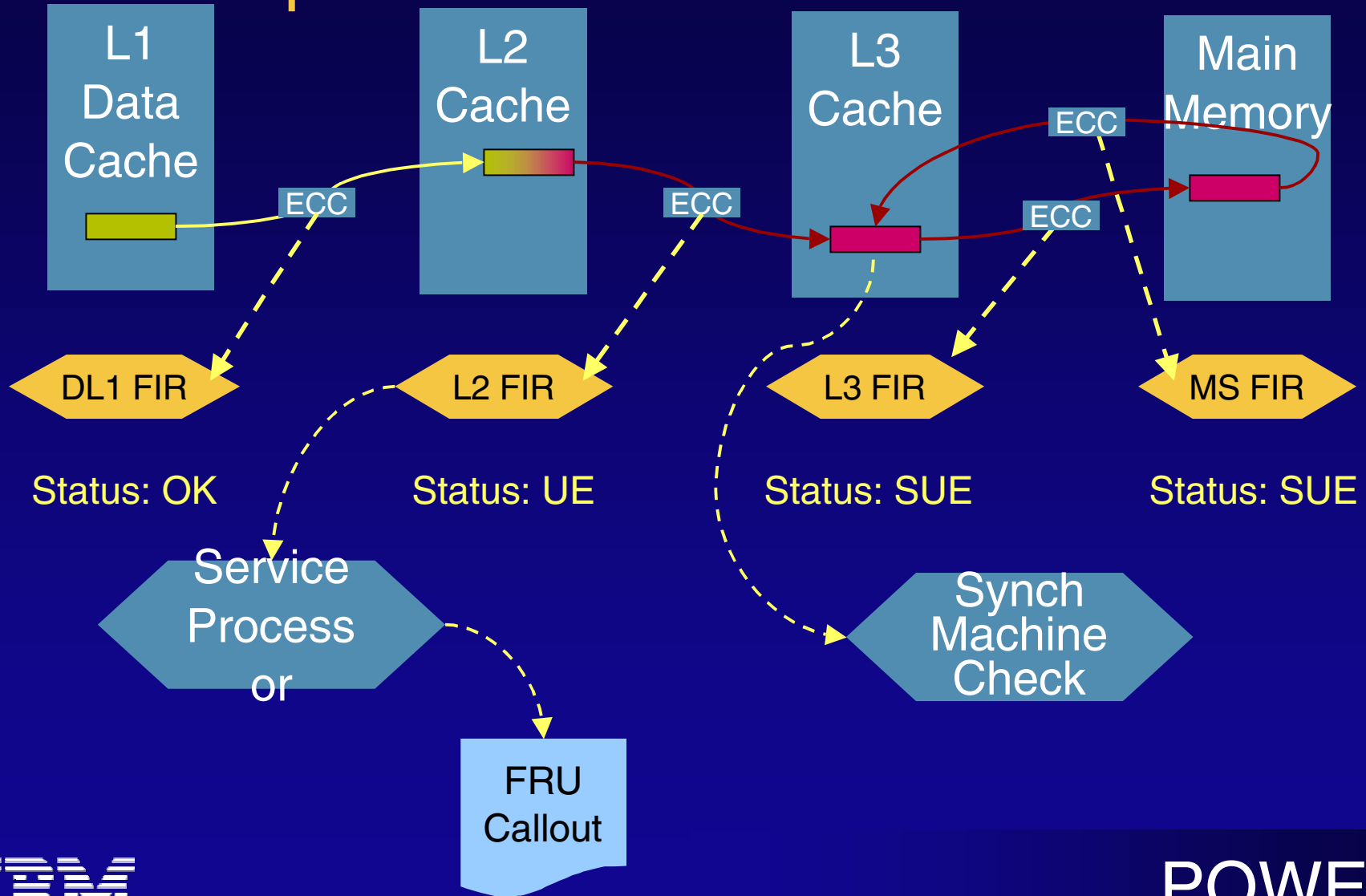
Memory scrubbing corrects soft single bit errors in background while memory is idle preventing multiple bit errors



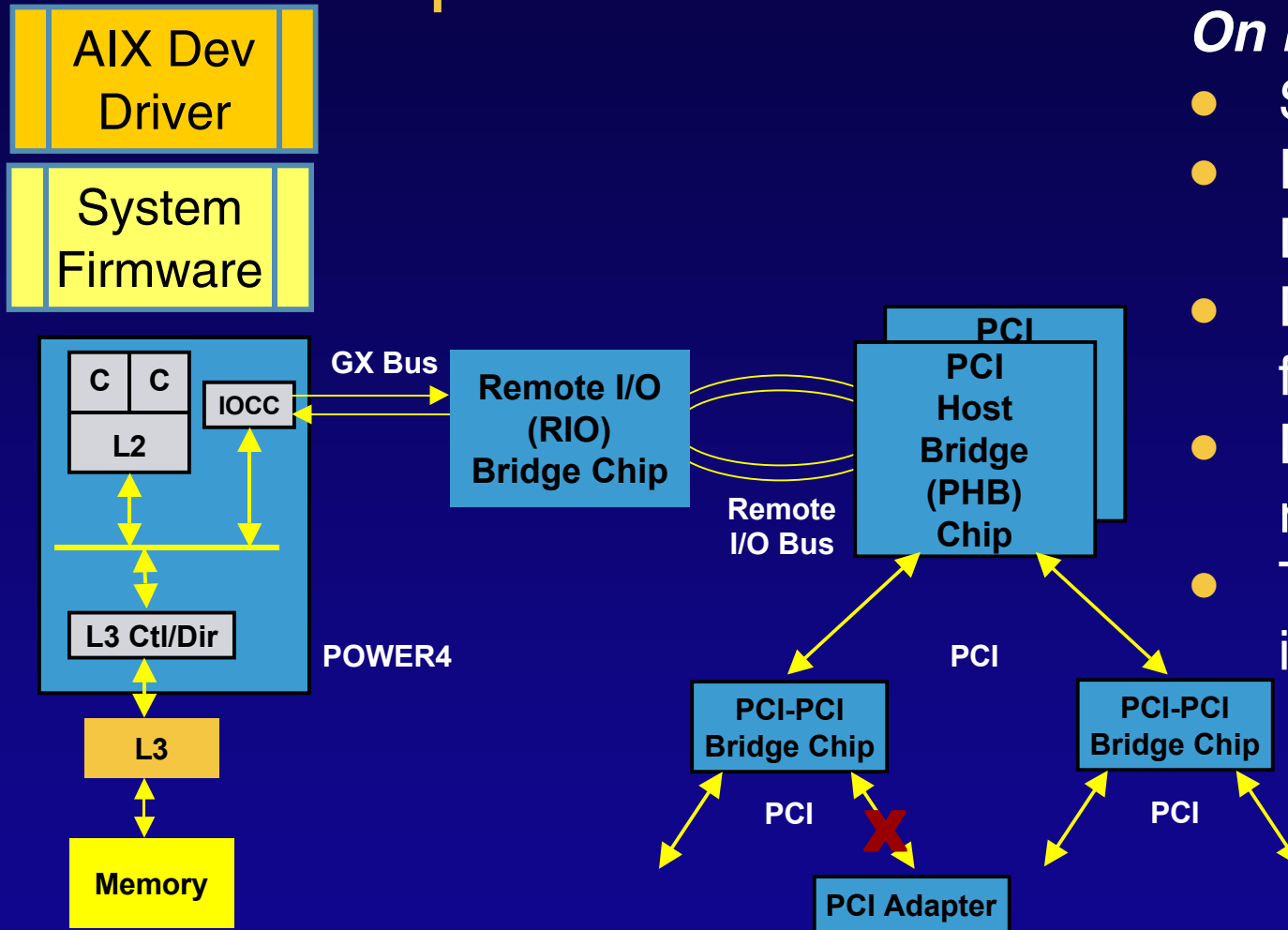
Bit scattering allows normal single bit ECC error processing to function even with a **chip kill** failure by scattering memory chip bits across separate ECC words

Bit steering dynamically reassigns memory I/O if error threshold is reached on same bit

# Recovery: Additional Logic to Avoid Checkstops



# Recovery: More Logic to Avoid Checkstops



## On PCI Error Detect:

- Slot freeze on error
- Return all 1's to Device Driver
- Driver calls firmware, reset slot
- Driver recovery, retry operation
- Threshold, fault isolation logged

## *Diagnosis & Reconfiguration: IPL RAS Design*

POWER4 Chip	<ul style="list-style-type: none"><li>➤ Built-in Self Test (BIST)</li><li>➤ Chips / single core can be deconfigured</li><li>➤ Spare bits switched in for single cell failures in L1, L2 caches, and in L2, L3 directories</li></ul>
L3	<ul style="list-style-type: none"><li>➤ Line delete capability maps out bad bits</li><li>➤ L3 cache bypassed for more serious failures</li></ul>
Main Memory	<ul style="list-style-type: none"><li>➤ ECC and redundant bit steering</li><li>➤ Entire memory card deconfigured for serious failure</li></ul>
I/O	<ul style="list-style-type: none"><li>➤ Redundant Remote I/O link takes over in case of failure</li><li>➤ I/O drawer deconfigured during boot if failure detected to allow IPL to proceed</li><li>➤ PCI adapter marked unavailable if error detected and can be replaced via Hot Swap later</li></ul>

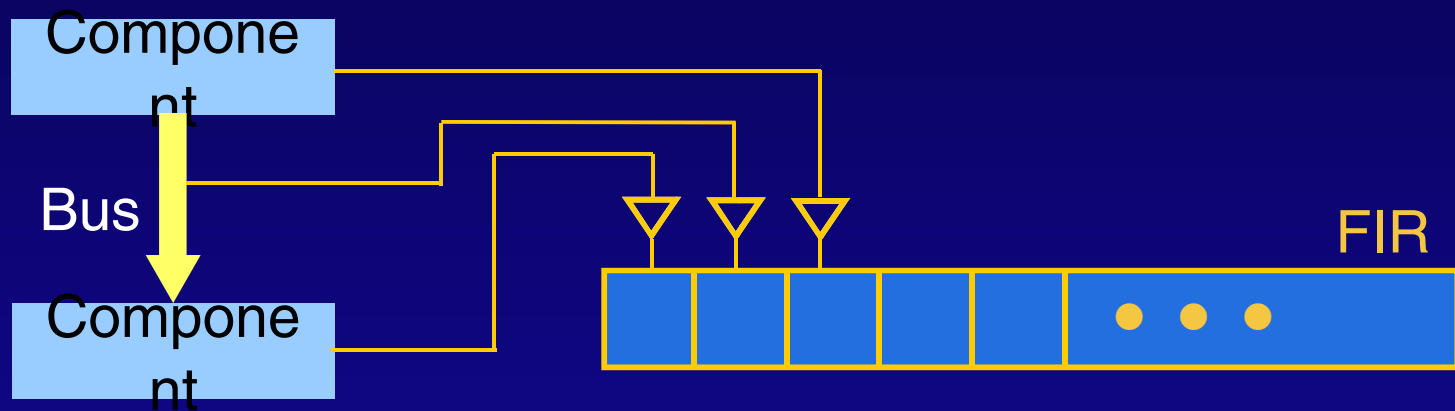
## Diagnosis & Reconfiguration: Run Time RAS

POWER4 Chip	<ul style="list-style-type: none"><li>➤ Internal checkers, parity, ECC, UE &amp; Special UE handling</li><li>➤ FIR error capture, <i>Who's on First</i> logic</li><li>➤ Spare bits / quiesce mode</li></ul>
L3	<ul style="list-style-type: none"><li>➤ Internal checkers, parity, ECC, UE &amp; Special UE handling</li><li>➤ FIR error capture, <i>Who's on First</i> logic</li><li>➤ Cache line delete</li><li>➤ Bypass on next boot following UE</li></ul>
Main Memory	<ul style="list-style-type: none"><li>➤ ECC, address checks, UE &amp; Special UE handling</li><li>➤ FIR error capture, <i>Who's on First</i> logic</li><li>➤ Memory scrubbing, chip kill, bit steering</li><li>➤ Card deconfigure on next boot following UE</li></ul>
I/O	<ul style="list-style-type: none"><li>➤ GX bus error checkers, UE &amp; Special UE handling</li><li>➤ FIR error capture, <i>Who's on First</i> logic</li><li>➤ Remote I/O link hardware failover</li><li>➤ PCI device detect, recover, isolate, deconfigure</li></ul>



# Diagnosis & Reconfiguration: Fault Isolation Registers

- FIR design guidelines:
  - Capture error at source
  - Identify component showing error symptoms



- Leads to:
  - FIR bits can be OR'd but carefully*

## *Diagnosis & Reconfiguration: Who's on First Logic*

- Requirement:
  - Separate cause from effect to isolate faulty component from components propagating fault
- Solution:
  - Each FIR starts a timer when it detects an error condition
  - FIRs freeze timer when checkstop finally occurs
  - FIR measuring longest elapsed time identifies causing component
- Used for:
  - FRU callout
  - Reconfigure system

## *Diagnosis & Reconfiguration: Role of Service Processor*

- Controls First Failure Data Capture
  - FIR and *Who's on First* logic enable capability
- Responsible for reconfiguring system
  - Use spares for dynamic reconfiguration where possible
  - Deconfigure failing part and bypass otherwise
  - FRU callout
- Allows for component de-allocation before hard failures occur in conjunction with Operating System
  - Processor
  - Chip
  - L3 line deletes
  - L3 and memory (on next IPL)
  - PCI adaptor and I/O devices

# Repair Policy: Maximize System Availability

POWER4 Chip	<ul style="list-style-type: none"><li>➤ Internal array spare bits eliminate single bit error causes allowing continued operation without repair</li><li>➤ Core / chip deconfiguration support deferred repair for other errors</li></ul>
L3	<ul style="list-style-type: none"><li>➤ Cache line delete eliminates single bit error causes allowing continued operation without repair</li><li>➤ L3 bypass supports deferred repair for other problems</li></ul>
Main Memory	<ul style="list-style-type: none"><li>➤ Bit steering eliminates single bit error causes allowing continued operation without repair</li><li>➤ Chip kill allows non-degraded operation and deferred repair for other DRAM failures</li><li>➤ Card deconfiguration allows deferred repair for card logic failures</li></ul>
I/O	<ul style="list-style-type: none"><li>➤ Remote I/O link hardware failover allows continued operation</li></ul>
Power and Cooling	<ul style="list-style-type: none"><li>➤ PCI device hot plug for concurrent repair</li><li>➤ Redundancy and hot plug to allow for concurrent repair</li></ul>



POWER4

# POWER4 RAS Design

- Focus on maximizing system operation
  - Avoid faults through masking and recovery
  - Dynamically bypass faulty componentry
  - Allow for concurrent repair where possible, deferred repair otherwise
- Requires special handling to uniquely identify failure source
- Total system design encompassing all hardware components, system firmware and operating system code
- Mainframe RAS attributes in a UNIX server



POWER4