IBM POWER6 Processor and Systems

IBM POWER6 Fault-Tolerant Design

Presenter: Natalya Kostenko

WHAT'S IBM POWER 6 MICROPOCESSOR

- **POWER** is a RISC instruction set architecture designed by IBM. (POWER is ***P****erformance* ***O****ptimization* ***W****ith* ***E****nhanced* ***R****ISC**)

- It's based on IBM POWER5 microprocessor technology (SMT, Dual Core) plus some extensions in order to increase performances.

- Its core is fabricated in 65-nm silicon-on-insulator (SOI) technology and operates at frequencies of more than 4 GHz.

- The microprocessor is a 13-FO4** design containing more than 790 million transistors, 1,953 signal I/Os, and more than 4.5 km of wire on ten copper metal layers.

*** reduced instruction set computing**

**** FO4 is a process independent delay metric used in digital CMOS technologies.**

Achieving High Frequency: POWER6 13FO4 Challenge Example

Circuit Design

- 1 FO4 = delay of 1 inverter that drives 4 receivers
- 1 Logical Gate = 2 FO4
- 1 cycle = Latch + function + wire
  - ★ 1 cycle = 3 FO4 + function + 4 FO4
- Function = 6 FO4 = 3 Gates

Integration

- It takes 6 cycles to send a signals across the core
- Communication between units takes 1 cycle using good wire
- Control across a 64-bit data flow takes a cycle

ARCHITECTURE OF POWER6 MICROPROCESSOR

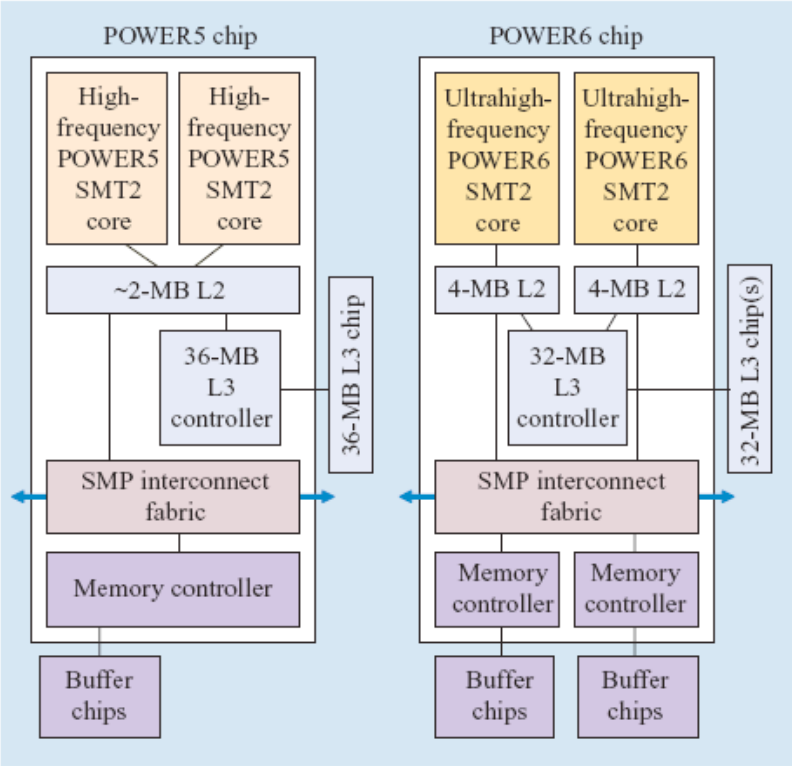


Figure 1

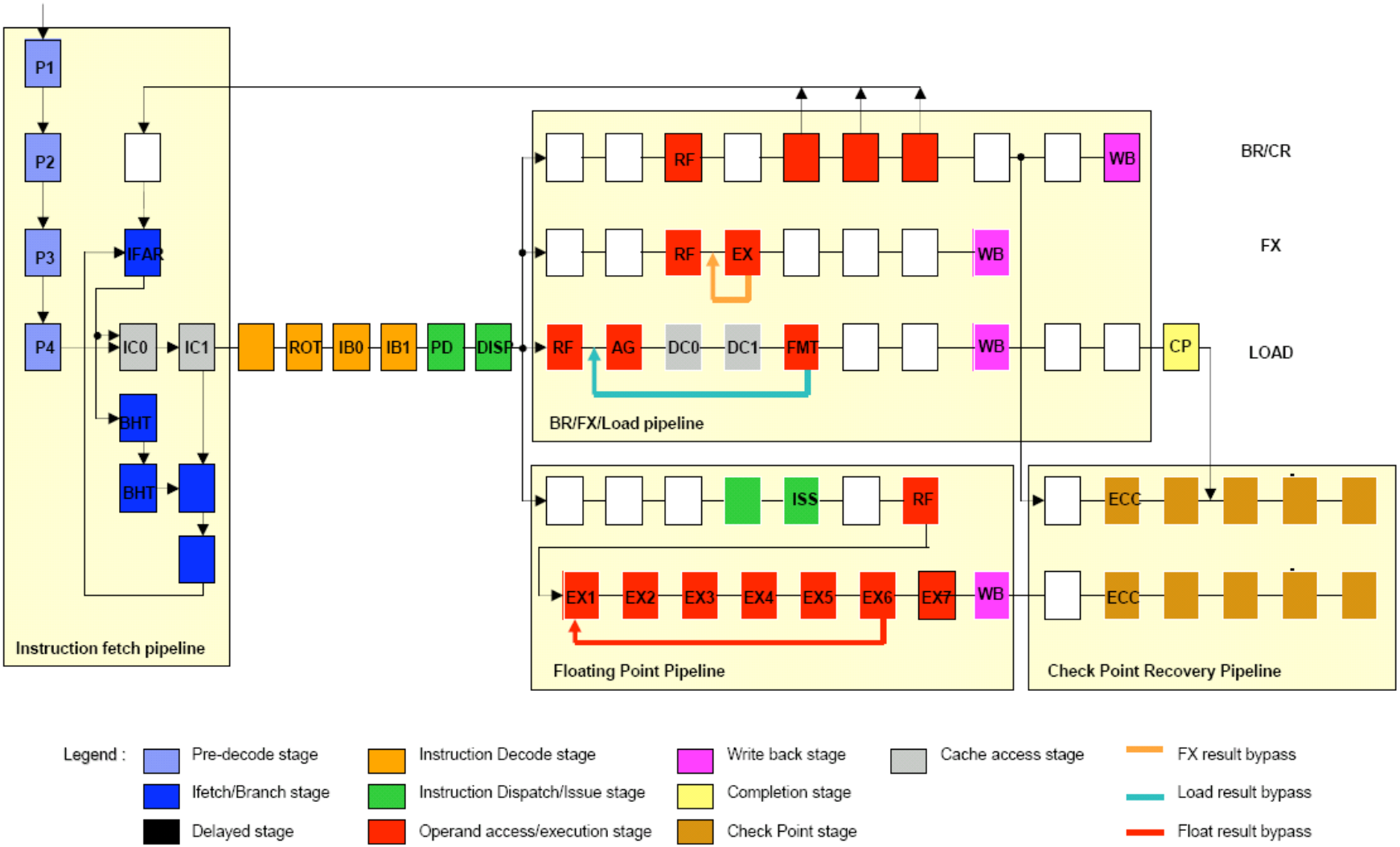
Evolution of the POWER6 chip structure. (SMT2: a dual-threaded simultaneous multithread.)

- The Power6 Chip operates at twice the frequency of Power5
- In place of speculative out-of-order execution that requires costly circuit renaming, the design concentrates on providing data prefetch.
- Limited out-of-order execution is implemented for FP instructions.
- Improvement of the Dispatch and Completion: 7 intr from both cores simultaneously
- Better SMT speed up due to increased cache size, associativity
- Designed to consume less power

THE PROCESSOR CORE

- Structured in Pipeline
- Developed to minimize logic content in the pipeline Stages
- Introduction of Decimal arithmetic as well as Vector Multimedia arithmetic
- Implement action of the Checkpoint Retry and Processor Sparing
- Instruction fetching and branch handling are performed in the instruction fetch pipe.
- Instructions from the L2 cache are decoded in pre-code stages P1 through P4 before they are written into the L1 I-cache.
- Branch prediction is performed using a branch history table (BHT) that contains 2 bits to indicate the direction of the branch.

THE PROCESSOR PIPELINE



USAGE OF THE PIPELINE

- Branch and logical condition instructions are executed in the branch and conditional pipeline
- FX (Fixed Point) instructions are executed in the FX pipeline, load/store instructions are executed in the load pipeline, FP instructions are executed in the FP pipeline, and decimal and vector multimedia extension instructions are executed in the decimal and vector multimedia execution unit.
- Data generated by the execution units is staged through the checkpoint recovery (CR) pipeline and saved in an errorcorrection code (ECC)-protected buffer for recovery
- The FX unit (FXU = Fixed Point Unit) is designed to execute dependent instructions back to back.

USAGE OF THE PIPELINE

- **Instruction fetching and branch handling**: a dedicated 64-KB four-way set-associative L1 I-cache => Fast address translation. The POWER6 processor also recodes some of the instructions in the pre-decode stages to help optimize the implementation of later pipeline stages.

- **Instruction sequencing**: handled by the IDU. For high dispatch bandwidth, the IDU employs two parallel instruction dataflow paths, one for each thread. Both threads can be dispatched simultaneously.

- **FX instruction execution**: The core implements two FXUs to handle FX instructions and generate addresses for the LSUs. The most signature features of these FXUs is that they support back-to-back execution of dependent instructions with no intervening cycles required to bypass the data to the dependent instruction.

- **Binary FP instruction execution**: The core includes two BFUs, essentially mirrored copies, which have their register files next to each other to reduce wiring. In general, the POWER6 processor is an in-order machine, but the BFU instructions can execute slightly out of order due of multiples empty slots in FP instructions (divide, Square root). The BFU notifies the IDU when these slots will occur, and the IDU can dispatch in the middle of these slots.

USAGE OF THE PIPELINE

- **Data fetching**: performed by the LSU. The LSU contains two load/store execution pipelines, with each pipeline able to execute a load or store operation in each cycle. The LSU contains several subunits: the load/store address generation and execution; the L1 D-cache array and the cache array supporting set-predict and directory arrays, address translation, store queue, load miss queue (LMQ), and data pre-fetch engine.
- **Accelerator**: The POWER6 core implements a vector unit to support the PowerPC VMX instruction set architecture (ISA) and a decimal execution unit to support the decimal ISA.
- **Cache Hierarchy:** Has 3 levels of caches

Cache attribute	*L1 instruction*	*L1 data*	*L2*	*L3*
Capacity	64 KB	64 KB	4 MB	32 MB
Shares (cores)	1	1	1	2
Location	Within core	Within core	On-chip	Off-chip
Line size (bytes)	128	128	128	128
Associativity	4 way	8 way	8 way	16 way
Update policy	Read only	Store through	Store in	Victim
Line inclusion rules	Resides in L2	Resides in L2	None	None
Snooped	No	No	Yes	Yes
Error protection	Parity	Parity	ECC	ECC

Symmetric Multithreading (SMT)

- P6 operates in two modes, ST (single thread) and SMT (multithreaded)
- In SMT mode two independent threads execute simultaneously, possibly from the same parallel program
- Instructions from both threads can dispatch in the same group, subject to unit availability
- This is highly profitable on P6 – it is a good way to fill otherwise empty machine cycles and achieve better resource usage

MEMORY & I/O SUBSYSTEMS

MEMORY :

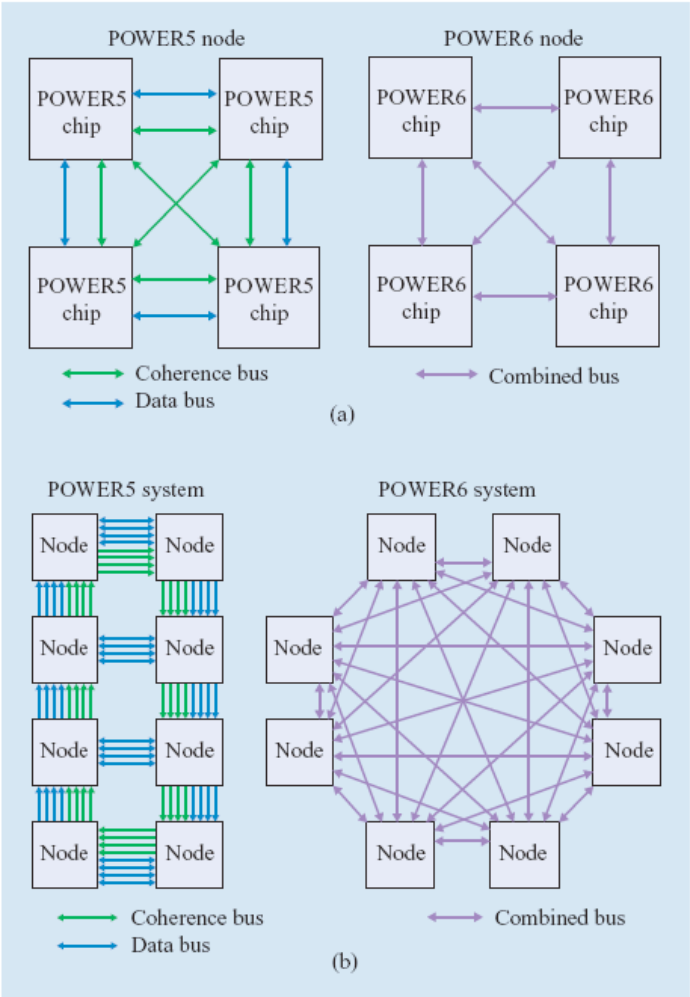
- Each POWER6 chip includes two integrated memory controllers that each of them commands up to 4 parallel channels.
- A channel supports a 2-byte read data path, a 1-byte write data path, and a command path that operates four times faster than the DRAM frequency
- Each memory controller is divided into two regions that operate at different frequencies: ***asynchronous region*** *(*four times the frequency of the attached DRAM*),* and the ***synchronous region*** (Half of the core frequency).
- Memory is protected by **SECDED ECCs***. **Scrubbing** is employed to find and correct soft, correctable errors.

I/O FEATURES :

- I/O Controller: 4-byte off-chip read/write interfaces are connected to I/O hub chip.
- A pipelined I/O high throughput mode was added whereby DMA write operations initiated by the I/O controller are speculatively pipelined. => This ensures that in the largest systems, inbound I/O throughput is not limited by the tenure of the coherence phase of the DMA write operations.

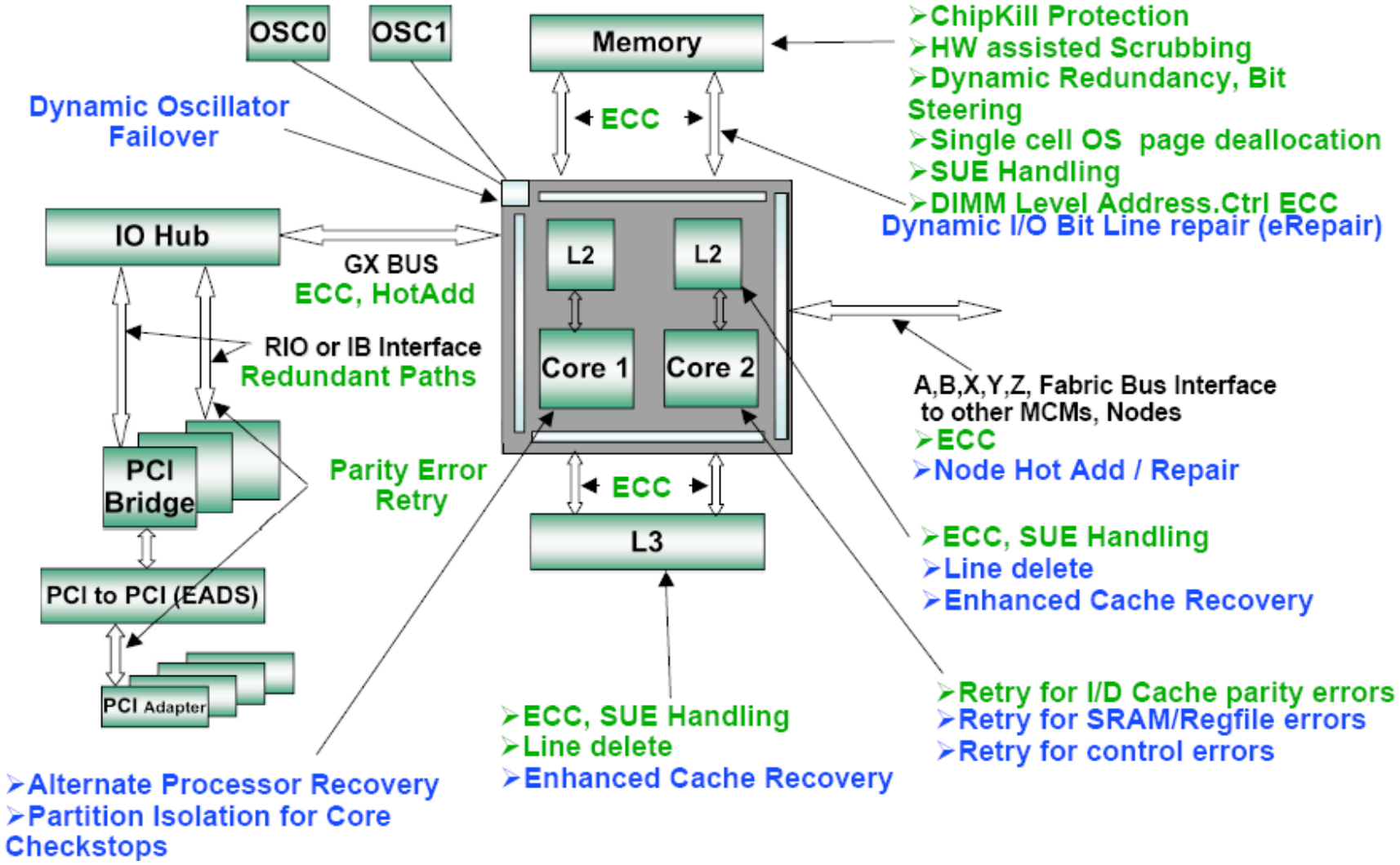
* **SECDED ECC** - single error correction, double error detection, error-correcting code

SMP INTERCONNECTION (Symmetric Processors)



- SMP interconnect fabric is built on the nonblocking broadcoast transport approach
- Relying on the traffic reduction, coherence and data traffic share the same physical links by using a time-division-multiplexing (TDM) approach.
- the ring-based topology is ideal for facilitating a non-blocking broadcast coherence-transport mechanism since it involves every node in the operation of all the others.

POWER6 RAS EXECUTION



POWER6 RAS EXECUTION

- Error Detection and Recovery requirements were specified during the High Level Design phase
- Firmware Recovery assists specified early
- Instruction Retry
- Alternate Processor Recovery
- Core checkstop isolation

Functions to protect against Core errors

- Processor Instruction retry
  - Retries instructions that were affected by hardware errors
  - Protects against intermittent errors
- Alternate Processor Recovery
  - If instruction retry encounters a second occurrence of the error. (i.e., Solid defect)
  - Moves workload over to an alternate/spare processor
- Processor contained checkstops
  - Limits impact of many processor logic/cmd/ctrl errors to just the processor executing the instruction

Error Detection is first step to Recovery

- 100% ECC protection for caches and interfaces
- >99% of small SRAMs and Register Files parity protected
- Dataflow protection
- Protocol checking between functional units
- Control logic protected by parity and consistency checking
- Floating Point Residue Checking
- Queue management (Underflow/Overflow)
- Architected Registers
- Store Data

Core Checkstop

- High levels of error detection and isolation were specified early in the design cycle
- Core checkstops fall into two categories:
  - ★ Recoverable
    - Core Sparing moves the work to another processor
  - ★ Non Recoverable
    - The partition running on the core at the time of the fault is terminated
    - Other partitions are not affected
    - Policy is set by the Hypervisor

Enhanced Cache Recovery

Single bit errors

- Soft errors are purged from the cache to force a refresh of the cell
- Hard errors will result in line delete. Reduces the risk of a double bit error

Multi bit errors

- Hardware will purge and delete the damaged location
- Firmware will dynamically de-configure the core attached to the defective cache

System Recovery of Cache UEs

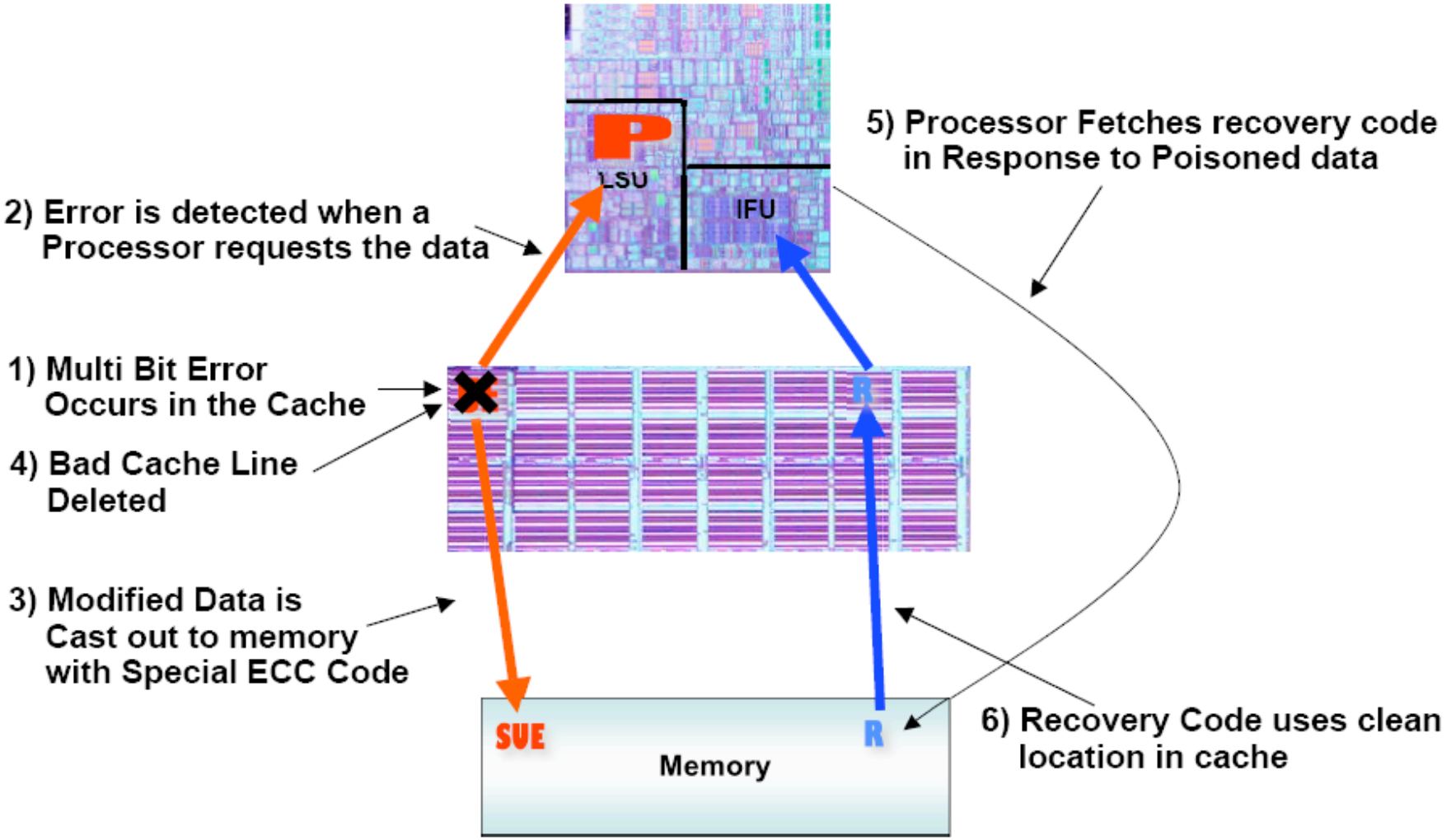
Problem

- POWER6 systems employ System Recovery Code for Uncorrectable Errors detected in the Cache Hierarchy
- If cache location is damaged, the same code being used to recover the initial error could be damaged as well

Solution

- POWER6 has automatic purge and delete for L2 and L3 Cache UEs
- Non-modified lines are re-fetched from Main Store and recovered transparently
- Modified lines are frequently contained to affected application, occasionally resulting in partition outage.

Enhanced Cache Recovery



POWER6 Summary

Extends POWER leadership for both Technical and Commercial Computing

- Approx. twice the frequency of P5+, with similar instruction pipeline length and dependency
  - ★ 6 cycle FP to FP use, same as P5+
  - ★ Mostly in-order but with OOO-like features (LLA)
  - ★ 5 instruction dispatch group from a thread, up to 7 for both threads, with one branch at any position
- Significant improvement in cache-memory-latency profile
  - ★ More cache with higher associativity
  - ★ Lower latency to all levels of cache and to memory
  - ★ Enhanced datastream prefetching with adjustable depth, store-stream prefetching
  - ★ Load look-ahead data prefetching
- Advanced simultaneous multi-threading gives added dimension to chip level performance

Enhanced Power Management

Advanced server virtualization

Advanced RAS: robust error detection, hardware checkpoint and recovery