TwinCastle: A Multi-processor North Bridge Server Chipset

Debendra Das Sharma,
Ashish Gupta, Gordon Kurpanek, Dean Mulla, Bob Pflederer, Ram Rajamani

Advanced Components Division,
Intel Corporation

Agenda

- Twin Castle Platform Overview
- Speeds and Feeds
- Twincastle Feature Set
- Twincastle North Bridge (TNB) Micro-architecture
- TNB Chip Statistics
- Summary

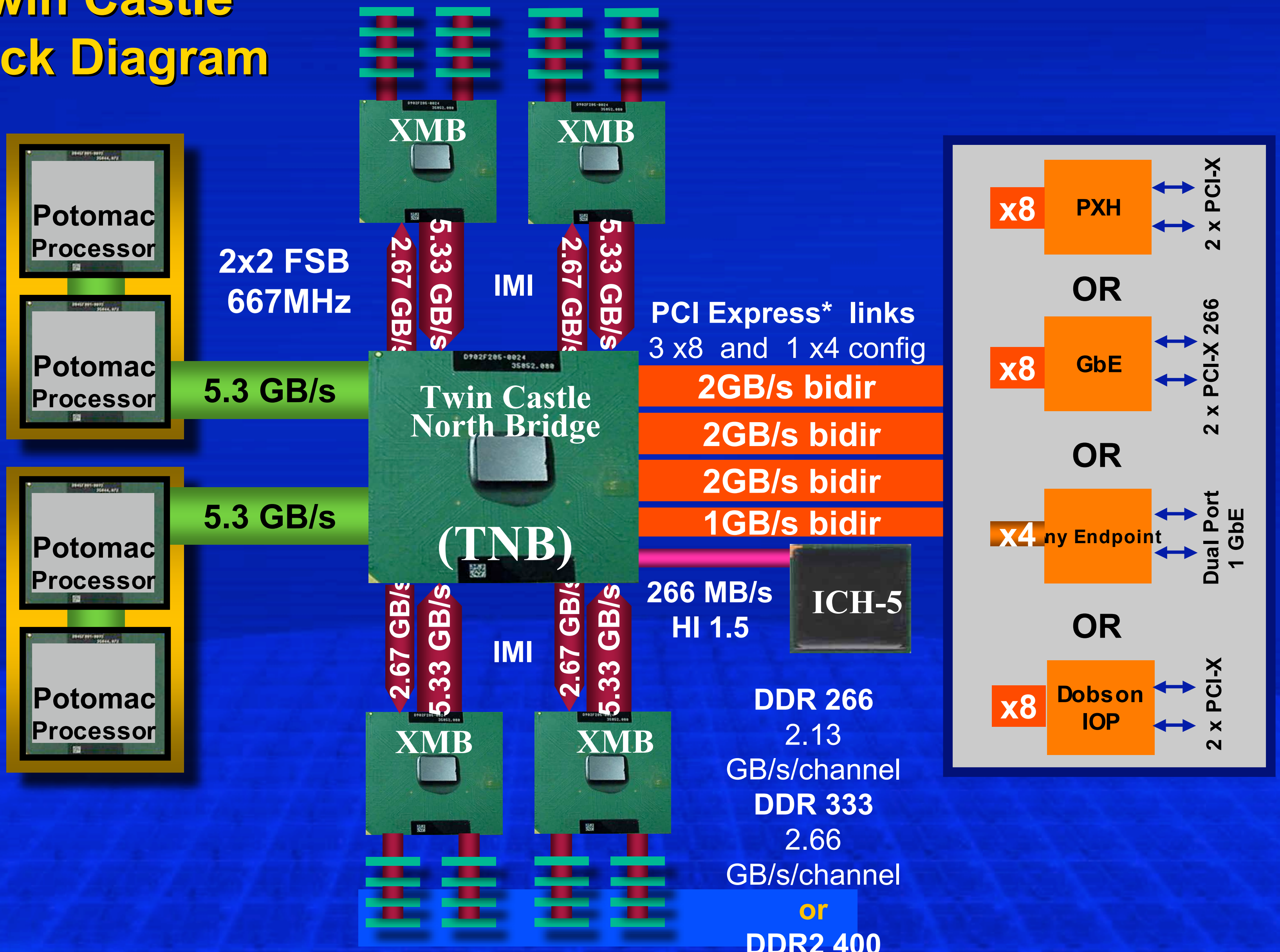
TwinCastle Overview

- Multiprocessor Platform: up to 4 CPUs
  - Support for multi-core CPUs.
- Platform support for multiple processor upgrades
  - Cranford, Potomac
  - Dual Core: Tulsa, Paxville
- Two chips: Twin Castle North bridge (TNB), and External Memory Bridge (XMB)
- Intel's x86 based MP chipset
  - Leadership technology such as PCI-E*, and IMI
  - Leadership RAS features

TNB Interconnects

- Two Front-side Buses to connect to CPUs
  - Capable of delivering 5.3 GB/sec per bus
  - Stream MP Triad: 4.35 GB/sec (read) and 5.8 GB/sec (read/ write combined)
  - Measured I/O system bandwidth: 5.02 GB/sec
  - Measurement is about 95% of theoretical
- Four IMI links
  - Each IMI connects to an XMB (memory)
    - Memory: DDR 266/ 333, DDR II 400 (up to 32 DIMMs)
  - Aggregate B/W: 21.33 GB/sec inbound , 10.67 GB/sec outbound
- I/O: Up to seven PCI-Express* links (28 lanes)
  - Aggregate B/W: 6.2 GB/sec inbound, 6.2 GB/sec outbound
  - Available with “no snoop” attribute (FSB b/w not bottleneck)
- Hub Interface (HI) to connect to ICH (south bridge)
  - 266 MB/sec of aggregate bandwidth
- Internal data path can sustain 42 GB/sec of data b/w

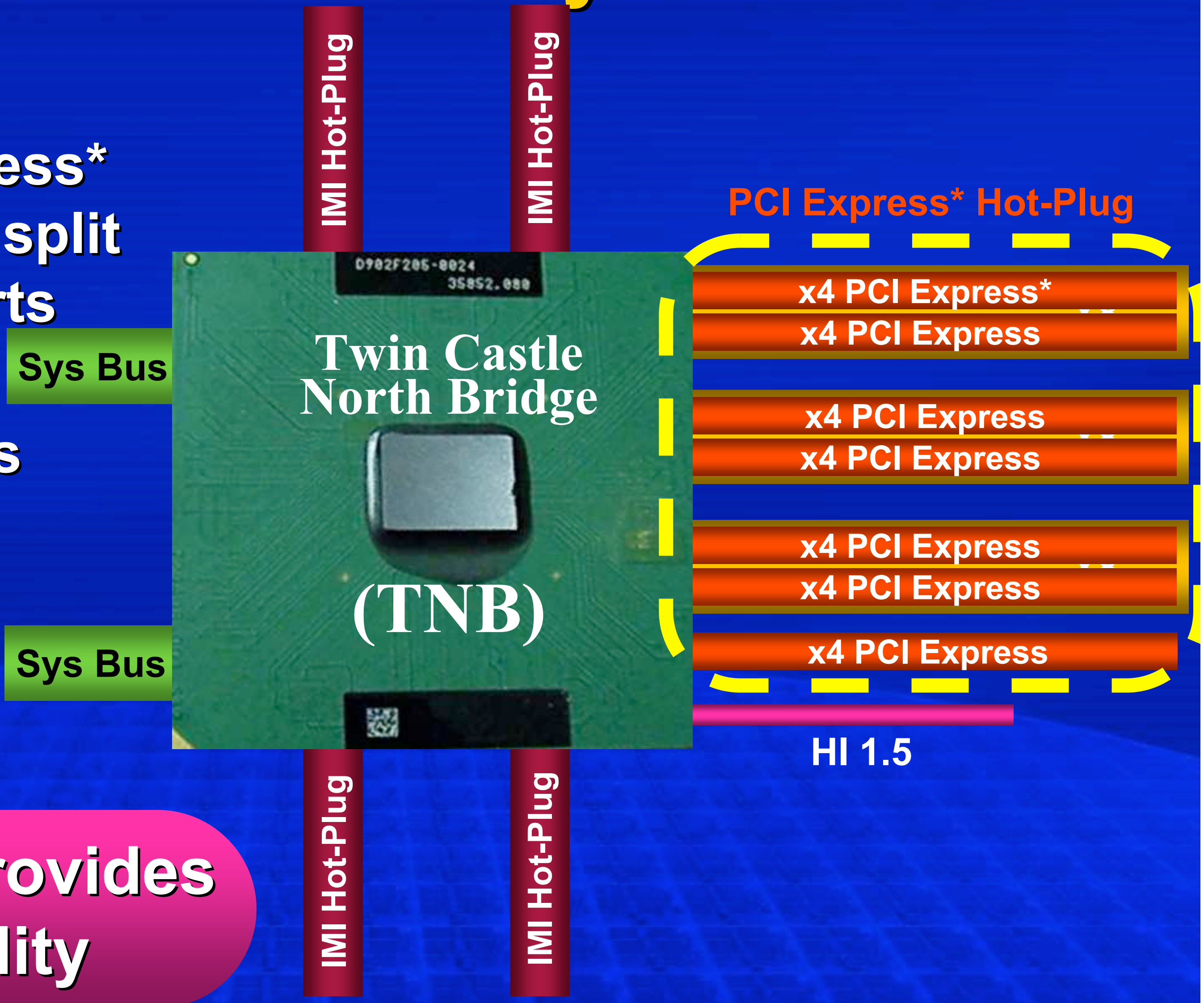
Twin Castle Block Diagram



DDR 266
2.13
GB/s/channel
DDR 333
2.66
GB/s/channel
or
DDR2 400
3.2GB/s/channel

TNB I/O Flexibility

- Each PCI Express* x8 port can be split into two x4 ports
- All PCI Express ports are Hot-Pluggable



Twin Castle provides I/O flexibility

Front-Side Bus

- 40-bit address bus. Double pumped
- 64 bit data bus. Quad pumped
- Independent Control Signals
- SEC-DED data ECC. Parity on address/ response
- Up to 3 loads per bus
- 167 MHz frequency => 5.3 GB/s B/W
- TNB ensures cache coherency across both FSBs
- Modified Enhanced Defer Protocol for improved pipelining

Independent Memory Interface Link

- Four IMI links. Connects up to 32 DIMMs through XMBs
- High-speed serial link operating at 2.67 GHz
- 21 bit wide incoming.10 bit wide outgoing
- Sustainable data bandwidth in each IMI link
  - 5.33 GB/s inbound
  - 2.37 GB/s outbound
- Data Integrity
  - Inbound data from DDR protected by a x8 SDDC ECC
  - Control and O/B data protected by CRC
- Data from memory sent directly to FSB for lower latency
  - ECC decode done off-line. Results sent to FSB and CDC
- Connects to XMB
  - Supports two 72-bit wide DDR channels
  - Each DDR capable of supporting up to 4 DIMMs/8 ranks
- End to end retry by TNB to correct uncorrectable errors

PCI-Express* Links

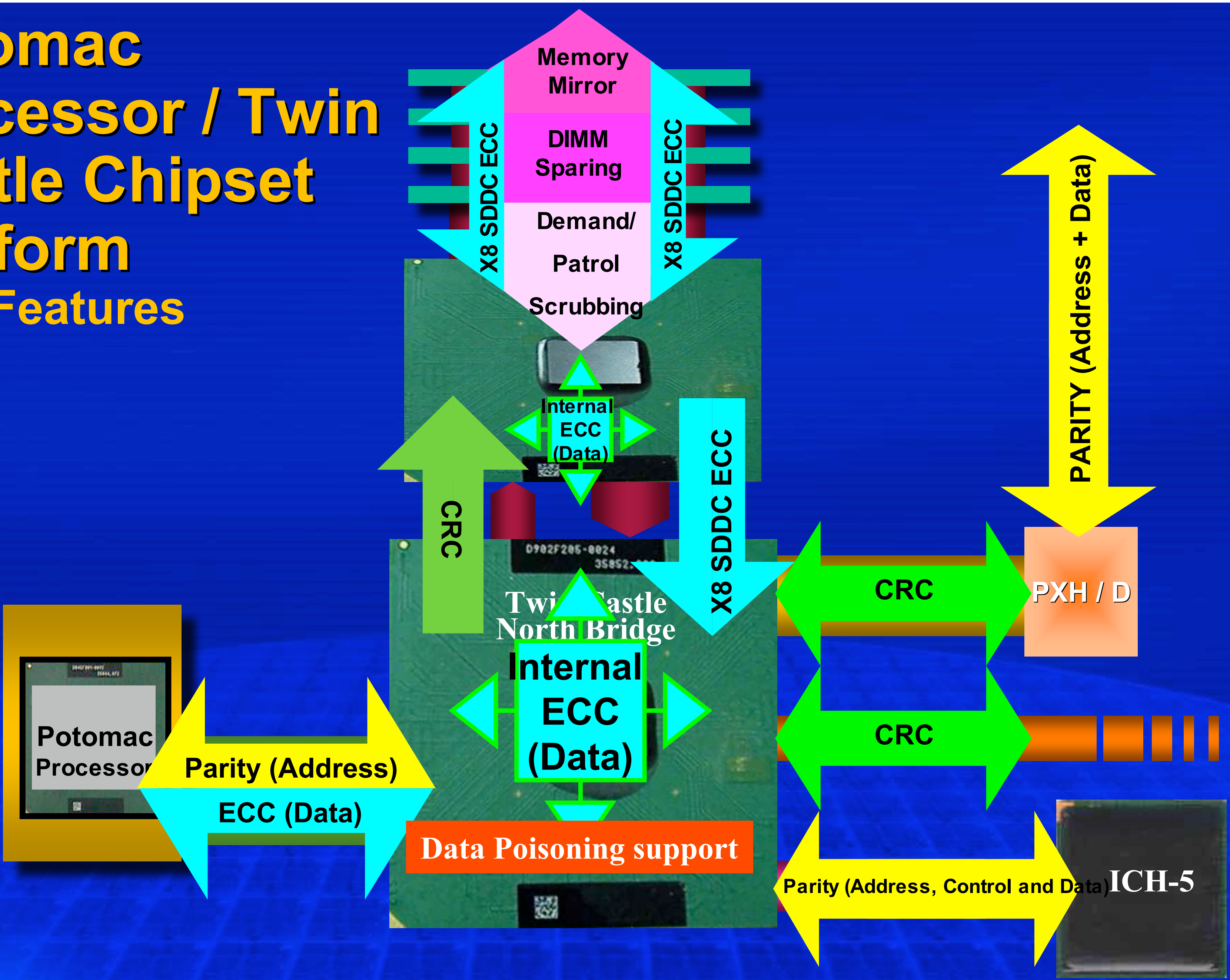
- Up to 7 PCI-Express* links with 28 Lanes
- Gen 1: Differential links at 2.5 GHz. 8b/10b
- Each x8 link partitionable to two independent x4s
- Lane Reversal Support
- Remembers prior failed lane assignment
  - Adopts a different assignment
- CRC-32 for TLPs. Retry. CRC-16 for DLLPs
- Degradation support: training and run-time
  - x4 supported on any two sets (includes reversal)
  - x2 supported on (0,1) and (4,5) including reversal
  - x1 anywhere
- Support for max 256 endpoints
- Max payload size: 256B. Read request size: 4KB

Other Interconnects

- Hub Interface
  - 66 MHz. 8 bit data. Quad data rate. => 266 MB/s
  - Parity Protected
  - Connects to ICH for access to legacy I/O
- JTAG: To access internal config registers and for production testing
- SM Bus: Two sets
  - One set for accessing internal config registers
  - Second set for Hot Plug through Phillips 9555 parts both for IMI and PCIE

Potomac Processor / Twin Castle Chipset Platform

RAS Features



RAS Features

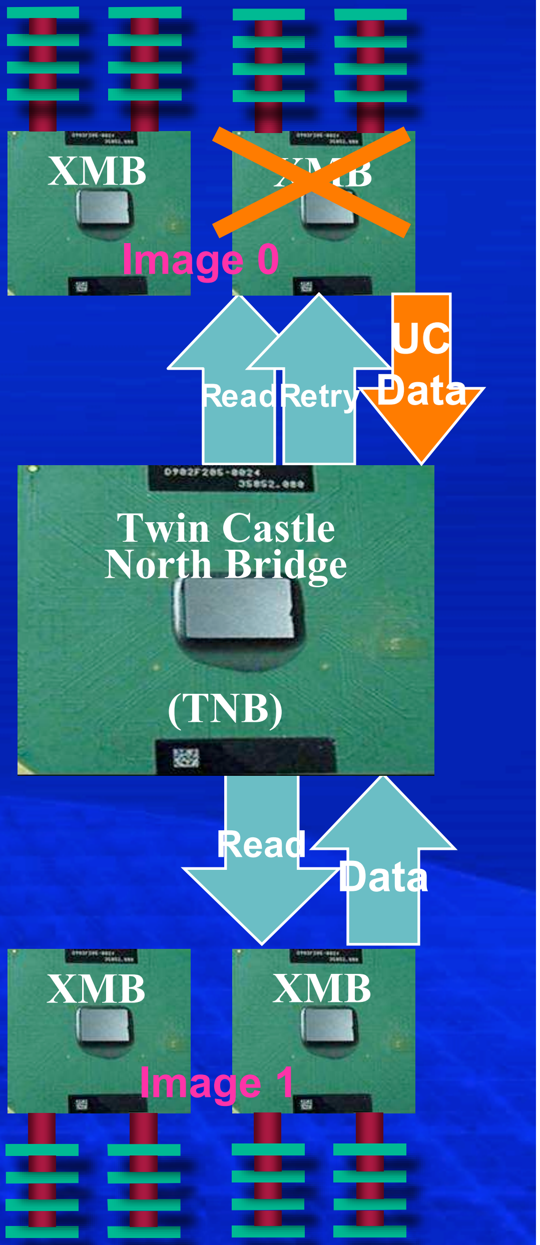
- Major interfaces: ECC/CRC/Parity protected
- Internal data path is ECC protected.
- Poison Support throughout : memory, PCI-E*, HI, and FSB
- Leadership Memory RAS
  - Single DRAM error correct (x8 or x4 device)
  - Double error detect
  - Demand and Patrol (proactive) scrubbing by H/W
  - Memory Mirroring
  - Memory RAID (level 5)
  - DIMM Sparing to survive DIMM failure (by XMB)
  - Hot Plug Memory board (XMB + DIMMs)

Memory Mirroring

- Memory replicated behind another XMB
- Write: TNB Updates both
- TNB interleaves between the two XMBs
  - Better performance
- Read failure (Uncorrectable) causes retry
  - If retry fails that copy is marked bad
  - Good data read from the other XMB
- TNB has a re-silvering engine to copy data when bad XMB is hot replaced
- Mirroring provides higher system reliability

Memory Mirroring

- Normal Operation
  - Correctable Error
- Uncorrectable error
  - Interface is disqualified after retry
  - New transactions directed to image1

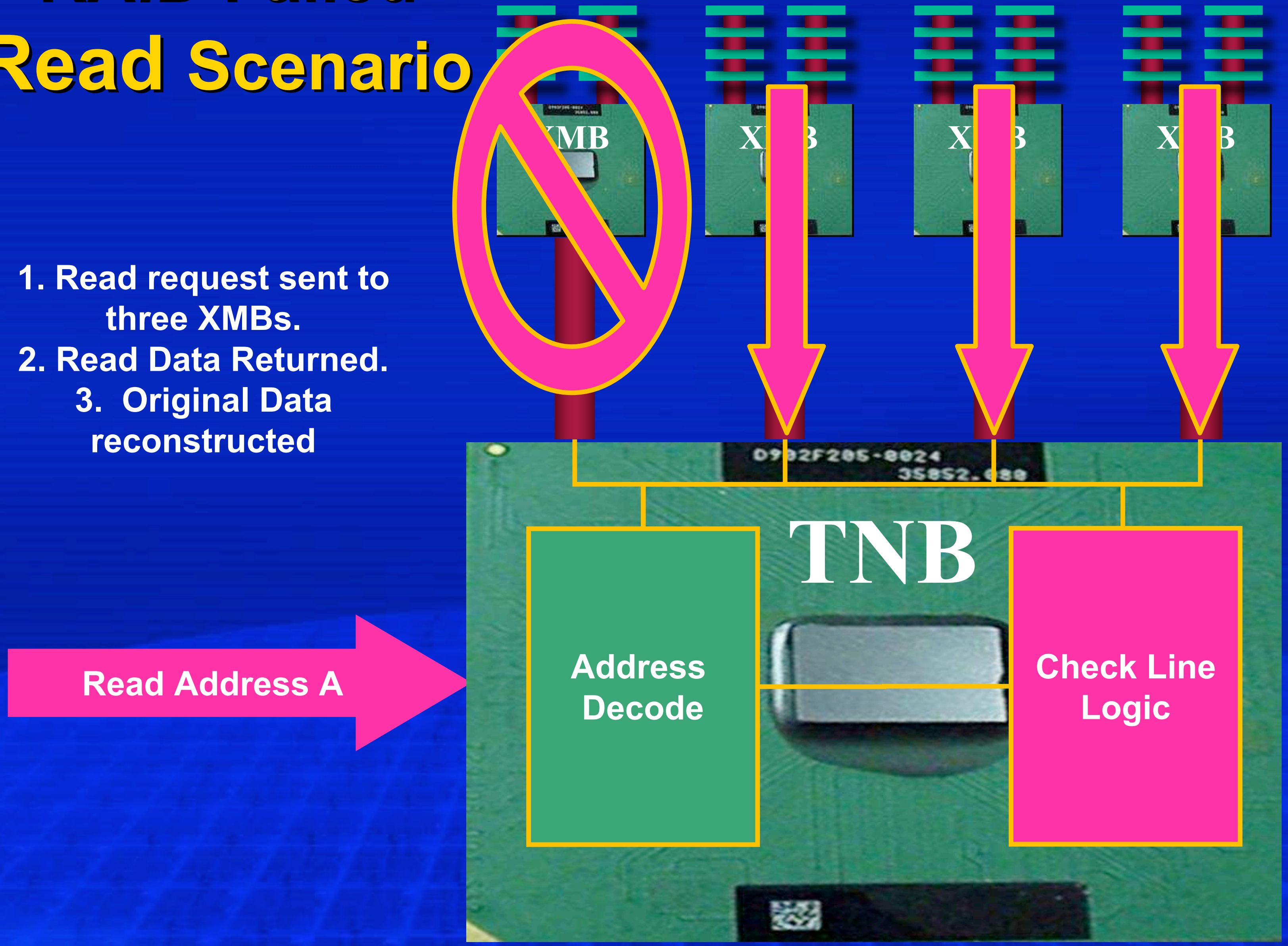


Memory RAID

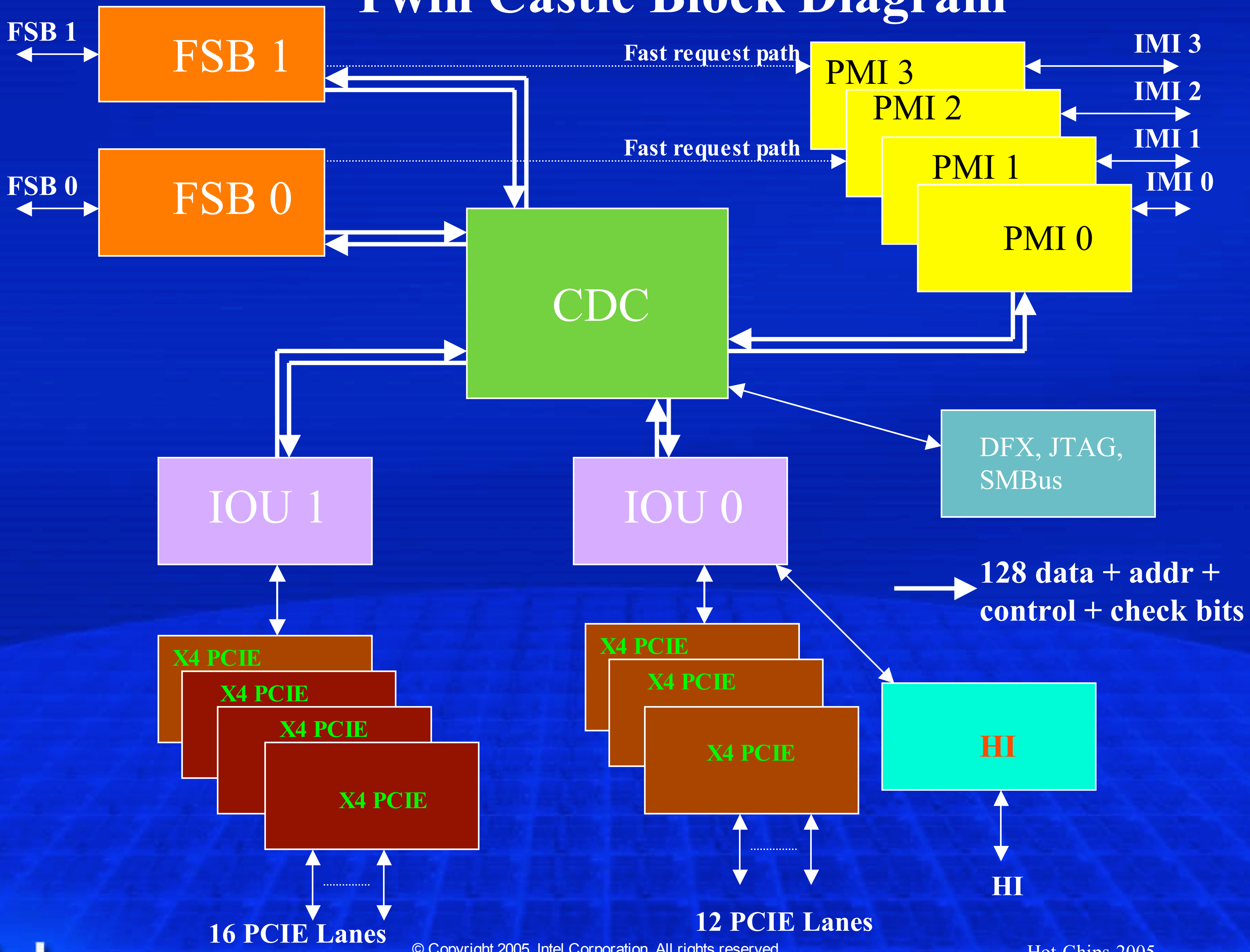
- Need 4 XMBs. Parity striped across
- Read sent to appropriate XMB
- Write operation
  - Read the parity XMB and the XMB with data
  - XOR and write back to both data and parity XMB
- Read from a failed XMB
  - Data read from three other XMBs
  - Data of the failed XMB reconstructed by XOR
- Write to failed XMB
  - Reconstruct old data by reading from 3 other XMBs
  - Update the 3 XMBs
- Resilvering engine in TNB reconstructs contents following a hot replace of the failed XMB.
- Provides higher system reliability and availability

RAID Failed Read Scenario

1. Read request sent to three XMBs.
2. Read Data Returned.
3. Original Data reconstructed



Twin Castle Block Diagram



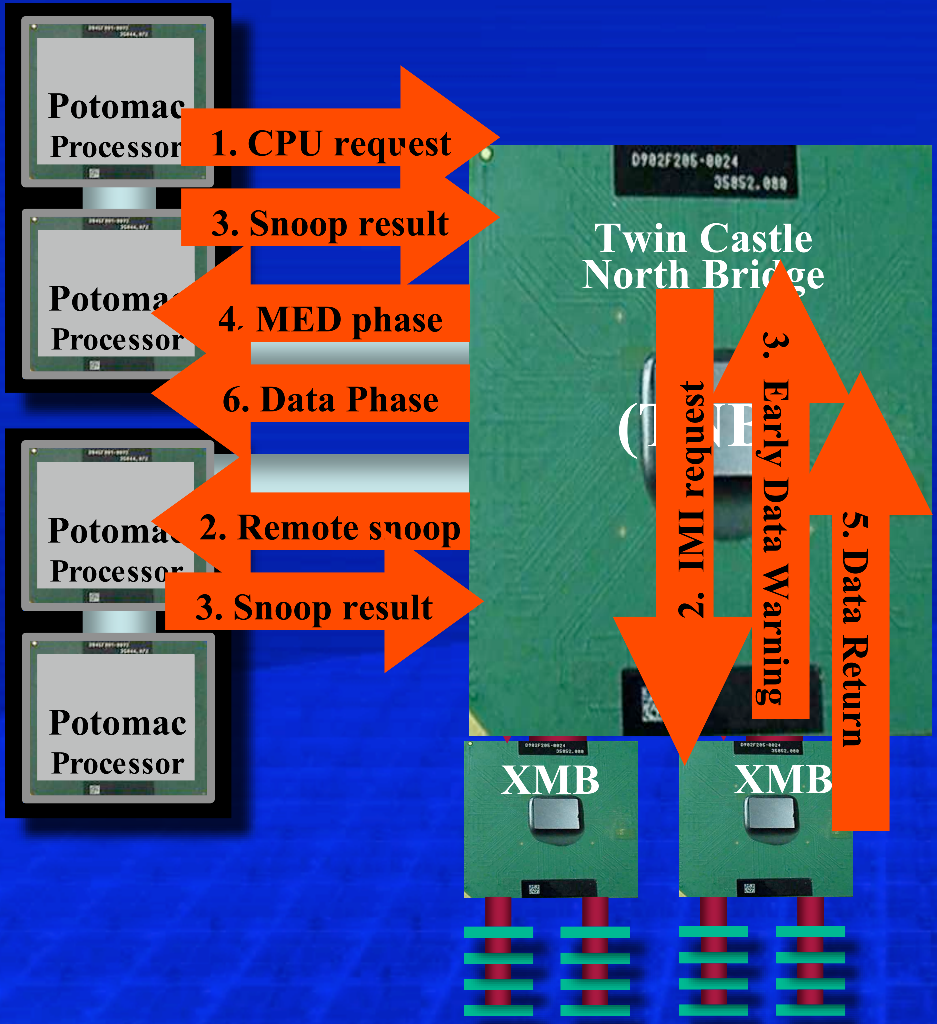
Common Data Cache (CDC)

- Coordinates FSB and I/O accesses to memory
- Connects to FSB, IMI and IO Unit clusters
  - 16B wide data path in each direction with ECC
  - Separate address and control buses in each direction
- Aggregate data bandwidth of more than 42 GB/s
- Core frequency: 333 MHz
- Stores coherent and non-coherent data
- Performance boost with hot cache lines
  - Latency: 72 ns with CDC hit vs 134 ns to memory (idle)
- Cache lines help I/O performance
  - IO Unit prefetches cache lines for DMA reads and writes
  - Masks I/O latency
- Performs write coalescing

Front Side Bus Cluster

- Interface logic to the CPU busses
- Tracks up to 12 in-order requests and 32 deferred requests
- Requests to memory and IO are deferred
  - Completed using Defer Reply transaction or Modified Enhanced Defer Phase
  - Defer Reply provides maximum CPU compatibility
  - Modified Enhanced Defer provides maximum performance
    - Supported by Potomac processor
    - Lower latency
    - Lower bandwidth consumption
- Logic support for two multi-core CPUs per bus
- Arbitration optimizations minimize latency and maintain fairness across the two front-side busses

Cache Coherency Flow



1. CPU issues request on FSB.
2. TNB forwards request through fast path to IMI. TNB checks CDC state and forwards snoop to remote FSB.
3. TNB collects snoop results from FSBs and early data warning from XMB.
4. TNB initiates modified enhanced defer phase on FSB. CPU prepares for data return.
5. Data arrives from IMI.
6. TNB forwards data through CDC fast path to FSB.

IO Unit

- Two independent IO Units in the chip
  - IOU 0: HI and 12 PCI-E* lanes
  - IOU 1: 16 PCI-E lanes
- Novel scalable design
  - Basic building block: x4
  - Can put two of them to work as a x8
    - e.g., can cascade two x4 CRC blocks to form x8 CRC block
  - Transaction layer FIFOs shared
    - A x8 link gets twice the storage of a x4 link
- PCI-E*: Physical, Link, and Transaction layers
- Transaction layer shared with HI
- HI has separate link and phy layers.

IO Unit

- Support for full peer-to-peer transactions
- PCI/E ordering handled in IOU only
- IO Unit requests CDC to prefetch cache lines for DMA read and writes.
  - Masks memory (and snoop) latency
  - NP requests prefetched even when ordering is not ok
    - CDC gets the data coherent into its cache
    - IOU fetches data when ordering is ok
  - IO Unit sends prefetches for posted transactions in order
    - Enables CDC to get data / ownership
    - Fetches are serialized to meet ordering rules
    - Writes follow fetches
- Performance features
  - Out of order reads
  - Completion coalescing
- Inbound completion credits: Infinite

Performance Enhancements

- FSB - IMI bypass paths to lower latency:
  - On memory requests from FSB to IMI
  - On data returns from memory to FSB. Late ECC indication to FSB
  - CDC updated along with bypass.
- Memory accesses from CPU deferred (Modified Enhanced Defer protocol) to effectively have a split transaction protocol with better pipelining.   MED reduces reduces FSB address bus utilization
- CDC caches data and state of recently accessed cache lines
- Conflicting resolution using conflict queue
  - Provides more consistent and predictable conflict resolution
  - Reduces bandwidth consumed retried requests
  - Eliminates need for performance-throttling forward progress
- Memory Interleaving
  - IMI level (Interleave across 4, then 2, then 1)
  - Rank level (Interleave across 8 then 4 then 2)

I/O Performance

- Prefetches for reads and Writes.
  - Hides latency and sustains bandwidth
- Support for `no snoop' attribute
  - Allows I/O to achieve line rate
  - Not limited by system coherent bandwidth (5.3 GB/s)
- Completion Coalescing for DMA reads
  - Sustainable DMA read b/w of 1.8 GB/sec on a x8 link
    - 256 B cache line coalescing with large read requests
    - B/W improvement exceeds 20% with this new feature
  - DMA Write bandwidth about 1.8 GB/s on x8 link
- No head of the line blocking in NP (reads)
  - Round-robin for top 8 (4 for a x4) entries which satisfy ordering
  - Data sent out in 256 B chunk for each request
  - NP entry popped from head of queue when data returned
  - Provides QoS and equitable bandwidth for multiple devices
    - Also helps achieve line rate since large requests from slower devices / interconnects will not block smaller requests from relatively faster devices/interconnects

Twin Castle Error Infrastructure

- Rich set of error logging and reporting features
  - allows error source and cause to be identified
  - first error, next Error, and syndrome logged
  - error logging / reporting can be disabled
  - Programmable severity with different interrupt types
- Error pollution and error containment enforced
  - Recoverable errors (CRC with retry and ECC)
  - Unrecoverable data errors can be poisoned.
  - Error data not allowed to propagate without poisoning
- PCI-E* Advanced Error Reporting support
- PCI-E* errors reporting through MSI or legacy
- Downstream PCI-E* errors are reported through in-band messages.

Hot Plug

- Each IMI link (memory board) hot-pluggable
- Each PCI-E* slot is hot pluggable
- Serial SMBus interface to 9555 chip for hotplug command and control (attention led, power led, etc)
- Interrupt mechanism for hot plug: SMI as well as MSI (for PCI-E*)

Power Management

- PCI-E*: ASPM Rx.L0s and L1
- PCI-E*: PCI-PM mechanisms
- PM messages
- Sideband interrupt (SMI) as well as MSI support for PM events
- System wide power management modes orchestrated in conjunction with ICH

Debug

- Rich debug features provides TTM advantages
  - Interconnect BIST (IBIST) on PCIE and IMI
  - Helps with board qualification
  - Margining hooks on all major interfaces
  - Transaction BIST for HVM testing
  - Performance Monitors for performance measurements

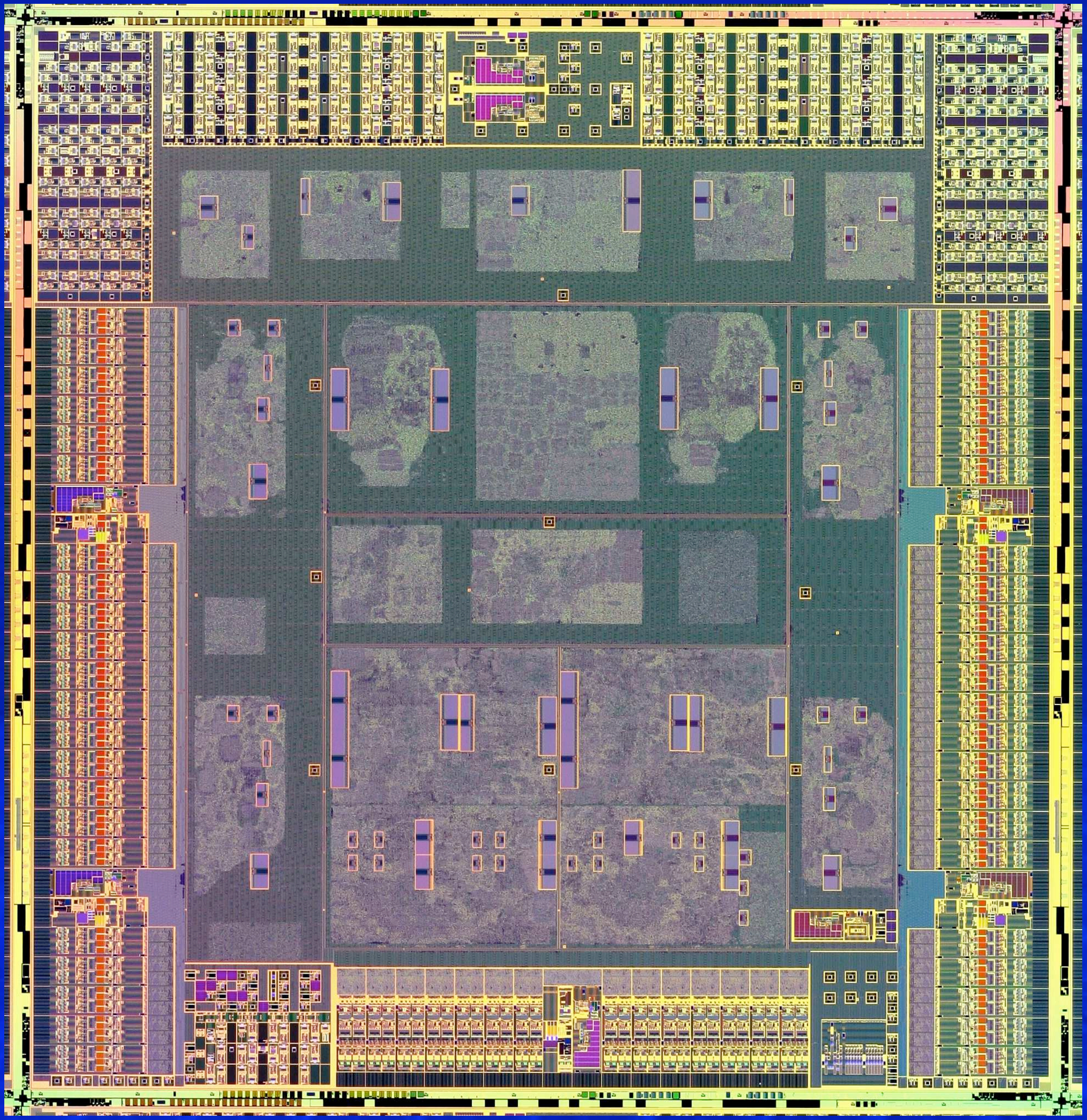
  - Event monitoring and triggering
    - Address and data matching: FSB
    - Header packet matching: PCI Express* and IMI
    - In-band messages: PCI Express and IMI
  - Response functions
    - Interface Throttle to stress interfaces
    - Chip Freeze
    - Error Injection to validate error correction/ reporting/ logging

  - JTAG/SMBus: access registers and inject transactions

TNB Chip Statistics

Process:	0.13 µ
Package Size:	42.5 mm
Transistors:	38.7 million
Cells:	2.2 million
Frequency:	2.67 GHz memory link
	2.5 GHz PCIE
	667 MT/s FSB
	333 MHz core

TNB Die



Summary

- Twincastle is an x86 based MP chipset.
- Leadership technology with multi-core multiple-CPU support, memory, and I/O to protect customer investments.
- Leadership RAS features.
- Several performance features to enable high performance at low cost
- Rich set of debug features to enable faster customer deployment