Montecito –
The next product in the Itanium® Processor Family

Cameron McNairy (Intel)
Rohit Bhatia (HP)

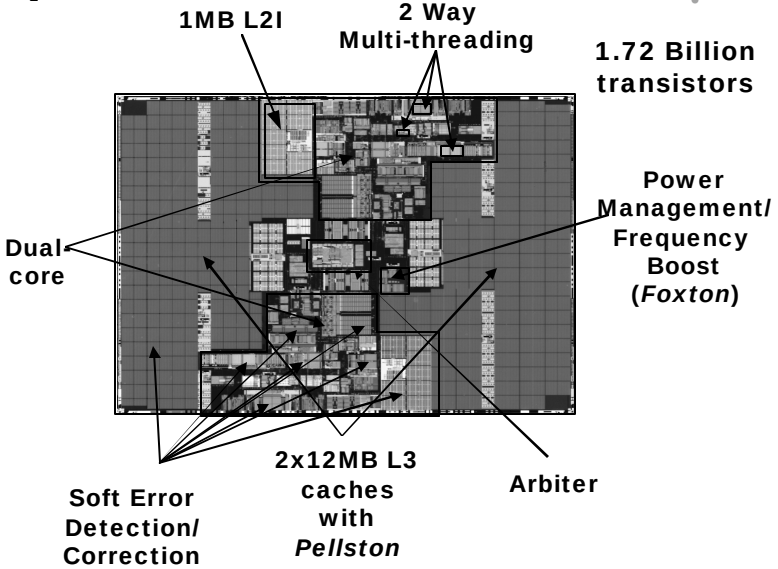
Agenda

- Montecito Overview
- Improving Instruction Level Parallelism
  - Cache, functional units, instructions
- Exploiting Thread Level Parallelism
  - Montecito Multi-threading
    - Sharing the core and caches
  - Montecito Multi-core
    - A single interface for two cores
- Power and thermal management
  - Exposing the entire power and thermal envelope
- Enterprise RAS+M
- Conclusion

Montecito Overview

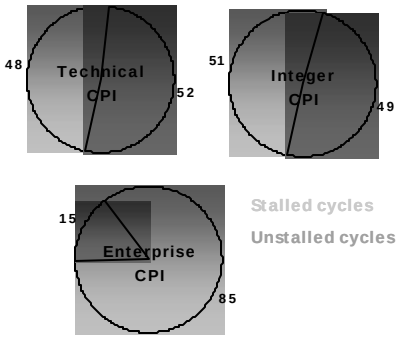
- Maintains full Itanium® 2 processor compatibility
  - System Interface compatible but enhanced
  - Power envelope is lowered to 100W from 130W
  - Thermal envelope is compatible
  - Software compatible but enhanced with new instructions
- Optimized for the enterprise
  - Performance
    - 2 cores and 2 threads/core
    - High frequency system interface
    - 27+ MByte of cache total
  - Power efficiency
    - 1.72 billion transistors @ 100W
  - Reliability, Availability, Serviceability, Manageability

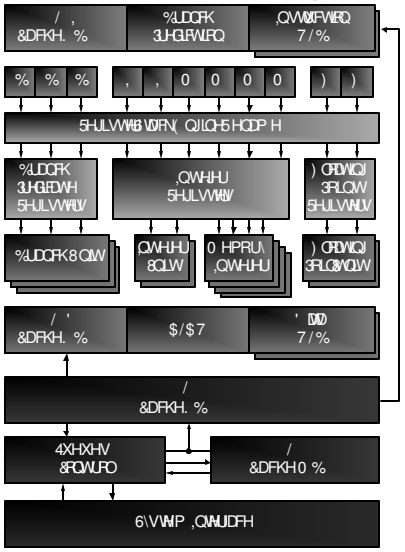
Key Features



Memory as the limiter

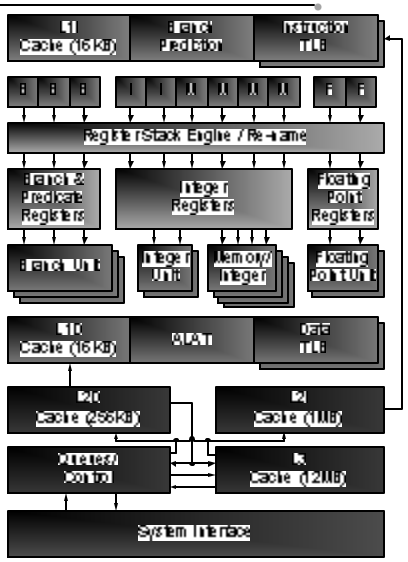
- EPIC approach works well
  - Many functional units available
- Dynamic behaviors reduce theoretical maximum
  - Stalls/misses from memory hierarchy
  - Functional unit asymmetries





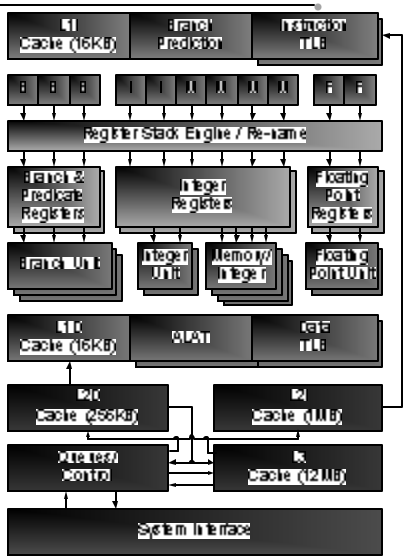
Improving ILP

- Improving the execution resources
  - New instructions
  - Additional shifter and popcount
- Improving the memory hierarchy
  - Split the L2 cache
    - Dedicated 1 MByte L2 instruction cache
    - Effectively grow L2 data cache
  - Larger L3
    - Grow L3 to 12 MByte
    - Maintain L3 latency of Itanium® 2 processor 6M and 9M
  - Queues and Control
    - Additional L3 and L2 victim buffers
    - More efficient control of queues

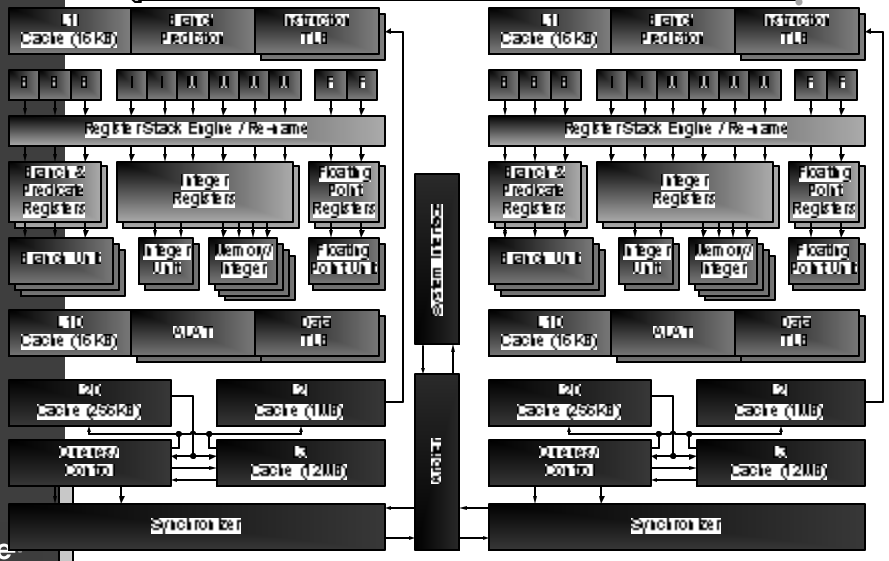


Exploiting TLP

- Interleave orthogonal threads to hide memory latency
  - Multiple cores in each socket
  - Multiple threads in each core
- Montecito Multi-Threading
  - Dynamically allocate resources based on effective use
    - Long latency events determine if a thread can effectively use execution resources
  - Some resources are competitively shared
  - Some resources are mutually exclusively shared



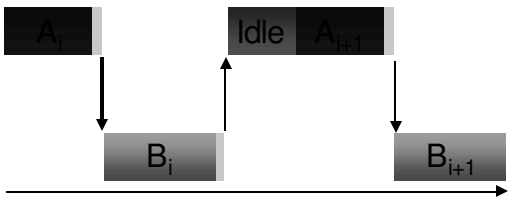
Multiple Cores



Montecito Multi-Threading

Serial Execution

Montecito Multi-threaded Execution



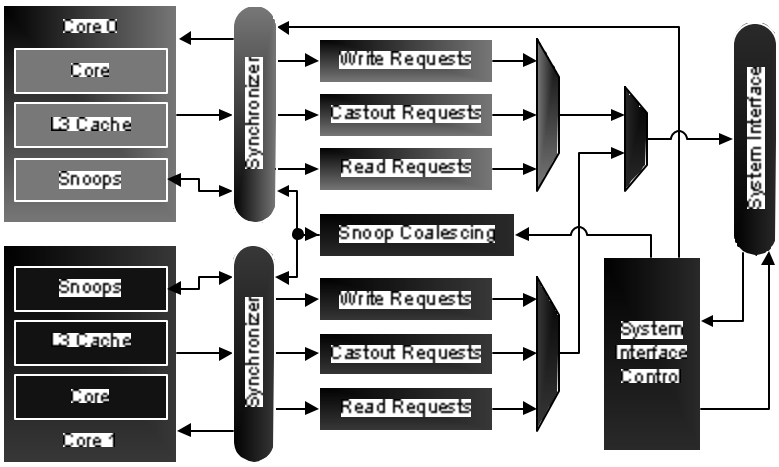
intel

Switching Threads

- Assume long latency event will stall execution
  - L3 misses or uncacheable accesses
  - Time slice expiration to ensure fairness
  - Maximum switch thresholds for forward progress
- Hysteresis
  - Urgency indicates a thread's ability to execute
  - Compare urgency at miss events
  - Latency from miss event to switch enables miss clustering
- Low power HALT enables dynamic allocation of CPU resource
- `hint@pause` gives software control

intel

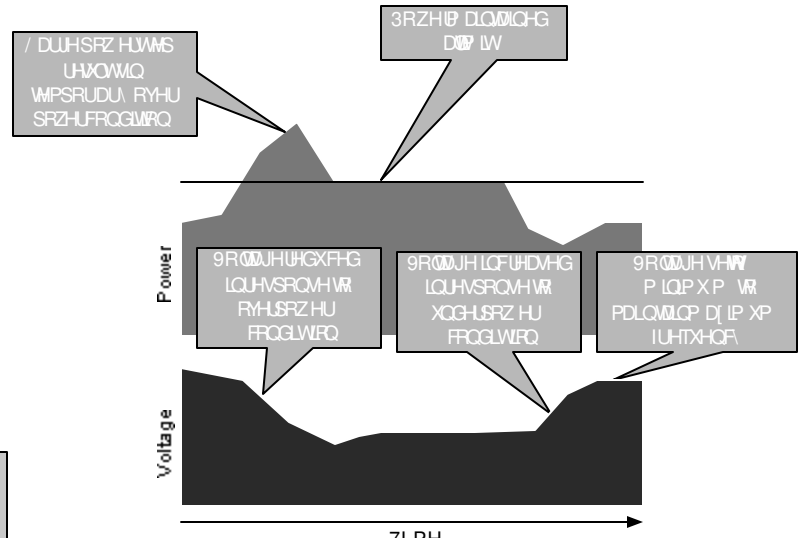
Montecito Multi-Core



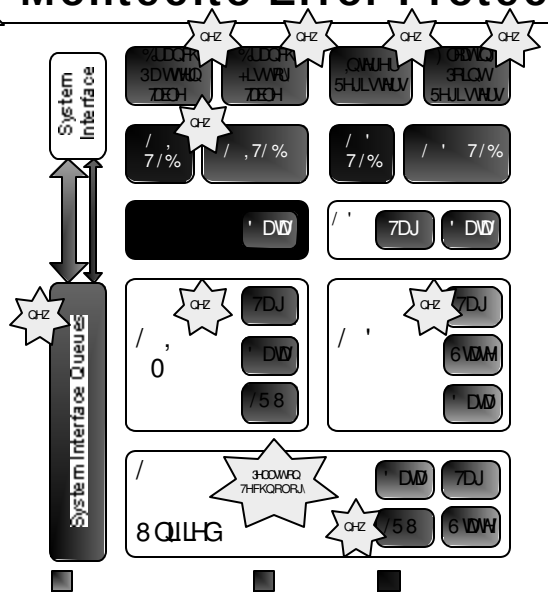
Foxton Technology

- Dynamically adjusting Voltage (V) and frequency (f) allows applications to exploit full power envelope
  - Monitor/calculate power and temperature
    - Set V to minimum value needed to support highest frequency
    - Over power and/or temperature results in voltage change
    - Frequency responds to global and local voltage
  - Large power change ? small frequency change from $P=CV^2f$
    - 3% power change with only 1% frequency change
    - Optimized around a TPCC activity factor
  - Activity factors (AF) – the switching power an application is likely to experience
    - Enterprise < Integer < Technical < Any known application
- Supports Demand Based Switching and OEM selectable power envelopes

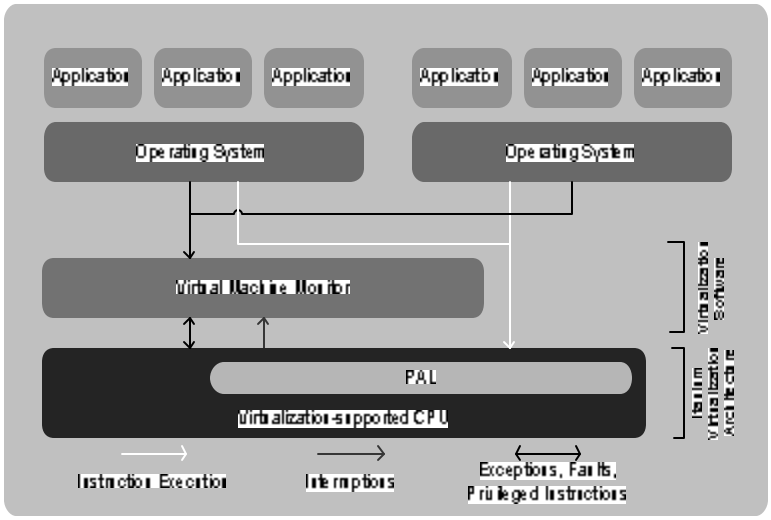
Power/Voltage Interaction



Montecito Error Protection



Silvervale Technology



Conclusion

- Montecito introduces several changes, innovations, and technologies
  - ILP improvement through execution and cache
  - TLP exploited through dynamically managed multiple thread and multiple cores
  - Foxton Technology provides power efficiency, frequency opportunity, and deployment flexibility
  - RAS improved through additional protection, logging, and Pellston Technology
  - Silvervale Technology provides system manageability
- 1.72 Billion transistors at 100W providing the industry with performance, capability, and flexibility