Tukwila – a Quad-Core Intel® Itanium® Processor

Eric DeLano
Intel® Corporation
Agenda

• Tukwila Overview
• Exploiting Thread Level Parallelism
• Improving Instruction Level Parallelism
• Scalability and Headroom
• Power and Frequency management
• Enterprise RAS and Manageability
• Conclusion
# Intel® Itanium® Processor Family Roadmap

## Processor Generation

### Intel® Itanium® Processor 9000, 9100 Series

<table>
<thead>
<tr>
<th>Dual Core</th>
<th>Quad Core (2 Billion Transistors)</th>
<th>Poulson</th>
<th>Kittson</th>
</tr>
</thead>
<tbody>
<tr>
<td>• 24MB L3 cache&lt;br&gt;• Hyper-Threading Technology&lt;br&gt;• Intel® Virtualization Technology&lt;br&gt;• Intel® Cache Safe Technology&lt;br&gt;• Lock-step data integrity technologies (9100 series)&lt;br&gt;• DBS Power Management Technology (9100 series)</td>
<td>• 30MB On-Die Cache, Hyper-Threading Technology&lt;br&gt;• QuickPath Interconnect&lt;br&gt;• Dual Integrated Memory Controllers, 4 Channels&lt;br&gt;• Mainframe-Class RAS&lt;br&gt;• Enhanced Virtualization&lt;br&gt;• Common chipset w/ Next Gen Xeon® processor MP&lt;br&gt;• Voltage Frequency Mgmt&lt;br&gt;• Up to 2x Perf Vs 9100 Series*</td>
<td>• Advanced multi-core architecture&lt;br&gt;• Hyper-threading enhancements&lt;br&gt;• Instruction-level advancements&lt;br&gt;• 32nm process technology&lt;br&gt;• Large On-Die Cache&lt;br&gt;• New RAS features&lt;br&gt;• Compatible with Tukwila platforms</td>
<td>9th Itanium® Product</td>
</tr>
</tbody>
</table>

### Targeted Segments

- Enterprise Business (Database, Business Intelligence, ERP, HPC, …)

### Availability

- **Poulson**: End 2008
- **Kittson**: Future
- **Tukwila**: Dual Core<br>**2006-07**

---

Intel and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. * Other names and brands may be claimed as the property of others. All products, dates, and figures are preliminary and are subject to change without notice.

Copyright © 2008, Intel Corp.

*Source: Intel performance projections
Tukwila Overview

- **Performance**
  - 4 cores with 2 threads/core
  - High memory and interconnect bandwidth
  - ~30 MBytes of total cache circuitry
  - Other ILP and TLP improvements

- **Scalability**
  - Directory coherency with ~2MB of directory cache
  - Up to 8 socket glueless
  - Higher scalability with OEM chipsets.

- **System integration and new system interfaces**
  - Integrated Memory Controllers and Router.
  - Intel QuickPath® Interconnect and new memory interconnect

- **Power efficiency and performance balance**
  - Dynamic Voltage/Frequency management

- **Reliability, Availability, Serviceability, Manageability**
  - Processor and platform level RAS features, plus virtualization, and partitioning features for improved resource utilization.

*World’s first 2 Billion transistor processor helps deliver improved performance & capabilities*
Die Photo

Dual Integrated Memory Controllers

Quad Multi-Threaded Cores

Power/Thermal/Freq Management

Memory interconnect

CrossBar Router

QuickPath® interconnect (QPI)
Block Diagram

QuickPath interconnect coherency engine for core caches

“Core” ↑

QuickPath interconnect coherency engine for a section of the physical memory

“Sysint: System Interface” ↓

Provides access to configuration resources and misc. external interfaces

Memory Controller

Home Agent

Crossbar Router

Memory Physical Layer

Processor Core 0+
6MB L3 Cache
Home Agent
Directory Cache
Intel® QPI Ports

Processor Core 1+
6MB L3 Cache
Caching Agent

Processor Core 2+
6MB L3 Cache
Caching Agent

Processor Core 3+
6MB L3 Cache
Caching Agent

Processor Core 0
6MB L3 Cache
Home Agent
Directory Cache

Processor Core 1
6MB L3 Cache
Caching Agent

Processor Core 2
6MB L3 Cache
Caching Agent

Processor Core 3
6MB L3 Cache
Caching Agent

Memory Controller

Tracks cache line owner and sharer information

Route and Link Layers

Physical Layer(x6)
4 full width and 2 half-width QPI Links

Configuration

Intel® QPI = Intel® QuickPath Interconnect

* Intel® QPI = Intel® QuickPath Interconnect

26 August 2008 – HotChips 20, Copyright 2008 Intel Corporation
Exploiting TLP

- 4 cores: 2x from previous generation
- 2 threads per core: Same as the previous generation
- Hyper-Threading Enhancements to hide more stall cycles
  - Improved thread switch events
    - Allow thread switches on pipeline stalls that are not necessarily L3 cache demand misses (e.g. secondary misses after a prefetch).
    - Switch on semaphore release
  - Improved thread switch algorithms
    - Thread switch decision is based on “urgency counters” which are based on hardware events. Changes to urgency update logic have improved multi-threading performance.

- Multi-Core implementation
  - Dedicated caches provide the lowest latency, highest bandwidth, and best Quality of Service characteristics.
  - Each core has its own high bandwidth interface directly to the on-die router.

2x TLP plus Hyper-Threading improvements
Improving ILP

- Same high ILP core as on the 9000/9100 series processor
  - WIDE execution: 6 wide instruction fetch and issue
    - 6 wide integer units, 2 wide FP units, 4 wide ld/st, 3 wide branch units
    - 1 cycle L1 data cache
    - Separate L1I, L1D, L2I, and L2D caches
  - Short 8 stage pipeline – HIGH performance and energy efficiency
  - 50bit physical addressing
- Increased frequency
- Improving the memory hierarchy
  - Increased memory bandwidth
  - 4k/8k/16k page size support in first level data TLB.
Scalability and Headroom

- Large caches and high bandwidth provides scalability and platform/application headroom
  - 30MBytes of total cache (vs. 26MB on previous generation)
  - 96GB/s total Intel QuickPath Interconnect (9x increase)
  - 34GB/s total memory bandwidth (6x increase)
  - Bandwidth increase >> core count increase

- Directory coherency for efficient scaling
  - Directory tracks cache line owners and sharers. Only owners and sharers are snooped.
  - Snoop traffic scales sub-linearly with system size for directory coherency rather than with the square of system size (for snoopy coherency).

- Supports up to 10 concurrent system global TLB purge broadcast transactions on QuickPath Interconnect.

- Supports glueless 8 socket systems.
  - Larger systems will be built using a hierarchy of multi-socket Nodes and OEM provided Node Controllers.

**Good performance across different system sizes, topologies & across applications. Includes scalable system interconnect**
Example Glueless topologies

4 CPU topology  8 CPU topology (memory not shown)

20 lane uni-directional channel (@4.8GT/s)
10 lane uni-directional channel (@4.8GT/s)
IOH: I/O Hub (e.g. bridge to PCIe)
Power/Frequency Management

- Dynamically adjusts voltage (V) and frequency (F) to maximize performance for a given power envelope and thermal envelope
  - Different applications may run at different frequencies
  - Optimized around a TPCC activity factor
  - Adjustments are transparent to the OS
  - Deterministic estimated power leads to deterministic voltage/frequency changes

- Power saved by deconfigured cores, halted threads, or threads with a low activity factor will automatically be reallocated to the remaining threads that can take advantage of the power

- Supports Demand Based Switching (DBS)
  - For reduced power states, lowering both V and F results in greater (~cubic with frequency) power savings compared to the previous generation.

- Supports multiple power envelopes

- Lowering V/F for thermal events helps guarantee data integrity

Allows balancing performance & power savings as needed
**Processor RAS**

- All major structures are protected.
  - Same extensive ECC and parity protection as on the 9000/9100 series processor core.
  - Extensive ECC and parity protection on new Sysint structures and datapaths

- Extensive use of Soft Error hardened (SE-hardened) latches and registers
  - SE-hardened latches/registers are 100x/80x (respectively) less susceptible to soft error events (due to alpha and cosmic particles).
  - More than 99% of Sysint latches are SE-hardened
  - More than 33% of Core latches are SE-hardened
  - 100% of Sysint register files are SE-hardened or have ECC

- Intel® Cache Safe Technology added to L2I, L2D, and Directory Cache (in addition to L3)
  - Cache lines that show too many errors in the field are mapped out by the processor.

- Core De-configuration
  - Cores that show too many errors in the field can be mapped out.

**Resiliency against soft and hard errors**
Memory RAS Features (1/3)

• Memory channel protection
  • Transient errors: Memory replay (N times) on CRC error
  • Persistent errors: Channel reset of Physical layer
  • Hard errors: Lane failover on channel reset

• Memory thermal management
  • Thermal sensor on DIMM provides trip point indication to memory controller command throttling logic via the memory interconnect
  • Several modes implemented to provide a flexible memory throttling response.

• Patrol and demand scrubbing
  – Prevent uncorrectable errors by writing good data/ECC when encountering a correctable error.
    • In background (patrol), or due to a request (demand)
Memory RAS Features (2/3)

Most processors cannot correct data when 2 different DRAM devices fail, which could result in memory loss and a fatal system crash.

- DRAM Protection – Double Device Data Correction (DDDC)
- Tukwila can fix both single and double device memory hard-errors and still correct an additional single bit error.
  - No performance penalty for mapped out devices
- Tukwila DDDC can improve system uptime and reduce DIMM replacement rates lowering overall service costs
Memory RAS Features (3/3)

- Memory channel Hot-Plug Support
  - Can be used for upgrades or to replace a faulty component without bringing down the system.

- Memory Migration and DIMM sparing
  - Copies memory from one physical location to another.
  - Using predictive failure analysis, this can avoid a system crash if memory goes bad.
  - Prepare for memory or field replaceable unit servicing
  - Migration provides more protection but requires more memory overhead (two memory channels), compared to DIMM sparing which only requires an extra DIMM.

Multiple levels of memory protection
QuickPath Interconnect RAS

- QuickPath Interconnect protection
  - Transient errors: Link Level Retry (N times) on CRC error
    - Two error detection “strengths”: 8 bit CRC across one flit (80 bits) or 16 bit CRC across two flits
  - Persistent errors: Reset of Physical layer
  - Hard errors: “Self-Healing” which means mapping out bad data or clock lanes on Physical layer reset

- Hot-Plug support
  - Can be used for upgrades or to replace a faulty component without bringing down the system

- Timeout mechanisms
  - To aid faulty component identification

- Error Containment mechanisms
  - Error signaling indicates scope of error. Eg. Data chunk, packet, or system
  - Provides opportunity to recover by killing the process instead of the whole machine

Tukwila implements all QuickPath Interconnect RAS features
Intel® Itanium® Virtualization Hierarchy

Physical Partitions
- Full electrical fault isolation, dedicated resources
- Coarse grain, more static control

Virtual Partitions
- Multiple OS/apps environments
- Dedicated or shared resources
- Finer grain, more dynamic control
- Supports one or more virtual partitions on one or more HW thread

OS Virtualization
- Allocation of single OS resources to specific users/apps
- Highly dynamic & flexible

Intel® Virtualization Technology for Itanium® architecture (Intel® VT-i)

Multiple Levels Of Granularity Provide Maximum Flexibility
Manageability Features

- Virtualization Enhancements: Vt-i extensions
  - Reduce latency by reducing instruction emulation code and reducing virtualization faults. This is achieved by:
    - New architecture state
    - New Guest copies of architecture state
    - New conditions using (selective) disable of virtualization faults
  - No Virtual Machine Monitor (VMM) modification required to receive the benefits of these extensions
- Dynamic Partitioning Support
  - Links can be enabled/disabled dynamically and snoop control and packet routing changed dynamically.
- Reconfiguration Support
  - Configuration registers can be dynamically changed via QuickPath Interconnect or via system management interface.

Improvements to Physical and Virtual Partitioning
Tukwila Status

- Multiple OSes have booted: Linux, Windows (3 versions), and HP-UX.
  - More OSes expected to boot in the future!
- A wide range of platform topologies are in volume testing
  - 4 and 8 socket (32-64 threads) glueless and hierarchical SMP with OEM Node Controller
  - Common chipset with Xeon® MP.
- Multiple OEM platforms have powered on and are in validation.
- On-track to achieve 2x performance compared to the previous generation.

Targeted to strengthen the Itanium ecosystem comprising multiple OSes, hardware platforms, and enterprise applications.
Conclusion

- Improved performance: Better TLP and ILP
- Improved scalability, platform flexibility, and headroom
- Power and frequency management
- Mission Critical RAS: Processor RAS, Memory RAS, QuickPath Interconnect RAS
- Manageability and Virtualization for resource management and efficiency

*Designed for a leap in mission critical capabilities*
Q&A
Tukwila Caches

L2-Instruction Cache  L2-Data Cache  L3 Tag

Directory Cache  system interface  L3 Data/Instruction Cache
# Cache Summary

<table>
<thead>
<tr>
<th></th>
<th>L1 Inst Cache per core</th>
<th>L1 Data Cache per core</th>
<th>L2 Inst Cache per core</th>
<th>L2 Data Cache per core</th>
<th>L3 Cache per core</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>16 KB</td>
<td>16 KB</td>
<td>512 KB</td>
<td>256 KB</td>
<td>6 MB</td>
</tr>
<tr>
<td>Line Size</td>
<td>64-byte</td>
<td>64-byte</td>
<td>128-byte</td>
<td>128-byte</td>
<td>128-byte</td>
</tr>
<tr>
<td>Ways</td>
<td>4-way</td>
<td>4-way</td>
<td>8-way</td>
<td>8-way</td>
<td>12-way</td>
</tr>
<tr>
<td>Write Policy</td>
<td>-</td>
<td>WT</td>
<td>-</td>
<td>WB</td>
<td>WB</td>
</tr>
<tr>
<td>Latency</td>
<td>1 clk</td>
<td>1 clk</td>
<td>7 clks</td>
<td>5,7,9 clks(^1)</td>
<td>15+ clks(^1)</td>
</tr>
<tr>
<td>Protection</td>
<td>SEC(^2) via Parity</td>
<td>SEC(^2) via Parity</td>
<td>SEC(^2) via Parity</td>
<td>SEC(^2) via ECC</td>
<td>SEC(^2) via ECC</td>
</tr>
</tbody>
</table>

\(^{1}\)Add 1 clk for FP loads  
\(^{2}\)SEC: Single Bit Error Correction  

Note: Higher level caches do not force inclusion

Tukwila’s native cache line size is 128 bytes - Cache fills, writebacks & snoops converted into 2 QuickPath Interconnect 64 byte line transfers
Tukwila Die and Package details

- 65nm bulk CMOS, 8 layer Cu interconnect
- 2.05 billion transistors
- 32.5mm X 21.5mm = 700 mm² die size
- 811 signal package pins in a 1248 pin LGA socket
- 5 major voltage and frequency domains
Example Hierarchical SMP

OEM interconnect

20 lane channel per direction(@4.8GT/s)
Number, type, and size are OEM dependent.
RAS and Manageability Overview

• Tukwila RAS and Manageability features fall into these categories.
  - Processor (Core + Sysint) RAS
  - Memory RAS
  - QuickPath Interconnect RAS
  - Manageability and Virtualization

• Tukwila supports a large set of RAS+M features providing value across the entire platform:
  - Reliability:
    • Assurance that computational results are correct.
  - Availability:
    • Assurance that the system is available to perform computations.
  - Serviceability:
    • Assurance that the error is reported and that the faulty component can be identified and replaced.
  - Manageability and Virtualization:
    • Ability to configure and/or share resources to handle error events or to increase system utilization and efficiency
Read to Idle/Shared Line: Directory

(A, B, C, … are caching agents)

No Snoops are issued because the directory knows no one has a copy of the line. This is the flow used by Tukwila.
All caching agents are snooped for all requests, consuming more interconnect bandwidth.

For comparison purposes only. Tukwila does not use this snoopy flow.

(not to scale)