Intel® 5520 Chipset: An I/O Hub Chipset for Server, Workstation, and High End Desktop

Debendra Das Sharma
Principal Engineer, Digital Enterprise Group
Intel Corporation
Contributors

• Nilesh Bhagat
• Rob Blankenship
• Celeste Brown
• Sam Chiang
• Ken Creta
• Debendra Das Sharma
• Hanh Hoang
• Siva Gadey Prasad
• S. Jayakrishna
• Michelle Jen
• Daniel Joe
• Chandra P Joshi
• Lily Looi
• Dean Mulla
• Sridhar Muthrasanallur
• K. Pattabhiraman
• Guru Rajamani
• Bill Rash
• Aquiles Saenz
• Rajesh Sankaran
• Miles Schwartz
• Mark R Swanson
• Patrick Tsui
• Cyprian Woo
• Robin Zhang
• and many others..
Agenda

- Platform Overview
- Feature Set
- Micro-architecture and Transaction Flows
- Performance
- Chip Statistics
- Summary
Server Chipset Evolution

**Trends:**
- Increased integration => fewer chips
- Increased bandwidth for coherent traffic: FSB -> multiple FSBs -> Links
- Increased memory capacity and bandwidth (memory buses -> FBD -> Integrated memory controller)
- Increased I/O connectivity and bandwidth: PCI bus /bridges to PCIe* links

FSB: Front Side Bus. *: Published in Hotchips
Platform Overview

- Platform Transition:
  - Front Side Bus to Intel® Quick Path Interconnect Link
  - Memory controller integrated to CPU

- Intel® 5520 Chipset is an I/O Hub
  - Bridge between QPI and I/O
  - First server chipset with PCIe*2.0
  - Flexible I/O with 36 PCIe* 2.0 Lanes (3 to 10 Root ports of different widths)
  - One or two CPU socket connection
  - Server and Workstation platforms
  - Customized to High-End Desktop (X58)
  - Multi-generational CPU upgrades

* Other names and brands may be claimed as the property of others. Copyright © 2009, Intel Corporation. Hotchips 2009
Dual IOH Platform

- Increased I/O Connectivity
  - 72 PCIe* 2.0 Lanes
  - Maintains I/O flexibility
  - Workstation and server
- Two IOHs coordinate their accesses to appear as a single IOH to the CPU(s)
  - Enhanced QPI protocol in the IOH-IOH Link (e.g. Recall cache lines from other IOH to ensure forward progress)
  - Dynamic adjustment of CPU HOME resources (trackers) to ensure performance
- One or two CPU sockets

* Other names and brands may be claimed as the property of others. Copyright © 2009, Intel Corporation. Hotchips 2009
**Intel® Quick Path Interconnect (QPI)**

- Two QPI links, each connecting to a CPU socket (or IOH)
- Differential links with forwarded clock
- Supports up to 17” channel length with two connectors
- Each QPI is 20 bits wide, up to 6.4 GT/s Data Rate
- 51.2 GB/s raw bandwidth and 31.5 GB/s of sustained real data transfer with two QPI Links at 6.4 GT/s
- CRC protection on every 80 bits along with Link level retry
- Unordered fabric for performance
- Snoopless IOH ensures CPU does not snoop IOH
  - QPI connecting IOH only used for I/O accesses
  - IOH does not remain in snoop path of CPU through upgrades

* Other names and brands may be claimed as the property of others. Copyright © 2009, Intel Corporation. Hotchips 2009
**Intel® 5520 I/O Virtualization Features**

- First server chipset shipping with PCIe* I/O Virtualization
  - Multiple VMM vendor (Xen, KVM, Citrix, Parallels, VMWare, etc) products enabled

- All inbound memory requests undergo VT-d (IOMMU) address translation and access privilege check

- Translation based on the requestor’s BDF <Bus, Device, Function> and page address
  - Each BDF may belong to a different virtual machine (VM)

- Context Entry (2-level) and a 4-level address translation structure resident in memory

- Multi-level caching structures inside IOH
  - Context entry cache, L1/L2, L3 and IOTLB caches
  - Invalidation: Register-based and Queue-based from memory

- Supports PCI-SIG defined ATS/ACS for end-point caching and access rights check

[Assigned I/O for VMs:
VM1: (P1,0), (P2,2)
VM2: (P1, 1), (P2, 0), P3
VM3: (P1, 2), (P2, 1)]
Reliability, Availability, Serviceability (RAS)

- High speed interfaces (ESI, PCIe*, QPI) are CRC protected with Link level retry
- Poison support throughout (PCIe*, ESI, QPI, internal paths)
- Internal data path mostly ECC/CRC protected
  - Configuration registers are parity protected
- Detailed error logs in each interface for each error type
- Advanced Error Reporting Structures
  - Both MSI-based as well as interrupt pin based notification
- Hot-plug support on all PCIe* Links
- Lane degradation and reversal support on PCIe*
- Live Error Recovery for guaranteed error containment
  - Can program each error type in to take the PCIe* link down
  - Example: If a device can not handle poison data, we can take that Link down rather than propagate poison to the device
Intel® 5520 Chipset Features

- **Isochronous** support on ESI for Quality of Service
  - Separate Virtual Channel for HD Audio (latency & bandwidth)
  - Separate Virtual Channel for USB (bandwidth guarantee)

- **IOAPI C**: I/O Advanced Programmable Interrupt Controller
  - Converts legacy interrupts to Message Signaled Interrupt
  - Avoids interrupt sharing => better for performance

- **QuickData Acceleration** for CPU off-load
  - DMA Move Engine w/ CRC capability: 5GB/s of bandwidth
  - Eight functions for better virtualization support
  - Direct Cache Access to CPU cores from PCIe*

- **Integrated Manageability Engine** for System management
  - Embedded microcontroller with encryption engine
  - Sideband paths to components
  - Inband PCI ports (serial port, DMA, emulated IDE, HECI)
  - Separate power well

* Other names and brands may be claimed as the property of others.  Copyright © 2009, Intel Corporation. Hotchips 2009
**Transaction Flow Example: DMA Read**

1. Memory Read (MRd) loaded to Queue
2. VTd translation and access right check
3. Ordering check. Packet broken to Cache line(s). Request sent to QPI0 (home in CPU0)
4. QPI 0: Conflict check; Check trackers; Consume tracker; Send request to CPU0
5. QPI 1 sends snoop request to CPU 1
6. CPU 1 sends snoop response to CPU0
7. CPU 0 sends Data Return to IOH. QPI 0 releases the tracker on receipt of Data Return
8. Data loaded to outbound PCIe* queue
9. Data completion sent out on PCIe*
DMA Write: Request for Ownership

1. Memory Write (MWr) loaded to queue
2. VTd translation and access right check. Page Walk on a miss.
3. Packet broken to Cache line(s). Request for Ownership (RFO) sent to QPI0 (home in CPU0). No Ordering check to pipeline RFOs
4. QPI 0: Conflict check; Check trackers; Consume tracker; Send request to CPU0
5. QPI 1 sends snoop request to CPU 1
6. CPU 1 sends snoop response to CPU0
7. CPU 0 returns the (Exclusive) Ownership of the Cache Line (without Data) to IOH
DMA Write: Writeback Phase

- **8. QPI 0**: ownership notification to CDP so that it can process DMA Write
- **9. CDP** waits till the posted transaction gets to the top of the posted queue, per PCIe* Ordering rules
- **10. CDP**: Check with QPI to ownership still there; perform write if there; else request line again
- **11. QPI 0** performs Writeback of Data and relinquishes ownership
- **12. CPU 0** sends completion for the Writeback Transaction. Tracker released for subsequent reuse

* Other names and brands may be claimed as the property of others. Copyright © 2009, Intel Corporation. Hotchips 2009
Performance

<table>
<thead>
<tr>
<th>Configuration (Single IOH)</th>
<th>100% Rd (GB/s)</th>
<th>100% Wr (GB/s)</th>
<th>50-50 Rd/Wr (GB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMA w/ 3 PCIe 2.0* cards (2 x16, 1 x4)</td>
<td>15.9</td>
<td>12.4</td>
<td>15.2</td>
</tr>
<tr>
<td>NUMA w/ 4 x8 PCIe 2.0* Cards</td>
<td>13.9</td>
<td>12.3</td>
<td>14.7</td>
</tr>
<tr>
<td>Interleaved w/ 4 x8 PCIe 2.0* Cards</td>
<td>14.1</td>
<td>12.1</td>
<td>14.5</td>
</tr>
</tbody>
</table>

- Measured results on single IOH at launch. More than 2X previous generations due to QPI as well as PCIe 2.0*
- 100% Rd B/W PCIe* limited
- 100% Wr and 50-50 RW B/W is tracker entry limited
  - Writes occupy tracker entry longer since there are two round-trips on QPI Link
  - Bandwidth expected to scale with compute capability in subsequent CPU generations due to more tracker entries from CPU
- Other details: IO Meter benchmark, 2.93 GHz 5500, QPI at 6.4 GT/s, 1333MHz RDIMM DDR3(6 x2 GB, 2 x8 channel), Request Size: 4KB, Max payload: 256B

* Other names and brands may be claimed as the property of others. Copyright © 2009, Intel Corporation. Hotchips 2009
Power

• PCIe* and ESI:
  – Active State Power Management puts idle link to low power L1 state
  – PCI* Power Management mechanisms to allow system software to manage System Sleep states (entry as well as exit)
• QPI: Support for low power L1 state on idle link
• Several power savings measures in the design (e.g., fine-grain and coarse-grain clock gating)
• System wide sleep state orchestrated with South Bridge
• Power numbers:
  – TDP: 27.1 W
    – All Links working on full speed (e.g., QPI at 6.4 GT/s and PCIe* at 5 GT/s)
    – All features and internal devices enabled
    – No active state power management benefit assumed
    – Accounts for worst possible combination of process, voltage, temperature
  – Idle power: 10 W (through system low power state)
Chip Statistics

- 65 nm process technology
- Die Size: 13.6 mm X 10.4 mm
- ~100 M transistors
  - 33x original Pentium
- Package: FCBGA 37.5 x 37.5 mm, 1.067 pin pitch, 10 layer
- Signal Pins: 570; total pins on package: 1295
Summary

• Intel® 5520 is first QPI-based chipset with PCIe® 2.0

• Leadership features
  – I/O bandwidth with flexible I/O Connectivity (36 or 72 PCIe 2.0* Lanes) for various segments
  – I/O Virtualization
  – QuickData for I/O Acceleration
  – Manageability
  – Isochrony for Quality of Service

• Designed to last multiple CPU generations on the same platform to protect customer investments

* Other names and brands may be claimed as the property of others.  Copyright © 2009, Intel Corporation. Hotchips 2009