AMD Opteron™ Multicore Processors

Brian Waldecker | February 1, 2009
Senior Member of Technical Staff
System Optimization Engineering Group
Processor Solutions Engineering, AMD
Austin TX

Pat Conway | Presenter
Principal Member of Technical Staff
Unified North Bridge team
Processor Solutions Engineering, AMD
Sunnyvale, CA
Outline

- AMD Roadmaps and Decoder Rings
- Hardware Overview
- Software Overview
- Questions
Roadmap
(and Decoder Rings)
# Planned Server Platform Roadmap

<table>
<thead>
<tr>
<th>Platform Type</th>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2/4-way Enterprise Platform</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>“Maranello”</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td>Socket G34 with AMD SR56x0 and SP5100</td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td>Magny-Cours New Architecture</td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Six-Core AMD Opteron™ Processor with AMD Chipset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td>Socket F(1207) with AMD SR56x0 and SP5100</td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td>Shanghai/Istanbul</td>
</tr>
<tr>
<td><strong>“Socket F (1207)”</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td>Socket F(1207) with Nvidia and Broadcom chipsets</td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td>Rev F, Barcelona, Shanghai, Istanbul</td>
</tr>
<tr>
<td><strong>1/2-way Power Efficient Platform</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>“San Marino”</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td>Socket C32+AMD SR56x0 and SP5100</td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td>Lisbon, New Architecture</td>
</tr>
<tr>
<td><strong>“Adelaide”</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td>Socket C32 EE+AMD SR5650 and SP5100</td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td>Lisbon, New Architecture</td>
</tr>
<tr>
<td><strong>1-way Platform</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>“Buenos Aires”</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td>Socket AM3 with AMD SR56x0 and SP5100</td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td>Suzuka</td>
</tr>
<tr>
<td><strong>“Socket AM2+”</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td>Socket AM2+ with Nvidia and Broadcom chipsets</td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td>Rev F, Budapest, Suzuka</td>
</tr>
</tbody>
</table>
AMD Multi-core Processor trends

Timeline: 2007 2008 2009 2010 2011

- 2FLOP/clk
- 4FLOP/clk
- 8FLOP/clk

- DDR2 533, 667, 800
- DDR3 1333, 1600

- 60 nm → 45 nm → 32 nm

- 16\(^w\) 32\(^e\)

- GFLOP/s
- GB/s
- Watts

* More computation while using less power per core
# x86 64-bit Architecture Evolution

<table>
<thead>
<tr>
<th>Year</th>
<th>AMD Opteron™</th>
<th>“Barcelona”</th>
<th>“Shanghai”</th>
<th>“Istanbul”</th>
<th>“Magny-Cours”</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>90nm SOI</td>
<td>65nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
</tr>
<tr>
<td>2005</td>
<td>90nm SOI</td>
<td>65nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
</tr>
<tr>
<td>2007</td>
<td>65nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
</tr>
<tr>
<td>2008</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
</tr>
<tr>
<td>2009</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
</tr>
<tr>
<td>2010</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
<td>45nm SOI</td>
</tr>
</tbody>
</table>

- **Mfg. Process**
  - 90nm SOI
  - 65nm SOI
  - 45nm SOI
  - 45nm SOI
  - 45nm SOI

- **CPU Core**
  - K8
  - Greyhound
  - Greyhound+
  - Greyhound+
  - Greyhound+

- **L2/L3**
  - 1MB/0
  - 512kB/2MB
  - 512kB/6MB
  - 512kB/6MB
  - 512kB/12MB

- **HyperTransport™ Technology**
  - 3x 1.6GT/s
  - 3x 2GT/s
  - 3x 4.0GT/s
  - 3x 4.8GT/s
  - 4x 6.4GT/s

- **Memory**
  - 2x DDR1 300
  - 2x DDR1 400
  - 2x DDR2 667
  - 2x DDR2 800
  - 2x DDR2 1066
  - 4x DDR3 1333

---

**Max Power Budget Remains Consistent**
A Strong Partnership in HPC: Past, Present, and Future

Cray XT5 & XT5a

2nd Gen. Opteron™ (K8 based)

3rd Gen. Opteron™ (Barcelona, Shanghai, Istanbul)

Designed for Upgradeability (G34+)

CX1

2010

2009

2008

“Baker”+ “Granite”

“Baker”

“Marble”

Cray XT4

Cray XMT

Scalar

Vector

Multithreaded

7 AMD Hex-Core Processors | Nersc/OLCF/NICS Cray XT5 Workshop | February 2010
“Istanbul” Silicon
Shanghai to Istanbul

- 6 cores (~1.5X flops)
  - Same per core L1 & L2
  - Same shared L3
  - NB & Xbar upgrades (going from 4 to 6 cores)
- HT Assist – provides 3 probe scenarios
  - No probe needed
  - Directed probe
  - Broadcast probe
- Memory BW and latency improvement
  - Amount depends on platform and configuration
- Socket Compatibility
Cache Hierarchy

Dedicated L1 cache
- 2 way associativity.
- 8 banks.
- 2 128-bit loads per cycle.

Dedicated L2 cache
- 16 way associativity.

Shared L3 cache
- 32 way (Barcelona),
  48 way (Shanghai and
  Istanbul) associativity.
- fills from L3 leave
  likely shared lines in
  L3.
- sharing aware
  replacement policy.
Core Micro Architecture
FastPath? Macro-Ops? Micro-Ops?

Notes / Considerations

Avoid having more than 2 or 3 branches per 16B of instructions.

Three Decode Categories (FastPath also called DirectPath)
- DirectPath Single - best
- DirectPath Double - better
- VectorPath (microcode) - good

Improved Out-of-Order Load Execution
In-Order Address Generation (per AGU)
Store-to-Load Forwarding Support
**TLB Review (Barcelona, Shanghai, Istanbul, Magny-Cours)**

- Support for 1GB pagesize (4k, 2M, 1G)
- 48 bit physical addresses = 256TB (increase from 40bits on K8)

**Data TLB**
- **L1 Data TLB**
  - 48 entries, fully associative
  - all 48 entries support any pagesize
- **L2 TLB**
  - 512 4k entries, and
  - 128 2M entries

**Instruction TLB**
- **L1 Instruction TLB**
  - fully associative
  - support for 4k or 2M pagesizes
- **L2 Instruction TLB**
Data Prefetch: Review of Options

- **Hardware prefetching**
  - DRAM prefetcher
    - tracks positive, negative, non-unit strides.
    - dedicated buffer (in NB) to hold prefetched data.
    - Aggressively use idle DRAM cycles.
  - Core prefetchers
    - Does hardware prefetching into L1 Dcache.

- **Software prefetching instructions**
  - MOV (prefetch via load / store)
  - prefetcht0, prefetcht1, prefetcht2 (currently all treated the same)
  - prefetchw = prefetch with intent to modify
  - prefetchnta = prefetch non-temporal (favor for replacement)
Six-Core AMD Opteron™ Processor

Performance
- Six-Core AMD Opteron™ Processor
  - 6M Shared L3 Cache
  - North Bridge enhancements (PF + prefetch)
  - 45nm Process Technology
- DDR2-800 Memory
- HyperTransport-3 @ 4.8 GT/sec

Reliability/Availability
- L3 Cache Index Disable
- HyperTransport Retry (HT-3 Mode)
- x8 ECC (Supports x4 Chipkill in unganged mode)

Virtualization
- AMD-V™ with Rapid Virtualization Indexing

Manageability
- APML Management Link

Scalability
- 48-bit Physical Addressing (256TB)
- HT Assist (Cache Probe Filter)

Continued Platform Compatibility
- Nvidia/Broadcom-based F/1207 platforms

4 socket example block diagram

SE: 2.8GHz
Std: 2.6GHz
HE: 2.1GHz
EE: 1.8GHz
Six-Core AMD Opteron™ Processor

Performance
- Six-Core AMD Opteron™ Processor
- 6M Shared L3 Cache
- North Bridge enhancements (PF + prefetch)
- 45nm Process Technology
- DDR2-800 Memory
- HyperTransport-3 @ 4.8 GT/sec

Reliability/Availability
- L3 Cache Index Disable
- HyperTransport Retry (HT-3 Mode)
- x8 ECC (Supports x4 Chipkill in unganged mode)

Virtualization
- AMD-V™ with Rapid Virtualization Indexing

Manageability
- APML Management Link

Scalability
- 48-bit Physical Addressing (256TB)
- HT Assist (Cache Probe Filter)

Continued Platform Compatibility
- Nvidia/Broadcom-based F/1207 platforms

<table>
<thead>
<tr>
<th>Product, Freq, Dram</th>
<th>STREAM Bandwidth (GB/s)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Barcelona,” 2.3/2.0, RDDR2-667</td>
<td>17.2 20.5</td>
</tr>
<tr>
<td>“Shanghai,” 2.7/2.2, RDDR2-800</td>
<td>21.4 24</td>
</tr>
<tr>
<td>“Istanbul,” 2.4/2.2, RDDR2-800</td>
<td>22 42 81.5</td>
</tr>
</tbody>
</table>

*Based on measurements at AMD performance labs. See backup slide for configuration information.
Significantly Higher Performance in Same Power Envelope

Two-socket servers using Six-Core AMD Opteron™ processors significantly outperform two-socket servers using Dual-Core and Quad-Core AMD Opteron™ processors without consuming significantly more power.
Istanbul to Magny-Cours

- 12 cores per socket (2 Istanbul die MCM)
  - Same per core L1 & L2
  - Same shared L3
  - NB & Xbar upgrades (going from 4 to 6 cores)
- DDR3 dimms (up to DDR3-1333)
  - 4 memory channels/socket (2 channels/die)
  - “local” memory refers to die, not socket
- Memory BW improvements
  - Same Probe filters as Istanbul
  - DDR3
“Magny-Cours” utilizes a Directly Connected MCM.

Package has 12 cores, 4 HT ports, & 4 memory channels.

Die (Node) has 6 cores, 4 HT ports & 2 memory channels.
(**) XFIRE BW is the maximum available coherent memory bandwidth if the HT links were the only limiting factor. Each node accesses its own memory and that of every other node in an interleaved fashion.
Cray XT5 Blade and Compute Node

- Eight Opteron™ cpus
- Seastar2+ chips form 3-D torus interconnect
- Four Compute Nodes per Blade (2 cpus per node)
- Memory DIMMS
- High Velocity Airflow
- Low Velocity Airflow
HT Assist Feature (Probe Filters)
Inter-socket Probes and Probe Responses travel:
SRI -> XBAR -> cHT -> cHT -> XBAR -> SRI

Probes Requests initiate at home memory node, but return directly to node making initial memory request.

**key:**
cHT = coherent HyperTransport
ncHT = non-coherent HyperTransport
XBAR = crossbar switch
SRI = system request interface (memory access, cache probes, etc.)
MCT = memory controller
HB = host bridge (e.g. HT to PCI, SeaStar, etc.)
HT Assist and Memory Latency

With “old” broadcast coherence protocol, the latency of a memory access is the longer of 2 paths:
- time it takes to return data from DRAM and
- the time it takes to probe all caches

With HT Assist, local memory latency is significantly reduced as it is not necessary to probe caches on other nodes.

Several server workloads naturally have ~100% local accesses
- SPECInt®, SPECfp®
- VMMARK™ typically run with 1 VM per core
- SPECpower_ssj® with 1 JVM per core
- STREAM

Probe Filter amplifies benefit of any NUMA optimizations in OS/application which make memory accesses local
HyperTransport™ Technology HT Assist (Probe Filter)

Key enabling technology on “Istanbul” and “Magny-Cours”

HT Assist is a sparse directory cache
- Associated with the memory controller on the home node
- Tracks all lines cached in the system from the home node

Eliminates most probe broadcasts (see diagram)
- Lowers latency
  - local accesses get local DRAM latency, no need to wait for probe responses
  - less queuing delay due to lower HT traffic overhead
- Increases system bandwidth by reducing probe traffic
Cache Coherence Protocol

- Track lines in M, E, O or S state in probe filter
- PF is fully inclusive of all cached data in system
  - if a line is cached, then a PF entry must exist.
- Presence of probe filter entry says line in M, E, O or S state
- Absence of probe filter entry says line is uncached
- New messages
  - Directed probe on probe filter hit
  - Replacement notification E -> I (clean VicBlk)
**Q:** Where do we store probe filter entries without adding a large on-chip probe filter RAM which is not used in a 1P desktop system?

**A:** Steal 1MB of 6MB L3 cache per die in “Magny-Cours” systems

<table>
<thead>
<tr>
<th></th>
<th>way 0</th>
<th>way 1</th>
<th>way 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3 cache</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dir</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>set</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Implementation in fast SRAM (L3) minimizes
- Access latency
- Port occupancy of read-modify-write operations
- Indirection latency for cache-to-cache transfers
Format of a Probe Filter Entry

- 16 probe filter entries per L3 cache line (64B), 4B per entry, 4-way set associative
- 1MB of a 6MB L3 cache per die holds 256k probe filter entries and covers 16MB of cache
Software
Compiler Options

- The Portland Group (PGI) family of compilers
  - Support for Linux and Windows.
  - Debuggers and Profilers for OpenMP and MPI.
- GCC and GNU Tools for AMD (gcc, glibc, binutils/gdb)
  - AMD actively contributes improvements targeting AMD cpus.
- C/C++/Fortran compilers based on Open64 technology
  - Available for download on AMD Developer Central at [http://developer.amd.com/cpu/open64](http://developer.amd.com/cpu/open64) in source and binary forms.
Some Notable PGI Flags

A Good Base Set of Flags

C/C++
-ffastsse -Mipa=fast -Mipa=inline -tp shanghai-64

Fortran
-ffastsse -Mvect=short -Mipa=fast -Mipa=inline -tp shanghai-64

(note: “-tp shanghai-64” is also appropriate for Istanbul)

And Some More to Usually Consider

<table>
<thead>
<tr>
<th>Flag</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Msmartalloc</td>
<td>Use smart malloc routines.</td>
</tr>
<tr>
<td>-Msmartalloc=huge</td>
<td>Use smart malloc routines with large pages (depends on amount of OS allocated large pages and the PGI_HUGE_PAGES environment variable).</td>
</tr>
<tr>
<td>-Mvect=fuse</td>
<td>Enables vectorizer to fuse loops.</td>
</tr>
<tr>
<td>-Mpf1=indirect (first pass) -Mpf0=indirect (second pass)</td>
<td>Use profile feedback optimizations. (requires compiling, doing training run(s), and recompiling).</td>
</tr>
<tr>
<td>-Msafeptr (C/C++ only)</td>
<td>Says arrays don’t overlap and different pointers point to distinct locations. (has many sub-options).</td>
</tr>
<tr>
<td>-Mfprelaxed</td>
<td>Allows relaxed precision for certain Intrinsic functions (sqrt, rsqrt, order, div).</td>
</tr>
</tbody>
</table>
ACML 4.3.0

- L3 BLAS improvements
  - SGEMM for Six-Core AMD Opteron™ Processor
  - New Intel DGEMM and SGEMM kernels
    - Supporting Woodcrest, Penryn, Nehalem
    - Competitive with MKL
  - New DGEMM "fast memory allocation" scheme
    - allows improved performance of other routines (such as LAPACK) which make heavy use of DGEMM

- Six-Core AMD Opteron™ processor tuning for Level 1 BLAS
  - xDOT, xCOPY, xAXPY, and xSCALE

- 3DFFT performance improvements
  - Now outperforming MKL

- AMD Family 10h tuning for real-complex FFTs
  - csfft, dzfft, scfft and zdfft have been re-tuned for FP128
New Memory Allocation Scheme

Timing dpotrf

MFlops

Problem size

AMD Hex-Core Processors | Cray XT5 Hex-core Workshop | December 2009
3DFFT Performance Improvement

Linux 4 threads

Lower is better

Lower is better
ACML 4.3.0 Supported Compilers

- PGI
  - 8.0-6
- GCC/GFORTRAN
  - 4.3.2
  - Now backwards compatible with GCC 4.1.2 and 4.2
- Open64
  - 4.2.1
- Intel Fortran 11.0
- Microsoft Visual Studio 2008
ACML 4.4
(Just Out)

- Ensure proper scaling with Istanbul/Magny-Cours
- Resolve scaling issues with small problems
- Family 10h optimized ZGEMM, associated L3 routines
- Double Complex 2D/3D-FFT improvements
AMD Developer Central
(for more info and downloads)

Downloads
- ACML and AMD LIBM are offered as free downloads to registered members of AMD Developer Central.

Forums
- The ACML forum is an excellent place to find answers.

Blogs
- Watch for blog entries by members of the ACML and LIBM team.

http://developer.amd.com
(A bit about Stream Computing)
ATI Radeon™ HD 5870 Graphics Architecture

2.72 Teraflops Single Precision
544 Gigaflops Double Precision

• Full Hardware Implementation of DirectCompute 11 and OpenCL™ 1.0

• IEEE754-2008 Compliance Enhancements

• Additional Compute Features:
  • 32-bit Atomic Operations
  • Flexible 32kB Local Data Shares
  • 64kB Global Data Share
  • Global synchronization
  • Append/consume buffers

• ATI Stream SDK beta
  • http://developer.amd.com/streambeta
Some GPGPU thoughts

• Massively Parallel Compute capability
  • FLOPS per watt and per mm² are truly impressive.

• Programmability
  • Improving but not painless.

• Moving Data onto and off of the GPGPU
  • Must do enough computation to amortize this.

• RAS
  • Not the same level as general purpose CPUs yet.
ACML-GPU 1.0
Released March 2009

Selected BLAS routines enabled for GPU

- DGEMM, SGEMM will run on GPU if present
- Small problems (N,M,K < 200) run on CPU
- Supports Quad- and Six-Core AMD Opteron™ processors
Questions
Backup
MOESI Cache Coherency Protocol

• "Read" and "Write" are by this core.
• "Probe Read" and "Probe Write" are reads and writes by others, that must probe this core's caches.