Commercially Available Chip Multiprocessors for Research

Bruce Childers
University of Pittsburgh
http://www.cs.pitt.edu/~childers

http://www.cs.pitt.edu  CACAO  http://www.cacao-team.org  PCM@PITI

Welcome to the Multi-core Era

Chip multiprocessors are everywhere!

– Cellular phone
– Tablets
– Netbooks
– Laptops
– Desktops
– Servers
Welcome to the Multi-core Era

Chip multiprocessors are everywhere!

– Cellular phone
– Tablets
– Netbooks
– Laptops
– Desktops
– Servers

Up to 100 giga-operations per second
App + media + radio operation
Increasing by 10x every 5 years
1W available (from total) for computing
Battery power determines limits
May be single to multiple chips

Qualcomm MSM8660
Dual 1.5 GHz Scorpion
GPU & cellular modem

Welcome to the Multi-core Era

Chip multiprocessors are everywhere!

– Cellular phone
– Tablets
– Netbooks
– Laptops
– Desktops
– Servers

Powerful applications from consumer to
science to business
Single processor (“socket”)
Moving toward high integration
Moving more toward heterogeneous

Intel Sandy Bridge-NB
4 cores, 8 HW threads
Integrated GPU, MC
Welcome to the Multi-core Era

Chip multiprocessors are **everywhere**!

- Cellular phone
- Tablets
- Netbooks
- Laptops
- Desktops
- Servers

AMD Opelon 6100
12 cores, 6MB L3
4 sockets, HyperTransport

---

Important Attributes

The **uncore** is what can matter for multi-core
It may also soon be the graphics processing capabilities
**Intel Processors**

- **NetBurst**
- **Core**
- **Nehalem**  
  - Nehalem (45nm)
  - Westmere (32nm)
  - Westmere EX
- **Sandy Bridge**  
  - Sandy Bridge (32nm)
  - Ivy Bridge (22nm)

### Sandy Bridge

- **Desktop, mobile & server variants**
- **Features**
  - Enhanced core microarchitecture
  - More closely coupled & integrated components
  - Hyper-threading with up to 8 cores (16 threads)
  - **On-chip shared L3 cache**
  - **Turbo Boost power/speed management**
  - Later server versions will feature improved QuickPath Interconnect
Intel Sandy Bridge

4 cores with L1, L2, L3 cache
Hyper-threaded: 8 logical cores
Advanced vector extensions (256-bit SIMD)
Micro-architecture changes
(Improved branch predictor, changed register renaming for AVX, 2x load ports)

4/2/11

L1 instruction cache
32KB L1 I-cache
Decode 4 x86 instr/cycle
Converted to u-ops
1.5K entry (L0) u-op cache
(just caches – not trace cache)
Gain is power
**Intel Sandy Bridge**

- **32KB L1 data cache**
- **256KB L2 cache (unified, private)**

---

**Intel Sandy Bridge**

- **8MB L3 cache (shared)**
- Designed for high bandwidth
- Shared by cores + GPU
- 435 GB/sec B/W @ 3.4 GHz*

---

* Source: Sandy Bridge Spans Generations, Linley Gwennap, MPR, Sept. 2010
L3 cache

Composed of 4 rings
32-byte data
Request
Acknowledgement
Snooping
Up to 26-31 clock traversal
Distributed coherence

Intel Sandy Bridge

Graphics processing unit
Integrated on-chip
More closely coupled with cores
(via L3 cache)
New FUs & video codec

Source: Sandy Bridge Spans Generations,
Linley Gwennap, MPR, Sept. 2010
Intel Sandy Bridge

Source: Sandy Bridge Spans Generations, Linley Gwennap, MPR, Sept. 2010

Intel Sandy Bridge

Platform Controller Hub (PCH)
Connects to I/O devices
E.g., SATA disk, USB, PCI Express, etc

Source: Sandy Bridge Spans Generations, Linley Gwennap, MPR, Sept. 2010
Turbo Boost Power Management

• Thermal design point (TDP)
  – Maximum power dissipated
  – Baseline: Consider impact of all cores
  – But not all cores are always active

• Change power allocation
  – Introduced in Nehalem
  – Shift available “budget” (under TDP) to “boost” speed of cores based on workload

Feedback Loop

[Diagram showing the feedback loop with nodes labeled for temperature, power, estimated current, OS state change, and core state (Active, Inactive). The diagram also shows the power manager adjusting voltage, frequency, and speed setting to maintain the thermal design point (TDP).]
Feedback Loop

Baseline frequency
Cores are active/inactive
Frequency with four cores
OS state change trigger

Core state (Active, Inactive)

OS state change
Temperature
Power
Estimated current

Power Manager

Speed setting

2000

2000

2000

2000

Inactive cores
Cores moved to inactive (C3/C6)
Leaves "headroom" in TDP
Spend on other cores

Core state (Active, Inactive)

OS state change
Temperature
Power
Estimated current

Power Manager

Boost cores

Speed setting

2000

0

0

0
Feedback Loop

Adjust speed upward
Change in small steps (100 MHz)
Up to maximum speed
Stay under TDP

Power Manager
Boost cores

Core state (Active, Inactive)

OS state change
Temperature
Power
Estimated current

Speed setting

Adjust speed upward
Change in small steps (100 MHz)
Up to maximum speed
Stay under TDP

Power Manager
Boost cores

Core state (Active, Inactive)

OS state change
Temperature
Power
Estimated current

Speed setting
Feedback Loop

Adjust speed downward
Move back under TDP
Temporarily exceeds b/c
thermals change slowly

Feedback Loop

Adjust speed downward
Move back under TDP
Temporarily exceeds b/c
thermals change slowly

Power Manager
Reduce cores

OS state change

Temperature
Power
Estimated current

Core state (Active, Inactive)

Speed setting

Power Manager
Reduce cores

OS state change

Temperature
Power
Estimated current

Core state (Active, Inactive)

Speed setting
Feedback Loop

Core i7-2920XM, 4 cores, base 2.5 GHz

<table>
<thead>
<tr>
<th>Active Cores</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Speed</td>
<td>3.2 GHz</td>
<td>3.3 GHz</td>
<td>3.4 GHz</td>
<td>3.5 GHz</td>
</tr>
</tbody>
</table>

Power Manager

OS state change

Temperature

Power

Estimated current

Speed setting

AMD Processors

- Shanghi (2008)
- Istanbul (2009)
- Magny-Cours (2010)
- Bulldozer (2011?)

November 2008

<table>
<thead>
<tr>
<th>Processor</th>
<th>Cores</th>
<th>Frequency</th>
<th>Technology</th>
<th>Cache</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shanghi</td>
<td>4 cores</td>
<td>2.9 GHz</td>
<td>45nm</td>
<td>6MB L3, DDR2</td>
<td></td>
</tr>
<tr>
<td>Istanbul</td>
<td>6 cores</td>
<td>2.8 GHz</td>
<td>45nm</td>
<td>6MB L3, DDR2, HT-assist</td>
<td></td>
</tr>
<tr>
<td>Magny-Cours</td>
<td>12 cores</td>
<td>2.6 GHz</td>
<td>45nm</td>
<td>12MB L3, DDR3, HT-assist</td>
<td></td>
</tr>
<tr>
<td>Bulldozer</td>
<td>16 cores</td>
<td>Tightly coupled cores, separate sched &amp; FUs (HWT like), 16 cores, 32nm, 16 MB L3, 256-bit FPU</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

January 2011
AMD Opteron 6100

- 12-core x86 processor, Istanbul core architecture

**Per package (Multi-chip Module)**
- 12 Istanbul cores
- 2 dies (nodes), 6 core ea, 45nm

**Per node**
- 6 MB shared L3
- Memory controller
- 2x memory channels
- 4x HyperTransport links
**AMD Opteron 6100**

- 12-core x86 processor, Istanbul core architecture

---

**HyperTransport Links**

- **HyperTransport**
  - Point-to-point interconnect (LVDS)
  - Arranged as multiple links (e.g., x16 links)
  - Up to 25.6 GB/second (x32 links)

- 4 x16 HT ports/processor allocated for within-package communication, cross processor communication & I/O
Interconnection (2 processors)

4 x16 HyperTransport links
x16 adjacent off-package nodes
x8 diagonal off-package nodes
x16 + x8 on-package nodes
x16 noncoherent I/O

Interconnection (4 processors)

4 x16 HyperTransport links
x8 between off-package nodes
x16 + x8 on-package nodes
x16 noncoherent I/O
Coherence Traffic

- Explosion in coherence traffic
  - 4 processors, 48 cores!
- Coherence
  - Data may reside in multiple caches
  - Need to keep it consistent
  - Single writer, multiple readers
- Broadcast
  - Request which core has most recent data
  - Clearly, doesn’t scale well

HT Assist

"Home" is location where memory address resides
Data can be cached anywhere, though
Need to find the location
Reader: Deliver potentially most recent copy
Writer: Get exclusive ownership to update data
HT Assist

1) P2 requests X data from P0 (home)
2) P0 broadcasts for most recent copy
3) Wait for reply from each processor
4) Data forwarded from P1 to P2
1) P2 requests data from P0 (home)
2) P0 broadcasts for most recent copy
3) Wait for reply from each processor
4) Data forwarded from P1 to P2

Yes!

HT Assist

1) P2 requests data from P0 (home)
2) P0 broadcasts for most recent copy
3) Wait for reply from each processor
4) Data forwarded from P1 to P2

HT Assist
HT Assist

1) P2 requests data from P0 (home)
2) P0 broadcasts for most recent copy
3) Wait for reply from each processor
4) Data forwarded from P1 to P2

Maintain directory of data location
1MB of L3 dedicated to directory
Reduces traffic (location known)
1) P2 requests data from P0 (home)
2) P0 broadcasts for most recent copy
3) Wait for reply from each processor
4) Data forwarded from P1 to P2

1) P2 requests data from P0 (home)
2) P0 broadcasts for most recent copy
3) Data forwarded from P1 to P2
HT Assist

1) P2 requests data from P0 (home)
2) P0 broadcasts for most recent copy
3) Wait for reply from each processor
4) Data forwarded from P1 to P2

Only makes sense for >2 nodes
Avoids most broadcasts
Reduces L3 cache capacity

HT Assist: Where to Keep Directory?

6 MB L3 cache with 16 ways

64-byte line with 16 directory entries

4-byte directory entry (probe filter)

Same processor in 1P and 4P systems
Reduce costs by reusing the L3 cache for directory
16-ways, 4 ways dedicated to directory
Sparse directory structure

Same processor in 1P and 4P systems
Reduce costs by reusing the L3 cache for directory
16-ways, 4 ways dedicated to directory
Sparse directory structure

Source: HotChips 2009
### AMD Opteron 6100

<table>
<thead>
<tr>
<th>Model</th>
<th>Speed</th>
<th>Cores</th>
<th>ACP</th>
<th>TDP</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>6180 SE</td>
<td>2.5 GHz</td>
<td>12</td>
<td>105W</td>
<td>140 W</td>
<td>$1514 *</td>
</tr>
<tr>
<td>6176</td>
<td>2.3 GHz</td>
<td>12</td>
<td>80 W</td>
<td>115 W</td>
<td>$1265 *</td>
</tr>
<tr>
<td>6172</td>
<td>2.1 GHz</td>
<td>12</td>
<td>80 W</td>
<td>115 W</td>
<td>$989</td>
</tr>
<tr>
<td>6168</td>
<td>1.9 GHz</td>
<td>12</td>
<td>80 W</td>
<td>115 W</td>
<td>$744</td>
</tr>
<tr>
<td>6140</td>
<td>2.6 GHz</td>
<td>8</td>
<td>80 W</td>
<td>115 W</td>
<td>$989</td>
</tr>
<tr>
<td>6128</td>
<td>2.0 GHz</td>
<td>8</td>
<td>80 W</td>
<td>115 W</td>
<td>$744</td>
</tr>
<tr>
<td>6166 HE</td>
<td>1.8 GHz</td>
<td>12</td>
<td>65 W</td>
<td>85 W</td>
<td>$873</td>
</tr>
<tr>
<td>6164 HE</td>
<td>1.7 GHz</td>
<td>12</td>
<td>65 W</td>
<td>85 W</td>
<td>$744</td>
</tr>
<tr>
<td>6132 HE</td>
<td>2.2 GHz</td>
<td>8</td>
<td>65 W</td>
<td>85 W</td>
<td>$591</td>
</tr>
<tr>
<td>6124 HE</td>
<td>1.8 GHz</td>
<td>8</td>
<td>65 W</td>
<td>85 W</td>
<td>$455</td>
</tr>
</tbody>
</table>

SE optimized for performance
HE optimized for low power
ACP average CPU power (workload derived power)

All have 12 MB L3 (2x 6 MB), HT3, AMD-V
Introduced March 29, 2010
* Introduced February 14, 2011

### What’s available?

- **AVA Direct Supermicro SuperServer $5087**
  - Quad AMD Opteron 6128 8-core 2.0 GHz (32C)
  - 64 GB memory, 500 GB SATA drive

- **Dell PowerEdge R415 $2457**
  - Dual AMD Opteron 4170HE (6C), 2.1 GHz (12C)
  - 16 GB memory, 250 GB SATA drive

- **Dell XPS 8300 (desktop) $1453**
  - Intel Core i7-2600 (8MB, 3.4 GHz)
  - 16 GB memory, 1TB SATA drive
Summary

• Multi-core is certainly here!
• Significant research challenges

• Platform infrastructure
  – Core architecture
  – Cache architecture
  – Interconnection
  – Power management
  – Integration and fusing of CPU+GPU

Today’s processors offer many of these capabilities for research!