Many Cores ➔ Disruptive Change

- Moore’s Law is enabling chaotic change
  - Double transistor density every 18 months
  - Creating need for new software models

- Power/Thermal Constrained World
  - Energy/Cooling Cost Exceed IT equipment Cost
  - Frequency alone is not our friend

- The Complexity Wall/Frequency Wall
  - Single threaded performance maxed

- Customers expect added Visible Value
Models for the Future of Computers

Heterogeneous Systems

- Parallel signal processing of sensor data
- Parallel Classification and Recognition
  Vision, Sound, Contact, Patterns, Object
- Serial Cognitive and Reasoning
- Parallel Search, Scan, Sort

Flame, the robot, walks the way humans walk. (Credit: Image courtesy of Delft University of Technology)
Multi-Core CPUs

**Advantages**

Optimized for execution of sequential program
- Complex Pipelines to achieve max frequency
- Out Of Order, Super Scalar to achieve max ILP
- Branch Predictors, Speculative execution
- Register Renaming
- Large Caches optimized for low latency access

Multi-Core enables parallel multi-tasks/threads
- Improved user response
- Background task such as OS chores, virus scan, etc
Multi-Core CPUs

Disadvantage

- Fine grain sharing of work between cores/caches
  - OS Overhead – spawn, communication, fork
  - Finding large number of task to Multi-task is hard
- Embarrassing parallel apps
  - Limited Speed up (ALU & Bandwidth)
  - Limited Power/Flop advantage
Advantages

- Optimized for structured parallel execution
  - Extensive ALU counts & Memory Bandwidth
  - Cooperative multi-threading hides latency
- Shared Instruction Resources
- Fixed function units for parallel workloads dispatch
- Extensive exploitation of Locality
Disadvantages

Ineffective for Single threaded execution

- Divergence
  - Parallelism lost opportunity
  - Loss of efficiency
- Upload/download of data cost
- Requires large workload to fully utilize
- Small Caches are optimized for locality/throughput
Amdahl’s Law

\[ \text{Speed-up} = \frac{1}{S_w + \frac{(1 - S_w)}{N}} \]

- \( S_w \): % Serial Work
- \( N \): Number of processors
Amdahl’s Law – zoom out a bit

- “Everyone knows Amdahl’s Law, but quickly forgets”
- Dr. Tom Puzak, IBM Research, 2007

Amdahl’s Law seriously inhibits unstructured parallelism …
<table>
<thead>
<tr>
<th>Relative Performance per core</th>
<th>0.25</th>
<th>0.5</th>
<th>1.0</th>
<th>1.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>mm²/Core</td>
<td>1.56 mm²</td>
<td>6.25 mm²</td>
<td>25 mm²</td>
<td>50 mm²</td>
</tr>
<tr>
<td>Power/Core</td>
<td>0.6 W</td>
<td>2.5 W</td>
<td>10 W</td>
<td>20 W</td>
</tr>
<tr>
<td># of Cores (200 mm² System)</td>
<td>128</td>
<td>32</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>System Performance for parallel Apps</td>
<td>32</td>
<td>16</td>
<td>8</td>
<td>5.6</td>
</tr>
<tr>
<td>Relative System Performance/mm²</td>
<td>16%</td>
<td>8%</td>
<td>4%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Relative System Performance/Watt</td>
<td>40%</td>
<td>20%</td>
<td>10%</td>
<td>7%</td>
</tr>
</tbody>
</table>

Structured Parallelism enables solutions for more flops less watts.
Heterogeneous Cores (Mixed CPU/GPU)

- Lowest Power & Highest Performance Solution
- Serial Single threaded – Leverage Fast OOCs
- Task Parallel – Leverage multi-CPU or GPU cores
- Data Parallel - Leverage GPU or Application Specific Cores
Unified Memory Architectures

• Remove large data copies of workload and results
  • Reduce power consumption per computation
  • Enable small job offload
• Remove OS Overhead and latency of communication
• Fast Synchronization Primitives
• Increasing capable memory systems & BW
AMD Stream SDK Software Development Stack

Applications

Compilers
- Brook+
- & Other App Specific

Libraries
- OpenCL
- ACM
- Rapidmind

3rd Party Tools
- Graphics API
- DirectX®
- OpenGL

AMD Runtime
- Multi-Core AMD CPUs

Compute Abstraction Layer (CAL)
- AMD Stream Processors
- AMD Opteron

AMD CPUs
Software Solution (Real Challenge)

Development of Software often exceeds that of hardware
- Determine and implement long term solutions
- Enable highly coherent machines with heterogeneous accelerators
- Enable control of memory hierarchies for system scaling
- Communication/Messaging ➔ Producer/Consumer relationships

Simplifying Programming Model
- Build on emerging multi-core models
- Enable the Masses to Programming with parallel abilities
- Enable Application Specific Libraries
- Enable open standards
Processing Efficiency*

GigaFLOPS per Watt
GigaFLOPS per $

AMD FireStream 9250 Stream Processor

*Source: internal AMD test results
Disclaimer and Attribution

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

© 2008 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD Opteron, ATI, the ATI logo, Radeon, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other names are for informational purposes only and may be trademarks of their respective owners.