FreeBSD and NUMA

John Baldwin
NYC*BUG
June 3, 2015
What is NUMA

- Non-Uniform Memory Architecture
- “Slow” vs “Fast” Memory
  - From CPUs
  - From I/O Devices
- Present on x86 starting with AMD Opterons (HyperTransport) and Intel Nehalem (QPI)
Front Side Bus (FSB)
Sandy Bridge (Romley)
PCI-e Transactions

- Memory Read / Write Initiated by Device (DMA)
- Memory Read / Write Initiated by CPU (PIO)
  - Managed by the I/O hub / MCH
- Memory Address Space
  - RAM (via MC)
  - Device Registers (via I/O Hub)
DMA & Cache Snooping

- Red = DMA Request
- Blue = DMA Reply

Diagram showing components:
- CPU
- LLC
- MCH
- RAM
- NIC
DMA & Cache Snooping

Red = DMA Request
Blue = DMA Reply

What if data is dirty in cache?

Data in RAM will be stale.

Stale data on wire
DMA & Cache Snooping

Red = DMA Request
Blue = DMA Reply
Yellow = Snooping
DDIO (Romley)

These are optional

Red = DMA Request
Blue = DMA Reply
Haswell EP

Haswell EP Die Configurations

<table>
<thead>
<tr>
<th>Chop</th>
<th>Columns</th>
<th>Home Agents</th>
<th>Cores</th>
<th>Power (W)</th>
<th>Transitors (B)</th>
<th>Die Area (mm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCC</td>
<td>4</td>
<td>2</td>
<td>14-18</td>
<td>110-145</td>
<td>5.69</td>
<td>662</td>
</tr>
<tr>
<td>MCC</td>
<td>3</td>
<td>2</td>
<td>6-12</td>
<td>65-160</td>
<td>3.84</td>
<td>492</td>
</tr>
<tr>
<td>LCC</td>
<td>2</td>
<td>1</td>
<td>4-8</td>
<td>55-140</td>
<td>2.60</td>
<td>354</td>
</tr>
</tbody>
</table>

Source:
http://www.anandtech.com/show/8423/intel-xeon-e5-version-3-up-to-18-haswell-ep-cores-/4
NUMA Implications / Tradeoffs

• Local vs Remote CPU Accesses
• Local vs Remote I/O Accesses
  – Maximize DDIO
  – Except When You Don't?
• Problems are Akin to SMP Scaling
  – (We Know How Well That's Working Out)
• “Soft” Partitioning
NUMA Support in FreeBSD 9

- Hackish “first-touch” Policy
- Not Enabled by Default
- Not Very General Purpose
- No I/O Awareness
NUMA Support in FreeBSD 10

- Start on a More Mature Framework...
- … But Mostly Out of Tree
  - At Least Three Variants
- Stock Tree Only Has “round-robin”
- Not Enabled By Default
- No I/O Awareness
NUMA Support in FreeBSD 11+

- More Work from More Folks
- Goal is to Permit Tuning
  - Not Trying to be Automagical
- Will Include (Some) I/O Awareness
  - Interrupts

- [http://wiki.freebsd.org/NUMA](http://wiki.freebsd.org/NUMA)
  - Not Set in Stone
- Merge to 10?
- Enabled in GENERIC?