EE482S
Lecture 2
Discussion of 2 Papers
Conclusion of Introductory Material

April 9, 2002

William J. Dally
Computer Systems Laboratory
Stanford University
billd@csl.stanford.edu
Today’s Class Meeting

• Discuss two papers
  – Imagine: Media Processing with Streams
    • Khailany et al., IEEE Micro, March-April 2001
  – Polygon Rendering on a Stream Architecture
    • Owens et al., Eurographics HWWS, 2000.

• Conclude introductory discussion on Stream Architecture
  – What is a stream processor
Discussion of Imagine Paper
Discussion of Polygon Rendering Paper
What is a Stream Processor?

• A processor that is optimized to execute a stream program

• Features include
  – Exploit parallelism
    • TLP with multiple processors
    • DLP with multiple clusters within each processor
    • ILP with multiple ALUs within each cluster
  – Exploit locality with a bandwidth hierarchy
    • Kernel locality within each cluster
    • Producer-consumer locality within each processor

• Many different possible architectures
The Imagine Stream Processor
Arithmetic Clusters

Local Register File

To SRF

From SRF

Intercluster Network

Cross Point

CU
A Bandwidth Hierarchy exploits locality and concurrency

- VLIW clusters with shared control
- 41.2 32-bit floating-point operations per word of memory BW
A Bandwidth Hierarchy exploits kernel and producer-consumer locality

<table>
<thead>
<tr>
<th></th>
<th>Memory BW</th>
<th>Global RF BW</th>
<th>Local RF BW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Depth Extractor</strong></td>
<td>0.80 GB/s</td>
<td>18.45 GB/s</td>
<td>210.85 GB/s</td>
</tr>
<tr>
<td><strong>MPEG Encoder</strong></td>
<td>0.47 GB/s</td>
<td>2.46 GB/s</td>
<td>121.05 GB/s</td>
</tr>
<tr>
<td><strong>Polygon Rendering</strong></td>
<td>0.78 GB/s</td>
<td>4.06 GB/s</td>
<td>102.46 GB/s</td>
</tr>
<tr>
<td><strong>QR Decomposition</strong></td>
<td>0.46 GB/s</td>
<td>3.67 GB/s</td>
<td>234.57 GB/s</td>
</tr>
</tbody>
</table>
Producer-Consumer Locality in the Depth Extractor

<table>
<thead>
<tr>
<th>Memory/Global Data</th>
<th>SRF/Streams</th>
<th>Clusters/Kernels</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Memory/Global Data Image]</td>
<td>![SRF/Streams Diagram]</td>
<td>![Clusters/Kernels Diagram]</td>
</tr>
</tbody>
</table>

- row of pixels
- previous partial sums
- new partial sums
- blurred row
- previous partial sums
- new partial sums
- sharpened row
- filtered row segment
- filtered row segment
- previous partial sums
- new partial sums
- depth map row segment

Processing:
- Convolution (Gaussian)
- Convolution (Laplacian)
- SAD
Bandwidth Demand of Applications
Die Photos

- 21 M transistors / TI 0.15µm 1.5V CMOS / 16mm x 16mm
- 300 MHz TTTT, hope for 400 MHz in lab
- Chips arrived 4/1/02, no fooling!
Performance demonstrated on signal and image processing

16-bit applications
16-bit kernels
Floating-point application
Floating-point kernel

GOPS

depth  mpeg  qrd  dct  convolve  fft
12.1   19.8  11.0  23.9  25.6   7.0
Initial studies indicate that it also applies to solving PDEs and ODEs
Architecture of a Streaming Supercomputer

- **Stream Processor**: 64 FPUs, 64 GFLOPS
- **Node**: 16 x DRDRAM, 2 GBytes
- **Board**: 16 Nodes, 1K FPUs, 1 TFLOPS, 32 GBytes
- **Node**: 32 + 32 pairs
- **Board 2**: 16 Nodes, 1K FPUs, 1 TFLOPS, 32 GBytes
- **Node**: 32 + 32 pairs
- **On-Board Network**: 38 GBytes/s, 20 GBytes/s, 32 + 32 pairs
- **Intra-Cabinet Network**: 160 GBytes/s, 256 + 256 pairs, 10.5" Teradyne GbX
- **E/O**: 5 TBytes/s, 8K + 8K links, Ribbon Fiber
- **Inter-Cabinet Network**: Bisection 64 TBytes/s
- All links 5 Gb/s per pair or fiber
- All bandwidths are full duplex
Streaming processor
Rough per-node budget

<table>
<thead>
<tr>
<th>Item</th>
<th>Cost</th>
<th>Per Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor chip</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Router chip</td>
<td>200</td>
<td>50</td>
</tr>
<tr>
<td>Memory chip</td>
<td>20</td>
<td>320</td>
</tr>
<tr>
<td>Board/Backplane</td>
<td>3000</td>
<td>188</td>
</tr>
<tr>
<td>Cabinet</td>
<td>50000</td>
<td>49</td>
</tr>
<tr>
<td>Power</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>Per-Node Cost</td>
<td></td>
<td>976</td>
</tr>
<tr>
<td>$/GFLOPS (64/node)</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td>$/M-GUPS (250/node)</td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

Preliminary numbers, parts cost only, no I/O included.
Many open problems

• A small sampling
  • Software
    – Program transformation
    – Program mapping
    – Bandwidth optimization
    – Conditionals
    – Irregular data structures
  • Hardware
    – Alternative stream models
    – Register organization
    – Bandwidth hierarchies
    – Memory organization
    – Short stream issues
    – ISA design
    – Cluster organization
    – Processor organization
Next Time

• Walk through a streaming application