Exotic Technologies Panel and Time Capsule Submission for Most Exciting Architecture at SC 20

Erik P. DeBenedictis

SAND2006-7724C
Approved for Unclassified Unlimited Release
<table>
<thead>
<tr>
<th></th>
<th>Red Storm (Historical)</th>
<th>μP part only</th>
<th>My Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cores</td>
<td>13,000×2</td>
<td>50,000×4</td>
<td>50,000×40</td>
</tr>
<tr>
<td>Node Type</td>
<td>μP</td>
<td>μP</td>
<td>μP &amp; macro function</td>
</tr>
<tr>
<td>Clock</td>
<td>2.5 GHz</td>
<td>20 GHz</td>
<td>20 GHz</td>
</tr>
<tr>
<td>Flops/chip</td>
<td>5×2 GF</td>
<td>50×4 GF</td>
<td>1.6 TF</td>
</tr>
<tr>
<td>Sys. Peak</td>
<td>125 TF</td>
<td>80 PF</td>
<td>800 PF</td>
</tr>
<tr>
<td>Maximum MPI Latency</td>
<td>10 μS</td>
<td>100 ns</td>
<td>100 ns</td>
</tr>
<tr>
<td>Power</td>
<td>2 MW</td>
<td>2 MW</td>
<td>2 MW</td>
</tr>
</tbody>
</table>
Packaging for a Spatial Locality

• Basic Module
  – 2 Chips
  – Each node 4 core conventional CPU plus
  – 36 accelerator cores
  – 1 GB+ on chip RAM
  – 100 GB memory on bottom of module
  – Each module includes a power unit
  – Six optical interconnect channels, 3D mesh
Packaging for a Spatial Locality

- Entire supercomputer is a single structure
- All mesh network connections are of constant length (8” max)
- Air flows front to back
  - General approach will work for liquid cooling as well

This region would be filled with heat sink
Design minimizes signal travel distance while maximizing use of surface area for cooling.
Outline

• Degree of Innovation
• Non-Architecture Projections
• Architecture Projections
• Programming
• Architecture Summary
• Current Activities to Watch and Why
• Conclusions
Perspective on Innovation

• 1992 + 14 = 2006; 2006 + 14 = 2020
• If rate of innovation stays the same, we should see as big an advance to 2020 as we saw from “late nCUBE” through now
• However, I think SC is maturing. I think the community will only accept innovations backwards compatible with what we have now. If there is major innovation, I think it will be best represented in a new conference, say “I Robot 2020.”
Outline

- Degree of Innovation
- Non-Architecture Projections
- Architecture Projections
- Programming
- Architecture Summary
- Current Activities to Watch and Why
- Conclusions
Scaling Implications for CPUs

• \(5\times\) performance increase for a single core
  – Burger and Keckler study, slide follows
  – NOTE: Integrated RAM will increase this another \(2\times\)

• 64 cores of today’s complexity
  – 90 nm \(\rightarrow\) 18 nm is \(5\times\). Dual core \(\times 5^2 \rightarrow 50 \approx 64\)

• I think we’ll see a hybrid – to be discussed later
UT Austin Study (2000)

• The Study
  – Clock Rate versus IPC: The End of the Road for Conventional Microarchitectures, Vikas Agarwal, M.S. Hrishikesh, Stephen W. Keckler, Doug Burger. 27th Annual International Symposium on Computer Architecture

• Conclusions (to be Explained)
  – Modified ITRS roadmap predictions to be more friendly to architectures
  – Concluded there would be a 12%/year growth…
  – However, recent growth has been ~30%, with industry’s maneuver to cheat the analysis instructive
## Critical Evaluation Memory

For each Technology Entry (e.g. 1D Structures), sum horizontally over the 8 Criteria

Max Sum = 24

Min Sum = 8

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nano Floating Gate Memory</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.5</td>
<td>2.2</td>
<td>2.7</td>
<td>2.7</td>
<td>3.0</td>
</tr>
<tr>
<td>Engineered Tunnel Barrier Memory</td>
<td>2.2</td>
<td>2.3</td>
<td>2.3</td>
<td>2.3</td>
<td>2.4</td>
<td>2.8</td>
<td>2.8</td>
<td>3.0</td>
</tr>
<tr>
<td>Ferroelectric FET Memory</td>
<td>1.9</td>
<td>2.3</td>
<td>2.5</td>
<td>2.2</td>
<td>2.0</td>
<td>3.0</td>
<td>2.6</td>
<td>3.0</td>
</tr>
<tr>
<td>Insulator Resistance Change Memory</td>
<td>2.5</td>
<td>2.5</td>
<td>2.0</td>
<td>2.2</td>
<td>1.9</td>
<td>2.8</td>
<td>2.6</td>
<td>2.8</td>
</tr>
<tr>
<td>Polymer Memory</td>
<td>2.1</td>
<td>1.5</td>
<td>2.3</td>
<td>2.2</td>
<td>1.6</td>
<td>2.9</td>
<td>2.3</td>
<td>2.5</td>
</tr>
<tr>
<td>Molecular Memory</td>
<td>2.3</td>
<td>1.5</td>
<td>2.4</td>
<td>1.6</td>
<td>1.4</td>
<td>2.6</td>
<td>1.9</td>
<td>2.3</td>
</tr>
</tbody>
</table>

4 good options
## Critical Evaluation

**Logic**

For each Technology Entry (e.g. 1D Structures, sum horizontally over the 8 Criteria
Max Sum = 24
Min Sum = 8

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1D Structures (CNTs &amp; NWs)</td>
<td>2.4</td>
<td>2.5</td>
<td>2.3</td>
<td>2.3</td>
<td>2.1</td>
<td>2.8</td>
<td>2.3</td>
<td>2.8</td>
</tr>
<tr>
<td>Resonant Tunneling Devices</td>
<td>1.5</td>
<td>2.2</td>
<td>2.1</td>
<td>1.7</td>
<td>1.7</td>
<td>2.5</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>SETs</td>
<td>1.9</td>
<td>1.5</td>
<td>2.6</td>
<td>1.4</td>
<td>1.2</td>
<td>1.9</td>
<td>2.1</td>
<td>2.1</td>
</tr>
<tr>
<td>Molecular Devices</td>
<td>1.6</td>
<td>1.8</td>
<td>2.2</td>
<td>1.5</td>
<td>1.6</td>
<td>2.3</td>
<td>1.7</td>
<td>1.8</td>
</tr>
<tr>
<td>Ferromagnetic Devices</td>
<td>1.4</td>
<td>1.3</td>
<td>1.9</td>
<td>1.5</td>
<td>2.0</td>
<td>2.5</td>
<td>1.7</td>
<td>1.7</td>
</tr>
<tr>
<td>Spin Transistor</td>
<td>2.2</td>
<td>1.3</td>
<td>2.4</td>
<td>1.2</td>
<td>1.2</td>
<td>2.4</td>
<td>1.5</td>
<td>1.7</td>
</tr>
</tbody>
</table>

1 good option, and it is not a change for SC
Outline

- Degree of Innovation
- Non-Architecture Projections
- **Architecture Projections**
  - Programming
  - Architecture Summary
  - Current Activities to Watch and Why
- Conclusions
Commodity μP Architecture in 2020

• Industry is now ramping the number of cores per die
• Intel and AMD are making serious noises about integrating graphics processors into CPU die
• I have special information that upcoming ITRS direction will advocate “macro functions” (to be explained later)
• These are self-confirming data points that answer the commodity μP architecture question
  – Note: this answer could be wrong…
Emerging Research Logic Technologies

Traditional Goal

Logic technology that is scaleable beyond CMOS, high-speed, and low-power.

I do not have mystical clairvoyance, but I do have a VG set from an influential meeting that hasn’t occurred yet…
• Current CPU style
• New direction proposed to industry will be to keep CPU but augment it with “macro functions.”

• Macro functions may include non-CMOS logic devices specialized to nontraditional functions, such as speech recognition, etc.
Outline

• Degree of Innovation
• Non-Architecture Projections
• Architecture Projections
  • Programming
• Architecture Summary
• Current Activities to Watch and Why
• Conclusions
Programmability Considerations

• Has code changed since the “late nCUBE” era?
  – MPI replaced proprietary message passing
  – We have a huge code base of math code (at Sandia)
  – We have frameworks (at Sandia)

• Conclusion
  – A lot of code written and put into reusable form, but little change in underlying programming method

• Implication
  – Further migration towards putting code into libraries, but the code will have the same basis
Programming

• Industry will integrate the following macro functions:
  – Graphics processors
  – Speech recognition
  – Visual recognition
• However, the hardware will be sufficiently general purpose to be used for supercomputing
• Still CMOS in this timeframe

• A small number of super-duper programming jocks will write supercomputing code for the macro functions
  – LAPACK
  – FEM meshing
  – Etc.
• Regular programmers will write C++/Fortran code interfacing like DirectX (Microsoft’s GPU API)
• I went by the PeakStream booth yesterday and see that they have a scientific programming library for graphics processors. I’ve never used it, but I think the approach might work with hardware up to 2020.
Outline

• Degree of Innovation
• Non-Architecture Projections
• Architecture Projections
• Programming

• Architecture Summary

• Current Activities to Watch and Why
• Conclusions
Processor Chip Prediction

- ¼ of chip to be four CPUs each with 10× throughput of today’s cores
- ¾ of chip to be a new Macro Function
- Layered nano memory

- Macro Function will be developed by industry and repurposed for supercomputing, originally
  - speech recognition
  - vision for robots

![Diagram showing Core of Today and Macro Function]
CPU Detail

• Entry
  – Four cores at 50 GF Linpack-peak each, total 200 GF
  – 36 macro functions of 440 GF each, total 15.8 TF total
    • graphics, speech, vision, repurposed to scientific kernels
  – 16 TF per chip

• Each chip to have 1 GB+ layered nano memory
• As much external memory as you like (not a limit)
• 50,000 chips in a 2 MW system → 800 Petaflops
Memory Story – No Memory Wall

- I predict one of the 4+ nano memory options will succeed
- 1 GB+ memory will be integrated onto the CPU
  - I don’t care if you call it cache, main memory, etc.
- Memory will be non-volatile
- This will boost CPU performance quite a bit over the 5× predicted by architecture study

![Diagram showing nano memory layer and super high density interconnect]
Interconnect

• Interconnect is likely to be optics, but not necessarily fiber
  – Free space
  – Waveguides
• Luxtera comes up often in discussions of optical interconnect. The Luxtera approach works with Si by having external lasers.
Outline

• Degree of Innovation
• Non-Architecture Projections
• Architecture Projections
• Programming
• Architecture Summary
• Current Activities to Watch and Why
• Conclusions
Current Activities to Watch and Why

• Cyclops – highly multicore architecture that could (with suitable systems software) blend legacy code compatibility with efficient use of multiple cores
  – Memory hierarchy is where the action is
  – I predict future will hold Cyclops + layered memory
• Layered memory (Nantero?)
• Optical interconnect (Luxtera?)
• Programming (PeakStream?)
Outline

• Degree of Innovation
• Non-Architecture Projections
• Architecture
• Programming
• Architecture Summary
• Current Activities to Watch and Why
• Conclusions
Conclusion I

• Industry is now putting additional resources created by Moore’s Law into more cores and is talking about the same for graphics chips and Macro Functions
• Coders are getting further away from programming the bare hardware
• My solution has the following properties:
Conclusion II

• The majority of users will program the conventional cores. They will see a fairly flat parallel Von Neumann computer. Of course, they are accustomed to using libraries for inner loops.

• A small number of users will optimize low level code (libraries) for edge of the envelope hardware where the programmers need to be cognizant of data and operation placement.

• I believe this is the most likely to happen, even if it does not make for the most exciting computer architecture research.