Computing Systems: Research Challenges Ahead
The HiPEAC Vision 2011/2012

M. DURANTON, D. BLACK-SHAFFER, S. YEHIA, K. DE BOSSCHERE

http://www.hipeac.net/roadmap
HiPEAC Roadmaps

http://www.hipeac.net/roadmap
(These slides are a short version of the HiPEAC presentation)
Part of the HiPEAC roadmap is required reading for AVDARK

You will find it and other required reading in the ”Extra course papers” directory.

As specified in the ”Reading instructions”:  
• page 2-33 are required reading  
• page 34-40 read-through (RT)
Impact on Society

HiPEAC Research objectives

Efficiency
Complexity
Dependability

Big data meets energy in an intelligent connected physical world

Application pull

Business trends

Technological constraints

Technological opportunities

HiPEAC Strengths

HiPEAC Weaknesses

7 HiPEAC Research objectives

Impact on Society

Technological constraints

Technological opportunities
## Trends influencing Computing Systems

<table>
<thead>
<tr>
<th>Application Pull</th>
<th>Business Trends</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data Deluge</td>
<td>• Convergence</td>
</tr>
<tr>
<td>• Intelligent Processing</td>
<td>• Specialization</td>
</tr>
<tr>
<td>• Ubiquitous Communication</td>
<td>• Post-PC Devices</td>
</tr>
</tbody>
</table>
Data Deluge

The "data deluge" gap

Data Growth vs. Moore's Law from 2006 to 2010.
Growth of data storage in Exabytes

Intelligent processing of “natural” data

More and more applications are not only “number crunching”

Recognition, Mining, Synthesis

Implicit and natural computing

Posture: Lying Down

Source: “The Landscape of Parallel Computing Research: A View from Berkeley”

Krste Asanovic et al.
Ubiquitous computing in a connected world

Sensory swarm, actuators and real world data

Smart house cities, …

Mobile access

Infrastructure Core (cloud)
Trends influencing Computing Systems

<table>
<thead>
<tr>
<th>Application Pull</th>
<th>Business Trends</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Data Deluge</td>
<td>• Convergence/standards</td>
</tr>
<tr>
<td>• Intelligent Processing</td>
<td>• Specialization</td>
</tr>
<tr>
<td>• Ubiquitous Communication</td>
<td>• Post-PC Devices</td>
</tr>
</tbody>
</table>
Convergence

- Business models
  - Standard(s)
  - Interoperability
  - ...

Broadcast

IP, Internet

Telecom
Post-PC devices

Ubiquitous access
## PC Market

### Western Europe: PC Vendor Unit Shipment Estimates for 2Q11 (Thousands of Units)

<table>
<thead>
<tr>
<th>Vendor</th>
<th>2Q11 Shipments</th>
<th>2Q11 Market Share (%)</th>
<th>2Q10 Shipments</th>
<th>2Q10 Market Share (%)</th>
<th>2Q11-2Q10 Growth (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP</td>
<td>3,171</td>
<td>25.1</td>
<td>3,376</td>
<td>21.6</td>
<td>-6.1</td>
</tr>
<tr>
<td>Acer Group</td>
<td>2,046</td>
<td>16.2</td>
<td>3,696</td>
<td>23.7</td>
<td>-44.6</td>
</tr>
<tr>
<td>Dell</td>
<td>1,371</td>
<td>10.8</td>
<td>1,571</td>
<td>10.1</td>
<td>-12.7</td>
</tr>
<tr>
<td>Asus</td>
<td>1,021</td>
<td>8.1</td>
<td>1,324</td>
<td>8.5</td>
<td>-22.9</td>
</tr>
<tr>
<td>Apple</td>
<td>879</td>
<td>7.0</td>
<td>875</td>
<td>5.6</td>
<td>0.5</td>
</tr>
<tr>
<td>Others</td>
<td>4161</td>
<td>32.8</td>
<td>4751</td>
<td>30.5</td>
<td>-12.4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12,649</strong></td>
<td><strong>100</strong></td>
<td><strong>15,593</strong></td>
<td><strong>100</strong></td>
<td><strong>-18.9</strong></td>
</tr>
</tbody>
</table>

Note: Data includes desk-based PCs and mobile PCs. Media tablets are excluded. Source: Gartner (August 2011) (from [http://www.gartner.com/it/page.jsp?id=1769215](http://www.gartner.com/it/page.jsp?id=1769215))
Computing Systems: Drivers

Big data meets energy in an intelligent connected physical world

Application pull

Business trends
Big data meets energy in an intelligent connected physical world

Application pull

Technology push

Business trends

Technological constraints

Technological opportunities

Efficiency
Complexity
Dependability
## Technological trends influencing Computing Systems

<table>
<thead>
<tr>
<th>Constraints</th>
<th>Opportunities</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Frequency Limits</td>
<td>• CMOS Phonotic</td>
</tr>
<tr>
<td>• Power Limits</td>
<td>• Non-volatile memories</td>
</tr>
<tr>
<td>• Dark Silicon</td>
<td>• 3D Stacking</td>
</tr>
<tr>
<td></td>
<td>• New paradigms</td>
</tr>
</tbody>
</table>
Technological constraints

*We are at a turning point*

- Continuation of Moore’s Law
- Power limits

Dark silicon
Moore's law: increase in transistor density

Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanovic.
Limited frequency increase $\Rightarrow$ more cores

Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanovic
Limitation by power density and dissipation

2009: GP CPU = 130 W (45 nm)
2009: Consumer SoC = 10W
2009: Mobile SoC = 1 W

Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanović
Dark Silicon

<table>
<thead>
<tr>
<th>Node</th>
<th>45nm</th>
<th>22nm</th>
<th>11nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Transistors scaling</td>
<td>1</td>
<td>x4</td>
<td>x16</td>
</tr>
<tr>
<td>Peak Freq. scaling</td>
<td>1.6</td>
<td>2.4</td>
<td></td>
</tr>
</tbody>
</table>

Power scaling:
- @45nm freq: 100%
- @peak freq: 2/3, 1, 2/3

Exploitable silicon:
- 100%
- 40%: 25%
- 20%: 10%

\[ P = CV^2f \]

\[ \text{power} = (\text{power of 1 transistor}) \times (\# \text{transistors used}) \]

- Won’t be able to use all transistors simultaneously
- Serious problem for many-cores...

Source: Krisztián Flautner “From niche to mainstream: can critical systems make the transition?”
Specialization leads to more efficiency

**GPU**

*200pJ/Instruction*

- Optimized for Throughput
- Explicit Management of On-chip Memory

**CPU**

*2nJ/Instruction*

- Optimized for Latency
- Caches

Source: Bill Dally, «To ExaScale and Beyond»

www.nvidia.com/content/PDF/sc_2010/theater/Dally_SC10.pdf
Locality and communications management

- In 22 nm, swapping 1 bit in a transistor has an energy cost:
  \[ \sim 1 \text{ attojoule (10}^{-18} \text{ J)} \]

- Moving a 1-bit data on the silicon cost:
  \[ \sim 1 \text{ picojoule/mm (10}^{-12} \text{ j/mm)} \]

- Moving a data \( 10^9 \) per second (1 GHz) in silicon has a cost:
  \[ 1 \text{ pJ/mm x } 10^9 \text{ s}^{-1} = \sim 1 \text{ milliwatt/mm} \]

- 64 bit bus @ 1 GHz: \( \sim 64 \text{ milliwatts/mm} \) (with 100% activity)

- For 1 cm of 64 bit bus @ 1 GHz: 0,64 W/cm

- On modern chips, there are about several km of wires on chip, even with low toggle rate, this leads to several Watt/cm\(^2\)
Technological consequences

Efficiency $\rightarrow$ locality
Frequency limit $\rightarrow$ parallelism
Energy efficiency $\rightarrow$ specialization

Ease of programming
## Technological trends influencing Computing Systems

<table>
<thead>
<tr>
<th>Constraints</th>
<th>Opportunities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency Limits</td>
<td>CMOS Phonotic</td>
</tr>
<tr>
<td>Power Limits</td>
<td>Non-volatile memories</td>
</tr>
<tr>
<td>Dark Silicon</td>
<td>3D Stacking</td>
</tr>
<tr>
<td></td>
<td>New paradigms</td>
</tr>
</tbody>
</table>
Optical interconnects

CMOS photonic is the integration of a photonic layer with an electronic circuit.

Advantages of CMOS photonic are:

- Use of standard tools and foundry, wafer scale co-integration
- **Lower energy** (~100 fJ/bit), (wire: ~1 pJ/mm)
- High bandwidth (10 Gbps), Low latency (~10 ps/mm)

Source: CEA, Ahmed Jerraya
Non-volatile memories...

Example: Memristive Devices Principle

\[ v = R(x, i) \cdot i \quad \frac{dx}{dt} = f(x, i) \]

- Metal (\(M_{x+1}\) layer)
- Metal (\(M_x\) layer)
- Insulator
  - Oxide
  - Solid electrolytic
  - Organic material

Source: CEA, C. Gamrat

Multiple integration with 3D stacking...
Technology also drives us to think differently…

- Stochastic computing
- Biologically inspired computing
- Organic Computing
- Autonomous computing, Self-*
- 3D stacking
- Photonic interconnect
- Non-volatile memories
- Molecular computing
- More-than-Moore
- Spintronics
- Chemical computing
- Biologically inspired cells
- Memristors
- ...
- Also silicon based!
Big data meets energy in an intelligent connected physical world

Efficiency
Complexity
Dependability

Application pull

Technology push

Business trends

Technological constraints

Technological opportunities

Application pull

Technology push

Business trends

Technological constraints

Technological opportunities

Efficiency
Complexity
Dependability
Core Computing Systems Challenges

<table>
<thead>
<tr>
<th>Efficiency</th>
<th>Complexity</th>
<th>Dependability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power, Performance</td>
<td>Parallelism, Heterogeneity</td>
<td>Reliability, Privacy</td>
</tr>
</tbody>
</table>
Improving efficiency

- Multiple performance metrics
- Power defines performance
- Communication defines performance
- Heterogeneity and accelerators to the rescue
Managing complexity

- The reign of legacy code
- Parallelism seems to be too complex for humans
- Hardware complexity

(4G is 500x more complex than 2G)
Improving dependability

- Worst case design is not an option anymore
- Systems must be built from unreliable components
- Safety and security!
Impact on Society

HiPEAC Research objectives

7 HiPEAC Research objectives

Impact on Society

Efficiency Complexity Dependability

HiPEAC Strengths

HiPEAC Weaknesses

Technological constraints

Technological opportunities

Big data meets energy in an intelligent connected physical world

Application pull

Business trends

Technological constraints

Technological opportunities

Big data meets energy in an intelligent connected physical world

Application pull

Business trends

Technological constraints

Technological opportunities

Big data meets energy in an intelligent connected physical world

Application pull

Business trends

Technological constraints

Technological opportunities

Big data meets energy in an intelligent connected physical world

Application pull

Business trends

Technological constraints

Technological opportunities

Big data meets energy in an intelligent connected physical world

Application pull

Business trends

Technological constraints

Technological opportunities

Big data meets energy in an intelligent connected physical world

Application pull

Business trends

Technological constraints

Technological opportunities

Big data meets energy in an intelligent connected physical world
Impact on Society

HiPEAC Research objectives

Big data meets energy in an intelligent connected physical world

Efficiency Complexity Dependability

HiPEAC Strengths

Application pull

Business trends

Technological constraints

Technological opportunities

HiPEAC Weaknesses

Big data meets energy in an intelligent connected physical world

7 HiPEAC Research objectives

Impact on Society

Technological constraints

Technological opportunities

Efficiency Complexity Dependability

HiPEAC Strengths

Application pull

Business trends

7 HiPEAC Research objectives

Impact on Society
Derived HiPEAC Research Objectives

- Cost-effective software for heterogeneous multicores
- Cross-component/cross-layer optimization for design integration
- Next-generation processor cores

- Architectures for the Data Deluge
- Reliable systems for Ubiquitous Computing
- Heterogeneous computing systems
- Locality and communications management
Cost-effective software for heterogeneous multicores

Frequency limit ➔ parallelism
Energy efficiency ➔ heterogeneity

Ease of programming
## Detailed HiPEAC Research areas

<table>
<thead>
<tr>
<th>Parallelism and Programming Models</th>
<th>Cross-component/computing cores</th>
<th>Reliable systems for Ubiquitous Computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1.1. Locality Management</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.1.2. Optimizations programmer hints, tuning</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.1.3. Runtime Systems and Adaptivity</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.2.1. Processors, Accelerators, Heterogeneity</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.2.2. Memory Architectures</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.2.3. Interconnection Architectures</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.2.4. Reconfigurability</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Compilers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.3.1. Automatic Parallelization</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.3.2. Adaptive Compilation</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>9.3.3. Intelligent Optimization</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Systems Software and Tools</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.4.1. Virtualization</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.4.2. Input, Output, Storage, and Networking</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.4.3. Simulation and Design Automation Tools</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>9.4.4. Deterministic Performance Tools</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Note: The table represents the extent to which each research area is addressed in the HiPEAC project, with 'x' indicating a focus area.
Global optimization

- Efficiency
  - Heterogeneous Computing
  - Locality & communication
  - Cost-effective software
  - Cross component optimization
- System Complexity
  - Next generation computing
- Dependability
  - Architecture for Data Deluge
  - Reliable ubiquitous systems

Data Deluge
Energy Wall, Connected, Real world data

*Turning point for Moore’s law*

Exiting new opportunities are ahead of us!