Plenary Panel: Rebooting Computing
Low Power Multicores with Accelerators and Automatic Parallelizing and Power Reducing Compiler for Exponential Performance Scaling

Hironori Kasahara
Professor, Dept. of Computer Science & Engineering
Director, Advanced Multicore Processor Research Institute
Waseda University (早稲田大学), Tokyo, Japan
IEEE Computer Society Multicore STC Chair
URL: http://www.kasahara.cs.waseda.ac.jp/
Multicores for Performance and Low Power

Power consumption is one of the biggest problems for performance scaling from smartphones to cloud servers and supercomputers (“K” more than 10MW).

\[ \text{Power} \propto \text{Frequency} \times \text{Voltage}^2 \]

(Voltage \(\propto\) Frequency)

\(\Rightarrow\) \[ \text{Power} \propto \text{Frequency}^3 \]

If **Frequency** is reduced to \(1/4\) (Ex. 4GHz \(\rightarrow\) 1GHz),

**Power** is reduced to \(1/64\) and

**Performance** falls down to \(1/4\).

<**Multicores**>

If **8cores** are integrated on a chip,

**Power** is still \(1/8\) and

**Performance** becomes \(2\) times.

---

IEEE ISSCC08: Paper No. 4.5, M. ITO, … and H. Kasahara, “An 8640 MIPS SoC with Independent Power-off Control of 8 CPUs and 8 RAMs by an Automatic Parallelizing Compiler”
For Performance Scaling and Power Reduction on Multicores

- Efficient parallel software is necessary though very difficult
- Development Cost & Period are crucial problems
- Compiler generating low power parallel software is required

<Key Technologies for Compilers>

Multigrain Parallelization

Hierarchical coarse-grain task parallelization among loops and subroutines, in addition to the traditional loop parallelization

Data Localization

Automatic data management for distributed shared memory, cache and local memory over the whole program

Data Transfer Overlapping

Data transfer overlapping using Data Transfer Controllers (DMAs)

Power Reduction

Compiler controls DVFS and Clock and Power gating automatically
Save lives from natural disasters

211 Times Speedup against the Sequential Processing using Sun Studio Compiler for GMS Earthquake Wave Propagation Simulation on Fujitsu M9000 Sparc 128 core SMP

- 100 times speedup on 128 cores against one core using OSCAR compiler
- 211 times speedup on 128 cores against original GMS program on one core using OSCAR Sun Studio Compiler
110 Times Speedup against the Sequential Processing for GMS Earthquake Wave Propagation Simulation on Hitachi SR16000 (Power7 Based 128 Core Linux SMP)
Cancer Treatment
Carbon Ion Radiotherapy

(Previous best was 2.5 times speedup on 16 processors with hand optimization)

8.9 times speedup by 12 processors
Intel Xeon X5670 2.93GHz 12 core SMP (Hitachi HA8000)

55 times speedup by 64 processors
IBM Power 7 64 core SMP (Hitachi SR16000)
Engine Control by multicore with Denso

Though so far parallel processing of the engine control on multicore has been very difficult, Denso and Waseda succeeded 1.95 times speedup on 2core V850 multicore processor.

Hard real-time automobile engine control by multicore

Automatic Parallelization of MATLAB/Simulink
Power Reduction of MPEG2 Decoding to 1/4 on Solar Powered 8 Core Multicore RP-2 by OSCAR Parallelizing Compiler

MPEG2 Decoding with 8 CPU cores

Without Power Control (Voltage: 1.4V)

With Compiler Automatic Power Control:
Clock & Power Gating and DVFS (1.4V-1.0V)

Avg. Power

Without Control: 5.73 [W]
With Compiler: 1.52 [W]

73.5% Power Reduction

2008 Demo Prime Minister
Power on 4 cores ARM CortexA9 with Android

http://www.youtube.com/channel/UCS43IYEIKqC8i_KIqFZYQBQ

H.264 decoder & Optical Flow (Using 3 cores)

ODROID X2
Samsung Exynos4412 Prime, ARM Cortex-A9 Quad core
1.7GHz～0.2GHz, used by Samsung's Galaxy S3

On the same 3 cores, the power control reduced the power to 1/5～1/7 against no power control.

The power control reduced the power to 1/2～1/3 compared with the ordinary sequential execution on 1 core without power control.
Automatic Power Reduction by OSCAR Compiler on Intel Haswell 4 Core Multicore

- Power Consumption for real-time face detection was reduced to 2/5 -

**Parallel processing of face detection program on Intel Haswell 4 cores**

- Average power consumption when automatic power reduction is applied

<table>
<thead>
<tr>
<th>Number of core</th>
<th>Without power control</th>
<th>With power control</th>
<th>Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.61</td>
<td>21.01</td>
<td>60.37%</td>
</tr>
<tr>
<td>3</td>
<td>39.24</td>
<td>15.55</td>
<td>60.37%</td>
</tr>
</tbody>
</table>

- Reduced to 3/5 (60.37%)

**Parallelization flow of OpenCV face detection program**

1. **Input**
2. **Camera**
3. **Face detection processing**
4. **Loop for searching changing sizes**
5. **Searching loop along x and y directions**
6. **Automatic parallelization by OSCAR compiler**
7. **Drawing**
8. **Output**
9. **Display**
10. **Next frame**

**Power measurement on Intel Haswell board**

- CPU: Intel Core i7 4770K
- Number of core: 4
- Clock frequency: 3.5GHz～0.8GHz
- Mother board: ASUS H81M-A

**Average Power Consumption [W]**

<table>
<thead>
<tr>
<th>Number of core</th>
<th>Speedup ratio</th>
<th>Speedup time [msec]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.44 times</td>
<td>93.06</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>48.80</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>38.08</td>
</tr>
</tbody>
</table>

**Inserting power measurement circuit between PMIC and CPU**
OSCAR Vector Multicore and Compiler for Solar Powered Disaster Survival Server

Vector supercomputers will evolve to low power embedded systems

Target: > 100GFLOPS/W

- Solar Powered with compiler power reduction.
- Fully automatic parallelization and vectorization including local memory management and data transfer.
Summary

- For Exponential Performance Scaling, a compiler and application cooperative low power multicore architecture having vector accelerators will be a key technology.
- For industry competitiveness, an automatic parallelizing and power reducing compiler is necessary since the software development cost and period are getting dominant factors.
- The multicore with the automatic parallelizing and power reducing compiler will be used in variety of application areas including smartphones, self-driving cars, cancer treatment systems, drinkable inner cameras, cloud servers, supercomputers and so on.
- Already the compiler gave us 110 times speedup for “Earthquake Wave Propagation Simulation” on 128 cores of IBM Power 7 against 1 core, 1.95 times for MATLAB/Simulink “Automobile Engine Control” on Renesas 2 cores using SH4A and V850 and so on.
- In automatic power reduction, consumed powers for H.264 and optical flow were reduced to 1/2 or 1/3 using 3 cores of ARM Cortex A9 and Intel Haswell against ordinary single core execution and real-time Human face detection to 3/5 using 3 cores of Haswell.