OSCAR Automatic Parallelization and Power Reduction Compiler for Homogeneous and Heterogeneous Multicores

Hironori Kasahara
Professor, Dept. of Computer Science & Engineering
Director, Advanced Multicore Processor Research Institute
Waseda University, Tokyo, Japan
IEEE Computer Society Multicore STC Chair
IEEE Computer Society 2015 Election President Candidate
URL: http://www.kasahara.cs.waseda.ac.jp/
Green Computing Systems R&D Center  
Waseda University  
Supported by METI (Mar. 2011 Completion)

<R & D Target>
Hardware, Software, Application for Super Low-Power Manycore Processors
- More than 64 cores
- Natural air cooling (No fan)  
  Cool, Compact, Clear, Quiet
- Operational by Solar Panel

<Industry, Government, Academia>
Hitachi, Fujitsu, NEC, Renesas, Olympus, Toyota, Denso, Mitsubishi, OSCAR Tech.

<Ripple Effect>
- Low CO₂ (Carbon Dioxide) Emissions
- Creation Value Added Products
  - Consumer Electronics, Automobiles, Servers
Industry-government-academia collaboration in R&D and target practical applications

Solar Powered Smart phones
Cameras
Robots
Cool desktop servers

Industry
Supercomputers and servers

For smart life

Protect environment

Waseda University : R&D
Many-core system technologies with ultra-low power consumption
OSCAR many-core chip

Compilers, API

OSC AR

Many-core Chip

Consumer electronic
Internet TV/DVD
Camcorders
Capsule inner cameras

Camer as

On-board vehicle technology (navigation systems, integrated controllers, infrastructure coordination)

OSCAR

Green supercomputers
Stock trading

Super real-time disaster simulation (tectonic shifts, tsunami, tornado, flood, fire spreading)

Green cloud servers

Heavy particle radiation planning, cerebral infarction)
Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Protect environment

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Ind ustry

Camcorders

Operation/recharging by solar cells

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR

Intelligent home appliances

Protect lives

Try Recharging by solar cells

Solar Powered Smart phones

Medical servers

Heavy particle radiation planning, cerebral infarction) Non-fan, cool, quiet servers designed for server

National Institute of Radiological Sciences

OSCAR
To improve effective performance, cost-performance and software productivity and reduce power.

**Multigrain Parallelization**

coarse-grain parallelism among loops and subroutines, near fine grain parallelism among statements in addition to loop parallelism

**Data Localization**

Automatic data management for distributed shared memory, cache and local memory

**Data Transfer Overlapping**

Data transfer overlapping using Data Transfer Controllers (DMAs)

**Power Reduction**

Reduction of consumed power by compiler control DVFS and Power gating with hardware supports.
Earliest Executable Condition Analysis for Coarse Grain Tasks (Macro-tasks)

A Macro Flow Graph

A Macro Task Graph
MTG of Su2cor-LOOPS-DO400

- Coarse grain parallelism PARA_ALD = 4.3
Data-Localization: Loop Aligned Decomposition

• Decompose multiple loop (Doall and Seq) into CARs and LRs considering inter-loop data dependence.
  – Most data in LR can be passed through LM.
  – LR: Localizable Region, CAR: Commonly Accessed Region
Multicore Program Development Using OSCAR API V2.0

Sequential Application Program in Fortran or C
(Consumer Electronics, Automobiles, Medical, Scientific computation, etc.)

Manual parallelization / power reduction

Accelerator Compiler/ User
Add “hint” directives before a loop or a function to specify it is executable by the accelerator with how many clocks

Waseda OSCAR Parallelizing Compiler
- Coarse grain task parallelization
- Data Localization
- DMAC data transfer
- Power reduction using DVFS, Clock/ Power gating

Low Power Heterogeneous Multicore Code Generation

Low Power Homogeneous Multicore Code Generation

Parallelized API F or C program

Proc0
Code with directives
Thread 0

Proc1
Code with directives
Thread 1

Accelerator 1
Code

Accelerator 2
Code

Server Code Generation
OpenMP Compiler

OSCAR: Optimally Scheduled Advanced Multiprocessor API: Application Program Interface

Hitachi, Renesas, NEC, Fujitsu, Toshiba, Denso, Olympus, Mitsubishi, Esol, Cats, Gaio, 3 univ.

Generation of parallel machine codes using sequential compilers

Homegeneous Multicores from Vendor A (SMP servers)

Heterogeneous Multicores from Vendor B

Executable on various multicores

Shred memory servers
#pragma omp parallel sections
{
  #pragma omp section
  main_vpc0();
  #pragma omp section
  main_vpc1();
  #pragma omp section
  main_vpc2();
  #pragma omp section
  main_vpc3();
}

Parallel Execution
(A thread and a core are bound one-by-one)

VPC: Virtual Processor Core
Memory Mapping

• Placing variables on an onchip centralized shared memory (onchipCSM)
  • #pragma oscar onchipshared (C)
  • !$oscar onchipshared (Fortran)

• Placing variables on a local data memory (LDM)
  • #pragma omp threadprivate (C)
  • !$omp threadprivate (Fortran)
  • This directive is an extension to OpenMP

• Placing variables on a distributed shared memory (DSM)
  • #pragma oscar distributedshared (C)
  • !$oscar distributedshared (Fortran)
Data Transfer

• Specifying data transfer lists
  – #pragma oscar dma_transfer (C)
  – !$oscar dma_transfer (Fortran)
  – Containing following parameter directives
• Specifying a contiguous data transfer
  – #pragma oscar dma_contiguous_parameter (C)
  – !$oscar dma_contiguous_parameter (Fortran)
• Specifying a stride data transfer
  – #pragma oscar dma_stride_parameter
  – !$oscar dma_stride_parameter
  – This can be used for scatter/gather data transfer
• Data transfer synchronization
  – #pragma oscsar dma_flag_check
  – !$oscar dma_flag_check
Hierarchical Barrier Synchronization

- Specifying a hierarchical group barrier
  - `#pragma oscar group_barrier` (C)
  - `!$oscar group_barrier` (Fortran)
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)
110 Times Speedup against the Sequential Processing for GMS Earthquake Wave Propagation Simulation on Hitachi SR16000 (Power7 Based 128 Core Linux SMP)
9.6 Times Speedup on 12 cores Power 8 against the Sequential Processing for GMS Earthquake Wave Propagation Simulation

- **IBM S812L**
  - CPU: POWER8
    - 12 cores
    - Clock Frequency: 3.026GHz
  - Memory: 60GB
  - OS: Redhat Linux 7.1
  - Backend Fortran compiler: IBM xlf 15.1.1
  - Evaluation using medium size input data (12GB)
Performance of OSCAR Compiler on IBM p6 595 Power6 (4.2GHz) based 32-core SMP Server

Compile Option:
(*2) Sequential: -O5 -q64 –qarch=pwr6, XLF: -O5 –q64 –qarch=pwr6 –qsmp=auto, OSCAR: -O5 –q64 –qarch=pwr6 –qsmp=noauto
(Others) Sequential: -O5 –qarch=pwr6, XLF: -O5 –qarch=pwr6 –qsmp=auto, OSCAR: -O5 –qarch=pwr6 –qsmp=noauto

OpenMP codes generated by OSCAR compiler accelerate IBM XL Fortran for AIX Ver.12.1 about 3.3 times on the average
Performance of OSCAR Compiler for Fortran Programs on a IBM p550q 8core Deskside Server

- 2.7 times speedup against loop parallelizing compiler on 8 cores

- Loop parallelization
- Multigrain parallelization

![Bar chart showing speedup ratio for various benchmarks](image-url)
Performance for C programs on IBM p5 550Q

5.8 times speedup against one processor on average
Performance of OSCAR Compiler on IBM pSeries690 (Power 4) RegattaH 16 Processors High-end Server

IBM XL Fortran for AIX Version 8.1

3.5 times speedup in average
Parallel Processing of Face Detection on Manycore, Highend and PC Server

- OSCAR compiler gives us 11.55 times speedup for 16 cores against 1 core on SR16000 Power7 highend server.

Automatic Parallelization of Face Detection
Speedup with 2 cores for Engine Crankshaft Handwritten Program on Renesas RPX Multi-core Processor

Macrotask graph with a lot of conditional branches

Branches are fused to macrotasks for static scheduling

Macrotask graph after task fusion

Grain is too fine (μs) for dynamic scheduling.

1.6 times Speed up by 2 cores against 1 core

Graph showing speedup comparison:
- 1.60 speedup against 1 core
- 1.00 speedup against 2 cores

1core 2core
OSCAR Compile Flow for Simulink Applications

1. Generate MTG → Parallelism

2. Generate gantt chart → Scheduling in a multicore

3. Generate parallelized C code using the OSCAR API → Multiplatform execution (Intel, ARM and SH etc)
Road Tracking, Image Compression, Buoy Detection, Image Compression, Color Edge Detection, Optical Flow, Vessel Detection

(Intel Xeon, ARM Cortex A15 and Renesas SH4A)

Road Tracking, Image Compression: [http://www.mathworks.co.jp/jp/help/vision/examples](http://www.mathworks.co.jp/jp/help/vision/examples)
OSCAR Heterogeneous Multicore

- **DTU**
  - Data Transfer Unit

- **LPM**
  - Local Program Memory

- **LDM**
  - Local Data Memory

- **DSM**
  - Distributed Shared Memory

- **CSM**
  - Centralized Shared Memory

- **FVR**
  - Frequency/Voltage Control Register
An Image of Static Schedule for Heterogeneous Multi-core with Data Transfer Overlapping and Power Control

<table>
<thead>
<tr>
<th></th>
<th>CPU0</th>
<th></th>
<th>CPU1</th>
<th></th>
<th>CPU2</th>
<th></th>
<th>CPU3</th>
<th></th>
<th>DRP0</th>
</tr>
</thead>
<tbody>
<tr>
<td>CORE</td>
<td></td>
<td>CORE</td>
<td></td>
<td>CORE</td>
<td></td>
<td>CORE</td>
<td></td>
<td>CORE</td>
<td></td>
</tr>
<tr>
<td>DTU</td>
<td></td>
<td>DTU</td>
<td></td>
<td>DTU</td>
<td></td>
<td>DTU</td>
<td></td>
<td>DTU</td>
<td></td>
</tr>
</tbody>
</table>

**MTG1**
- MT1-1
  - LOAD
  - LOAD
  - SEND
- MT1-2
  - LOAD
  - LOAD
- MT1-3
  - SEND
- MT1-4
  - SEND

**MTG2**
- MT2-1
  - LOAD
  - LOAD
  - SEND
- MT2-2
  - OFF
  - OFF
  - OFF
- MT2-3
  - LOAD
  - LOAD
  - SEND

**MTG3**
- MT3-1
  - LOAD
  - LOAD
  - SEND
- MT3-2
  - LOAD
  - LOAD
  - SEND
- MT3-3
  - LOAD
  - LOAD
  - SEND
- MT3-4
  - SEND
  - SEND
  - SEND
- MT3-5
  - SEND
  - SEND
  - SEND
- MT3-6
  - SEND
  - SEND
  - SEND
- MT3-7
  - STORE
  - STORE
  - STORE
- MT3-8
  - STORE
  - STORE
  - STORE

**DRP0**
- OFF
  - OFF
  - OFF

**TIME**
- 25
33 Times Speedup Using OSCAR Compiler and OSCAR API on RP-X
(Optical Flow with a hand-tuned library)

Y. Yuyama, et al., “A 45nm 3.73GOPS/W Heterogeneous Multi-Core SoC”, ISSCC2010

1 SH  2 SH  4 SH  8 SH  2 SH+1 FE  4 SH+2 FE  8 SH+4 FE

Speedups against a single SH processor

111 [fps]
Power Reduction in a real-time execution controlled by OSCAR Compiler and OSCAR API on RP-X (Optical Flow with a hand-tuned library)

Without Power Reduction

Average: 1.76[W]

1 cycle: 33[ms] → 30[fps]

With Power Reduction by OSCAR Compiler

Average: 0.54[W]

70% of power reduction
Low-Power Optimization with OSCAR API

Scheduled Result by OSCAR Compiler

VC0
MT1
MT2
Sleep
MT3
MT4

VC1

Generate Code Image by OSCAR Compiler

void main_VC0() {
    MT2
    #pragma oscar fvcontrol ¥
    ((OSCAR_CPU(),0))
    Sleep
    MT1
    MT3
    MT4
}

void main_VC1() {
    MT2
    #pragma oscar fvcontrol ¥
    (1,(OSCAR_CPU(),100))
    Sleep
    MT1
    MT3
    MT4
}
Power Reduction of MPEG2 Decoding to 1/4 on 8 Core Homogeneous Multicore RP-2 by OSCAR Parallelizing Compiler

Without Power Control (Voltage : 1.4V)

With Power Control (Frequency, Resume Standby: Power shutdown & Voltage lowering 1.4V-1.0V)

 Avg. Power  
5.73 [W]  73.5% Power Reduction  Avg. Power  
1.52 [W]
Low Power High Performance Multicore Computer with Solar Panel

- Clean Energy Autonomous
- Servers operational in deserts
Power on 4 cores ARM CortexA9 with Android

http://www.youtube.com/channel/UCS43INYEIkC8i_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

<table>
<thead>
<tr>
<th></th>
<th>H.264</th>
<th>Optical flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 core</td>
<td>1.07</td>
<td>0.72</td>
</tr>
<tr>
<td>2 cores</td>
<td>1.69</td>
<td>0.79</td>
</tr>
<tr>
<td>3 cores</td>
<td>2.45</td>
<td>0.95</td>
</tr>
</tbody>
</table>

- 86.5% (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.

- On the same 3 cores, the power control reduced the power to 1/5 ～ 1/7 against no power control.
Power Reduction on Intel Haswell 3cores

H.264 decoder & Optical Flow

H81M-A, Intel Core i7 4770k
Quad core, 3.5GHz〜0.8GHz

<table>
<thead>
<tr>
<th></th>
<th>1 core</th>
<th>2 cores</th>
<th>3 cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.264</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>without power control</td>
<td>29.67</td>
<td>37.11</td>
<td>41.81</td>
</tr>
<tr>
<td>with power control</td>
<td>17.37</td>
<td>16.15</td>
<td>12.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optical flow</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>without power control</td>
<td>29.29</td>
<td>36.59</td>
<td>41.58</td>
</tr>
<tr>
<td>with power control</td>
<td>12.21</td>
<td>9.60</td>
<td></td>
</tr>
</tbody>
</table>

- The power consumption was reduced to 1/3〜1/4 by OSCAR power control against no power control on the same 3 processor cores.
- The power reduced to 2/5〜1/3 by the compiler power control against 1 core no 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>Average Power Consumption [W]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>without power control: 27.61 W, with power control: 21.01 W (reduced to 3/5, -43.68%)</td>
</tr>
<tr>
<td>3</td>
<td>without power control: 39.24 W, with power control: 15.55 W (reduced to 2/5, -60.37%)</td>
</tr>
</tbody>
</table>

Average Power Consumption

- Number of core:
  - 1: without power control: 27.61 W, with power control: 21.01 W (reduced to 3/5, -43.68%)
  - 3: without power control: 39.24 W, with power control: 15.55 W (reduced to 2/5, -60.37%)

Speedups against sequential execution at the fastest execution mode

<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>93.06 [msec]</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>48.80 [msec]</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.44 times speedup</td>
<td>38.08 [msec]</td>
</tr>
</tbody>
</table>

Parallelization flow of OpenCV face detection program

Input → Camera → Loop for searching changing sizes → Face detection processing → Drawing → Display

Automatic parallelization by OSCAR compiler

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

Inserting power measurement circuit between PMIC and CPU
OSCAR Vector Multicore with Power Processor Core and Compiler for Embedded to Servers with OSCAR Technology

Target:

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

- OSCAR Automatic Parallelizing and Power Reducing Compiler has succeeded speedup and/or power reduction of scientific applications including medical applications and natural disaster simulation, and real-time applications like Automobile Engine Control and MATLAB/Simulink, and media applications codec and face detection on various homogeneous and heterogeneous multicores.

- In automatic parallelization on Power Architectures: OSCAR Compiler has been developed on Power architectures like Power 4, 5, 5+, 6, 7 and 8.
  - for “Earthquake Wave Propagation Simulation”, 110 times speedup on 128 cores of IBM Power 7 and 9.6 times speedup on 12 cores IBM Power8 against 1 core,
  - for “Cancer Treatment Using Carbon Ion”, 55 times speedup on 64 cores of IBM Power 7 against 1 core

- In automatic power reduction, consumed powers were reduced to 1/2 or 1/3 using 3 cores on ARM Cortex A9 and Intel Haswell.

- We are planning to realize automatic power reduction using “On-Chip Regulator” on Power Processors with accelerators.