OSCAR Parallelizing and Power Reducing Compiler for Multicores

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 President Elect 2017, President 2018 URL: http://www.kasahara.cs.waseda.ac.jp/

Waseda Univ. GCSC

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).



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" Power ∝ Frequency * Voltage² (Voltage ∝ Frequency)

▶ Power ∝ Frequency³

If <u>Frequency</u> is reduced to <u>1/4</u> (Ex. 4GHz→1GHz), Power is reduced to 1/64 and Performance falls down to <u>1/4</u>. <<u>Multicores</u>> If <u>8cores</u> are integrated on a chip, Power is still <u>1/8</u> and

<u>Performance</u> becomes <u>2 times</u>.





original (sun studio)





With 128 cores, OSCAR compiler gave us 100 times speedup against 1 core execution and 211 times speedup against 1 core using Sun (Oracle) Studio compiler.



Demo of NEDO Multicore for Real Time Consumer Electronics at the Council of Science and Engineering Policy on April 10, 2008

第74回総合科学技術会議【平成20年4月10日】



第74回総合科学技術会議の様子(1)







第74回総合科学技術会議の様子(4)

CSTP Members **Prime Minister:** Mr. Y. FUKUDA **Minister of State for** Science, Technology and Innovation **Policy:** Mr. F. KISHIDA **Chief Cabinet** Secretary: Mr. N. MÁCHIMURA **Minister of Internal** Affairs and **Communications** : Mr. H. MASUDA **Minister of Finance :** Mr. F. NUKAGA **Minister of Education**, Culture, **Sports, Science and Technology:** Mr. K. TOKAI **Minister** of **Economy, Trade and Industry:** Mr. A. AMARI

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, Toshiba, etc

<Ripple Effect>

>Low CO₂ (Carbon Dioxide) Emissions Creation Value Added Products

Consumer Electronics, Automobiles,

Servers





SPARC VII 256 core SMP



Beside Subway Waseda Station, Near Waseda Univ. Main Campus



OSCAR Parallelizing Compiler

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.



Generation of Coarse Grain Tasks

Macro-tasks (MTs)

- Block of Pseudo Assignments (BPA): Basic Block (BB)
- Repetition Block (RB) : natural loop

Subroutine Block (SB): subroutine



Earliest Executable Condition Analysis for Coarse Grain Tasks (Macro-tasks)



Automatic processor assignment in 103.su2cor

• Using 14 processors

Coarse grain parallelization within DO400



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







Data Layout for Removing Line Conflict Misses by Array Dimension Padding Declaration part of arrays in spec95 swim

after padding before padding PARAMETER (N1=513, N2=544) PARAMETER (N1=513, N2=513) COMMON U(N1,N2), V(N1,N2), P(N1,N2), COMMON U(N1,N2), V(N1,N2), P(N1,N2), * UNEW(N1,N2), VNEW(N1,N2), UNEW(N1,N2), VNEW(N1,N2), * PNEW(N1,N2), UOLD(N1,N2), PNEW(N1,N2), UOLD(N1,N2), 1 1 VOLD(N1,N2), POLD(N1,N2), * VOLD(N1,N2), POLD(N1,N2), * CU(N1,N2), CV(N1,N2), CU(N1,N2), CV(N1,N2), 2 2 Z(N1,N2), H(N1,N2) Z(N1,N2), H(N1,N2) * 4MB 4MB padding Box: Access range of DLG0



Parallel Execution

- Start of parallel execution
 - #pragma omp parallel sections (C)
 - !\$omp parallel sections (Fortran)
- Specifying critical section
 - #pragma omp critical (C)
 - !\$omp critical (Fortran)
- Enforcing an order of the memory operations
 - #pragma omp flush (C)
 - !\$omp flush (Fortran)
- These are from **OpenMP**.

Thread Execution Model



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

Power Control

- Making a module into specifying frequency and voltage state
 - #pragma oscar fvcontrol (C)
 - !\$oscar fvcontrol (Fortran)
 - state examples
 - 100: max frequency
 - 50: half frequency
 - 0: clock off
 - -1: power off



- Getting a frequency and voltage state of a module
 - #pragma oscar get_fvstatus (C)
 - !\$oscar get_fvstatus (Fortran)

Cancer Treatment Carbon Ion Radiotherapy

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



8.9times 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)

Renesas-Hitachi-Waseda Low Power 8 core RP2 Developed in 2007 in METI/NEDO project

Core#0	Core#1		Process Technology	90nm, 8-layer, triple- Vth, CMOS
Core#2	Core#3		Chip Size	104.8mm ² (10.61mm x 9.88mm)
Core#6	WY Core#7		CPU Core Size	6.6mm ² (3.36mm x 1.96mm)
Core#4	Core#5		Supply Voltage	1.0V–1.4V (internal), 1.8/3.3V (I/O)
DDRPAD	GCPG		Power Domains	17 (8 CPUs, 8 URAMs, common)

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"

8 Core RP2 Chip Block Diagram



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.



Macro Task Fusion for Static Task Scheduling



Evaluation Environment : Embedded Multi-core Processor RPX



□ SH-4A 648MHz * 8

 As a first step, we use just two SH-4A cores because target dual-core processors are currently under design for nextgeneration automobiles



Evaluation of Crankshaft Program with Multi-core Processors



- □ Attain 1.54 times speedup on RPX
 - There are no loops, but only many conditional branches and small basic blocks and difficult to parallelize this program
- This result shows possibility of multi-core processor for engine control programs

OSCAR Compile Flow for Simulink Applications



Speedups of MATLAB/Simulink Image Processing on Various 4core Multicores

(Intel Xeon, ARM Cortex A15 and Renesas SH4A)



to-grayscale-/

Vessel Detection : <u>http://www.mathworks.co.jp/matlabcentral/fileexchange/24990-retinal-blood-vessel-extraction/</u>

OSCAR Compiler Accelerates Various Multicores' Performance Several Times Including Intel and IBM



Performance of OSCAR Compiler on Intel Core i7 Notebook PC



• OSCAR Compiler accelerate Intel Compiler about 2.0 times on average

Parallelization of 2D Rendering Engine SKIA on 3 cores of Google NEXUS7

http://www.youtube.com/channel/UCS43INYEIkC8i_KIgFZYQBQ

DrawRect : FPS

DrawImage : FPS

110 Times Speedup against the Sequential Processing for GMS Earthquake Wave Propagation Simulation on Hitachi SR16000 (Power7 Based 128 Core Linux SMP)

Automatic Parallelization of Still Image Encoding Using JPEG-XR for the Next Generation Cameras and Drinkable Inner Camera

Waseda U. & Olympus

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

OSCAR Heterogeneous Multicore

An Image of Static Schedule for Heterogeneous Multicore with Data Transfer Overlapping and Power Control

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

Low-Power Optimization with OSCAR API

- On 3 cores, Automatic Power Reduction control successfully reduced power to 1/7 against without Power Reduction control.
- 3 cores with the compiler power reduction control reduced power to 1/3 against ordinary 1 core execution.

Power was reduced to 1/4 (9.6W) by the compiler power optimization on the same 3 cores (41.6W).

Power with 3 core was reduced to 1/3 (9.6W) against 1 core (29.3W).

Performance of OSCAR Compiler Software Coherence Control

- Faster or Equal Processing Performance up to 4cores with hardware coherent mechanism on RP2.
- Software Coherence gives us correct execution without hardware coherence mechanism on 8 cores.

OSCAR Technology

Started up on Feb.28, 2013: Licensing the all patents and OSCAR compiler from Waseda Univ.

CEO: Dr. T. Ono (Ex- CEO of First Section-listed Company, VP of National Univ., Invited Prof. of Waseda U.) Executives: Mr. T. Ito (Visiting Prof. Tokyo Agricult. and Eng. U.) Prof. K. Shirai (Ex-President of Waseda U Chairman of Japanese Open Univ.) CTO: Mr. M. Takamura (Ex-Fellow Fujitsu Lab., Fujitsu VPP500, 5000 & NWT Development Leader) Mr. K. Ashida (Ex-VP Sumitomo Trading, Ashida Consult. CEO, A leader of Business World Auditor: Dr. S. Matsuda (Prof. Emeritus Waseda U. **Ex-President Ventures and Entrepreneurs Society**) Advisors: Dr. T. Sato (Patent Attorney, Ex-President of Patent Attorneys Assoc., Gov. IP Committee) Fujitsu VPP5000 Ms. K. Ishiguro (Lawyer, Supreme Court Trainer) Mr. A. Fukuda (Leader of Alumni Assoc.) Prof. K. Kimura (Waseda Univ.) PE / memory uni Prof. H. Kasahara (Waseda Univ.)

OSCARTECHNOLOGY CORPORATION

Copyright 2015

OSCAR Vector Multicore and Compiler for Embedded to Severs with OSCAR Technology

Target:

Solar Powered with

compiler power reduction.

Fully automatic

parallelization and

vectorization including

local memory management

and data transfer.

Fujitsu VPP500/NWT: PE Unit

49

Summary

- Waseda University Green Computing Systems R&D Center supported by METI has been researching on low-power high performance Green Multicore hardware, software and application with government and industry including Hitachi, Fujitsu, NEC, Renesas, Denso, Toyota, Olympus and OSCAR Technology.
- Socar Automatic Parallelizing and Power Reducing Compiler has succeeded speedup and/or power reduction of scientific applications including "Earthquake Wave Propagation", medical applications including "Cancer Treatment Using Carbon Ion", and "Drinkable Inner Camera", industry application including "Automobile Engine Control", "Smartphone", and "Wireless communication Base Band Processing" on various multicores from different vendors including Intel, ARM, IBM, AMD, Qualcomm, Freescale, Renesas and Fujitsu.
- In automatic parallelization, 110 times speedup for "Earthquake Wave Propagation Simulation" on 128 cores of IBM Power 7 against 1 core, 55 times speedup for "Carbon Ion Radiotherapy Cancer Treatment" on 64cores IBM Power7, 1.95 times for "Automobile Engine Control" on Renesas 2 cores using SH4A or V850, 55 times for "JPEG-XR Encoding for Capsule Inner Cameras" on Tilera 64 cores Tile64 manycore.
 - > The compiler will be available on market from OSCAR Technology.
- In automatic power reduction, consumed powers for real-time multi-media applications like Human face detection, H.264, mpeg2 and optical flow were reduced to 1/2 or 1/3 using 3 cores of ARM Cortex A9 and Intel Haswell and 1/4 using Renesas SH4A 8 cores against ordinary single core execution.