

IBM XL C/C++ for Linux, V13.1.6



Optimization and Programming Guide for Little Endian Distributions

Version 13.1.6

IBM XL C/C++ for Linux, V13.1.6



Optimization and Programming Guide for Little Endian Distributions

Version 13.1.6

Note

Before using this information and the product it supports, read the information in “Notices” on page 117.

First edition

This edition applies to IBM XL C/C++ for Linux, V13.1.6 (Program 5765-J08; 5725-C73) and to all subsequent releases and modifications until otherwise indicated in new editions. Make sure you are using the correct edition for the level of the product.

© Copyright IBM Corporation 1996, 2017.

US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

About this document	v
Who should read this document.	v
How to use this document.	v
How this document is organized	v
Conventions	vi
Related information.	ix
Available help information	ix
Standards and specifications	xi
Other IBM information.	xi
Other information	xi
Technical support	xi
How to send your comments	xii
Chapter 1. Using XL C/C++ with Fortran	1
Identifiers	1
Corresponding data types	2
Character and aggregate data.	4
Function calls and parameter passing	4
Pointers to functions.	5
Sample program: C/C++ calling Fortran	5
Chapter 2. Aligning data.	7
Using alignment modes.	7
Alignment of aggregates	8
Alignment of bit-fields	9
Using alignment modifiers	10
Chapter 3. Handling floating-point operations	13
Floating-point formats.	13
Handling multiply-and-add operations	13
Compiling for strict IEEE conformance	14
Handling floating-point constant folding and rounding	14
Matching compile-time and runtime rounding modes	15
Handling floating-point exceptions	16
Chapter 4. Constructing a library	17
Compiling and linking a library	17
Compiling a static library.	17
Compiling a shared library	17
Linking a library to an application.	17
Linking a shared library to another shared library	18
Initializing static objects in libraries (C++)	18
Chapter 5. Optimizing your applications 21	
Distinguishing between optimization and tuning	21
Steps in the optimization process	22
Basic optimization	22
Optimizing at level 0	22
Optimizing at level 2	23
Advanced optimization	24
Optimizing at level 3	25

An intermediate step: adding -qhot suboptions at level 3	26
Optimizing at level 4	26
Optimizing at level 5	27
Tuning for your system architecture	28
Getting the most out of target machine options	28
Using high-order loop analysis and transformations	29
Getting the most out of -qhot	30
Using shared-memory parallelism (SMP)	31
Getting the most out of -qsmp	32
Using interprocedural analysis	33
Getting the most from -qipa	35
Using profile-directed feedback.	36
Compiling with -qpdf1	36
Training with typical data sets	37
Recompiling or linking with -qpdf2	40
Marking variables as local or imported	43
Getting the most out of -qdatalocal	44
Using compiler reports to diagnose optimization opportunities	45
Parsing compiler reports with development tools	47
Other optimization options	47
Chapter 6. Debugging optimized code 49	
Detecting errors in code	49
Understanding different results in optimized programs	50
Debugging in the presence of optimization	51
Chapter 7. Coding your application to improve performance	53
Finding faster input/output techniques	53
Reducing function-call overhead	53
Managing memory efficiently (C++ only)	55
Optimizing variables	55
Manipulating strings efficiently.	56
Optimizing expressions and program logic	56
Optimizing operations in 64-bit mode	57
The C++ template model	58
Using delegating constructors (C++11)	58
Using rvalue references (C++11)	59
Using visibility attributes (IBM extension)	61
Types of visibility attributes	62
Rules of visibility attributes	63
Propagation rules (C++ only)	69
Specifying visibility attributes using the -fvisibility option	71
Specifying visibility attributes using pragma preprocessor directives	71
Chapter 8. Exploiting POWER9 technology	75
POWER9 compiler options	75
High-performance libraries that are tuned for POWER9	75

POWER9 built-in functions 75

Chapter 9. Using the high performance libraries 81

Using the Mathematical Acceleration Subsystem (MASS) libraries 81

- Using the scalar library 82
- Using the vector libraries 84
- Using the SIMD libraries 88
- Compiling and linking a program with MASS. . . 91

Using the Basic Linear Algebra Subprograms - BLAS 92

- BLAS function syntax 93
- Linking the libxlopt library 95

Chapter 10. Using vector technology 97

Chapter 11. Parallelizing your programs 99

Countable loops 99

Enabling automatic parallelization 101

Data sharing attribute rules. 101

Using OpenMP directives 103

Shared and private variables in a parallel environment. 104

Reduction operations in parallelized loops. . . . 105

Chapter 12. Offloading computations to the NVIDIA GPUs 107

System prerequisites for offloading to the NVIDIA GPUs 107

Programming with OpenMP device constructs . . 107

- Extensions from OpenMP Technical Reports . . 108

Using IBM XL C/C++ for Linux with NVCC . . . 110

Limitations 110

Chapter 13. Vector element order toggling. 113

Notices 117

Trademarks 119

Index 121

About this document

This guide discusses advanced topics related to the use of IBM® XL C/C++ for Linux, V13.1.6, with a particular focus on program portability and optimization. The guide provides both reference information and practical tips for getting the most out of the compiler's capabilities through recommended programming practices and compilation procedures.

Who should read this document

This document is addressed to programmers building complex applications, who already have experience compiling with XL C/C++ and would like to take further advantage of the compiler's capabilities for program optimization and tuning, support for advanced programming language features, and add-on tools and utilities.

How to use this document

This document uses a task-oriented approach to present the topics by concentrating on a specific programming or compilation problem in each section. Each topic contains extensive cross-references to the relevant sections of the reference guides in the IBM XL C/C++ for Linux, V13.1.6 documentation set, which provides detailed descriptions of compiler options, pragmas, and specific language extensions.

How this document is organized

This guide includes the following chapters:

- Chapter 1, "Using XL C/C++ with Fortran," on page 1 discusses considerations for calling Fortran code from XL C/C++ programs.
- Chapter 2, "Aligning data," on page 7 discusses options available for controlling the alignment of data in aggregates, such as structures and classes.
- Chapter 3, "Handling floating-point operations," on page 13 discusses options available for controlling how floating-point operations are handled by the compiler.
- Chapter 4, "Constructing a library," on page 17 discusses how to compile and link static and shared libraries and how to specify the initialization order of static objects in C++ programs.
- Chapter 5, "Optimizing your applications," on page 21 discusses various options provided by the compiler for optimizing your programs, and it provides recommendations on how to use these options.
- Chapter 6, "Debugging optimized code," on page 49 discusses the potential usability problems of optimized programs and the options that can be used to debug optimized code.
- Chapter 7, "Coding your application to improve performance," on page 53 discusses recommended programming practices and coding techniques to enhance program performance and compatibility with the compiler's optimization capabilities.
- Chapter 9, "Using the high performance libraries," on page 81 discusses two performance libraries that are shipped with XL C/C++: the Mathematical Acceleration Subsystem (MASS), which contains tuned versions of standard

math library functions, and the Basic Linear Algebra Subprograms (BLAS), which contains basic functions for matrix multiplication.

- Chapter 10, “Using vector technology,” on page 97 provides an overview of the vector technology, which can be used to accelerate the performance-driven, high-bandwidth communications and computing applications.
- Chapter 11, “Parallelizing your programs,” on page 99 provides an overview of options offered by XL C/C++ for creating multi-threaded programs, including OpenMP language constructs.
- Chapter 12, “Offloading computations to the NVIDIA GPUs,” on page 107 provides an overview of the methods to exploit the NVIDIA GPUs by using XL C/C++.
- Chapter 13, “Vector element order toggling,” on page 113 discusses that if users want to consistently use the instructions generated by vector built-in functions, users need to make all existing Vector Multimedia Extension (VMX) and Vector Scalar Extension (VSX) load and store built-in functions operate on the vectors in registers in the same vector element order, either little-endian or big-endian element order.

Conventions

Typographical conventions

The following table shows the typographical conventions used in the IBM XL C/C++ for Linux, V13.1.6 information.

Table 1. *Typographical conventions*

Typeface	Indicates	Example
bold	Lowercase commands, executable names, compiler options, and directives.	The compiler provides basic invocation commands, xl c and xl C (xl c++), along with several other compiler invocation commands to support various C/C++ language levels and compilation environments.
<i>italics</i>	Parameters or variables whose actual names or values are to be supplied by the user. Italics are also used to introduce new terms.	Make sure that you update the <i>size</i> parameter if you return more than the <i>size</i> requested.
<u>underlining</u>	The default setting of a parameter of a compiler option or directive.	nomaf <u>maf</u>
monospace	Programming keywords and library functions, compiler builtins, examples of program code, command strings, or user-defined names.	To compile and optimize myprogram.c, enter: <code>xl c myprogram.c -O3</code> .

Qualifying elements (icons)

Most features described in this information apply to both C and C++ languages. In descriptions of language elements where a feature is exclusive to one language, or where functionality differs between languages, this information uses icons to delineate segments of text as follows:

Table 2. Qualifying elements

Qualifier/Icon	Meaning
C only begins   C only ends	The text describes a feature that is supported in the C language only; or describes behavior that is specific to the C language.
C++ only begins   C++ only ends	The text describes a feature that is supported in the C++ language only; or describes behavior that is specific to the C++ language.
C11 begins   C11 ends	The text describes a feature that is introduced into standard C as part of C11.
C++11 begins   C++11 ends	The text describes a feature that is introduced into standard C++ as part of C++11.
C++14 begins   C++14 ends	The text describes a feature that is introduced into standard C++ as part of C++14.
IBM extension begins   IBM extension ends	The text describes a feature that is an IBM extension to the standard language specifications.
GPU begins   GPU ends	The text describes the information that is relevant to offloading computations to the NVIDIA GPUs.

Syntax diagrams

Throughout this information, diagrams illustrate XL C/C++ syntax. This section helps you to interpret and use those diagrams.

- Read the syntax diagrams from left to right, from top to bottom, following the path of the line.

The  symbol indicates the beginning of a command, directive, or statement.

The  symbol indicates that the command, directive, or statement syntax is continued on the next line.

The \blacktriangleright — symbol indicates that a command, directive, or statement is continued from the previous line.

The — \blacktriangleleft symbol indicates the end of a command, directive, or statement.

Fragments, which are diagrams of syntactical units other than complete commands, directives, or statements, start with the \mid — symbol and end with the — \mid symbol.

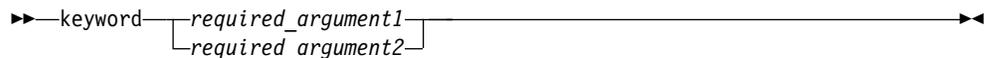
- Required items are shown on the horizontal line (the main path):



- Optional items are shown below the main path:



- If you can choose from two or more items, they are shown vertically, in a stack. If you *must* choose one of the items, one item of the stack is shown on the main path.



If choosing one of the items is optional, the entire stack is shown below the main path.



- An arrow returning to the left above the main line (a repeat arrow) indicates that you can make more than one choice from the stacked items or repeat an item. The separator character, if it is other than a blank, is also indicated:



- The item that is the default is shown above the main path.



- Keywords are shown in nonitalic letters and should be entered exactly as shown.
- Variables are shown in italicized lowercase letters. They represent user-supplied names or values.
- If punctuation marks, parentheses, arithmetic operators, or other such symbols are shown, you must enter them as part of the syntax.

Example of a syntax statement

EXAMPLE *char_constant* {*a|b*}[*c|d*]*e*[,*e*]*...* *name_list*{*name_list*}*...*

The following list explains the syntax statement:

- Enter the keyword EXAMPLE.

- Enter a value for *char_constant*.
- Enter a value for *a* or *b*, but not for both.
- Optionally, enter a value for *c* or *d*.
- Enter at least one value for *e*. If you enter more than one value, you must put a comma between each.
- Optionally, enter the value of at least one *name* for *name_list*. If you enter more than one value, you must put a comma between each *name*.

Note: The same example is used in both the syntax-statement and syntax-diagram representations.

Examples in this information

The examples in this information, except where otherwise noted, are coded in a simple style that does not try to conserve storage, check for errors, achieve fast performance, or demonstrate all possible methods to achieve a specific result.

The examples for installation information are labelled as either *Example* or *Basic example*. *Basic examples* are intended to document a procedure as it would be performed during a default installation; these need little or no modification.

Related information

The following sections provide related information for XL C/C++:

Available help information IBM XL C/C++ information

XL C/C++ provides product information in the following formats:

- Quick Start Guide

The Quick Start Guide (`quickstart.pdf`) is intended to get you started with IBM XL C/C++ for Linux, V13.1.6. It is located by default in the XL C/C++ directory and in the `\quickstart` directory of the installation DVD.

- README files

README files contain late-breaking information, including changes and corrections to the product information. README files are located by default in the XL C/C++ directory, and in the root directory and subdirectories of the installation DVD.

- Installable man pages

Man pages are provided for the compiler invocations and all command-line utilities provided with the product. Instructions for installing and accessing the man pages are provided in the *IBM XL C/C++ for Linux, V13.1.6 Installation Guide*.

- Online product documentation

The fully searchable HTML-based documentation is viewable in IBM Knowledge Center at http://www.ibm.com/support/knowledgecenter/SSXVZZ_13.1.6/com.ibm.compilers.linux.doc/welcome.html.

- PDF documents

PDF documents are available on the web at https://www.ibm.com/support/knowledgecenter/SSXVZZ_13.1.6/com.ibm.compilers.linux.doc/download_pdf.html.

The following files comprise the full set of XL C/C++ product information.

Note: To ensure that you can access cross-reference links to other XL C/C++ PDF documents, download and unzip the .zip file that contains all the product documentation files, or you can download each document into the same directory on your local machine.

Table 3. XL C/C++ PDF files

Document title	PDF file name	Description
<i>What's New for IBM XL C/C++ for Linux, V13.1.6, GC27-8035-00</i>	whats_new.pdf	Provides an executive overview of new functions in the IBM XL C/C++ for Linux, V13.1.6 compiler, with new functions categorized according to user benefits.
<i>Getting Started with IBM XL C/C++ for Linux, V13.1.6, GI13-2875-05</i>	getstart.pdf	Contains an introduction to XL C/C++, with information about setting up and configuring your environment, compiling and linking programs, and troubleshooting compilation errors.
<i>IBM XL C/C++ for Linux, V13.1.6 Installation Guide, GC27-6540-05</i>	install.pdf	Contains information for installing XL C/C++ and configuring your environment for basic compilation and program execution.
<i>IBM XL C/C++ for Linux, V13.1.6 Migration Guide, GC27-8036-00</i>	migrate.pdf	Contains migration considerations for using XL C/C++ to compile programs that were previously compiled on different platforms, by previous releases of XL C/C++, or by other compilers.
<i>IBM XL C/C++ for Linux, V13.1.6 Compiler Reference, SC27-6570-05</i>	compiler.pdf	Contains information about the various compiler options, pragmas, macros, environment variables, and built-in functions.
<i>IBM XL C/C++ for Linux, V13.1.6 Language Reference, SC27-6550-05</i>	langref.pdf	Contains information about language extensions for portability and conformance to nonproprietary standards.
<i>IBM XL C/C++ for Linux, V13.1.6 Optimization and Programming Guide, SC27-6560-05</i>	proguide.pdf	Contains information about advanced programming topics, such as application porting, interlanguage calls with Fortran code, library development, application optimization, and the XL C/C++ high-performance libraries.

To read a PDF file, use Adobe Reader. If you do not have Adobe Reader, you can download it (subject to license terms) from the Adobe website at <http://www.adobe.com>.

More information related to XL C/C++, including IBM Redbooks® publications, white papers, and other articles, is available on the web at <http://www.ibm.com/support/docview.wss?uid=swg27036675>.

For more information about the compiler, see the XL compiler on Power® community at <http://ibm.biz/xl-power-compilers>.

Other IBM information

- *ESSL product documentation* available at http://www.ibm.com/support/knowledgecenter/SSFHY8/essl_welcome.html?lang=en

Other information

- *Using the GNU Compiler Collection* available at <http://gcc.gnu.org/onlinedocs>

Standards and specifications

XL C/C++ is designed to support the following standards and specifications. You can refer to these standards and specifications for precise definitions of some of the features found in this information.

- *Information Technology - Programming languages - C, ISO/IEC 9899:1990*, also known as C89.
- *Information Technology - Programming languages - C, ISO/IEC 9899:1999*, also known as C99.
- *Information Technology - Programming languages - C, ISO/IEC 9899:2011*, also known as C11.
- *Information Technology - Programming languages - C++, ISO/IEC 14882:1998*, also known as C++98.
- *Information Technology - Programming languages - C++, ISO/IEC 14882:2003*, also known as C++03.
- *Information Technology - Programming languages - C++, ISO/IEC 14882:2011*, also known as C++11.
- *Information Technology - Programming languages - C++, ISO/IEC 14882:2014*, also known as C++14 (Partial support).
- *AltiVec Technology Programming Interface Manual*, Motorola Inc. This specification for vector data types, to support vector processing technology, is available at http://www.freescale.com/files/32bit/doc/ref_manual/ALTIVECPIM.pdf.
- *ANSI/IEEE Standard for Binary Floating-Point Arithmetic, ANSI/IEEE Std 754-1985*.
- *OpenMP Application Program Interface Version 3.1 (full support), OpenMP Application Program Interface Version 4.0 (partial support), and OpenMP Application Program Interface Version 4.5 (partial support)*, available at <http://www.openmp.org>

Other IBM information

- *ESSL product documentation* available at http://www.ibm.com/support/knowledgecenter/SSFHY8/essl_welcome.html?lang=en

Other information

- *Using the GNU Compiler Collection* available at <http://gcc.gnu.org/onlinedocs>

Technical support

Additional technical support is available from the XL C/C++ Support page at https://www.ibm.com/support/home/product/Q833644Y89702U61/XL_C/C++_for_Linux. This page provides a portal with search capabilities to a large selection of Technotes and other support information.

If you cannot find what you need, you can send an email to compinfo@cn.ibm.com.

For the latest information about XL C/C++, visit the product information site at <https://www.ibm.com/us-en/marketplace/xl-cpp-linux-compiler-power>.

How to send your comments

Your feedback is important in helping us to provide accurate and high-quality information. If you have any comments about this information or any other XL C/C++ information, send your comments to compinfo@cn.ibm.com.

Be sure to include the name of the manual, the part number of the manual, the version of XL C/C++, and, if applicable, the specific location of the text you are commenting on (for example, a page number or table number).

Chapter 1. Using XL C/C++ with Fortran

With XL C/C++, you can call functions written in Fortran from your C and C++ programs. This section discusses some programming considerations for calling Fortran code in the following areas:

- “Identifiers”
- “Corresponding data types” on page 2
- “Character and aggregate data” on page 4
- “Function calls and parameter passing” on page 4
- “Pointers to functions” on page 5

In topic “Sample program: C/C++ calling Fortran” on page 5, an example of a C program that calls a Fortran subroutine is provided.

For more information about language interoperability, see the information about the BIND attribute and the interoperability of procedures in the *XL Fortran Language Reference*.

Related information:



Calling Fortran code

Identifiers

C++ functions callable from Fortran should be declared with extern "C" to avoid name mangling. For details, see the appropriate section about options and conventions for mixing Fortran with C/C++ code in the *XL Fortran Optimization and Programming Guide*.

You need to follow these recommendations when writing C and C++ code to call functions that are written in Fortran:

- Avoid using uppercase letters in identifiers. Although XL Fortran folds external identifiers to lowercase by default, the Fortran compiler can be set to distinguish external names by case.
- Avoid using long identifier names. The maximum number of significant characters in XL Fortran identifiers is 250¹.

Note:

1. The Fortran 90 and 95 language standards require identifiers to be no more than 31 characters; the Fortran 2003 and the Fortran 2008 standards require identifiers to be no more than 63 characters.

Corresponding data types

The following table shows the correspondence between the data types available in C/C++ and Fortran. Several data types in C have no equivalent representation in Fortran. Do not use them when you program for interlanguage calls.

Table 4. Correspondence of data types between C/C++ and Fortran

C and C++ data types	Fortran data types	
	Types	Types with kind type parameters from the ISO_C_BINDING module
bool (C++) _Bool (C)	LOGICAL*1 or LOGICAL(1)	LOGICAL(C_BOOL)
char	CHARACTER	CHARACTER(C_CHAR)
signed char	INTEGER*1 or INTEGER(1)	INTEGER(C_SIGNED_CHAR)
unsigned char	LOGICAL*1 or LOGICAL(1)	
signed short int	INTEGER*2 or INTEGER(2)	INTEGER(C_SHORT)
unsigned short int	LOGICAL*2 or LOGICAL(2)	
int	INTEGER*4 or INTEGER(4)	INTEGER(C_INT)
unsigned int	LOGICAL*4 or LOGICAL(4)	
signed long int	INTEGER*8 or INTEGER(8)	INTEGER(C_LONG)
unsigned long int	INTEGER*8 or INTEGER(8)	
signed long long int	INTEGER*8 or INTEGER(8)	INTEGER(C_LONG_LONG)
unsigned long long int	LOGICAL*8 or LOGICAL(8)	
size_t	INTEGER*8 or INTEGER(8)	INTEGER(C_SIZE_T)
intptr_t	INTEGER*8 or INTEGER(8)	INTEGER(C_INTPTR_T)
intmax_t	INTEGER*8 or INTEGER(8)	INTEGER(C_INTMAX_T)
int8_t	INTEGER*1 or INTEGER(1)	INTEGER(C_INT8_T)
int16_t	INTEGER*2 or INTEGER(2)	INTEGER(C_INT16_T)
int32_t	INTEGER*4 or INTEGER(4)	INTEGER(C_INT32_T)
int64_t	INTEGER*8 or INTEGER(8)	INTEGER(C_INT64_T)
int_least8_t	INTEGER*1 or INTEGER(1)	INTEGER(C_INT_LEAST8_T)

Table 4. Correspondence of data types between C/C++ and Fortran (continued)

C and C++ data types	Fortran data types	
	Types	Types with kind type parameters from the ISO_C_BINDING module
int_least16_t	INTEGER*2 or INTEGER(2)	INTEGER(C_INT_LEAST16_T)
int_least32_t	INTEGER*4 or INTEGER(4)	INTEGER(C_INT_LEAST32_T)
int_least64_t	INTEGER*8 or INTEGER(8)	INTEGER(C_INT_LEAST64_T)
int_fast8_t	INTEGER, INTEGER*4, or INTEGER(4)	INTEGER(C_INT_FAST8_T)
int_fast16_t	INTEGER*4 or INTEGER(4)	INTEGER(C_INT_FAST16_T)
int_fast32_t	INTEGER*4 or INTEGER(4)	INTEGER(C_INT_FAST32_T)
int_fast64_t	INTEGER*8 or INTEGER(8)	INTEGER(C_INT_FAST64_T)
float	REAL, REAL*4, or REAL(4)	REAL(C_FLOAT)
double	REAL*8, REAL(8), or DOUBLE PRECISION	REAL(C_DOUBLE)
long double	REAL*16 or REAL(16)	REAL(C_LONG_DOUBLE)
float _Complex	COMPLEX*4, COMPLEX(4), COMPLEX*8, or COMPLEX(8)	COMPLEX(C_FLOAT_COMPLEX)
double _Complex	COMPLEX*8, COMPLEX(8), COMPLEX*16, or COMPLEX(16)	COMPLEX(C_DOUBLE_COMPLEX)
long double _Complex	COMPLEX*16, COMPLEX(16), COMPLEX*32, or COMPLEX(32)	COMPLEX(C_LONG_DOUBLE_COMPLEX)
struct or union	derived type	
enum	INTEGER*4 or INTEGER(4)	
char[n]	CHARACTER*n or CHARACTER(n)	
array pointer to type, or type []	Dimensioned variable (transposed)	
pointer to function	Functional parameter	
struct with -fpack-struct (-qalign)	Sequence derived type	

Related information in the *XL C/C++ Compiler Reference*

 -fpack-struct (-qalign)

Character and aggregate data

Most numeric data types have counterparts across C/C++ and Fortran. However, character and aggregate data types require special treatment:

- C character strings are delimited by a '\0' character. In Fortran, all character variables and expressions have a length that is determined at compile time. Whenever Fortran passes a string argument to another routine, it appends a hidden argument that provides the length of the string argument. This length argument must be explicitly declared in C. The C code should not assume a null terminator; the supplied or declared length should always be used.
- An n-element C/C++ array is indexed with 0...n-1, whereas an n-element Fortran array is typically indexed with 1...n. In addition, Fortran supports user-specified bounds while C/C++ does not.
- C stores array elements in row-major order (array elements in the same row occupy adjacent memory locations). Fortran stores array elements in ascending storage units in column-major order (array elements in the same column occupy adjacent memory locations). The following table shows how a two-dimensional array declared by A[3][2] in C and by A(3,2) in Fortran, is stored:

Table 5. Storage of a two-dimensional array

Storage unit	C and C++ element name	Fortran element name
Lowest	A[0][0]	A(1,1)
	A[0][1]	A(2,1)
	A[1][0]	A(3,1)
	A[1][1]	A(1,2)
	A[2][0]	A(2,2)
Highest	A[2][1]	A(3,2)

- In general, for a multidimensional array, if you list the elements of the array in the order they are laid out in memory, a row-major array will be such that the rightmost index varies fastest, while a column-major array will be such that the leftmost index varies fastest.

Function calls and parameter passing

Functions must be prototyped equivalently in both C/C++ and Fortran.

In C and C++, by default, all function arguments are passed by value, and the called function receives a copy of the value passed to it. In Fortran, by default, arguments are passed by reference, and the called function receives the address of the value passed to it. You can use the Fortran %VAL built-in function or the VALUE attribute to pass by value. Refer to the *XL Fortran Language Reference* for more information.

For call-by-reference (as in Fortran), the address of the parameter is passed in a register. When passing parameters by reference, if you write C or C++ functions that call a program written in Fortran, all arguments must be pointers, or scalars with the address operator.

For more information about interlanguage calls to functions or routines, see the information about interlanguage calls in the *XL Fortran Optimization and Programming Guide*.

Pointers to functions

A function pointer is a data type whose value is a function address. In Fortran, a dummy argument that appears in an EXTERNAL statement is a function pointer. Starting from the Fortran 2003 standard, Fortran variables of type C_FUNPTR are interoperable with function pointers. Function pointers are supported in contexts such as the target of a call statement or an actual argument of such a statement.

Sample program: C/C++ calling Fortran

The following example illustrates how program units written in different languages can be combined to create a single program. It also demonstrates parameter passing between C/C++ and Fortran subroutines with different data types as arguments. The example includes the following source files:

- The main program source file: `example.c`
- The Fortran add function source file: `add.f`

Main program source file: `example.c`

```
#include <stdio.h>
extern double add(int *, double [], int *, double []);

double ar1[4]={1.0, 2.0, 3.0, 4.0};
double ar2[4]={5.0, 6.0, 7.0, 8.0};

main()
{
  int x, y;
  double z;

  x = 3;
  y = 3;

  z = add(&x, ar1, &y, ar2); /* Call Fortran add routine */
  /* Note: Fortran indexes arrays 1..n */
  /* C indexes arrays 0..(n-1) */

  printf("The sum of %1.0f and %1.0f is %2.0f \n",
    ar1[x-1], ar2[y-1], z);
}
```

Fortran add function source file: `add.f`

```
REAL*8 FUNCTION ADD (A, B, C, D)
REAL*8 B,D
INTEGER*4 A,C
DIMENSION B(4), D(4)
ADD = B(A) + D(C)
RETURN
END
```

Compile the main program and Fortran add function source files as follows:

```
xlc -c example.c
xlf -c add.f
```

Link the object files from compile step to create executable `add`:

```
xlc -o add example.o add.o
```

Execute binary:

```
./add
```

The output is as follows:
The sum of 3 and 7 is 10

Chapter 2. Aligning data

XL C/C++ provides many mechanisms for specifying data alignment at the levels of individual variables, members of aggregates, entire aggregates, and entire compilation units. If you are porting applications between different platforms, or between 32-bit and 64-bit modes, you need to take into account the differences between alignment settings available in different environments, to prevent possible data corruption and deterioration in performance. In particular, vector types have special alignment requirements which, if not followed, can produce incorrect results. For more information, see the *AltiVec Technology Programming Interface Manual*.

XL C/C++ provides alignment modes and alignment modifiers for specifying data alignment. Using alignment modes, you can set alignment defaults for all data types for a compilation unit or subsection of a compilation unit by specifying a predefined suboption.

Using alignment modifiers, you can set the alignment for specific variables or data types within a compilation unit by specifying the exact number of bytes that should be used for the alignment.

“Using alignment modes” discusses the default alignment modes for all data types on different platforms and addressing models, the suboptions and pragmas that you can use to change or override the defaults, and rules for the alignment modes for simple variables, aggregates, and bit fields.

“Using alignment modifiers” on page 10 discusses the different specifiers, pragmas, and attributes you can use in your source code to override the alignment mode currently in effect, for specific variable declarations. It also provides the rules that govern the precedence of alignment modes and modifiers during compilation.

Related information in the XL C/C++ Compiler Reference

 -maltivec

Related external information

 AltiVec Technology Programming Interface Manual, available at http://www.freescale.com/files/32bit/doc/ref_manual/ALTIVECPIM.pdf

Using alignment modes

Each data type that is supported by XL C/C++ is aligned along byte boundaries according to platform-specific default alignment modes. The default alignment mode is **linuxppc**.

Each of the valid alignment modes is defined in Table 6, which provides the alignment value, in bytes, for scalar variables of all data types.

Table 6. Alignment settings (values given in bytes)

Data type	Storage	Alignment setting	
		linuxppc	bit_packed
_Bool (C), bool (C++)	1	1	1

Table 6. Alignment settings (values given in bytes) (continued)

Data type	Storage	Alignment setting	
		linuxppc	bit_packed
char, signed char, unsigned char	1	1	1
wchar_t	4	4	1
int, unsigned int	4	4	1
short int, unsigned short int	2	2	1
long int, unsigned long int	8	8	1
long long	8	8	1
float	4	4	1
double	8	8	1
long double	16	16	1
pointer	8	8	1
vector types	16	16	1

If you generate data with an application on one platform and read the data with an application on another platform, it is recommended that you use the **bit_packed** mode, which results in equivalent data alignment on all platforms.

Notes:

- Vectors in a bit-packed structure might not be correctly aligned unless you take extra action to ensure their alignment.
- Vectors might suffer from alignment issues if they are accessed through heap-allocated storage or through pointer arithmetic. For example, `double my_array[1000] __attribute__((aligned(16)))` is 16-byte aligned while `my_array[1]` is not. How `my_array[i]` is aligned is determined by the value of `i`.

“Alignment of aggregates” discusses the rules for the alignment of entire aggregates and provides examples of aggregate layouts. “Alignment of bit-fields” on page 9 discusses additional rules and considerations for the use and alignment of bit fields and provides an example of bit-packed alignment.

Related information in the XL C/C++ Compiler Reference

 `-fpack-struct (-qalign)`

Alignment of aggregates

The data contained in Table 6 on page 7 in “Using alignment modes” on page 7 apply to scalar variables, and variables that are members of aggregates such as structures, unions, and classes. The following rules apply to aggregate variables, namely structures, unions or classes, as a whole (in the absence of any modifiers):

- For all alignment modes, the size of an aggregate is the smallest multiple of its alignment value that can encompass all of the members of the aggregate.
-  Empty aggregates are assigned a size of zero bytes. As a result, two distinct variables might have the same address.
-  Empty aggregates are assigned a size of one byte. Note that static data members do not participate in the alignment or size of an aggregate; therefore, a structure or class containing only a single static data member has a size of one byte.

- For all alignment modes, the alignment of an aggregate is equal to the largest alignment value of any of its members. With the exception of packed alignment modes, members whose natural alignment is smaller than that of their aggregate's alignment are padded with empty bytes.
- Aligned aggregates can be nested, and the alignment rules applicable to each nested aggregate are determined by the alignment mode that is in effect when a nested aggregate is declared.

Notes:

- **C++** The C++ compiler might generate extra fields for classes that contain base classes or virtual functions. Objects of these types might not conform to the usual mappings for aggregates.
- The alignment of an aggregate must be the same in all compilation units. For example, if the declaration of an aggregate is in a header file and you include that header file into two distinct compilation units, choose the same alignment mode for both compilation units.

For rules on the alignment of aggregates containing bit fields, see “Alignment of bit-fields.”

Alignment of bit-fields

You can declare a bit-field as a **C** `_Bool`, **C++** `bool`, `char`, signed `char`, unsigned `char`, `short`, unsigned `short`, **C++** `int`, unsigned `int`, `long`, unsigned `long`, **C++** `long long`, or unsigned `long long` **C++** data type. The alignment of a bit-field depends on its base type and the compilation mode.

C The length of a bit-field cannot exceed the length of its base type. In extended mode, you can use the `sizeof` operator on a bit-field. The `sizeof` operator on a bit-field returns the size of the base type. **C**

C++ The length of a bit-field can exceed the length of its base type, but the remaining bits are used to pad the field and do not actually store any value.

C++

However, alignment rules for aggregates containing bit-fields are different depending on the alignment mode in effect. These rules are described below.

Rules for Linux PowerPC alignment

- Bit-fields are allocated from a bit-field container. The size of this container is determined by the declared type of the bit-field. For example, a `char` bit-field uses an 8-bit container, and an `int` bit-field uses 31 bits. The container must be large enough to contain the bit-field because the bit-field will not be split across containers.
- Containers are aligned in the aggregate as if they start on a natural boundary for that type of container. Bit-fields are not necessarily allocated at the start of the container.
- If a zero-length bit-field is the first member of an aggregate, it has no effect on the alignment of the aggregate and is overlapped by the next data member. If a zero-length bit-field is a non-first member of the aggregate, it pads to the next alignment boundary determined by its base declared type but does not affect the alignment of the aggregate.
- Unnamed bit-fields do not affect the alignment of the aggregate.

Rules for bit-packed alignment

- Bit-fields have an alignment of one byte and are packed with no default padding between bit-fields.
- A zero-length bit-field causes the next member to start at the next byte boundary. If the zero-length bit-field is already at a byte boundary, the next member starts at this boundary. A non-bit-field member that follows a bit-field is aligned on the next byte boundary.

Using alignment modifiers

XL C/C++ also provides alignment modifiers, with which you can exercise even finer-grained control over alignment, at the level of declaration or definition of individual variables or aggregate members. Available modifiers are as follows:

#pragma pack(...)

Valid application

The entire aggregate (as a whole) immediately following the directive.

Effect Sets the maximum alignment of the members of the aggregate to which it applies, to a specific number of bytes. Also allows a bit-field to cross a container boundary. Used to reduce the effective alignment of the selected aggregate.

Valid values

number

Is 1, 2, 4, 8, or 16. That is, structure members are aligned on *number*-byte boundaries or on their natural alignment boundary, whichever is less.

push When specified without a *number*, pushes whatever value is currently in effect to the top of the packing "stack". When used with a *number*, pushes that value to the top of the packing stack, and sets the packing value to that of *number* for structures that follow.

pop Removes the previous value added with #pragma pack.

empty brackets

Has the same functionality as pop.

__attribute__((aligned(*n*)))

Valid application

As a variable attribute, it applies to a single aggregate (as a whole), namely a structure, union, or class, or it applies to an individual member of an aggregate.¹ As a type attribute, it applies to all aggregates declared of that type. If it is applied to a typedef declaration, it applies to all instances of that type.²

Effect Sets the minimum alignment of the specified variable or variables to a specific number of bytes. Typically used to increase the effective alignment of the selected variables.

Valid values

n must be a positive power of two, or NIL. NIL can be specified as either `__attribute__((aligned()))` or `__attribute__((aligned))`; this is the same as specifying the maximum system alignment (16 bytes on all UNIX platforms).

`__attribute__((packed))`

Valid application

As a variable attribute, it applies to simple variables or individual members of an aggregate, namely a structure or class¹. As a type attribute, it applies to all members of all aggregates declared of that type.

Effect Sets the maximum alignment of the selected variable or variables, to which it applies, to the smallest possible alignment value, namely one byte for a variable and one bit for a bit field.

Notes:

1. In a comma-separated list of variables in a declaration, if the modifier is placed at the beginning of the declaration, it applies to all the variables in the declaration. Otherwise, it applies only to the variable immediately preceding it.
2. Depending on the placement of the modifier in the declaration of a struct, it can apply to the definition of the type, and hence applies to all instances of that type; or it can apply to only a single instance of the type. For details, see the information about type attributes in the *XL C/C++ Language Reference* and the C  and C++  language standards.

Related information in the *XL C/C++ Compiler Reference*

 #pragma pack

Related information in the *XL C/C++ Language Reference*

 The aligned type attribute (IBM extension)

 The packed type attribute (IBM extension)

 Type attributes (IBM extension)

 The aligned variable attribute (IBM extension)

 The packed variable attribute (IBM extension)

Chapter 3. Handling floating-point operations

The following sections provide reference information, portability considerations, and suggested procedures for using compiler options to manage floating-point operations:

- “Floating-point formats”
- “Handling multiply-and-add operations”
- “Compiling for strict IEEE conformance” on page 14
- “Handling floating-point constant folding and rounding” on page 14
- “Handling floating-point exceptions” on page 16

Floating-point formats

XL C/C++ supports the following binary floating-point formats:

- 32-bit single precision, with an approximate absolute normalized range of 0 and 10^{-38} to 10^{38} and with a precision of about 7 decimal digits
- 64-bit double precision, with an approximate absolute normalized range of 0 and 10^{-308} to 10^{308} and with a precision of about 16 decimal digits
- 128-bit extended precision, with slightly greater range than double-precision values, and with a precision of about 32 decimal digits

The 128-bit extended precision format of XL C/C++ is different from the **binary128** formats that are suggested by the IEEE standard. The IEEE standard suggests that extended formats use more bits in the exponent for greater range and the fraction for higher precision.

It is possible that special numbers, such as NaN, infinity, and negative zero, cannot be represented by the 128-bit extended precision values. Arithmetic operations do not necessarily propagate these numbers in extended precision.

Handling multiply-and-add operations

By default, the compiler generates a single non-IEEE 754 compatible multiply-and-add instruction for binary floating-point expressions, such as $a + b * c$, partly because one instruction is faster than two. Because no rounding occurs between the multiply and add operations, this might also produce a more precise result. However, the increased precision might lead to different results from those obtained in other environments, and might cause $x*y-x*y$ to produce a nonzero result. To avoid these issues, you can suppress the generation of multiply-add instructions by using the **-qfloat=nomaf** option.

Related information in the XL C/C++ Compiler Reference

 -qfloat

Compiling for strict IEEE conformance

By default, XL C/C++ follows most but not all of the rules in the IEEE standard. If you compile with the **-qnostrict** option, which is enabled by default at optimization level **-O3** or higher, some IEEE floating-point rules are violated in ways that can improve performance but might affect program correctness. To avoid this issue and to compile for strict compliance with the IEEE standard, use the following options:

- Use the **-qfloat=nomaf** compiler option.
- If the program changes the rounding mode at run time, use the **-qfloat=rrm** option.
- If the data or program code contains signaling NaN values (NaNs), use any of the following groups of options. (A signaling NaN is different from a quiet NaN; you must explicitly code it into the program or data, or create it by using the **-qinitauto** compiler option.)
 - The **-qfloat=nans** and **-qstrict=nans** options
 - The **-qfloat=nans** and **-qstrict** options
- If you compile with **-O3**, **-O4**, or **-O5**, include the option **-qstrict** after it. You can also use the suboptions of **-qstrict** to refine the level of control for the transformations performed by the optimizers.

Related information:

“Advanced optimization” on page 24

Related information in the XL C/C++ Compiler Reference

-  **-qfloat**
-  **-qstrict**
-  **-qinitauto**

Handling floating-point constant folding and rounding

By default, the compiler replaces most operations involving constant operands with their result at compile time. This process is known as constant folding. Additional folding opportunities might occur with optimization or with the **-qnostrict** option. The result of a floating-point operation folded at compile time normally produces the same result as that obtained at execution time, except in the following cases:

- The compile-time rounding mode is different from the execution-time rounding mode. By default, both are round-to-nearest; however, if your program changes the execution-time rounding mode, to avoid differing results, perform either of the following operations:
 - Change the compile-time rounding mode to match the execution-time mode, by compiling with the appropriate **-y** option. For more information and an example, see “Matching compile-time and runtime rounding modes” on page 15.
 - Suppress folding by compiling with the **-qfloat=nofold** option.
- Expressions like $a+b*c$ are partially or fully evaluated at compile time. The results might be different from those produced at execution time, because $b*c$ might be rounded before being added to a , while the runtime multiply-add instruction does not use any intermediate rounding. To avoid differing results, perform either of the following operations:

- Suppress the use of multiply-add instructions by compiling with the **-qfloat=nomaf** option.
- Suppress folding by compiling with the **-qfloat=nofold** option.
- An operation produces an infinite, NaN, or underflow to zero result. Compile-time folding prevents execution-time detection of an exception, even if you compile with the **-ftrapping-math (-qflttrap)** option. To avoid missing these exceptions, suppress folding with the **-qfloat=nofold** option.

Related information:

“Handling floating-point exceptions” on page 16

Related information in the *XL C/C++ Compiler Reference*

 -qfloat

 -qstrict

 -ftrapping-math (-qflttrap)

Matching compile-time and runtime rounding modes

The default rounding mode used at compile time and run time is round-to-nearest, ties to even. If your program changes the rounding mode at run time, the results of a floating-point calculation might be slightly different from those that are obtained at compile time. The following example illustrates this:

```
#include <float.h>
#include <fenv.h>
#include <stdio.h>

int main ( )
{
    volatile double one = 1.f, three = 3.f; /* volatiles are not folded */
    double one_third;

    one_third = 1. / 3.; /* folded */
    printf ("1/3 with compile-time rounding = %.17f\n", one_third);

    fesetround (FE_TOWARDZERO);
    one_third = one / three; /* not folded */

    printf ("1/3 with execution-time rounding to zero = %.17f\n", one_third);

    fesetround (FE_TONEAREST);
    one_third = one / three; /* not folded */

    printf ("1/3 with execution-time rounding to nearest = %.17f\n", one_third);

    fesetround (FE_UPWARD);
    one_third = one / three; /* not folded */

    printf ("1/3 with execution-time rounding to +infinity = %.17f\n", one_third);

    fesetround (FE_DOWNWARD);
    one_third = one / three; /* not folded */

    printf ("1/3 with execution-time rounding to -infinity = %.17f\n", one_third);

    return 0;
}
```

When compiled with the default options, this code produces the following results:

```
1/3 with compile-time rounding = 0.3333333333333331
1/3 with execution-time rounding to zero = 0.3333333333333331
1/3 with execution-time rounding to nearest = 0.3333333333333331
1/3 with execution-time rounding to +infinity = 0.3333333333333337
1/3 with execution-time rounding to -infinity = 0.3333333333333331
```

Because the fourth computation changes the rounding mode to round-to-infinity, the results are slightly different from the first computation, which is performed at compile time, using round-to-nearest. If you do not use the **-qfloat=nofold** option to suppress all compile-time folding of floating-point computations, it is recommended that you use the **-y** compiler option with the appropriate suboption to match compile-time and runtime rounding modes. In the previous example, compiling with **-yp** (round-to-infinity) produces the following result for the first computation:

```
1/3 with compile-time rounding = 0.3333333333333337
```

In general, if the rounding mode is changed to rounding to +infinity, -infinity, or zero, it is recommended that you also use the **-qfloat=rrm** option.

Related information in the XL C/C++ Compiler Reference

 **-qfloat**

 **-y**

Handling floating-point exceptions

By default, invalid operations such as division by zero, division by infinity, overflow, and underflow are ignored at run time. However, you can use the **-ftrapping-math (-qfltttrap)** option or call C or operating system functions to detect these types of exceptions. If you enable floating-point traps without using the **-ftrapping-math (-qfltttrap)** option, use the **-qfloat=fenv** option.

In addition, you can add suitable support code to your program to make program execution continue after an exception occurs and to modify the results of operations causing exceptions.

Because, however, floating-point computations involving constants are usually folded at compile time, the potential exceptions that would be produced at run time might not occur. To ensure that the **-ftrapping-math (-qfltttrap)** option traps all runtime floating-point exceptions, you can use the **-qfloat=nofold** option to suppress all compile-time folding.

Related information in the XL C/C++ Compiler Reference

 **-qfloat**

 **-ftrapping-math (-qfltttrap)**

Chapter 4. Constructing a library

You can include static and shared libraries in your C and C++ applications.

“Compiling and linking a library” describes how to compile your source files into object files for inclusion in a library, how to link a library into the main program, and how to link one library into another.

“Initializing static objects in libraries (C++)” on page 18 describes how to use priorities to control the order of initialization of objects across multiple files in a C++ application.

Compiling and linking a library

This section describes how to compile your source files into object files for inclusion in a library, how to link a library into the main program, and how to link one library into another.

Related information in the *Getting Started with XL C/C++*

 Dynamic and static linking

Compiling a static library

To compile a static library, follow this procedure:

1. Compile each source file to get an object file. For example:

```
xlc -c test.c example.c
```
2. Use the ar command to add the generated object files to an archive library file. For example:

```
ar -rv libex.a test.o example.o
```

Compiling a shared library

To compile a shared library, follow this procedure:

1. Compile your source files to get an object file. Note that in the case of compiling a shared library, you must use the **-fPIC (-qpPic)** compiler option. For example:

```
xlc -fPIC -c foo.c
```
2. Use the **-shared (-qmKshroBj)** compiler option to create a shared object from the generated object files. For example:

```
xlc -shared -o libfoo.so foo.o
```

Related information in the *XL C/C++ Compiler Reference*

 -fPIC (-qpPic)

 -shared (-qmKshroBj)

Linking a library to an application

You can use the following command string to link a static or shared library to your main program. For example:

```
xlc -o myprogram main.c -Ldirectory1:directory2 [-Rdirectory] -ltest
```

At compile time, you instruct the linker to search for `libtest.so` in the first directory specified by the `-L` option. If `libtest.so` is not found, the linker searches for `libtest.a`. If neither file is found, the search continues with the next directory specified by the `-L` option.

At run time, the runtime linker searches for `libtest.so` in the first directory specified by the `-R` option. If `libtest.so` is not found, the search continues with the next directory specified by the `-R` option. The path specified by the `-R` option can be overridden at run time by the `LD_LIBRARY_PATH` environment variable.

For additional linkage options, including options that modify the default behavior, see the operating system `ld` documentation .

Related information in the *XL C/C++ Compiler Reference*

 `-l`

 `-L`

 `-R`

Linking a shared library to another shared library

Just as you link modules into an application, you can create dependencies between shared libraries by linking them together. For example:

```
xlc -shared -o mylib.so myfile.o -ldirectory -Rdirectory -ltest
```

Related information in the *XL C/C++ Compiler Reference*

 `-shared` (`-qmkshrobj`)

 `-R`

 `-L`

Initializing static objects in libraries (C++)

The C++ language definition specifies that all non-local objects with constructors from all the files included in the program must be properly constructed before the `main` function in a C++ program is executed.

You can assign a priority level number to objects and files within a single library using the following approaches. The objects will be initialized at run time according to the order of priority level. In addition, because modules are loaded and objects are initialized differently on different platforms, you can choose an approach that fits the platform better.

Set the priority level for an entire file

To use this approach, specify the `-qpriority` compiler option during compilation. By default, all objects within a single file are assigned the same priority level; they are initialized in the order in which they are declared, and they are terminated in reverse declaration order.

Set the priority level for individual objects

To use this approach, use `init_priority` variable attributes in the source files. The `init_priority` attribute can be applied to objects in any declaration order. On Linux, the objects are initialized according to their priority and terminated in reverse priority across compilation units.

Priority numbers can range from 101 to 65535. The smallest priority number that you can specify, 101, is initialized first. The largest priority number, 65535, is initialized last. If you do not specify a priority level, the default priority is 65535.

Related information in the *XL C/C++ Compiler Reference*

 -qpriority

 -shared (-qmkshrobj)

Related information in the *XL C/C++ Language Reference*

 The `init_priority` variable attribute

Chapter 5. Optimizing your applications

The XL compilers enable development of high performance applications by offering a comprehensive set of performance enhancing techniques that exploit the multilayered PowerPC® architecture. These performance advantages depend on good programming techniques, thorough testing and debugging, followed by optimization and tuning.

Distinguishing between optimization and tuning

You can use optimization and tuning separately or in combination to increase the performance of your application. Understanding the difference between them is the first step in understanding how the different levels, settings, and techniques can increase performance.

Optimization

Optimization is a compiler-driven process that searches for opportunities to restructure your source code and give your application better overall performance at run time, without significantly impacting development time. The XL compiler optimization suite, which you control using compiler options and directives, performs best on well-written source code that has already been through a thorough debugging and testing process. These optimization transformations can bring the following benefits:

- Reduce the number of instructions that your application executes to perform critical operations.
- Restructure your object code to make optimal use of the PowerPC architecture.
- Improve memory subsystem usage.

Each basic optimization technique can result in a performance benefit, although not all optimizations can benefit all applications. Consult the “Steps in the optimization process” on page 22 for an overview of the common sequence of steps that you can use to increase the performance of your application.

Tuning

Tuning is a user-driven process where you experiment with changes, for example to source code or compiler options, to make the compiler better optimize your program. While optimization applies general transformations designed to improve the performance of any application in any supported environment, tuning offers you opportunities to adjust specific characteristics or target execution environments of your application to improve its performance. Even at low optimization levels, tuning for your application and target architecture can have a positive impact on performance. With proper tuning, the compiler can make the following improvements:

- Select more efficient machine instructions.
- Generate instruction sequences that are more relevant to your application.
- Select from more focussed optimizations to improve your code.

For instructions, see “Tuning for your system architecture” on page 28.

Steps in the optimization process

When you begin the optimization process, consider that not all optimization techniques suit all applications. Trade-offs sometimes occur between an increase in compile time, a reduction in debugging capability, and the improvements that optimization can provide.

Learning about and experimenting with different optimization techniques can help you strike the right balance for your XL compiler applications while achieving the best possible performance. Also, though it is unnecessary to hand-optimize your code, compiler-friendly programming can be extremely beneficial to the optimization process. Unusual constructs can obscure the characteristics of your application and make performance optimization difficult. Use the steps in this section as a guide for optimizing your application.

1. The Basic optimization step begins your optimization processes at levels 0 and 2.
2. The Advanced optimization step exposes your application to more intense optimizations at levels 3, 4, and 5.
3. The Using high-order loop analysis and transformations step can help you limit loop execution time.
4. The Using interprocedural analysis step can optimize your entire application at once.
5. The Using profile-directed feedback step focuses optimizations on specific characteristics of your application.
6. The Debugging optimized code step can help you identify issues and problems that can occur with optimized code.

Basic optimization

The XL compiler supports several levels of optimization, with each option level building on the levels below through increasingly aggressive transformations and consequently using more machine resources.

Ensure that your application compiles and executes properly at low optimization levels before you try more aggressive optimizations. This topic discusses two optimizations levels, listed with complementary options in Table 7. The table also includes a column for compiler options that can have a performance benefit at that optimization level for some applications.

Table 7. Basic optimizations

Optimization level	Additional options implied by default	Complementary options	Other options with possible benefits
-O0	None	-mcpu	None
-O2	-qmaxmem=8192	-mcpu -mtune	-qmaxmem=-1 -qhot=level=0

Optimizing at level 0

Benefits at level 0

- Provides minimal performance improvement with minimal impact on machine resources
- Exposes some source code problems that can be helpful in the debugging process

Begin your optimization process at **-O0**, which the compiler already specifies by default. This level performs basic analytical optimization by removing obviously redundant code, and it can result in better compile time. It also ensures your code is algorithmically correct so you can move forward to more complex optimizations. **-O0** also includes some redundant instruction elimination and constant folding. The **-qfloat=nofold** option can be used to suppress folding floating-point operations. Optimizing at this level accurately preserves all debugging information and can expose problems in existing code, such as uninitialized variables and bad casting.

Additionally, specifying **-mcpu** at this level targets your application for a particular machine and can significantly improve performance by ensuring that your application takes advantage of all applicable architectural benefits.

Note: For SMP programs, you need to add an additional option **-qsmp=noopt**.

For more information about tuning, see “Tuning for your system architecture” on page 28.

Related information in the XL C/C++ Compiler Reference

 **-mcpu**

Optimizing at level 2

Benefits at level 2

- Eliminates redundant code
- Performs basic loop optimization
- Structures code to take advantage of **-mcpu** and **-mtune** settings

After you successfully compile, execute, and debug your application using **-O0**, recompiling at **-O2** opens your application to a set of comprehensive low-level transformations that apply to subprogram or compilation unit scopes and can include some inlining. Optimizations at **-O2** attain a relative balance between increasing performance while limiting the impact on compilation time and system resources. You can increase the memory available to some of the optimizations in the **-O2** portfolio by providing a larger value for the **-qmaxmem** option. Specifying **-qmaxmem=-1** allows the optimizer to use memory as needed without checking for limits but does not change the transformations the optimizer applies to your application at **-O2**.

 In C, compile with **-qlibansi** unless your application defines functions with names identical to those of library functions. If you encounter problems with **-O2**, consider using **-qalias=noansi** rather than turning off optimization.

Also, ensure that pointers in your C code follow these type restrictions:

- Generic pointers can be `char*` or `void*`.
- Mark all shared variables and pointers to shared variables `volatile`.



Starting to tune at O2

Choosing the right hardware architecture target or family of targets becomes even more important at **-O2** and higher. By targeting the proper hardware, the optimizer can make the best use of the available hardware facilities. If you choose a family of hardware targets, the **-mtune** option can direct the compiler to emit code that is consistent with the architecture choice and that can execute optimally on the

chosen tuning hardware target. With this option, you can compile for a general set of targets and have the code run best on a particular target.

For details on the `-mcpu` and `-mtune` options, see “Tuning for your system architecture” on page 28.

The `-O2` option can perform a number of additional optimizations as follows:

- Common subexpression elimination: Eliminates redundant instructions
- Constant propagation: Evaluates constant expressions at compile time
- Dead code elimination: Eliminates instructions that a particular control flow does not reach or that generate an unused result
- Dead store elimination: Eliminates unnecessary variable assignments
- Global register allocation: Globally assigns user variables to registers
- Value numbering: Simplifies algebraic expressions by eliminating redundant computations
- Instruction scheduling for the target machine
- Loop unrolling and software pipelining
- Moving loop-invariant code out of loops
- Simplifying control flow
- Strength reduction and effective use of addressing modes
- Widening: Merges adjacent load/stores and other operations
- Pointer aliasing improvements to enhance other optimizations

Even with `-O2` optimizations, some useful information about your source code is made available to the debugger if you specify `-g`. Using a higher `-g` level increases the information provided to the debugger but reduces the optimization that can be done. Conversely, higher optimization levels can transform code to an extent to which debugging information is no longer accurate.

Advanced optimization

Higher optimization levels can have a tremendous impact on performance, but some trade-offs can occur in terms of code size, compile time, resource requirements, and numeric or algorithmic precision.

After applying “Basic optimization” on page 22 and successfully compiling and executing your application, you can apply more powerful optimization tools. The XL compiler optimization portfolio includes many options for directing advanced optimization, and the transformations that your application undergoes are largely under your control. The discussion of each optimization level in Table 8 includes information on the performance benefits and the possible trade-offs and information on how you can help guide the optimizer to find the best solutions for your application.

Table 8. Advanced optimizations

Optimization Level	Additional options implied	Complementary options	Options with possible benefits
<code>-O3</code>	<code>-qnostrict</code> <code>-qmaxmem=-1</code> <code>-qhot=level=0</code>	<code>-mcpu</code> <code>-mtune</code>	<code>-qpdf</code>

Table 8. Advanced optimizations (continued)

Optimization Level	Additional options implied	Complementary options	Options with possible benefits
-04	-qnostrict -qmaxmem=-1 -qhot -qipa -qarch=auto -qtune=auto -qcache=auto	-mcpu -mtune -qcache	-qpdf -qsmp=auto
-05	All of -04 -qipa=level=2	-mcpu -mtune -qcache	-qpdf -qsmp=auto

When you compile programs with any of the following sets of options:

- **-qhot -qignerrno -qnostrict**
- **-03 -qhot**
- **-04**
- **-05**

the compiler automatically attempts to vectorize calls to system math functions by calling the equivalent vector functions in the Mathematical Acceleration Subsystem libraries (MASS), with the exceptions of functions `vdnint`, `vdint`, `vcosin`, `vscosin`, `vqdr`, `vsqdr`, `vrqdr`, `vsrqdr`, `vpopcnt4`, and `vpopcnt8`. If the compiler cannot vectorize, it automatically tries to call the equivalent MASS scalar functions. For automatic vectorization or scalarization, the compiler uses versions of the MASS functions contained in the system library `libxlopt.a`.

In addition to any of the preceding sets of options, when the **-qipa** option is in effect, if the compiler cannot vectorize, it tries to inline the MASS scalar functions before deciding to call them.

Optimizing at level 3

Benefits at level 3

- In-depth memory access analysis
- Better loop scheduling
- High-order loop analysis and transformations (**-qhot=level=0**)
- Inlining of small procedures within a compilation unit by default
- Eliminating implicit compile-time memory usage limits

Specifying **-03** initiates more intense low-level transformations that remove many of the limitations present at **-02**. For instance, the optimizer no longer checks for memory limits, by setting the default to **-qmaxmem=-1**. Additionally, optimizations encompass larger program regions and attempt more in-depth analysis. Although not all applications contain opportunities for the optimizer to provide a measurable increase in performance, most applications can benefit from this type of analysis.

Potential trade-offs at level 3

With the in-depth analysis of **-03** comes a trade-off in terms of compilation time and memory resources. Also, because **-03** implies **-qnostrict**, the optimizer can

alter certain floating-point semantics in your application to gain execution speed. This typically involves precision trade-offs as follows:

- Reordering of floating-point computations
- Reordering or elimination of possible exceptions, such as division by zero or overflow
- Using alternative calculations that might give slightly less precise results or not handle infinities or NaNs in the same way

You can still gain most of the **-O3** benefits while preserving precise floating-point semantics by specifying **-qstrict**. Compiling with **-qstrict** is necessary if you require the same absolute precision in floating-point computational accuracy as you get with **-O0**, **-O2**, or **-qnoopt** results. The option **-qstrict=ieee** also ensures adherence to all IEEE semantics for floating-point operations. If your application is sensitive to floating-point exceptions or the order of evaluation for floating-point arithmetic, compiling with **-qstrict**, **-qstrict=exceptions**, or **-qstrict=order** helps to ensure accurate results. You should also consider the impact of the **-qstrict=precision** suboption group on floating-point computational accuracy. The precision suboption group includes the individual suboptions: **subnormals**, **operationprecision**, **association**, **reductionorder**, and **library** (described in the **-qstrict** option in the *XL C/C++ Compiler Reference*).

Without **-qstrict**, the difference in computation for any one source-level operation is very small in comparison to “Basic optimization” on page 22. Although a small difference can be compounded if the operation is in a loop structure where the difference becomes additive, most applications are not sensitive to the changes that can occur in floating-point semantics.

For information about the **-O** level syntax, see “-O -qoptimize” in the *XL C/C++ Compiler Reference*.

An intermediate step: adding **-qhot** suboptions at level 3

At **-O3**, the optimization includes minimal **-qhot** loop transformations at **level=0** to increase performance. To further increase your performance benefit from **-qhot**, increase the optimization aggressiveness by increasing the optimization level of **-qhot**. Try specifying **-qhot** without any suboptions or **-qhot=level=1**.

For more information about **-qhot**, see “Using high-order loop analysis and transformations” on page 29.

Conversely, if the application does not use loops processing arrays, which **-qhot** improves, you can improve compile speed significantly, usually with minimal performance loss by using **-qnohot** after **-O3**.

Optimizing at level 4

Benefits at level 4

- Propagation of global and argument values between compilation units
- Inlining code from one compilation unit to another
- Reorganization or elimination of global data structures
- An increase in the precision of aliasing analysis

Optimizing at **-04** builds on **-03** by triggering **-qipa=level=1**, which performs interprocedural analysis (IPA), optimizing your entire application as a unit. This option is particularly pertinent to applications that contain a large number of frequently used routines.

To make full use of IPA optimizations, you must specify **-04** on the compilation and link steps of your application build as interprocedural analysis occurs in stages at both compile time and link time.

Beyond **-qipa**, **-04** enables other optimization options:

- **-qhot**
Enables more aggressive HOT transformations to optimize loop constructs and array language.
- **-qarch=auto** and **-qtune=auto**
Optimizes your application to execute on a hardware architecture identical to your build machine. If the architecture of your build machine is incompatible with the execution environment of your application, you must specify a different **-qarch** suboption after the **-04** option. This overrides **-qtune=auto**.
- **-qcache=auto**
Optimizes your cache configuration for execution on specific hardware architecture. The **auto** suboption assumes that the cache configuration of your build machine is identical to the configuration of your execution architecture. Specifying a cache configuration can increase program performance, particularly loop operations by blocking them to process only the amount of data that can fit into the data cache at a time.

If you want to execute your application on a different machine, specify correct cache values.

Potential trade-offs at level 4

In addition to the trade-offs already mentioned for **-03**, specifying **-qipa** can significantly increase compilation time, especially at the link step.

The IPA process

1. At compile time optimizations occur on a file-by-file basis, as well as preparation for the link stage. IPA writes analysis information directly into the object files the compiler produces.
2. At the link stage, IPA reads the information from the object files and analyzes the entire application.
3. This analysis guides the optimizer on how to rewrite and restructure your application and apply appropriate **-03** level optimizations.

The “Using interprocedural analysis” on page 33 section contains more information about IPA including details on IPA suboptions.

Optimizing at level 5

Benefits at level 5

- Makes most aggressive optimizations available
- Makes full use of loop optimizations and interprocedural analysis

As the highest optimization level, **-05** includes all **-04** optimizations and deepens whole program analysis by increasing the **-qipa** level to 2. Compiling with **-05** also increases how aggressively the optimizer pursues aliasing improvements.

Additionally, if your application contains a mix of C/C++ and Fortran code that you compile using the XL compilers, you can increase performance by compiling and linking your code with the **-05** option.

Potential trade-offs at level 5

Compiling at **-05** requires more compilation time and machine resources than any other optimization levels, particularly if you include **-05** on the IPA link step. Compile at **-05** as the final phase in your optimization process after successfully compiling and executing your application at **-04**.

Tuning for your system architecture

You can instruct the compiler to generate code for optimal execution on a given microprocessor or architecture family. By selecting appropriate target machine options, you can optimize to suit the broadest possible selection of target processors, a range of processors within a given family of processor architectures, or a specific processor.

The following table lists the optimization options that affect individual aspects of the target machine. Using a predefined optimization level sets default values for these individual options.

Table 9. Target machine options

Option	Behavior
-mcpu	Selects a family of processor architectures for which instruction code should be generated. This option restricts the instruction set generated to a subset of that for the PowerPC architecture. See “Getting the most out of target machine options” for more information about this option.
-mtune	Biases optimization toward execution on a given microprocessor, without implying anything about the instruction set architecture to be used as a target. See “Getting the most out of target machine options” for more information about this option.
-qcache	Defines a specific cache or memory geometry. The defaults are determined through the setting of -mtune . See “Getting the most out of target machine options” for more information about this option.

Related information in the *XL C/C++ Compiler Reference*

-  [-mcpu](#)
-  [-qipa](#)
-  [-qcache](#)

Getting the most out of target machine options

Using the **-mcpu** (**-qarch**) option

Use the **-mcpu** (**-qarch**) compiler option to generate instructions that are optimized for a specific machine architecture. For example, if you want to generate an object code that contains instructions optimized for POWER8®, use **-mcpu=pwr8**. If your application runs on the same machine on which you compile it, use the **-qarch=auto** option, which automatically detects the specific architecture of the compiling machine and generates code to take advantage of instructions available only on that machine (or on a system that supports the equivalent processor

architecture). Otherwise, use the **-mcpu (-qarch)** option to specify the smallest possible family of the machines that can run your code reasonably well.

Using the **-mtune (-qtune)** option

Use the **-mtune (-qtune)** compiler option to control the scheduling of instructions that are optimized for your machine architecture. If you specify a particular architecture with **-mcpu (-qarch)**, **-mtune (-qtune)** automatically selects the suboption that generates instruction sequences with the best performance for that architecture.

If you need to create a single binary file that runs on a range of PowerPC hardware, you can use the **-qtune=balanced** option. With this option in effect, optimization decisions made by the compiler are not targeted to a specific version of hardware. Instead, tuning decisions try to include features that are generally helpful across a broad range of hardware and avoid those optimizations that might be harmful on some hardware.

Note: You must verify the performance of code compiled with the **-qtune=balanced** option before distributing it.

The difference between **-qtune=balanced** and other **-qtune** suboptions including **-qtune=auto** is as follows:

- With the **-qtune=balanced** option, the compiler generates instructions that perform reasonably well across a range of Power hardware.
- With other suboptions, the compiler generates instructions that are optimized for that specified versions of hardware architecture and might not perform well on others.

Using **-qcache** options

If you decide to specify your own **-qcache** suboptions, use **-qhot** or **-qsmp** along with it.

Related information in the *XL C/C++ Compiler Reference*

 **-qhot**

 **-qcache**

 **-mcpu**

 **-mtune**

Using high-order loop analysis and transformations

High-order transformations are optimizations that specifically improve the performance of loops through techniques such as interchange, fusion, and unrolling.

The goals of these loop optimizations include:

- Reducing the costs of memory access through the effective use of caches and address translation look-aside buffers
- Overlapping computation and memory access through effective utilization of the data prefetching capabilities provided by the hardware

- Improving the utilization of microprocessor resources through reordering and balancing the usage of instructions with complementary resource requirements
- Generating SIMD vector instructions to offer better program performance when **-qsimd=auto** is specified
- Generating calls to vector math library functions

To enable high-order loop analysis and transformations, use the **-qhot** option, which implies an optimization level of **-02**. The following table lists the suboptions available for **-qhot**.

Table 10. *-qhot* suboptions

Suboption	Behavior
level=0	Instructs the compiler to perform a subset of high-order transformations that enhance performance by improving data locality. This suboption implies -qhot=novector and -qhot=noarraypad . This level is automatically enabled if you compile with -03 .
level=1	This is the default suboption if you specify -qhot with no suboptions. This level is automatically enabled if you compile with -04 or -05 . This is equivalent to specifying -qhot=vector .
level=2	When used with -qsmp , instructs the compiler to perform the transformations of -qhot=level=1 plus some additional transformation on nested loops. The resulting loop analysis and transformations can lead to more cache reuse and loop parallelization.
vector	When specified with -qnostrict and -qignerrno , or -03 or a higher optimization level, instructs the compiler to transform some loops to use the optimized versions of various math functions contained in the MASS libraries, rather than use the system versions. The optimized versions make different trade-offs with respect to accuracy and exception-handling versus performance. This suboption is enabled by default if you specify -qhot with no suboptions. Also, specifying -qhot=vector with -03 implies -qhot=level=1 .
arraypad	Instructs the compiler to pad any arrays where it infers there might be a benefit and to pad by whatever amount it chooses.

Related information in the *XL C/C++ Compiler Reference*

 [-qhot](#)

 [-qstrict](#)

 [-qignerrno](#)

 [-mcpu](#)

 [-qsimd](#)

Getting the most out of **-qhot**

Here are some suggestions for using **-qhot**:

- Try using **-qhot** along with **-03** for all of your code. It is designed to have a neutral effect when no opportunities for transformation exist. However, it increases compilation time and might have little benefit if the program has no loop processing vectors or arrays. In this case, using **-03 -qnohot** might be better.

- If the runtime performance of your code can significantly benefit from automatic inlining and memory locality optimizations, try using **-O4** with **-qhot=level=0** or **-qhot=novector**.
- If you encounter unacceptably long compilation time (this can happen with complex loop nests), try **-qhot=level=0** or **-qnohot**.
- If your code size is unacceptably large, try reducing the inlining level or using **-qcompact** along with **-qhot**.
- You can compile some source files with the **-qhot** option and some files without the **-qhot** option, allowing the compiler to improve only the parts of your code that need optimization.
- Use **-qreport** along with **-qhot** to generate a loop transformation listing. The listing file identifies how loops are transformed in a section marked LOOP TRANSFORMATION SECTION. Use the listing information as feedback about how the loops in your program are being transformed. Based on this information, you might want to adjust your code so that the compiler can transform loops more effectively. For example, you can use this section of the listing to identify non-stride-one references that might prevent loop vectorization.
- Use **-qreport** along with **-qhot** or any optimization option that implies **-qhot** to generate information about nested loops in the LOOP TRANSFORMATION SECTION of the listing file. In addition, when you use **-qprefetch=assistthread** to generate prefetching assist threads, a message Assist thread for data prefetching was generated is also displayed in this section of the report. To generate a list of aggressive loop transformations and parallelizations performed on loop nests in the LOOP TRANSFORMATION SECTION of the listing file, use **-qhot=level=2** and **-qsmp** together with **-qreport**.

Related information in the *XL C/C++ Compiler Reference*

-  -qcompact
-  -qhot
-  -qsimd
-  -qprefetch
-  -qstrict

Using shared-memory parallelism (SMP)

Most IBM pSeries machines are capable of shared-memory parallel processing. You can compile with **-qsmp** to generate the threaded code needed to exploit this capability. The **-qsmp** option implies the **-qhot** option and an optimization level of **-O2** or higher.

The following table lists the most commonly used suboptions. Descriptions and syntax of all the suboptions are provided in **-qsmp** in the *XL C/C++ Compiler Reference*. An overview of automatic parallelization, as well as of OpenMP directives is provided in Chapter 11, “Parallelizing your programs,” on page 99.

Table 11. Commonly used `-qsmp` suboptions

suboption	Behavior
auto	Instructs the compiler to automatically generate parallel code where possible without user assistance. Any SMP programming constructs in the source code, including OpenMP directives, are also recognized. This is the default setting if you do not specify any <code>-qsmp</code> suboptions, and it also implies the <code>opt</code> suboption.
omp	Parallelizes only program code that is explicitly annotated with OpenMP directives. Note that <code>-qsmp=omp</code> is incompatible with <code>-qsmp=auto</code> .
opt	Enables optimization of parallelized program code. The optimization is equivalent to <code>-O2 -qhot</code> in the absence of other optimization options.
noopt	Performs the smallest amount of optimization that is required to parallelize the code. During development, it can be useful to turn off optimization to facilitate debugging.
<i>fine_tuning</i>	Other values for the suboption provide control over thread scheduling, locking, and so on.

Related information in the *XL C/C++ Compiler Reference*

 `-O, -qoptimize`

 `-qsmp`

 `-qhot`

Getting the most out of `-qsmp`

Here are some suggestions for using the `-qsmp` option:

- Before using `-qsmp` with automatic parallelization, test your programs using optimization and `-qhot` in a single-threaded manner.
- If you are compiling an OpenMP program and do not want automatic parallelization, use `-qsmp=omp:noauto`.
- By default, the runtime environment uses all available processors. Do not set the `XLSMPOPTS=PARTHDS` or `OMP_NUM_THREADS` environment variables unless you want to use fewer than the number of available processors. You might want to set the number of executing threads to a small number or to 1 to ease debugging.

Note: The `XLSMPOPTS=PARTHDS` environment variable is deprecated.

- If you are using a dedicated machine or node, consider setting the `SPINS` and `YIELDS` environment variables (suboptions of the `XLSMPOPTS` environment variable) to 0. Doing so prevents the operating system from intervening in the scheduling of threads across synchronization boundaries such as barriers.
- When debugging an OpenMP program, try using `-qsmp=noopt` (without `-O`) to make the debugging information produced by the compiler more precise.

Related information in the *XL C/C++ Compiler Reference*

 `-qsmp`

 `-qhot`

 Invoking the compiler

 `XLSMPOPTS`

 Environment variables for parallel processing

Using interprocedural analysis

Interprocedural analysis (IPA) enables the compiler to optimize across different files (whole-program analysis), and it can result in significant performance improvements.

You can specify interprocedural analysis on the compilation step only or on both compilation and link steps in whole program mode. Whole program mode expands the scope of optimization to an entire program unit, which can be an executable or a shared object. As IPA can significantly increase compilation time, you should limit using IPA to the final performance tuning stage of development.

You can enable IPA by specifying the **-qipa** option. The most commonly used suboptions and their effects are described in the following table. The full set of suboptions and syntax is described in the **-qipa** section of the *XL C/C++ Compiler Reference*.

The steps to use IPA are as follows:

1. Do preliminary performance analysis and tuning before compiling with the **-qipa** option, because the IPA analysis uses a two-pass mechanism that increases compilation time and link time. You can reduce some compilation and link overhead by using the **-qipa=noobject** option.
2. Specify the **-qipa** option on both the compilation and the link steps of the entire application, or as much of it as possible. Use suboptions to indicate assumptions to be made about parts of the program *not* compiled with **-qipa**.

Table 12. Commonly used **-qipa** suboptions

Suboption	Behavior
level=0	Program partitioning and simple interprocedural optimization, which consists of: <ul style="list-style-type: none">• Automatic recognition of standard libraries.• Localization of statically bound variables and procedures.• Partitioning and layout of procedures according to their calling relationships. (Procedures that call each other frequently are located closer together in memory.)• Expansion of scope for some optimizations, notably register allocation.
level=1	Inlining and global data mapping. Specifically: <ul style="list-style-type: none">• Procedure inlining.• Partitioning and layout of static data according to reference affinity. (Data that is frequently referenced together will be located closer together in memory.) <p>This is the default level if you do not specify any suboptions with the -qipa option.</p>

Table 12. Commonly used `-qipa` suboptions (continued)

Suboption	Behavior
<code>level=2</code>	Global alias analysis, specialization, interprocedural data flow: <ul style="list-style-type: none"> • Whole-program alias analysis. This level includes the disambiguation of pointer dereferences and indirect function calls, and the refinement of information about the side effects of a function call. • Intensive intraprocedural optimizations. This can take the form of value numbering, code propagation and simplification, moving code into conditions or out of loops, and elimination of redundancy. • Interprocedural constant propagation, dead code elimination, pointer analysis, code motion across functions, and interprocedural strength reduction. • Procedure specialization (cloning). • Whole program data reorganization.
<code>inline=suboptions</code>	Provides precise control over function inlining.
<code>fine_tuning</code>	Other values for <code>-qipa</code> provide the ability to specify the behavior of library code, tune program partitioning, read commands from a file, and so on.

Notes:

- XL C/C++ and XL Fortran provide backwards compatibility with IPA objects that are created by earlier compiler versions. If IPA object files that are compiled with newer versions of compilers are linked by an earlier version, errors occur during the link step. For example, if IPA object file `a.o` is compiled by XL C/C++, V13.1.3 and is to be linked with IPA object file `b.o` that is compiled by XL Fortran, V15.1.0, then you must use a compiler whose version is XL C/C++, V13.1.3 or later.
- XL C/C++ and XL Fortran versions released at the same time produce matching IPA level information. IPA object files with matching IPA level information can be linked together. For example, the IPA level for XL C/C++, V13.1.3 matches with the IPA for XL Fortran, V15.1.3. The following table lists some matching XL C/C++ and XL Fortran releases:

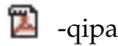
Table 13. Compiler versions and release dates

Compiler version	General availability (Release date)
XL C/C++ for Linux, V13.1.3	11-Dec-2015
XL Fortran for Linux, V15.1.3	
XL C/C++ for Linux, V13.1.0	06-Jun-2014
XL Fortran for Linux, V15.1.0	
XL C/C++ for Linux, V12.1.0	18-May-2012
XL Fortran for Linux, V14.1.0	

For more information about the release dates of compiler products, see the Support Lifecycle website at http://www-01.ibm.com/software/support/lifecycle/index_x.html.

If your compiler version has two release dates on the Support Lifecycle web site, determine the date based on your product ID.

Related information in the XL C/C++ Compiler Reference



-qipa

Getting the most from -qipa

It is not necessary to compile everything with **-qipa**, but try to apply it to as much of your program as possible. Here are some suggestions:

- Specify the **-qipa** option on both the compile and the link steps of the entire application. Although you can also use **-qipa** with libraries, shared objects, and executable files, be sure to use **-qipa** to compile the main and exported functions.
- When compiling and linking separately, use **-qipa=noobject** on the compile step for faster compilation.
- When specifying optimization options in a makefile, use the compiler driver (**xlc**) to link with all the compiler options on the link step included.
- As IPA can generate significantly larger object files than traditional compilations, ensure that there is enough space in the `/tmp` directory (at least 200 MB). You can use the `TMPDIR` environment variable to specify a directory with sufficient free space.
- Try varying the **level** suboption if link time is too long. Compiling with **-qipa=level=0** can still be very beneficial for little additional link time.
- Use **-qipa=list=long** to generate a report of functions that were previously inlined. If too few or too many functions are inlined, consider using **-finline-functions (-qinline)** or **-qnoinline**. To control the inlining of specific functions, use **-qinline+function_name** or **-qinline-function_name**.
- To generate data reorganization information in the listing file, specify the optimization level **-qipa=level=2** or **-05** together with **-qreport**. During the IPA link pass, the data reorganization messages for program variable data will be produced to the data reorganization section of the listing file with the label `DATA REORGANIZATION SECTION`. Reorganizations include array splitting, array transposing, memory allocation merging, array interleaving, and array coalescing.

Note: While IPA's interprocedural optimizations can significantly improve performance of a program, they can also cause incorrect but previously functioning programs to fail. Here are examples of programming practices that can work by accident without aggressive optimization but are exposed with IPA:

- Relying on the allocation order or location of automatic variables, such as taking the address of an automatic variable and then later comparing it with the address of another local variable to determine the growth direction of a stack. The C language does not guarantee where an automatic variable is allocated, or its position relative to other automatic variables. Do not compile such a function with IPA.
- Accessing a pointer that is either invalid or beyond an array's bounds. Because IPA can reorganize global data structures, a wayward pointer that might have previously modified unused memory might now conflict with user-allocated storage.
- Dereferencing a pointer that has been cast to an incompatible type.

Related information in the *XL C/C++ Compiler Reference*

 [-finline-functions](#)

 [-qlist](#)

 [-qipa](#)

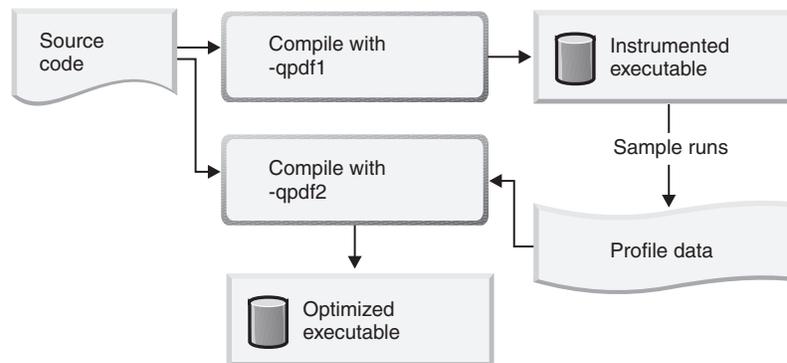
Using profile-directed feedback

You can use profile-directed feedback (PDF) to tune the performance of your application for a typical usage scenario. The compiler optimizes the application based on an analysis of how often branches are taken and blocks of code are run.

Use the PDF process as one of the last steps of optimization before you put the application into production. Optimization at all levels from **-O2** up can benefit from PDF. Other optimizations such as the **-qipa** option and optimization levels **-O4** and **-O5** can also benefit from PDF process.

The following diagram illustrates the PDF process.

Figure 1. Profile-directed feedback



To use the PDF process to optimize your application, follow these steps:

1. "Compiling with -qpdf1"
2. "Training with typical data sets" on page 37
3. "Recompiling or linking with -qpdf2" on page 40

Related information in the XL C/C++ Compiler Reference

 -qpdf1, -qpdf2

 -O, -qoptimize

 -qipa

Compiling with -qpdf1

Compile some or all of the source files in a program with the **-qpdf1** option. This step is called the PDF1 step.

Usage

You do not have to compile all of the files of the programs with the **-qpdf1** option. In a large application, you can concentrate on those areas of the code that can benefit most from the optimization.

When you compile multiple programs with **-qpdf1**, you must ensure that the PDF file that is to be generated in the PDF2 step has a different name for each program

because each PDF file can contain data for only one application. Otherwise, if the PDF file names of multiple programs are the same, you can run the first application in the PDF training step and write profile information into the PDF file successfully. However, if you try to run the second application, an error is issued to indicate that there is an existing incompatible PDF file and the program exits with a nonzero return code.

PDF optimization must be done at least at the **-02** optimization level and is recommended at **-04** and higher.

When you compile your program with **-qpdf1**, the **-qipa=level=0** option is enabled by default. When option **-04**, **-05**, or any level of option **-qipa** is in effect, and you specify the **-qpdf1** option at the link step but not at the compile step, the compiler issues a warning message. The message indicates that you must recompile your program to get all the profiling information.

A PDF map file is generated at this step. You can use the **showpdf** command to display part of the profiling information in text or XML format. If you do not need to view the profiling information, specify the **-qnoshowpdf** option at this step so that the PDF map file is not generated.

Example

You can compile the following source files with **-qpdf1** at **-03**.

```
xlc -qpdf1 -03 file1.c file2.c file3.c
```

Related information in the *XL C/C++ Compiler Reference*

 **-qpdf1, -qpdf2**

 **-O, -qoptimize**

 **showpdf**

 **-qshowpdf**

Training with typical data sets

Run the resulting application with typical data sets. This step is called the PDF training step.

You can train the resulting application multiple times with different typical data sets. The profiling information is accumulated into PDF files to provide a count of how often branches are taken and blocks of code are run, based on the input data used.

You can get PDF files in the following ways:

- Generate PDF files upon normal termination. For more information, see “Generating PDF files upon normal termination” on page 38.
- Dump snapshot PDF profiling information to files during execution. For more information, see “Dumping snapshot PDF profiling information during execution” on page 39.

Usage

You should use typical data sets for the PDF training step. Otherwise, the analysis of infrequently executed code paths might be distorted.

The PDF file is placed in the current working directory or the directory specified by the PDFDIR environment variable. To override the default file path or name, use the **-qpdf1=pdfname** or **-qpdf1=defname** option. If the PDFDIR environment variable is set but the specified directory does not exist, the compiler issues a warning message. To avoid wasting compile and run time, make sure that the PDFDIR environment variable is set to an absolute path. Otherwise, you might run the application from a wrong directory and the compiler cannot locate the profiling information files. When it happens, the program might not be optimized correctly or might be stopped by a segmentation fault. A segmentation fault might also happen if you change the value of the PDFDIR environment variable and run the application before the PDF process finishes.

If you have several PDF files, use the **mergepdf** command to combine these PDF files into one PDF file.

If you recompile your program with **-qpdf1**, the compiler removes the existing PDF file or files whose names and locations are the same as the file or files that are to be created in the training step before the compiler generates a new application.

Notes:

- You cannot mix PDF and PDF map files that are generated for different programs. Further, you cannot mix PDF and PDF map files that are generated from multiple compilation processes with different **-qpdf1** settings for the same program.
- You must use the same version and PTF level of the compiler to generate the PDF file and the PDF map file.
- You cannot edit PDF files that are generated by the resulting application. Otherwise, the performance or function of the generated executable application might be affected.

Related information in the *XL C/C++ Compiler Reference*

-  [-qpdf1, -qpdf2](#)
-  [-O, -qoptimize](#)
-  [mergepdf](#)
-  [Runtime environment variables](#)

Generating PDF files upon normal termination

When you run an application that was generated in the PDF1 step and end the application using normal methods, a standard PDF file is generated.

This PDF file is named `.<output_name>.pdf` by default, where `<output_name>` is the name of the output file that was generated in the PDF1 step.

Normal termination methods include reaching the end of the execution for the main function and calling the `exit()` function in `libc (stdlib.h)`.

Related information

-  [Using profile-directed feedback](#)
-  [-qpdf1, -qpdf2](#)
-  [Training with typical data sets](#)



Dumping snapshot PDF profiling information during execution

Dumping snapshot PDF profiling information during execution

Before you begin

Compile some or all of the source files in a program with the **-qpdf1** option.

About this task

In addition to getting PDF profiling information that is written to one or more PDF files upon normal termination, you can dump PDF profiling information to one or more PDF snapshot files during execution. This is especially useful when the application is to be terminated abnormally, for example, when you end a process by using SIGKILL or a system call `exit()`, `_Exit()`, or `abort()`.

Procedure

1. Define environment variable `PDF_SIGNAL_TO_DUMP` as a signal value that is an integer in the range of `SIGRTMIN` and `SIGRTMAX` inclusive. This value is read during program initialization.
2. Run the application that was generated in the PDF1 step.
3. Send the user-defined signal value as a trigger to dump PDF profiling information to PDF snapshot files. You can send the user-defined signal value multiple times for each process, and the corresponding PDF snapshot file is overwritten each time. The running application processes continue running until normal termination.

Results

A PDF snapshot file is created for each process with suffix `_snapshot.<pid>`, and it is named `.<output_name>_pdf_snapshot.<pid>` by default, where `<output_name>` is the name of the output file that is generated in the PDF1 step and `<pid>` is the ID of the process.

If the application creates more than one process, equivalent number of PDF data files are generated when each process receives the signal.

If multiple PDF snapshot files are generated, you must merge the PDF snapshot files for multiple processes by using the **mergepdf** command before linking or recompiling the program with **-qpdf2**.

If the application ends normally by executing exit handlers, a standard PDF file named `.<output_name>_pdf` is also generated.

Example

In this example, program `myprogram` has a parent process and a child process. You can dump snapshot PDF profiling information to files during execution and use the merged PDF file to fine-tune the optimizations as follows:

1. Compile `myprogram.c` with **-qpdf1** and name the executable file as `myprogram`.

```
xlc -O2 -qpdf1 -o myprogram myprogram.c
```
2. Define environment variable `PDF_SIGNAL_TO_DUMP` as 52, which is in the range of `SIGRTMIN` and `SIGRTMAX` inclusive.

```
export PDF_SIGNAL_TO_DUMP=52
```

3. Run the compiled application.

```
./myprogram < sample.data &
```

The following process ID is displayed:

```
[138705] ./myprogram &
```

4. Send the user-defined signal 52 to dump PDF profiling information by using the **kill** command.

Note: The **kill** command does not terminate the processes when the signal value is recognized.

```
kill -s 52 138705
```

Because the application creates a parent process and a child process, both processes receive the signal value. The following messages are displayed for both processes:

```
PDFRunTime: Caught user signal "52", dumping PDF data ... Done.  
PDFRunTime: Caught user signal "52", dumping PDF data ... Done.
```

Two PDF snapshot files that are suffixed with process ID, `.myprogram_pdf_snapshot.138705` and `.myprogram_pdf_snapshot.138706`, are created, where 138706 is a child process of 138705.

5. Merge PDF data for the two processes as `.myprogram_pdf`.
6. Recompile `myprogram.c` by using the same compiler options as step 1, but change **-qpdf1** to **-qpdf2**. `.myprogram_pdf` is used to fine-tune the optimizations.

```
xlc -O2 -qpdf2 -o myprogram myprogram.c
```

Related information in the XL C/C++ Compiler Reference

 `-qpdf1, -qpdf2`

 `mergepdf`

Recompiling or linking with **-qpdf2**

Recompile or link your program by using the same compiler options as before, but change **-qpdf1** to **-qpdf2**. This step is called the PDF2 step.

If you want to optimize the entire application, you can do the PDF2 step at executable level in the link step and create executable files for your programs. For details, see “Executable level profile-directed feedback” on page 41.

In addition to optimizing entire application, you can do the PDF2 step at object level in the compile time and create object files for your programs. This can be an advantage for applications where patches or updates are distributed as object files or libraries rather than as executables. In this case, you cannot enable interprocedural analysis (IPA) optimizations. For details, see “Object level profile-directed feedback” on page 42.

Usage

The accumulated profiling information is used to fine-tune the optimizations. The resulting program contains no profiling overhead and runs at full speed. You must ensure the PDF file that is fed into the PDF2 step is the one that is generated in the PDF training step. For example, when the **-qpdf1=pdfname=file_path** option is used during the PDF1 step, you must use the same **-qpdf2=pdfname=file_path**

option during the PDF2 step for the compiler to recognize the correct PDF file. This rule also applies to the **-qpdf[1|2]=exename** and **-qpdf[1|2]=defname** options.

You are highly recommended to use the same optimization level at all compilation steps for a particular program. Otherwise, the PDF process cannot optimize your program correctly and might even slow it down. All compiler settings that affect optimization must be the same, including any supplied by configuration files.

You can modify your source code and use the **-qpdf1** and **-qpdf2** options to compile your program. Old profiling information can still be preserved and used during the second stage of the PDF process. The compiler issues a list of warnings but the compilation does not stop. An information message is also issued with a number in the range of 0 - 100 to indicate how outdated the old profiling information is. If you have not changed your program between the **-qpdf1** and **-qpdf2** phases, the number is 100, which means that all the profiling information can be used to optimize the program. If the number is 0, it means that the profiling information is completely outdated, and the compiler cannot take advantage of any information. When the number is less than 100, you can choose to recompile your program with the **-qpdf1** option and regenerate the profiling information.

If the compiler cannot read any PDF files in this step, the compiler issues error message 1586-401 but continues the compilation.

If you want to erase the PDF information, use the **cleanpdf** command.

Related information in the *XL C/C++ Compiler Reference*

-  Using profile-directed feedback
-  **-qpdf1**, **-qpdf2**
-  **-O**, **-qoptimize**
-  **-qreport**
-  **cleanpdf**

Executable level profile-directed feedback

About this task

When you compile your program with **-qpdf2**, the **-qipa=level=0** option is enabled by default, so IPA and PDF optimization is done in the link step.

It is recommended that you use the **-qpdf2** option to link the object files that are created during the PDF1 step without recompiling your program. Using this approach, you can save considerable compilation time and achieve the same optimization result as if you had recompiled your program during the PDF2 step.

When option **-04**, **-05**, or any level of option **-qipa** is in effect, and you specify the **-qpdf1** or **-qpdf2** option at the link step but not at the compile step, the compiler issues a warning message. The message indicates that you must recompile your program to get all the profiling information.

Examples

1. Set the PDFDIR environment variable
`export PDFDIR=$HOME/project_dir`
2. Compile most of the files with **-qpdf1**

```
xlc -qpdf1 -O3 -c file1.c file2.c file3.c
```

3. Compile the file that does not need optimization without **-qpdf1**

```
xlc -c file4.c
```

4. Link the PDF object files (file1.o, file2.o, and file3.o) with the non-PDF object file (file4.o)

```
xlc -qpdf1 -O3 file1.o file2.o file3.o file4.o
```

5. Run the resulting application several times with different input data

```
./a.out < polar_orbit.data
```

```
./a.out < elliptical_orbit.data
```

```
./a.out < geosynchronous_orbit.data
```

6. Link all the object files into the final application with **-qpdf2**.

```
xlc -qpdf2 -O3 file1.o file2.o file3.o file4.o
```

Instead of step 6 in this example, you can recompile your program with **-qpdf2**.

```
xlc -qpdf2 -O3 file1.c file2.c file3.c file4.c
```

Both recompiling and linking your program with **-qpdf2** can use the accumulated profiling information to fine-tune the optimizations.

Related information



Using profile-directed feedback



-qpdf1, -qpdf2



Recompiling or linking with -qpdf2



Object level profile-directed feedback



-O, -qoptimize



-qipa

Object level profile-directed feedback

About this task

In addition to optimizing entire executables, profile-directed feedback (PDF) can also be applied to specific object files. This approach can be an advantage in applications where patches or updates are distributed as object files or libraries rather than as executables. Also, specific areas of functionality in your application can be optimized without the process of relinking the entire application. In large applications, you can save the time and trouble that otherwise need to be spent relinking the application.

The process for using object level PDF is essentially the same as the executable level PDF process but with a small change to the PDF2 step. For object level PDF, compile your program by using the **-qpdf1** option, run the resulting application with representative data, compile the program again with the **-qpdf2** option. You need to specify **-qnoipa** with **-qpdf2**, which means you cannot use interprocedural analysis (IPA) optimizations and object level PDF at the same time.

If you specify **-qpdf2 -qnoipa**, object level PDF optimization is done in the compile step. You must take either one of the following actions to ensure that the compiler can recognize the correct PDF file:

- In the PDF1 step, specify **-qpdf1=defname** to revert the PDF file name to `._pdf`. Thus, the compiler looks for `._pdf` in the PDF2 step.

- In the PDF2 step, specify **-qpdf2=pdfname=file_path**, where *file_path* is the path and name of the generated PDF file.

The following steps outline this process:

1. Compile your program by using the **-qpdf1** option. For example:

```
xlc -c -O3 -qpdf1 file1.c file2.c file3.c
```

In this example, the optimization level **-O3** is used, which is a moderate level of optimization.

2. Link the object files to get an instrumented executable:

```
xlc -O3 -qpdf1 file1.o file2.o file3.o
```

3. Run the instrumented executable with sample data that is representative of the data you want to optimize for.

```
a.out < sample_data
```

4. Compile the program again by using the **-qpdf2** option. Specify the **-qnoipa** option so that the linking step is skipped and PDF optimization is applied to the object files rather than to the entire executable.

```
xlc -c -O3 -qpdf2 -qnoipa file1.c file2.c file3.c
```

The resulting output of this step are object files optimized for the sample data processed by the original instrumented executable. In this example, the optimized object files would be `file1.o`, `file2.o`, and `file3.o`. These object files can be linked by using the system loader `ld` or by omitting the **-c** option in the PDF2 step.

Related information

-  Using profile-directed feedback
-  `-qpdf1`, `-qpdf2`
-  Recompiling or linking with `-qpdf2`
-  Executable level profile-directed feedback
-  `-O`, `-qoptimize`
-  `-qipa`

Marking variables as local or imported

The compiler assumes that all variables in applications are imported, but the use of `-qdatalocal` and `-qdataimported` can mark variables local or imported. The compiler optimizes applications that are based on the specification of static or dynamic binding for program variables.

-qdatalocal

Local variables are stored in a special segment of memory that is uniquely bound to a program or shared library. Specify the `-qdatalocal` option to identify variables to be treated as local to a compiled program or shared library. You can specify the option with no parameters to indicate that all appropriate variables are local. Alternatively, you can append a list of colon-separated names to the option to treat only a subset of the program arguments as local.

When it can be, a variable that is marked as local is embedded directly into a structure that is called the table of contents (TOC) instead of in a separate global piece of memory. The prerequisite is that the variable's storage must be no more than the pointer size for it to be embedded in the TOC. Usually, pointers to data are stored in the TOC. The `-qdatalocal` option allows storage of data directly in the TOC, hence reducing data accesses from two load instructions to one load instruction.

-qdataimported

Imported variables are stored according to the default memory allocation scheme. The `-qdataimported` option is the default data binding mechanism. Specifying the option implies that the data is visible to other program or shared library that is linked. As a result, specifying variable names as arguments to the `-qdataimported` option or compiling with the `-qdataimported` option without arguments in isolation has no effect.

The `-qdataimported` option is useful when you use it in combination with `-qdatalocal`. Because it is unlikely that you want to store all data in the TOC, the `-qdataimported` option can override `-qdatalocal` for variables external to a program or shared library. For example, the use of options `-qdatalocal -qdataimported=<variable>` stores all global data in the TOC except for `<variable>`.

Related information in the *XL C/C++ Compiler Reference*



`-qdataimported`, `-qdatalocal`, `-qtocdata`

Getting the most out of -qdatalocal

You can see some examples that illustrate the use of the `-qdatalocal` option.

In the source for the following program file, A1 and A2 are global variables:

```
int A1;
int A2;
int main(){
    A2=A1+1;
    return A2;
}
```

Here is an excerpt of the listing file that is created if you specify `-qlist` without `-qdatalocal`:

		000000			PDEF	main
4					PROC	
5		000000 lwz	80620004	1	L4A	gr3=.A1(gr2,0)
5		000004 lwz	80630000	1	L4A	gr3=A1(gr3,0)
5		000008 addi	38630001	1	AI	gr3=gr3,1
5		00000C lwz	80820008	1	L4A	gr4=.A2(gr2,0)
5		000010 stw	90640000	1	ST4A	A2(gr4,0)=gr3

Here is an excerpt of the listing file that is created if you specify `-qlist` with `-qdatalocal`:

		000000			PDEF	main
4					PROC	
5		000000 lwz	80620004	1	L4A	gr3=A1(gr2,0)
5		000004 addi	38630001	1	AI	gr3=gr3,1
5		000008 stw	90620008	1	ST4A	A2(gr2,0)=gr3

When you specify `-qdatalocal`, the data is accessed by a single load instruction because the A1 and A2 variables are embedded in the TOC. When you do not

specify `-qdatalocal`, A1 and A2 variables are accessed by two load instructions. In this example, you can use `-qdatalocal=A1:A2` to specify local variables individually.

You can always see the section that begins with `>>>> OPTIONS SECTION <<<<<` in the `.lst` file that is created by `-qlist` to confirm the use of these options. For example, you can view `DATALOCAL=<variables>` or `DATALOCAL` when the option is specified.

Notes:

- On 64-bit Linux, TOC entries are pointer size. When you specify `-qdatalocal` without arguments, the option is ignored for variables that are larger than the pointer size. Conversely, data smaller than pointer size is word-aligned. See the following example of an `objdump` excerpt that shows when a char (`r3`) is marked local. The offset between the byte and the next data (`r4`) is still 4 bytes. The data is accessed by a load byte instruction instead of a regular load.

```
10000380:      88 62 00 20      1bz      r3,32(r2)
10000384:      80 82 00 24      1        r4,36(r2)
r2 (base address of the TOC), r3 (char), r4 (int)
```
- If you specify an unsuitable variable as a parameter to `-qdatalocal`, `-qdatalocal` is ignored. Unsuitable variables can be data that exceeds pointer-size bytes or variables that do not exist. When you specify `-qdatalocal` for a variable that is not a TOC candidate, the default storage for that variable is set to `-qdataimported` and the variable is not stored in the TOC.
-  You must use the mangled names when you specify local variables. Otherwise, you might encounter an error message. 
- Mark variables as local with care. If you specify `-qdatalocal` without any arguments, expect all global variables to be candidates for TOC direct placement, even those variables that are marked as external. Variables with static linkage do not have the same issues.
- Since each TOC structure is unique to a module or shared library, the utility of the `-qdatalocal` option is limited to data within that module or shared library.
- For programs with multiple modules, switching between multiple TOC structures might dilute the speedup that is associated with this option.

Related information in the *XL C/C++ Compiler Reference*



`-qdataimported`, `-qdatalocal`, `-qtocdata`

Using compiler reports to diagnose optimization opportunities

You can use the `-qlistfmt` option to generate a compiler report in XML or HTML format. It provides information about how your program is optimized. You can also use the `genhtml` utility to convert an existing XML report to HTML format. This information helps you understand your application codes and tune codes for better performance.

The compiler report in XML format can be viewed in a browser that supports XSLT. If you compile with the `stylesheet` suboption, for example, `-qlistfmt=xml=all:stylesheet=xlstyle.xml`, the report contains a link to a stylesheet that renders the XML readable. By reading the report, you can detect opportunities to further optimize your code. You can also create tools to parse this information.

By default, the name of the report is `a.xml` for XML format, and `a.html` for HTML format. You can use the `-qlistfmt=xml=filename` or `-qlistfmt=html=filename` option to override the default name.

Inline reports

If you compile with `-finline-functions` and one of `-qlistfmt=xml=inlines`, `-qlistfmt=html=inlines`, `-qlistfmt=xml`, or `-qlistfmt=html`, the generated compiler report includes a list of inline attempts during compilation. The report also specifies the type of attempt and its outcome.

For each function that the compiler has attempted to inline, there is an indication of whether the inline was successful. The report might contain any number of reasons why a named function has not been successfully inlined. Some examples of these reasons are as follows:

- `FunctionTooBig` - The function is too big to be inlined.
- `RecursiveCall` - The function is not inlined because it is recursive.
- `ProhibitedByUser` - Inlining was not performed because of a user-specified pragma or directive.
- `CallerIsNoopt` - No inlining was performed because the caller was compiled without optimization.
- `WeakAndNotExplicitlyInline` - The calling function is weak and not marked as inline.

For a complete list of the possible reasons, see the `Inline optimization types` section of the XML schema help file named `XMLContent.html` in the `/opt/ibm/xlc/13.1.6/listings/` directory. The Japanese and Chinese versions of the help file, `XMLContent-Japanese.utf8.html` and `XMLContent-Chinese.utf8.html`, are included in this directory as well.

Loop transformations

If you compile with `-qhot` and one of `-qlistfmt=xml=transforms`, `-qlistfmt=html=transforms`, `-qlistfmt=xml` or `-qlistfmt=html`, the generated compiler report includes a list of the transformations performed on all loops in the file during compilation. The report also lists the reasons why transformations were not performed in some cases:

- Reasons why a loop cannot be automatically parallelized
- Reasons why a loop cannot be unrolled
- Reasons why SIMD vectorization failed

For a complete list of the possible transformation problems, see the `Loop transformation types` section of the XML schema help file named `XMLContent.html` in the `/opt/ibm/xlc/13.1.6/listings/` directory.

Data reorganizations

If you compile with `-qhot` and one of `-qlistfmt=xml=data`, `-qlistfmt=html=data`, `-qlistfmt=xml`, or `-qlistfmt=html`, the generated compiler report includes a list of data reorganizations performed on the program during compilation. Here are some examples of data reorganizations:

- Array splitting
- Array coalescing

- Array interleaving
- Array transposition
- Memory merge

For each of these reorganizations, the report contains details about the name of the data, file names, line numbers, and the region names.

Profile-directed feedback reports

If you compile with `-qpdf2` and one of `-qlistfmt=xml=pdf`, `-qlistfmt=html=pdf`, `-qlistfmt=xml`, or `-qlistfmt=html`, the generated compiler report includes the following information:

- Loop iteration counts
- Block and call counts
- Cache misses (if compiled with `-qpdf1=level=2`)

Related information:



`-qlistfmt`

Parsing compiler reports with development tools

You can write tools to parse the compiler reports produced in XML format to help you find opportunities to improve application performance.

The compiler includes an XML schema that you can use to create a tool to parse the compiler reports and display aspects of your code that might represent performance improvement opportunities. The schema, `xllisting.xsd`, is located in the `/opt/ibm/xlC/13.1.6/listings/` directory. This schema helps present the information from the report in a tree structure.

You can also find a schema help file named `XMLContent.html` that helps you understand the schema details. The Japanese and Chinese versions of the help file, `XMLContent-Japanese.utf8.html` and `XMLContent-Chinese.utf8.html`, are in the same directory.

Other optimization options

Options are available to control particular aspects of optimization. They are often enabled as a group or given default values when you enable a more general optimization option or level.

For more information about these options, see the heading for each option in the *XL C/C++ Compiler Reference*.

Table 14. Selected compiler options for optimizing performance

Option	Description
<code>-qignerrno</code>	Allows the compiler to assume that <code>errno</code> is not modified by library function calls, so that such calls can be optimized. Also allows optimization of square root operations, by generating inline code rather than calling a library function.
<code>-qsmallstack</code>	Instructs the compiler to compact stack storage. Doing so might increase heap usage, which might increase execution time. However, it might be necessary for the program to run or to be optimally multithreaded.

Table 14. Selected compiler options for optimizing performance (continued)

Option	Description
-finline-functions (-qinline)	Controls inlining.
-funroll-loops (-qunroll), -funroll-all-loops (-qunroll=yes)	Independently controls loop unrolling. -funroll-all-loops is implicitly activated under -O3 .
 -fno-exceptions (-qnoeh)	Informs the compiler that no C++ exceptions will be thrown and that cleanup code can be omitted. If your program does not throw any C++ exceptions, use this option to compact your program by removing exception-handling code.
-qnounwind	Informs the compiler that the stack will not be unwound while any routine in this compilation is active. This option can improve optimization of nonvolatile register saves and restores. In C++, the -qnounwind option implies the -fno-exceptions (-qnoeh) option. It should not be used if the program uses setjmp/longjmp or any other form of exception handling.
-qstrict	Disables all transformations that change program semantics. In general, compiling a correct program with -qstrict and any levels of optimization produces the same results as without optimization.
-qnostrict	Allows the compiler to reorder floating-point calculations and potentially excepting instructions. A potentially excepting instruction is one that might raise an interrupt due to erroneous execution (for example, floating-point overflow, a memory access violation). -qnostrict is used by default for the -O3 and higher optimization levels.
-qprefetch	Inserts prefetch instructions in compiled code to improve code performance. In situations where you are working with applications that generate a high cache-miss rate, you can use its suboption assistthread to generate prefetching assist threads (for example, -qprefetch=assistthread). -qnoprefetch is the default option.

Related information in the *XL C/C++ Compiler Reference*

-  -qignerrno
-  -qsmallstack
-  -finline-functions
-  -funroll-loops, -funroll-all-loops
-  -qinlglue
-  -fexceptions (-qeh) (C++ only)
-  -qunwind
-  -qstrict
-  -qprefetch

Chapter 6. Debugging optimized code

Debugging optimized programs presents special usability problems. Optimization can change the sequence of operations, add or remove code, change variable data locations, and perform other transformations that make it difficult to associate the generated code with the original source statements.

For example:

Data location issues

With an optimized program, it is not always certain where the most current value for a variable is located. For example, a value in memory might not be current if the most current value is being stored in a register. Most debuggers cannot follow the removal of stores to a variable, and to the debugger it appears as though that variable is never updated, or possibly even never set. This contrasts with no optimization where all values are flushed back to memory and debugging can be more effective and usable.

Instruction scheduling issues

With an optimized program, the compiler might reorder instructions. That is, instructions might not be executed in the order you would expect based on the sequence of lines in the original source code. Also, the sequence of instructions for a statement might not be contiguous. As you step through the program with a debugger, the program might appear as if it is returning to a previously executed line in the code (interleaving of instructions).

Consolidating variable values

Optimizations can result in the removal and consolidation of variables. For example, if a program has two expressions that assign the same value to two different variables, the compiler might substitute a single variable. This can inhibit debug usability because a variable that a programmer is expecting to see is no longer available in the optimized program.

There are a couple of different approaches you can take to improve debug capabilities while also optimizing your program:

Debug non-optimized code first

Debug a non-optimized version of your program first, and then recompile it with your desired optimization options. See “Debugging in the presence of optimization” on page 51 for some compiler options that are useful in this approach.

Use -g level

Use the **-g** level suboption to control the amount of debugging information made available. Increasing it improves debug capability but prevents some optimizations. For more information, see **-g**.

Detecting errors in code

The compiler provides environment variables and options that help with detecting errors in your source code.

OMP_DISPLAY_ENV

Setting the `OMP_DISPLAY_ENV` environment variable instructs the OpenMP runtime library to display the values of the internal control variables (ICVs) associated with OpenMP environment variables and the OpenMP runtime library. It also displays information about OpenMP version and build-specific information about the runtime library. You can use the environment variable in the following cases:

- When the runtime library is statically linked with an OpenMP program, use `OMP_DISPLAY_ENV=VERBOSE` to check the version of the library that is used during link time.
- When the runtime library is dynamically linked with an OpenMP program, use `OMP_DISPLAY_ENV=VERBOSE` to check the library that is used at run time.
- Use `OMP_DISPLAY_ENV=VERBOSE` or `OMP_DISPLAY_ENV=TRUE` to check the current setting of the runtime environment.

For more information on the environment variable, see `OMP_DISPLAY_ENV` in the *XL C/C++ Compiler Reference*.

-qcheck=bounds

The `-qcheck=bounds` option performs runtime checking of addresses for subscripting within an object of known size. The index is checked to ensure that it will result in an address that lies within the bounds of the object's storage. A trap will occur if the address does not lie within the bounds of the object. This suboption has no effect on accesses to a variable length array.

-qcheck=stackclobber

The `-qcheck=stackclobber` option detects stack corruption of nonvolatile registers in the save area in user programs. This type of corruption happens only if any of the nonvolatile registers in the save area of the stack is modified.

-qcheck=unset

The `-qcheck=unset` option checks for automatic variables that are used before they are set at run time.

The `-qinitauto` option initializes automatic variables. As a result, the `-qinitauto` option hides uninitialized variables from the `-qcheck=unset` option.

Understanding different results in optimized programs

Here are some reasons why an optimized program might produce different results from one that has not undergone the optimization process:

- Optimized code can fail if a program contains code that is not valid. The optimization process relies on your application conforming to language standards.
- If a program that works without optimization fails when you optimize, check the cross-reference listing and the execution flow of the program for variables that are used before they are initialized. Compile with the `-qinitauto=hex_value` option to try to produce the incorrect results consistently. For example, using `-qinitauto=FF` gives variables an initial value of "negative not a number" (-NAN). Any operations on these variables will also result in NAN values. Other bit patterns (*hex_value*) might yield different results and provide further clues as

to what is going on. Programs with uninitialized variables can appear to work properly when compiled without optimization because of the default assumptions the compiler makes, but such programs might fail when you optimize. Similarly, a program can appear to execute correctly after optimization, but it fails at lower optimization levels or when it is run in a different environment. You can also use the **-qcheck=unset** option to detect variables that are not or might not be initialized.

- Referring to an automatic-storage variable by its address after the owning function has gone out of scope leads to a reference to a memory location that can be overwritten as other auto variables come into scope as new functions are called.

Use with caution debugging techniques that rely on examining values in storage, unless the **-g8** or **-g9** option is in effect and the optimization level is **-O2**. The compiler might have deleted or moved a common expression evaluation. It might have assigned some variables to registers so that they do not appear in storage at all.

Debugging in the presence of optimization

Debug and compile your program with your desired optimization options. Test the optimized program before placing it into production. If the optimized code does not produce the expected results, you can attempt to isolate the specific optimization problems in a debugging session.

The following list presents options that provide specialized information, which can be helpful during the debugging of optimized code:

-qlist Instructs the compiler to emit an object listing. The object listing includes hex and pseudo-assembly representations of the generated instructions, traceback tables, and text constants.

-qreport

Instructs the compiler to produce a report of the loop transformations it performed, what inlining was done, and some other transformations. To generate a listing file, you must specify the **-qreport** option with at least one optimization option such as **-qhot**, **-qsmp**, **-finline-functions** (**-qinline**), or **-qsimd**.

-qipa=list

Instructs the compiler to emit an object listing that provides information for IPA optimization.

-qcheck

Generates code that performs certain types of runtime checking.

-qsmp=noopt

If you are debugging SMP code, **-qsmp=noopt** ensures that the compiler performs only the minimum transformations necessary to parallelize your code and preserves maximum debug capability.

-qkeepparm

Ensures that procedure parameters are stored on the stack even during optimization. This can negatively impact execution performance. The **-qkeepparm** option then provides access to the values of incoming parameters to tools, such as debuggers, simply by preserving those values on the stack.

-qinitauto

Instructs the compiler to emit code that initializes all automatic variables to a given value.

-g

Generates debugging information to be used by a symbolic debugger. You can use different **-g** levels to debug optimized code by viewing or possibly modifying accessible variables at selected source locations in the debugger. Higher **-g** levels provide a more complete debug support, while lower levels provide higher runtime performance. For details, see **-g**.

Chapter 7. Coding your application to improve performance

Chapter 5, “Optimizing your applications,” on page 21 discusses the various compiler options that the XL C/C++ compiler provides for optimizing your code with minimal coding effort. If you want to take your application a step further to complement and take the most advantage of compiler optimizations, the topics in this section discuss C and C++ programming techniques that can improve performance of your code.

Finding faster input/output techniques

There are a number of ways to improve your program's performance of input and output:

- If your file I/O accesses do not exhibit locality (that is truly random access such as in a database), implement your own buffering or caching mechanism on the low-level I/O functions.
- If you do your own I/O buffering, make the buffer a multiple of 4KB, which is the minimum size of a page.
- Use buffered I/O to handle text files.
- If you have to process an entire file, determine the size of the data to be read in, allocate a single buffer to read it to, read the whole file into that buffer at once using read, and then process the data in the buffer. This reduces disk I/O, provided the file is not so big that excessive swapping will occur. Consider using the mmap function to access the file.

Reducing function-call overhead

When you write a function or call a library function, consider the following guidelines:

- Call a function directly, rather than using function pointers.
- Use const arguments in inlined functions whenever possible. Functions with constant arguments provide more opportunities for optimization.
- Use the restrict keyword for pointers that can never point to the same memory.
- Use **#pragma disjoint** within functions for pointers or reference parameters that can never point to the same memory.
- Declare a nonmember function as static whenever possible. This can speed up calls to the function and increase the likelihood that the function will be inlined.
- **> C++** Usually, you should not declare all your virtual functions inline. If all virtual functions in a class are inline, the virtual function table and all the virtual function bodies will be replicated in each compilation unit that uses the class.
- **> C++** When declaring functions, use the const specifier whenever possible.
- **> C** Fully prototype all functions. A full prototype gives the compiler and optimizer complete information about the types of the parameters. As a result, promotions from unwidened types to widened types are not required, and parameters can be passed in appropriate registers.
- **> C** Avoid using unprototyped variable argument functions.
- Design functions so that they have few parameters and the most frequently used parameters are in the leftmost positions in the function prototype.

- Avoid passing by value large structures or unions as function parameters or returning a large structure or a union. Passing such aggregates requires the compiler to copy and store many values. This is worse in C++ programs in which class objects are passed by value because a constructor and destructor are called when the function is called. Instead, pass or return a pointer to the structure or union, or pass it by reference. Homogeneous structs, unions and arrays that meet one of the following conditions can be efficiently passed as value parameters or returned as function results:
 - Contain only up to eight floating-point values of the same type where complex is counted as two.
 - Contain up to eight vector values or values processed in vector registers.
- Pass non-aggregate types such as `int` and `short` or small aggregates by value rather than passing by reference, whenever possible.
- If your function exits by returning the value of another function with the same parameters that were passed to your function, put the parameters in the same order in the function prototypes. The compiler can then branch directly to the other function.
- Use the built-in functions, which include string manipulation, floating-point, and trigonometric functions, instead of coding your own. Intrinsic functions require less overhead and are faster than a function call, and they often allow the compiler to perform better optimization.
 - ▶ **C++** Many functions from the C++ standard libraries are mapped to optimized built-in functions by the compiler.
 - ▶ **C** Many functions from `string.h` and `math.h` are mapped to optimized built-in functions by the compiler.
- Selectively mark your functions for inlining using the `inline` keyword. An inlined function requires less overhead and is generally faster than a function call. The best candidates for inlining are small functions that are called frequently from a few places, or functions called with one or more compile-time constant parameters, especially those that affect `if`, `switch`, or `for` statements. You might also want to put these functions into header files, which allows automatic inlining across file boundaries even at low optimization levels. Be sure to inline all functions that only load or store a value, or use simple operators such as comparison or arithmetic operators. Large functions and functions that are called rarely are generally not good candidates for inlining. Neither are medium size functions that are called from many places.
- Avoid breaking your program into too many small functions. If you must use small functions, you can use the `-qipa` compiler option to automatically inline such functions and use other techniques to optimize calls between functions.
- ▶ **C++** Avoid virtual functions and virtual inheritance unless required for class extensibility. These language features are costly in object space and function invocation performance.

Related information in the *XL C/C++ Compiler Reference*

-  `-qisolated_call`
-  `#pragma disjoint`
-  `-qipa`

Managing memory efficiently (C++ only)

Because C++ objects are often allocated from the heap and have limited scope, memory use affects performance more in C++ programs than it does in C programs. For that reason, consider the following guidelines when you develop C++ applications:

- In a structure, declare the largest aligned members first. Members of similar alignment should be grouped together where possible.
- In a structure, place variables near each other if they are frequently used together.
- Ensure that objects that are no longer needed are freed or otherwise made available for reuse. One way to do this is to use an *object manager*. Each time you create an instance of an object, pass the pointer to that object to the object manager. The object manager maintains a list of these pointers. To access an object, you can call an object manager member function to return the information to you. The object manager can then manage memory usage and object reuse.
- Storage pools are a good way of keeping track of used memory (and reclaiming it) without having to resort to an object manager or reference counting. Do not use storage pools for objects with non-trivial destructors, because in most implementations the destructors cannot be run when the storage pool is cleared.
- Avoid copying large and complicated objects.
- Avoid performing a *deep copy* if you only need a *shallow copy*. For an object that contains pointers to other objects, a shallow copy copies only the pointers and not the objects to which they point. The result is two objects that point to the same contained object. A deep copy, however, copies the pointers and the objects they point to, as well as any pointers or objects that are contained within that object, and so on. A deep copy must be performed in multithreaded environments, because it reduces sharing and synchronization.
- Use virtual methods only when absolutely necessary.
- Use the "Resource Acquisition is Initialization" (RAII) pattern.
- Use `shared_ptr` and `weak_ptr`.

Optimizing variables

Consider the following guidelines:

- Use local variables, preferably automatic variables, as much as possible. The compiler must make several worst-case assumptions about global variables. For example, if a function uses external variables and also calls external functions, the compiler assumes that every call to an external function could use and change the value of every external variable. If you know that a global variable is not read or affected by any function call and this variable is read several times with function calls interspersed, copy the global variable to a local variable and then use this local variable.
- If you must use global variables, use static variables with file scope rather than external variables whenever possible. In a file with several related functions and static variables, the optimizer can gather and use more information about how the variables are affected.
- If you must use external variables, group external data into structures or arrays whenever it makes sense to do so. All elements of an external structure use the same base address. Do not group variables whose addresses are taken with variables whose addresses are not taken.

- Avoid taking the address of a variable. If you use a local variable as a temporary variable and must take its address, avoid reusing the temporary variable for a different purpose. Taking the address of a local variable can inhibit optimizations that would otherwise be done on calculations involving that variable.
- Use constants instead of variables where possible. The optimizer is able to do a better job reducing runtime calculations by doing them at compile time instead. For instance, if a loop body has a constant number of iterations, use constants in the loop condition to improve optimization (for (i=0; i<4; i++) can be better optimized than for (i=0; i<x; i++)). An enumeration declaration can be used to declare a named constant for maintainability.
- Use register-sized integers (long data type) for scalars to avoid sign extension instructions after each change. For large arrays of integers, consider using one-byte or two-byte integers or bit fields.
- Use the smallest floating-point precision appropriate to your computation.
- An extern variables that must be shared within a shared library but need not be accessed from outside the library must be declared with a visibility attribute or an option that limits its visibility to the library. This allows it to be accessed directly instead of via the TOC.

Related information in the XL C/C++ Compiler Reference

 -qisolated_call

Manipulating strings efficiently

The handling of string operations can affect the performance of your program.

- When you store strings into allocated storage, align the start of the string on an 8-byte or 16-byte boundary.
- Keep track of the length of your strings. If you know the length of a string, you can use mem functions instead of str functions. For example, memcpy is faster than strcpy because it does not have to search for the end of the string.
- If you are certain that the source and target do not overlap, use memcpy instead of memmove. This is because memcpy copies directly from the source to the destination, while memmove might copy the source to a temporary location in memory before copying to the destination, or it might copy in reverse order depending on the length of the string.
- When manipulating strings using mem functions, faster code can be generated if the *count* parameter is a constant rather than a variable. This is especially true for small count values.
- Make string literals read-only, whenever possible. When the same string is used multiple times, making it read-only improves certain optimization techniques, reduces memory usage, and shortens compilation time. You can explicitly set strings to read-only by using **-qro** (this is enabled by default  except when compiling with **cc** ) to avoid changing your source files.

Related information in the XL C/C++ Compiler Reference

 -qro

Optimizing expressions and program logic

Consider the following guidelines:

- If components of an expression are used in other expressions and they include function calls or there are function calls between the uses, assign the duplicated values to a local variable.

- Avoid forcing the compiler to convert numbers between integer and floating-point internal representations. For example:

```
float array[10];
float x = 1.0;
int i;
for (i = 0; i < 9; i++) {    /* No conversions needed */
    array[i] = array[i]*x;
    x = x + 1.0;
}
for (i = 0; i < 9; i++) {    /* Multiple conversions needed */
    array[i] = array[i]*i;
}
```

When you must use mixed-mode arithmetic, code the integer and floating-point arithmetic in separate computations whenever possible.

- Do not use global variables as loop indices or bounds.
- Avoid goto statements that jump into the middle of loops. Such statements inhibit certain optimizations.
- Improve the predictability of your code by making the fall-through path more probable. Code such as:

```
if (error) {handle error} else {real code}
```

should be written as:

```
if (!error) {real code} else {error}
```

- If one or two cases of a switch statement are typically executed much more frequently than other cases, break out those cases by handling them separately before the switch statement. If possible, replace the switch statement by checking whether the value is in range to be obtained from an array.
- **> C++** Use try blocks for exception handling only when necessary because they can inhibit optimization.
- Keep array index expressions as simple as possible.

Optimizing operations in 64-bit mode

The ability to handle larger amounts of data directly in physical memory rather than relying on disk I/O is perhaps the most significant performance benefit of 64-bit machines. However, some applications compiled in 32-bit mode perform better than when they are recompiled in 64-bit mode. Some reasons for this include:

- 64-bit programs are larger. The increase in program size places greater demands on physical memory.
- 64-bit long division is more time-consuming than 32-bit integer division.
- 64-bit programs that use 32-bit signed integers as array indexes or loop counts might require additional instructions to perform sign extension each time the array is referenced or the loop count is incremented.

Some ways to compensate for the performance liabilities of 64-bit programs include:

- Avoid performing mixed 32-bit and 64-bit operations. For example, adding a 32-bit data type to a 64-bit data type requires that the 32-bit be sign-extended to clear or set the upper 32-bit of the register. This slows the computation.

- Use long types instead of signed, unsigned, and plain int types for variables that will be frequently accessed, such as loop counters and array indexes. Doing so frees the compiler from having to truncate or sign-extend array references, parameters during function calls, and function results during returns.

The C++ template model

In C++, you can use a template to declare a set of related following entities:

- Classes (including structures)
- Functions
- Static data members of template classes

Each compiler implements templates according to a model that determines the meaning of a template at various stages of the translation of a program. In particular, the compiler determines what the various constructs in a template mean when the template is instantiated. Name lookup is an essential ingredient of the compilation model.

Template instantiation is a process that generates types and functions from generic template definitions. The concept of instantiation of C++ templates is fundamental but also intricate because the definitions of entities generated by a template are no longer limited to a single location in the source code. The location of the template, the location where the template is used, and the locations where the template arguments are defined all contribute to the meaning of the entity.

XL C/C++ supports *Greedy instantiation*. The compiler generates a template instantiation in each compilation unit that uses it. The linker discards the duplicates.

Related information in the XL C/C++ Compiler Reference



-qtmplinst (C++ only)

Using delegating constructors (C++11)

Before C++11, common initialization in multiple constructors of the same class cannot be concentrated in one place in a robust and maintainable manner. Starting from C++11, with the delegating constructors feature, you can concentrate common initialization in one constructor, which can make the program more readable and maintainable. Delegating constructors help reduce code size and collective size of object files.

Syntactically, *delegating constructors* and *target constructors* present the same interface as other constructors.

Consider the following points when you use the delegating constructors feature:

- Call the target constructor implementation in such a way that virtual bases, direct nonvirtual bases, and class members are initialized by the target constructor as appropriate.
- The feature has minimal impact on compile-time and runtime performance. However, use of default arguments with an existing constructor is recommended in place of a delegating constructor where possible. Without inlining and interprocedural analysis, runtime performance might degrade because of function call overhead and increased opacity.

Using rvalue references (C++11)

In C++11, you can overload functions based on the value categories of arguments and similarly have lvalueness detected by template argument deduction. You can also have an rvalue bound to an rvalue reference and modify the rvalue through the reference. This enables a programming technique with which you can reuse the resources of expiring objects and therefore improve the performance of your libraries, especially if you use generic code with class types, for example, template data structures. Additionally, the value category can be considered when writing a forwarding function.

Move semantics

When you want to optimize the use of temporary values, you can use a move operation in what is known as destructive copying. Consider the following string concatenation and assignment:

```
std::string a, b, c;  
c = a + b;
```

In this program, the compiler first stores the result of `a + b` in an internal temporary variable, that is, an rvalue.

The signature of a normal copy assignment operator is as follows:

```
string& operator = (const string&)
```

With this copy assignment operator, the assignment consists of the following steps:

1. Copy the temporary variable into `c` using a deep-copy operation.
2. Discard the temporary variable.

Deep copying the temporary variable into `c` is not efficient because the temporary variable is discarded at the next step.

To avoid the needless duplication of the temporary variable, you can implement an assignment operator that moves the variable instead of copying the variable. That is, the argument of the operator is modified by the operation. A move operation is faster because it is done through pointer manipulation, but it requires a reference through which the source variable can be manipulated. However, `a + b` is a temporary value, which is not easily differentiated from a const-qualified value in C++ before C++11 for the purposes of overload resolution.

With rvalue references, you can create a move assignment operator as follows:

```
string& operator= (string&&)
```

With this move assignment operator, the memory allocated for the underlying C-style string in the result of `a + b` is assigned to `c`. Therefore, it is not necessary to allocate new memory to hold the underlying string in `c` and to copy the contents to the new memory.

The following code can be an implementation of the `string` move assignment operator:

```
string& string::operator=(string&& str)  
{  
    // The named rvalue reference str acts like an lvalue  
    std::swap(_capacity, str._capacity);  
    std::swap(_length, str._length);  
}
```

```

    // char* _str points to a character array and is a
    // member variable of the string class
    std::swap(_str, str._str);
    return *this;
}

```

However, in this implementation, the memory originally held by the string being assigned to is not freed until `str` is destroyed. The following implementation that uses a local variable is more memory efficient:

```

string& string::operator=(string&& parm_str)
{
    // The named rvalue reference parm_str acts like an lvalue
    string sink_str;
    std::swap(sink_str, parm_str);
    std::swap(*this, sink_str);
    return *this;
}

```

In a similar manner, the following program is a possible implementation of a string concatenation operator:

```

string operator+(string&& a, const string& b)
{
    return std::move(a+=b);
}

```

Note: The `std::move` function only casts the result of `a+=b` to an rvalue reference, without moving anything. The return value is constructed using a move constructor because the expression `std::move(a+=b)` is an rvalue. The relationship between a move constructor and a copy constructor is analogous to the relationship between a move assignment operator and a copy assignment operator.

Perfect forwarding

The `std::forward` function is a helper template, much like `std::move`. It returns a reference to its function argument, with the resulting value category determined by the template type argument. In an instantiation of a forwarding function template, the value category of an argument is encoded as part of the deduced type for the related template type parameter. The deduced type is passed to the `std::forward` function.

The wrapper function in the following example is a forwarding function template that forwards to the `do_work` function. Use `std::forward` in forwarding functions on the calls to the target functions. The following example also uses the `decltype` and trailing return type features to produce a forwarding function that forwards to one of the `do_work` functions. Calling the wrapper function with any argument results in a call to a `do_work` function if a suitable overload function exists. Extra temporaries are not created and overload resolution on the forwarding call resolves to the same overload as it would if the `do_work` function were called directly.

```

struct s1 *do_work(const int&);           // #1
struct s2 *do_work(const double&);      // #2
struct s3 *do_work(int&&);              // #3
struct s4 *do_work(double&&);           // #4
template <typename T> auto wrapper(T && a)->
    decltype(do_work(std::forward<T>(*static_cast<typename std::remove_reference<T>
        ::type*>(0))))
{
    return do_work(std::forward<T>(a));
}
template <typename T> void tPtr(T *t);
int main()

```

```

{
  int x;
  double y;
  tPtr<s1>(wrapper(x));    // calls #1
  tPtr<s2>(wrapper(y));    // calls #2
  tPtr<s3>(wrapper(0));    // calls #3
  tPtr<s4>(wrapper(1.0));  // calls #4
}

```

Related information in the XL C/C++ Compiler Reference

 -qlanglvl

Using visibility attributes (IBM extension)

Visibility attributes describe whether and how an entity that is defined in one module can be referenced or used in other modules. Visibility attributes affect entities with external linkage only, and they cannot increase the visibility of other entities. By specifying visibility attributes for entities, you can export only the entities that are necessary to shared libraries. With this feature, you can get the following benefits:

- Decrease the size of shared libraries.
- Reduce the possibility of symbol collision.
- Allow more optimization for the compile and link phases.
- Improve the efficiency of dynamic linking.
- Within a shared library, allow direct access instead of via a TOC pointer.

Supported types of entities

> C++

The compiler supports visibility attributes for the following entities:

- Function
- Variable
- Structure/union/class
- Enumeration
- Template
- Namespace

C++ <

> C

The compiler supports visibility attributes for the following entities:

- Function
- Variable

Note: Data types in the C language do not have external linkage, so you cannot specify visibility attributes for C data types.

C <

Related information in the XL C/C++ Compiler Reference

-  -fvisibility
 -  -shared (-qmkshrobj)
 -  #pragma GCC visibility push, #pragma GCC visibility pop
- Related information in the *XL C/C++ Language Reference***
-  The visibility variable attribute (IBM extension)
 -  The visibility function attribute (IBM extension)
 -  The visibility type attribute (C++ only) (IBM extension)
 -  The visibility namespace attribute (C++ only) (IBM extension)

Types of visibility attributes

The following table describes different visibility attributes.

Table 15. Visibility attributes

Attribute	Description
default	Indicates that external linkage entities have the default attribute in object files. These entities are exported in shared libraries, and can be preempted.
protected	Indicates that external linkage entities have the protected attribute in object files. These entities are exported in shared libraries, but cannot be preempted.
hidden	Indicates that external linkage entities have the hidden attribute in object files. These entities are not exported in shared libraries, but their addresses can be referenced indirectly through pointers.
internal	Indicates that external linkage entities have the internal attribute in object files. These entities are not exported in shared libraries, and their addresses are not available to other modules in shared libraries.
Notes:	
<ul style="list-style-type: none"> • In this release, the hidden and internal visibility attributes are the same. The addresses of the entities that are specified with either of these visibility attributes can be referenced indirectly through pointers. 	

Example: Differences among the default, protected, hidden, and internal visibility attributes

```
//a.c
#include <stdio.h>
void __attribute__((visibility("default"))) func1(){
    printf("func1 in the shared library");
}
void __attribute__((visibility("protected"))) func2(){
    printf("func2 in the shared library");
}
void __attribute__((visibility("hidden"))) func3(){
    printf("func3 in the shared library");
}
void __attribute__((visibility("internal"))) func4(){
    printf("func4 in the shared library");
}

//a.h
extern void func1();
extern void func2();
extern void func3();
extern void func4();
```

```

//b.c
#include "a.h"
void temp(){
    func1();
    func2();
}

//b.h
extern void temp();

//main.c
#include "a.h"
#include "b.h"

void func1(){
    printf("func1 in b.c");
}
void func2(){
    printf("func2 in b.c");
}
void main(){
    temp();
    // func3(); // error
    // func4(); // error
}

```

You can use the following commands to create a shared library named `libtest.so`:

```

xlc -c -fPIC a.c b.c
xlc -shared -o libtest.so a.o b.o

```

Then, you can dynamically link `libtest.so` during run time by using the following commands:

```

xlc main.c -L. -ltest -o main
./main

```

The output of the example is as follows:

```

func1 in b.c
func2 in the shared library

```

The visibility attribute of function `func1()` is default, so it is preempted by the function with the same name in `main.c`. The visibility attribute of function `func2()` is protected, so it cannot be preempted. The compiler always calls `func2()` that is defined in the shared library `libtest.so`. The visibility attribute of function `func3()` is hidden, so it is not exported in the shared library. The compiler issues a link error to indicate that the definition of `func3()` cannot be found. The same issue is with function `func4()` whose visibility attribute is internal.

Rules of visibility attributes

Priority of visibility attributes

The visibility attributes have a priority sequence, which is default < protected < hidden < internal. You can see Example 3 and Example 9 for reference.

Rules of determining the visibility attributes



The visibility attribute of an entity is determined by the following rules:

1. If the entity has an explicitly specified visibility attribute, the specified visibility attribute takes effect.
2. Otherwise, if the entity has a pair of enclosing pragma directives, the visibility attribute that is specified by the pragma directives takes effect.
3. Otherwise, the setting of the **-fvisibility** option takes effect.

C

C++

The visibility attribute of an entity is determined by the following rules:

1. If the entity has an explicitly specified visibility attribute, the specified visibility attribute takes effect.
2. Otherwise, if the entity is a template instantiation or specialization, and the template has a visibility attribute, the visibility attribute of the entity is propagated from that of the template. See Example 1.
3. Otherwise, if the entity has any of the following enclosing contexts, the visibility attribute of this entity is propagated from that of the nearest context. See Example 2. For the details of propagation rules, see “Propagation rules (C++ only)” on page 69.
 - Structure/class
 - Enumeration
 - Namespace
 - Pragma directives

Restriction: Pragma directives do not affect the visibility attributes of class members and template specializations.

4. Otherwise, the visibility attribute of the entity is determined by the following visibility attribute settings. The visibility attribute that has the highest priority is the actual visibility attribute of the entity. See Example 3. For the priority of the visibility attributes, see Priority of visibility attributes.
 - The setting of the **-fvisibility** option.
 - The visibility attribute of the type of the entity, if the entity is a variable and its type has a visibility attribute.
 - The visibility attribute of the return type of the entity, if the entity is a function and its return type has a visibility attribute.
 - The visibility attributes of the parameter types of the entity, if the entity is a function and its parameter types have visibility attributes.
 - The visibility attributes of template arguments or template parameters of the entity, if the entity is a template and its arguments or parameters have visibility attributes.

Example 1

In the following example, template `template<typename T, typename U> B{}` has the protected visibility attribute. The visibility attribute is propagated to those of template specialization `template<> class B<char, char>{}`, partial specialization `template<typename T> class B<T, float>{}`, and all the types of template instantiations.

```
class __attribute__((visibility("internal"))) A{ vis_v_a; //internal
//protected
```

```

template<typename T, typename U>
class __attribute__((visibility("protected"))) B{
    public:
    void func(){}
};

//protected
template<>
class B<char, char>{
    public:
    void func(){}
};

//protected
template<typename T>
class B<T, float>{
    public:
    void func(){}
};

B<int, int> a;    //protected
B<A, int> b;     //protected
B<char, char> c; //protected
B<int, float> d; //protected
B<A, float> e;  //protected

int main(){
    a.func();
    b.func();
    c.func();
    d.func();
    e.func();
}

```

Example 2

In the following example, the nearest enclosing context of function `func()` is class `B`, so the visibility attribute of `func()` is propagated from that of class `B`, which is hidden. The nearest enclosing context of class `A` is the pragma directives whose setting is protected, so the visibility of class `A` is protected.

```

namespace __attribute__((visibility("internal"))) ns{
#pragma GCC visibility push(protected)
    class A{
        class __attribute__((visibility("hidden"))) B{
            int func(){};
        };
    };
#pragma GCC visibility pop
};

```

Example 3

In the following example, the visibility attribute specified by the **-fvisibility** option is protected. The type of variable `vis_v_d` is class `CD`, whose visibility attribute is default. The visibility attribute that has a higher priority of these two attributes is protected, so the actual visibility attribute of variable `vis_v_d` is protected. The same rule applies to the determination of the visibility attributes of variables `vis_v_p`, `vis_v_h`, and `vis_v_i`. For functions `vis_f_fun1`, `vis_f_fun2`, and `vis_f_fun3`, their visibility attributes are determined by those of their parameter types, return types, and the setting of the **-fvisibility** option. For template functions `vis_f_template1` and `vis_f_template2`, their visibility attributes are determined by those of their template arguments, template parameters, function

parameter types, return types, and the setting of the **-fvisibility** option. The visibility attribute that has the highest priority takes effect.

```
//The -fvisibility=protected option is specified
class __attribute__((visibility("default"))) CD {} vis_v_d; //protected
class __attribute__((visibility("protected"))) CP {} vis_v_p; //protected
class __attribute__((visibility("hidden"))) CH {} vis_v_h; //hidden
class __attribute__((visibility("internal"))) CI {} vis_v_i; //internal

void vis_f_fun1(CH a, CP b, CD c, CI d) {} //internal
void vis_f_fun2(CD a) {} //protected
CH vis_f_fun3(CI a, CP b) {} //internal

template<class T, class U> T vis_f_template1(T t, U u){}
template<class T, int N> void vis_f_template2(T t, int i){}

int main(){
    vis_f_template1<CD, CH>(vis_v_d, vis_v_p); //hidden
    vis_f_template2<CD, 10>(vis_v_p, 10); // protected
}
```

C++ <

Rules and restrictions of using the visibility attributes

When you specify visibility attributes for entities, consider the following rules and restrictions:

- You can specify visibility attributes only for entities that have external linkage. The compiler issues a warning message when you set the visibility attribute for entities with other linkages, and the specified visibility attribute is ignored. See Example 4.
- You cannot specify different visibility attributes in the same declaration or definition of an entity; otherwise, the compiler issues an error message. See Example 5.
- If an entity has more than one declaration that is specified with different visibility attributes, the visibility attribute of the entity is the first visibility attribute that the compiler processes. See Example 6.
- You cannot specify visibility attributes in the typedef statements. See Example 7.
- > C++ If type T has a visibility attribute, types T*, T&, and T&& have the same visibility attribute with that of type T. See Example 8.
- > C++ If a class and its enclosing classes do not have explicitly specified visibilities and the visibility attribute of the class has a lower priority than those of its nonstatic member types and its bases classes, the compiler issues a warning message. See Example 9. For the priority of the visibility attributes, see Priority of visibility attributes. > C++ <
- > C++ The visibility attribute of a namespace does not apply for the namespace with the same name. See Example 10. > C++ <
- > C++ If you specify a visibility attribute for a global new or delete operator, the compiler issues a warning message to ignore the visibility attribute unless the visibility attribute is default. See Example 11. > C++ <

Example 4

In this example, because m and i have internal linkage and j has no linkage, the compiler ignores the visibility attributes of variables m, i, and j.

```

static int m __attribute__((visibility("protected")));
int n __attribute__((visibility("protected")));

int main(){
    int i __attribute__((visibility("protected")));
    static int j __attribute__((visibility("protected")));
}

```

Example 5

In this example, the compiler issues an error message to indicate that you cannot specify two different visibility attributes at the same time in the definition of variable m.

```

//error
int m __attribute__((visibility("hidden"))) __attribute__((visibility("protected")));

```

Example 6

In this example, the first declaration of function fun() that the compiler processes is extern void fun() __attribute__((visibility("hidden"))), so the visibility attribute of fun() is hidden.

```

extern void fun() __attribute__((visibility("hidden")));
extern void fun() __attribute__((visibility("protected")));

int main(){
    fun();
}

```

Example 7

In this example, the visibility attribute of variable vis_v_ti is default, which is not affected by the setting in the typedef statement.

```

//The -fvisibility=default option is specified.
typedef int __attribute__((visibility("protected"))) INT;
INT vis_v_ti = 1;

```

> C++

Example 8

In this example, the visibility attribute of class CP is protected, so the visibility attribute of CP* and CP& is also protected.

```

class __attribute__((visibility("protected"))) CP {} vis_v_p;
class CP* vis_v_p_p = &vis_v_p; //protected
class CP& vis_v_l_r_p = vis_v_p; //protected

```

Example 9

In this example, the compiler accepts the default visibility attribute of class Derived1 because the visibility attribute is explicitly specified for class Derived1. The compiler also accepts the protected visibility attribute of class Derived2 because the visibility attribute is propagated from that of the enclosing class A. Class Derived3 does not have an explicitly specified visibility attribute or an enclosing class, and its visibility attribute is default. The compiler issues a warning message because the visibility attribute of class Derived3 has a lower priority than those of its parent class Base and the nonstatic member function fun().

```

//The -fvisibility=default option is specified.
//base class
struct __attribute__((visibility("hidden"))) Base{
    int vis_f_fun(){
        return 0;
    }
};

//Ok
struct __attribute__((visibility("default"))) Derived1: public Base{
    int vis_f_fun(){
        return Base::vis_f_fun();
    }
};
}vis_v_d;

//Ok
struct __attribute__((visibility("protected"))) A{
    struct Derived2: public Base{
        int vis_f_fun(){
            __attribute__((visibility("protected")))
        };
    }
};

//Warning
struct Derived3: public Base{
    //Warning
    int fun() __attribute__((visibility("protected")));
};

```

Example 10

In this example, the visibility attribute of the definition of namespace X does not apply to the extension of namespace X.

```

//The -fvisibility=default option is specified.
//namespace definition
namespace X __attribute__((visibility("protected"))){
    int a; //protected
    int b; //protected
}
//namespace extension
namespace X {
    int c; //default
    int d; //default
}
//equivalent to namespace X
namespace Y {
    int __attribute__((visibility("protected"))) a; //protected
    int __attribute__((visibility("protected"))) b; //protected
    int c; //default
    int d; //default
}

```

Example 11

In this example, the new and delete operators defined outside of class A are global functions, so the explicitly specified hidden visibility attribute does not take effect. The new and delete operations defined within class A are local ones, so you can specify visibility attributes for them.

```

#include <stddef.h>
//default
void* operator new(size_t) throw (std::bad_alloc) __attribute__((visibility("hidden")))
{
    return 0;
}

```

```

};
void operator delete(void*) throw () __attribute__((visibility("hidden"))){}

class A{
public:
//hidden
void* operator new(size_t) throw (std::bad_alloc) __attribute__((visibility("hidden")))
{
return 0;
};
void operator delete(void*) throw () __attribute__((visibility("hidden"))){}
};

```

C++ ◀

Propagation rules (C++ only)

Visibility attributes can be propagated from one entity to other entities. The following table lists all the cases for visibility propagation.

Table 16. Propagation of visibility attributes

Original entity	Destination entities	Example
Namespace	Named namespaces that are defined in the original namespace	<pre> namespace A __attribute__((visibility("hidden"))){ // Namespace B has the hidden visibility attribute, // which is propagated from namespace A. namespace B{} // The unnamed namespace does not have a visibility // attribute. namespace{} } </pre>
Namespace	Classes that are defined in the original namespace	<pre> namespace A __attribute__((visibility("hidden"))){ // Class B has the hidden visibility attribute, // which is propagated from namespace A. class B; // Object x has the hidden visibility attribute, // which is propagated from namespace A. class{} x; } </pre>
Namespace	Functions that are defined in the original namespace	<pre> namespace A __attribute__((visibility("hidden"))){ // Function fun() has the hidden visibility // attribute, which is propagated from namespace A. void fun(){}; } </pre>
Namespace	Objects that are defined in the original namespace	<pre> namespace A __attribute__((visibility("hidden"))){ // Variable m has the hidden visibility attribute, // which is propagated from namespace A. int m; } </pre>
Class	Member classes	<pre> class __attribute__((visibility("hidden"))) A{ // Class B has the hidden visibility attribute, // which is propagated from class A. class B{}; } </pre>

Table 16. Propagation of visibility attributes (continued)

Original entity	Destination entities	Example
Class	Member functions or static member variables	<pre>class __attribute__((visibility("hidden"))) A{ // Function fun() has the hidden visibility // attribute, which is propagated from class A. void fun(){}; // Static variable m has the hidden visibility // attribute, which is propagated from class A. static int m; }</pre>
Template	Template instantiations/ template specifications/ template partial specializations	<pre>template<typename T, typename U> class __attribute__((visibility("hidden"))) A{ public: void fun(){}; }; // Template instantiation class A<int, char> has the // hidden visibility attribute, which is propagated // from template class A(T,U). class A<int, char>{ public: void fun(){}; }; // Template specification // template<> class A<double, double> has the hidden // visibility attribute, which is propagated // from template class A(T,U). template<> class A<double, double>{ public: void fun(){}; }; // Template partial specification // template<typename T> class A<T, char> has the // hidden visibility attribute, which is propagated // from template class A(T,U). template<typename T> class A<T, char>{ public: void fun(){}; };</pre>
Template argument/ parameter	Template instantiations/ template specifications/ template partial specializations	<pre>template<typename T> void fun1(){} template<typename T> void fun2(T){} class M __attribute__((visibility("hidden"))){} m; // Template instantiation fun1<M>() has the hidden // visibility attribute, which is propagated from // template argument M. fun1<M>(); // Template instantiation fun2<M>(M) has the hidden // visibility attribute, which is propagated from // template parameter m. fun2(m); // Template specification fun1<M>() has the hidden // visibility attribute, which is propagated from // template argument M. template<> void fun1<M>();</pre>

Table 16. Propagation of visibility attributes (continued)

Original entity	Destination entities	Example
Inline function	Static local variables	<pre>inline void __attribute__((visibility("hidden"))) fun(){ // Variable m has the hidden visibility attribute, // which is propagated from inline function fun(). static int m = 4; }</pre>
Type	Entities of the original type	<pre>class __attribute__((visibility("hidden"))) A {}; // Object x has the hidden visibility attribute, // which is propagated from class A. class A x;</pre>
Function return type	Function	<pre>class __attribute__((visibility("hidden"))) A{}; // Function fun() has the hidden visibility attribute, // which is propagated from function return type A. A fun();</pre>
Function parameter type	Function	<pre>class __attribute__((visibility("hidden"))) A{}; // Function fun(class A) has the hidden visibility // attribute, which is propagated from function // parameter type A. void fun(class A);</pre>

Specifying visibility attributes using the `-fvisibility` option

You can use the `-fvisibility` option to globally set visibility attributes for external linkage entities in your program. The entities have the visibility attribute that is specified by the `-fvisibility` option if they do not get visibility attributes from pragma directives, explicitly specified attributes, or propagation rules.

Specifying visibility attributes using pragma preprocessor directives

You can selectively set visibility attributes for entities by using pairs of the `#pragma GCC visibility push` and `#pragma GCC visibility pop` preprocessor directives throughout your source program.

The compiler supports nested visibility pragma preprocessor directives. If entities are included in several pairs of the nested `#pragma GCC visibility push` and `#pragma GCC visibility pop` directives, the nearest pair of directives takes effect. See Example 1.

You must not specify the visibility pragma directives for header files. Otherwise, your program might exhibit undefined behaviors. See Example 2.

> C++ Visibility pragma directives `#pragma GCC visibility push` and `#pragma GCC visibility pop` affect only namespace-scope declarations. Class members and template specializations are not affected. See Example 3 and Example 4. **C++ <**

Examples

Example 1

In this example, the function and variables have the visibility attributes that are specified by their nearest pairs of pragma preprocessor directives.

```

#pragma GCC visibility push(default)
namespace ns
{
    void vis_f_fun() {} //default
# pragma GCC visibility push(internal)
    int vis_v_i; //internal
# pragma GCC visibility push(protected)
    int vis_v_j; //protected
# pragma GCC visibility push(hidden)
    int vis_v_k; //hidden
# pragma GCC visibility pop
# pragma GCC visibility pop
# pragma GCC visibility pop
}
#pragma GCC visibility pop

```

Example 2

In this example, the compiler issues a link error message to indicate that the definition of the `printf()` library function cannot be found.

```

#pragma GCC visibility push(hidden)
#include <stdio.h>
#pragma GCC visibility pop

int main(){
    printf("hello world!");
    return 0;
}

```

► C++

Example 3

In this example, the visibility attribute of class members `vis_v_i` and `vis_f_fun()` is hidden. The visibility attribute is propagated from that of the class, but is not affected by the `pragma` directives.

```

class __attribute__((visibility("hidden"))) A{
#pragma GCC visibility push(protected)
public:
    static int vis_v_i;
    void vis_f_fun() {}
#pragma GCC visibility pop
} vis_v_a;

```

Example 4

In this example, the visibility attribute of function `vis_f_fun()` is hidden. The visibility attribute is propagated from that of the template specialization or partial specialization, but is not affected by the `pragma` directives.

```

namespace ns{
#pragma GCC visibility push(hidden)
template <typename T, typename U> class TA{
public:
    void vis_f_fun(){}
};
#pragma GCC visibility pop

#pragma GCC visibility push(protected)
//The visibility attribute of the template specialization is hidden.
template <> class TA<char, char>{
public:

```

```
    void vis_f_fun(){}  
};  
#pragma GCC visibility pop  
  
#pragma GCC visibility push(default)  
//The visibility attribute of the template partial specialization is hidden.  
template <typename T> class TA<T, long>{  
    public:  
    void vis_f_fun(){}  
};  
#pragma GCC visibility pop
```

C++ ◀

Chapter 8. Exploiting POWER9 technology

Releases of IBM XL C/C++ for Linux, starting from V13.1.6, provide full exploitation of the POWER9 technology. The POWER9 exploitation features include compiler options, high-performance libraries, and built-in functions.

POWER9 compiler options

Use the `-mcpu` (`-qarch`) and `-mtune` (`-qtune`) options to utilize the POWER9 technology.

`-mcpu=pwr9` or `-mcpu=power9` (`-qarch=pwr9`)

Specifying `-mcpu=pwr9` or `-mcpu=power9` (`-qarch=pwr9`) produces object code that contains instructions that utilize the POWER9 technology.

For more information, see `-mcpu=pwr9` or `-mcpu=power9` (`-qarch=pwr9`).

`-mtune=pwr9` or `-mtune=power9` (`-qtune=pwr9`)

If you specify `-mtune=pwr9` or `-mtune=power9` (`-qtune=pwr9`), optimizations are tuned to utilize the POWER9 technology.

For more information, see `-mtune=pwr9` or `-mtune=power9` (`-qtune=pwr9`).

High-performance libraries that are tuned for POWER9

Use the MASS vector library and SIMD library to exploit the POWER9 architecture.

MASS libraries

The vector library `libmassvp9.a` and the SIMD library `libmass_simdp9.a` contain functions that are tuned for the POWER9 architecture.

For how to use MASS libraries, see Using the Mathematical Acceleration Subsystem (MASS) libraries.

POWER9 built-in functions

A number of built-in functions are available to exploit the POWER9 architecture.

Note: The following built-in functions are valid only when `-qarch(-mcpu)` is set to utilize the POWER9 technology.

Fixed-point built-in functions

`__builtin_addex`

Adds Extended using Overflow Bit.

`__builtin_cmpeqb`

Compares the corresponding bytes of the given parameters and returns the result.

`__builtin_cmprb`

Compares the ranged byte.

`__builtin_setb`

Sets boolean extension.

`__builtin_maddhd`

Fixed-point Multiply-Add signed high doubleword.

__builtin_maddhdu
Fixed-point Multiply-Add high doubleword unsigned.

__builtin_maddld
Fixed-point Multiply-Add low doubleword.

__builtin_darn, __builtin_darn_32, __builtin_darn_raw
Delivers a random number.

Binary floating-point built-in functions

**__builtin_compare_exp_uo, __builtin_compare_exp_lt,
__builtin_compare_exp_eq, __builtin_compare_exp_gt**
Compares the exponents of two parameters.

__builtin_extract_exp
Returns the exponent of the given parameter.

__builtin_extract_sig
Returns the significand of the given parameter.

__builtin_insert_exp
Replaces the exponent of the first parameter with the second parameter and returns the result.

__builtin_test_data_class
Determines the data class of the given parameter.

Binary-coded decimal built-in functions

__builtin_bcdcopysign
Returns the decimal value of the first parameter combined with the sign code of the second parameter.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_bcdsetsign
Set the sign code of the input parameter in packed decimal format.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_bcdshift
Decimal shift.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_bcdshiftround
Decimal shift and round.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_bcdtruncate
Decimal truncate.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_bcdunsignedshift
Decimal unsigned shift.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_bcdunsignedtruncate
Decimal unsigned truncate.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_national2packed

Transforms the national decimal value to packed decimal format.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_packed2national

Transforms the packed decimal value to national decimal format.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_packed2zoned

Transforms the packed decimal value to zoned decimal format.

This built-in function is valid only when you include the `altivec.h` file.

__builtin_zoned2packed

Transforms the zoned decimal value to packed decimal format.

This built-in function is valid only when you include the `altivec.h` file.

Vector built-in functions

vec_absd

Returns a vector that contains the absolute difference of the corresponding elements of the given vectors.

This built-in function is valid only when you include the `altivec.h` file.

vec_cmpne

Returns a vector containing the results of comparing each set of the corresponding elements of the given vectors for inequality.

This built-in function is valid only when you include the `altivec.h` file.

vec_cmpnez

Returns a vector that contains the results of comparing each set of the corresponding elements of the given vectors for inequality, or the results of testing the corresponding element of given vectors for the value of zero.

This built-in function is valid only when you include the `altivec.h` file.

vec_cntlz

Counts the most significant zero bits of each element of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_cntlz_1sbb

Counts the leading byte elements of the given vector that have a least significant bit of 0.

This built-in function is valid only when you include the `altivec.h` file.

vec_cnttz

Counts the least significant zero bits of each element of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_cnttz_1sbb

Counts the trailing byte elements of the given vector that have a least significant bit of 0.

This built-in function is valid only when you include the `altivec.h` file.

vec_extract_exp

Returns a vector that contains the exponent of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_extract_sig

Returns a vector that contains the significand of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_extsbd

Sign-extends the rightmost byte of each doubleword element of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_extsbw

Sign-extends the rightmost byte of each word element of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_extshd

Sign-extends the rightmost halfword of each doubleword element of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_extshw

Sign-extends the rightmost halfword of each word element of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_extswd

Sign-extends the rightmost word of each doubleword element of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_first_match_index

Compares each set of the corresponding elements of the given vectors and returns the first position of equality.

This built-in function is valid only when you include the `altivec.h` file.

vec_first_match_or_eos_index

Compares each set of the corresponding elements of the given vectors and returns the first position of equality or the position of the end-of-string terminator

This built-in function is valid only when you include the `altivec.h` file.

vec_first_mismatch_index

Compares each set of the corresponding elements of the given vectors and returns the first position of inequality.

This built-in function is valid only when you include the `altivec.h` file.

vec_first_mismatch_or_eos_index

Compares each set of the corresponding elements of the given vectors and returns the first position of inequality or the position of the end-of-string terminator

This built-in function is valid only when you include the `altivec.h` file.

vec_insert_exp

Returns a vector that combines the exponents of elements from one vector with the signs and the significands of elements from another vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_load_splats

Loads a 4-byte element from the memory address specified by the displacement

This built-in function is valid only when you include the `altivec.h` file.

vec_parity_lsbb

Returns a vector that computes parity on the least significant bit of each byte of each element of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_rlmi

Returns a vector that contains each element of the given vector rotated left and inserted under a mask into another vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_rlnm

Returns a vector that contains each element of the given vector rotated left and intersected with a mask.

This built-in function is valid only when you include the `altivec.h` file.

vec_slv

Left shifts the elements of a given vector by a given number of bits.

This built-in function is valid only when you include the `altivec.h` file.

vec_srv

Right-shifts the elements of a given vector by a given number of bits.

This built-in function is valid only when you include the `altivec.h` file.

vec_test_data_class

Determines the data class of the elements of the given vector.

This built-in function is valid only when you include the `altivec.h` file.

vec_xl_len

Returns a vector that loads a given number of bytes from the given address.

This built-in function is valid only when you include the `altivec.h` file.

vec_xl_len_r

Loads a string of bytes into vector register, right-justified. Sets the leftmost elements (*16-`cnt`*) to 0.

This built-in function is valid only when you include the `altivec.h` file.

vec_xst_len

Stores a given byte length of a vector to a given address.

This built-in function is valid only when you include the `altivec.h` file.

vec_xst_len_r

Stores a right-justified string of bytes.

This built-in function is valid only when you include the `altivec.h` file.

Chapter 9. Using the high performance libraries

IBM XL C/C++ for Linux, V13.1.6 is shipped with a set of libraries for high-performance mathematical computing:

- The Mathematical Acceleration Subsystem (MASS) is a set of libraries of tuned mathematical intrinsic functions that provide improved performance over the corresponding standard system math library functions. MASS is described in “Using the Mathematical Acceleration Subsystem (MASS) libraries.”
- The Basic Linear Algebra Subprograms (BLAS) are a set of routines that provide matrix/vector multiplication functions tuned for PowerPC architectures. The BLAS functions are described in “Using the Basic Linear Algebra Subprograms - BLAS” on page 92.

Using the Mathematical Acceleration Subsystem (MASS) libraries

XL C/C++ is shipped with a set of Mathematical Acceleration Subsystem (MASS) libraries for high-performance mathematical computing.

The MASS libraries consist of a library of scalar C/C++ functions described in “Using the scalar library” on page 82, a set of vector libraries tuned for specific architectures described in “Using the vector libraries” on page 84, and a set of SIMD libraries tuned for specific architectures described in “Using the SIMD libraries” on page 88. The functions contained in both scalar and vector libraries are automatically called at certain levels of optimization, but you can also call them explicitly in your programs. Note that accuracy and exception handling might not be identical in MASS functions and system library functions.

The MASS functions must run with the default rounding mode and floating-point exception trapping settings.

When you compile programs with any of the following sets of options:

- **-qhot -qignerrno -qnostrict**
- **-qhot -qignerrno -qstrict=no library**
- **-qhot -O3**
- **-O4**
- **-O5**

the compiler automatically attempts to vectorize calls to system math functions by calling the equivalent MASS vector functions (with the exceptions of functions `vdnint`, `vdint`, `vcosin`, `vscosin`, `vqdr`, `vsqdr`, `vrqdr`, `vsrqdr`, `vpopcnt4`, `vpopcnt8`, `vexp2`, `vexp2m1`, `vsexp2`, `vsexp2m1`, `vlog2`, `vlog21p`, `vslog2`, and `vslog21p`). If it cannot vectorize, it automatically tries to call the equivalent MASS scalar functions. For automatic vectorization or scalarization, the compiler uses versions of the MASS functions contained in the XLOPT library `libxlopt.a`.

In addition to any of the preceding sets of options, when the **-qipa** option is in effect, if the compiler cannot vectorize, it tries to inline the MASS scalar functions before deciding to call them.

“Compiling and linking a program with MASS” on page 91 describes how to compile and link a program that uses the MASS libraries, and how to selectively use the MASS scalar library functions in conjunction with the regular system libraries.

Related external information

 Mathematical Acceleration Subsystem website, available at <http://www.ibm.com/software/awdtools/mass/>

Using the scalar library

The MASS scalar library `libmass.a` contains an accelerated set of frequently used math intrinsic functions that provide improved performance over the corresponding standard system library functions. The MASS scalar functions are used when you explicitly link `libmass.a`.

If you want to explicitly call the MASS scalar functions, you can take the following steps:

1. Provide the prototypes for the functions by including `math.h` and `mass.h` in your source files.
2. Link the MASS scalar library with your application. For instructions, see “Compiling and linking a program with MASS” on page 91.

The MASS scalar functions accept double-precision parameters and return a double-precision result, or accept single-precision parameters and return a single-precision result, except `sincos` which gives 2 double-precision results. They are summarized in Table 17.

Table 17. MASS scalar functions

Double-precision function	Single-precision function	Description	Double-precision function prototype	Single-precision function prototype
<code>acos</code>	<code>acosf</code>	Returns the arccosine of x	<code>double acos (double x);</code>	<code>float acosf (float x);</code>
<code>acosh</code>	<code>acoshf</code>	Returns the hyperbolic arccosine of x	<code>double acosh (double x);</code>	<code>float acoshf (float x);</code>
	<code>anint</code>	Returns the rounded integer value of x		<code>float anint (float x);</code>
<code>asin</code>	<code>asinf</code>	Returns the arcsine of x	<code>double asin (double x);</code>	<code>float asinf (float x);</code>
<code>asinh</code>	<code>asinhf</code>	Returns the hyperbolic arcsine of x	<code>double asinh (double x);</code>	<code>float asinhf (float x);</code>
<code>atan2</code>	<code>atan2f</code>	Returns the arctangent of x/y	<code>double atan2 (double x, double y);</code>	<code>float atan2f (float x, float y);</code>
<code>atan</code>	<code>atanf</code>	Returns the arctangent of x	<code>double atan (double x);</code>	<code>float atanf (float x);</code>
<code>atanh</code>	<code>atanhf</code>	Returns the hyperbolic arctangent of x	<code>double atanh (double x);</code>	<code>float atanhf (float x);</code>
<code>cbrt</code>	<code>cbrtf</code>	Returns the cube root of x	<code>double cbrt (double x);</code>	<code>float cbrtf (float x);</code>
<code>copysign</code>	<code>copysignf</code>	Returns x with the sign of y	<code>double copysign (double x, double y);</code>	<code>float copysignf (float x);</code>
<code>cos</code>	<code>cosf</code>	Returns the cosine of x	<code>double cos (double x);</code>	<code>float cosf (float x);</code>

Table 17. MASS scalar functions (continued)

Double-precision function	Single-precision function	Description	Double-precision function prototype	Single-precision function prototype
cosh	coshf	Returns the hyperbolic cosine of x	double cosh (double x);	float coshf (float x);
cosisin		Returns a complex number with the real part the cosine of x and the imaginary part the sine of x.	double_Complex cosisin (double);	
dnint		Returns the nearest integer to x (as a double)	double dnint (double x);	
erf	erff	Returns the error function of x	double erf (double x);	float erff (float x);
erfc	erfcf	Returns the complementary error function of x	double erfc (double x);	float erfcf (float x);
exp	expf	Returns the exponential function of x	double exp (double x);	float expf (float x);
expm1	expm1f	Returns (the exponential function of x) - 1	double expm1 (double x);	float expm1f (float x);
hypot	hypotf	Returns the square root of $x^2 + y^2$	double hypot (double x, double y);	float hypotf (float x, float y);
lgamma	lgammaf	Returns the natural logarithm of the absolute value of the Gamma function of x	double lgamma (double x);	float lgammaf (float x);
log	logf	Returns the natural logarithm of x	double log (double x);	float logf (float x);
log10	log10f	Returns the base 10 logarithm of x	double log10 (double x);	float log10f (float x);
log1p	log1pf	Returns the natural logarithm of (x + 1)	double log1p (double x);	float log1pf (float x);
pow	powf	Returns x raised to the power y	double pow (double x, double y);	float powf (float x, float y);
rsqrt		Returns the reciprocal of the square root of x	double rsqrt (double x);	
sin	sinf	Returns the sine of x	double sin (double x);	float sinf (float x);
sincos		Sets *s to the sine of x and *c to the cosine of x	void sincos (double x, double* s, double* c);	
sinh	sinhf	Returns the hyperbolic sine of x	double sinh (double x);	float sinhf (float x);
sqrt		Returns the square root of x	double sqrt (double x);	
tan	tanf	Returns the tangent of x	double tan (double x);	float tanf (float x);
tanh	tanhf	Returns the hyperbolic tangent of x	double tanh (double x);	float tanhf (float x);

Notes:

- The trigonometric functions (`sin`, `cos`, `tan`) return NaN (Not-a-Number) for large arguments (where the absolute value is greater than $2^{50}\pi$).
- In some cases, the MASS functions are not as accurate as the ones in the `libm.a` library, and they might handle edge cases differently (`sqrt(Inf)`, for example).
- For accuracy comparisons with `libm.a`, see Product documentation (manuals) in the Product support content section of the Mathematical Acceleration Subsystem website.

Related external information

 Mathematical Acceleration Subsystem website, available at <http://www.ibm.com/software/awdtools/mass/>

Using the vector libraries

If you want to explicitly call any of the MASS vector functions, you can do so by including `massv.h` in your source files and linking your application with the appropriate vector library. Information about linking is provided in “Compiling and linking a program with MASS” on page 91.

The vector libraries shipped with XL C/C++ are listed below:

libmassv.a

The generic vector library that runs on any supported POWER[®] processor. Unless your application requires this portability, use the appropriate architecture-specific library below for maximum performance.

libmassvp8.a

Contains functions that are tuned for the POWER8 architecture.

libmassvp9.a

Contains functions that are tuned for the POWER9 architecture.

The single-precision and double-precision floating-point functions contained in the vector libraries are summarized in Table 18 on page 85. The integer functions contained in the vector libraries are summarized in Table 19 on page 87. Note that in C and C++ applications, only call by reference is supported, even for scalar arguments.

With the exception of a few functions (described in the following paragraph), all of the floating-point functions in the vector libraries accept three parameters:

- A double-precision (for double-precision functions) or single-precision (for single-precision functions) vector output parameter
- A double-precision (for double-precision functions) or single-precision (for single-precision functions) vector input parameter
- An integer vector-length parameter.

The functions are of the form

function_name (*y*,*x*,*n*)

where *y* is the target vector, *x* is the source vector, and *n* is the vector length. The parameters *y* and *x* are assumed to be double-precision for functions with the prefix *v*, and single-precision for functions with the prefix *vs*. As an example, the following code outputs a vector *y* of length 500 whose elements are `exp(x[i])`, where `i=0,...,499`:

```
#include <massv.h>

double x[500], y[500];
int n;
n = 500;
...
vexp (y, x, &n);
```

The functions `vdiv`, `vsincos`, `vpow`, and `vatan2` (and their single-precision versions, `vsdiv`, `vssincos`, `vspow`, and `vsatan2`) take four arguments. The functions `vdiv`, `vpow`, and `vatan2` take the arguments (z,x,y,n) . The function `vdiv` outputs a vector z whose elements are $x[i]/y[i]$, where $i=0,..,*n-1$. The function `vpow` outputs a vector z whose elements are $x[i]^{y[i]}$, where $i=0,..,*n-1$. The function `vatan2` outputs a vector z whose elements are $\text{atan}(x[i]/y[i])$, where $i=0,..,*n-1$. The function `vsincos` takes the arguments (y,z,x,n) , and outputs two vectors, y and z , whose elements are $\sin(x[i])$ and $\cos(x[i])$, respectively.

In `vcosisin(y,x,n)` and `vscosisin(y,x,n)`, x is a vector of n elements and the function outputs a vector y of n `_Complex` elements of the form $(\cos(x[i]),\sin(x[i]))$.

Table 18. MASS floating-point vector functions

Double-precision function	Single-precision function	Description	Double-precision function prototype	Single-precision function prototype
<code>vacos</code>	<code>vsacos</code>	Sets $y[i]$ to the arc cosine of $x[i]$, for $i=0,..,*n-1$	<code>void vacos (double y[], double x[], int *n);</code>	<code>void vsacos (float y[], float x[], int *n);</code>
<code>vacosh</code>	<code>vsacosh</code>	Sets $y[i]$ to the hyperbolic arc cosine of $x[i]$, for $i=0,..,*n-1$	<code>void vacosh (double y[], double x[], int *n);</code>	<code>void vsacosh (float y[], float x[], int *n);</code>
<code>vasin</code>	<code>vsasin</code>	Sets $y[i]$ to the arc sine of $x[i]$, for $i=0,..,*n-1$	<code>void vasin (double y[], double x[], int *n);</code>	<code>void vsasin (float y[], float x[], int *n);</code>
<code>vasinh</code>	<code>vsasinh</code>	Sets $y[i]$ to the hyperbolic arc sine of $x[i]$, for $i=0,..,*n-1$	<code>void vasin (double y[], double x[], int *n);</code>	<code>void vsasinh (float y[], float x[], int *n);</code>
<code>vatan2</code>	<code>vsatan2</code>	Sets $z[i]$ to the arc tangent of $x[i]/y[i]$, for $i=0,..,*n-1$	<code>void vatan2 (double z[], double x[], double y[], int *n);</code>	<code>void vsatan2 (float z[], float x[], float y[], int *n);</code>
<code>vatanh</code>	<code>vsatanh</code>	Sets $y[i]$ to the hyperbolic arc tangent of $x[i]$, for $i=0,..,*n-1$	<code>void vatanh (double y[], double x[], int *n);</code>	<code>void vsatanh (float y[], float x[], int *n);</code>
<code>vcbrt</code>	<code>vsqrt</code>	Sets $y[i]$ to the cube root of $x[i]$, for $i=0,..,*n-1$	<code>void vcbrt (double y[], double x[], int *n);</code>	<code>void vsqrt (float y[], float x[], int *n);</code>
<code>vcos</code>	<code>vscos</code>	Sets $y[i]$ to the cosine of $x[i]$, for $i=0,..,*n-1$	<code>void vcos (double y[], double x[], int *n);</code>	<code>void vscos (float y[], float x[], int *n);</code>
<code>vcosh</code>	<code>vscosh</code>	Sets $y[i]$ to the hyperbolic cosine of $x[i]$, for $i=0,..,*n-1$	<code>void vcosh (double y[], double x[], int *n);</code>	<code>void vscosh (float y[], float x[], int *n);</code>
<code>vcosisin</code>	<code>vscosisin</code>	Sets the real part of $y[i]$ to the cosine of $x[i]$ and the imaginary part of $y[i]$ to the sine of $x[i]$, for $i=0,..,*n-1$	<code>void vcosisin (double _Complex y[], double x[], int *n);</code>	<code>void vscosisin (float _Complex y[], float x[], int *n);</code>
<code>vdint</code>		Sets $y[i]$ to the integer truncation of $x[i]$, for $i=0,..,*n-1$	<code>void vdint (double y[], double x[], int *n);</code>	

Table 18. MASS floating-point vector functions (continued)

Double-precision function	Single-precision function	Description	Double-precision function prototype	Single-precision function prototype
vdiv	vsdiv	Sets $z[i]$ to $x[i]/y[i]$, for $i=0,..,*n-1$	void vdiv (double z[], double x[], double y[], int *n);	void vsdiv (float z[], float x[], float y[], int *n);
vdnint		Sets $y[i]$ to the nearest integer to $x[i]$, for $i=0,..,*n-1$	void vdnint (double y[], double x[], int *n);	
verf	vserf	Sets $y[i]$ to the error function of $x[i]$, for $i=0,..,*n-1$	void verf (double y[], double x[], int *n)	void vserf (float y[], float x[], int *n)
verfc	vserfc	Sets $y[i]$ to the complementary error function of $x[i]$, for $i=0,..,*n-1$	void verfc (double y[], double x[], int *n)	void vserfc (float y[], float x[], int *n)
vexp	vsexp	Sets $y[i]$ to the exponential function of $x[i]$, for $i=0,..,*n-1$	void vexp (double y[], double x[], int *n);	void vsexp (float y[], float x[], int *n);
vexp2	vsexp2	Sets $y[i]$ to 2 raised to the power of $x[i]$, for $i=1,..,*n-1$	void vexp2 (double y[], double x[], int *n);	void vsexp2 (float y[], float x[], int *n);
vexpm1	vsexpm1	Sets $y[i]$ to (the exponential function of $x[i]$)-1, for $i=0,..,*n-1$	void vexpm1 (double y[], double x[], int *n);	void vsexpm1 (float y[], float x[], int *n);
vexp2m1	vsexp2m1	Sets $y[i]$ to (2 raised to the power of $x[i]$) - 1, for $i=1,..,*n-1$	void vexp2m1 (double y[], double x[], int *n);	void vsexp2m1 (float y[], float x[], int *n);
vhypot	vshypot	Sets $z[i]$ to the square root of the sum of the squares of $x[i]$ and $y[i]$, for $i=0,..,*n-1$	void vhypot (double z[], double x[], double y[], int *n);	void vshypot (float z[], float x[], float y[], int *n);
vlog	vslog	Sets $y[i]$ to the natural logarithm of $x[i]$, for $i=0,..,*n-1$	void vlog (double y[], double x[], int *n);	void vslog (float y[], float x[], int *n);
vlog2	vslog2	Sets $y[i]$ to the base-2 logarithm of $x[i]$, for $i=1,..,*n-1$	void vlog2 (double y[], double x[], int *n);	void vslog2 (float y[], float x[], int *n);
vlog10	vslog10	Sets $y[i]$ to the base-10 logarithm of $x[i]$, for $i=0,..,*n-1$	void vlog10 (double y[], double x[], int *n);	void vslog10 (float y[], float x[], int *n);
vlog1p	vslog1p	Sets $y[i]$ to the natural logarithm of $(x[i]+1)$, for $i=0,..,*n-1$	void vlog1p (double y[], double x[], int *n);	void vslog1p (float y[], float x[], int *n);
vlog21p	vslog21p	Sets $y[i]$ to the base-2 logarithm of $(x[i]+1)$, for $i=1,..,*n-1$	void vlog21p (double y[], double x[], int *n);	void vslog21p (float y[], float x[], int *n);
vpow	vspow	Sets $z[i]$ to $x[i]$ raised to the power $y[i]$, for $i=0,..,*n-1$	void vpow (double z[], double x[], double y[], int *n);	void vspow (float z[], float x[], float y[], int *n);
vqdrft	vsqdrft	Sets $y[i]$ to the fourth root of $x[i]$, for $i=0,..,*n-1$	void vqdrft (double y[], double x[], int *n);	void vsqdrft (float y[], float x[], int *n);

Table 18. MASS floating-point vector functions (continued)

Double-precision function	Single-precision function	Description	Double-precision function prototype	Single-precision function prototype
vrcbrt	vsrbrt	Sets $y[i]$ to the reciprocal of the cube root of $x[i]$, for $i=0,\dots,*n-1$	void vrcbrt (double y[], double x[], int *n);	void vsrbrt (float y[], float x[], int *n);
vrec	vsrec	Sets $y[i]$ to the reciprocal of $x[i]$, for $i=0,\dots,*n-1$	void vrec (double y[], double x[], int *n);	void vsrec (float y[], float x[], int *n);
vrqdrft	vsrqdrft	Sets $y[i]$ to the reciprocal of the fourth root of $x[i]$, for $i=0,\dots,*n-1$	void vrqdrft (double y[], double x[], int *n);	void vsrqdrft (float y[], float x[], int *n);
vrsqrt	vsrsqrt	Sets $y[i]$ to the reciprocal of the square root of $x[i]$, for $i=0,\dots,*n-1$	void vrsqrt (double y[], double x[], int *n);	void vsrsqrt (float y[], float x[], int *n);
vsin	vssin	Sets $y[i]$ to the sine of $x[i]$, for $i=0,\dots,*n-1$	void vsin (double y[], double x[], int *n);	void vssin (float y[], float x[], int *n);
vsincos	vssincos	Sets $y[i]$ to the sine of $x[i]$ and $z[i]$ to the cosine of $x[i]$, for $i=0,\dots,*n-1$	void vsincos (double y[], double z[], double x[], int *n);	void vssincos (float y[], float z[], float x[], int *n);
vsinh	vssinh	Sets $y[i]$ to the hyperbolic sine of $x[i]$, for $i=0,\dots,*n-1$	void vsinh (double y[], double x[], int *n);	void vssinh (float y[], float x[], int *n);
vsqrt	vssqrt	Sets $y[i]$ to the square root of $x[i]$, for $i=0,\dots,*n-1$	void vsqrt (double y[], double x[], int *n);	void vssqrt (float y[], float x[], int *n);
vtan	vstan	Sets $y[i]$ to the tangent of $x[i]$, for $i=0,\dots,*n-1$	void vtan (double y[], double x[], int *n);	void vstan (float y[], float x[], int *n);
vtanh	vstanh	Sets $y[i]$ to the hyperbolic tangent of $x[i]$, for $i=0,\dots,*n-1$	void vtanh (double y[], double x[], int *n);	void vstanh (float y[], float x[], int *n);

Integer functions are of the form *function_name* ($x[], *n$), where $x[]$ is a vector of 4-byte (for vpopcnt4) or 8-byte (for vpopcnt8) numeric objects (integral or floating-point), and $*n$ is the vector length.

Table 19. MASS integer vector library functions

Function	Description	Prototype
vpopcnt4	Returns the total number of 1 bits in the concatenation of the binary representation of $x[i]$, for $i=0,\dots,*n-1$, where x is a vector of 32-bit objects.	unsigned int vpopcnt4 (void *x, int *n)
vpopcnt8	Returns the total number of 1 bits in the concatenation of the binary representation of $x[i]$, for $i=0,\dots,*n-1$, where x is a vector of 64-bit objects.	unsigned int vpopcnt8 (void *x, int *n)

Overlap of input and output vectors

In most applications, the MASS vector functions are called with disjoint input and output vectors; that is, the two vectors do not overlap in memory. Another common usage scenario is to call them with the same vector for both input and output parameters (for example, `vsin (y, y, &n)`). Other kinds of overlap (where

input and output vectors are neither disjoint nor identical) should be avoided, since they might produce unexpected results:

- For calls to vector functions that take one input and one output vector (for example, `vsin (y, x, &n)`):
The vectors `x[0:n-1]` and `y[0:n-1]` must be either disjoint or identical, or unexpected results might be obtained.
- For calls to vector functions that take two input vectors (for example, `vatan2 (y, x1, x2, &n)`):
The previous restriction applies to both pairs of vectors `y,x1` and `y,x2`. That is, `y[0:n-1]` and `x1[0:n-1]` must be either disjoint or identical; and `y[0:n-1]` and `x2[0:n-1]` must be either disjoint or identical.
- For calls to vector functions that take two output vectors (for example, `vsincos (y1, y2, x, &n)`):
The above restriction applies to both pairs of vectors `y1,x` and `y2,x`. That is, `y1[0:n-1]` and `x[0:n-1]` must be either disjoint or identical; and `y2[0:n-1]` and `x[0:n-1]` must be either disjoint or identical. Also, the vectors `y1[0:n-1]` and `y2[0:n-1]` must be disjoint.

Alignment of input and output vectors

To get the best performance from the vector libraries, align the input and output vectors on 8-byte (or better, 16-byte) boundaries.

Consistency of MASS vector functions

All the functions in the MASS vector libraries are consistent, in the sense that a given input value will always produce the same result, regardless of its position in the vector, and regardless of the vector length.

Related information in the *XL C/C++ Compiler Reference*



-D

Related external information



Mathematical Acceleration Subsystem website, available at <http://www.ibm.com/software/awdtools/mass/>

Using the SIMD libraries

The MASS SIMD library contains a set of frequently used math intrinsic functions that provide improved performance over the corresponding standard system library functions.

The SIMD libraries shipped with XL C/C++ are listed below:

libmass_simdp8.a

Contains functions that are tuned for the POWER8 architecture.

libmass_simdp9.a

Contains functions that are tuned for the POWER9 architecture.

If you want to use the MASS SIMD functions, follow the procedure below:

1. Provide the prototypes for the functions by including `mass_simd.h` in your source files.
2. Link the MASS SIMD library `libmass_simdp8.a` with your application. For instructions, see “Compiling and linking a program with MASS” on page 91.

The single-precision MASS SIMD functions accept single-precision arguments and return single-precision results. Likewise, the double-precision MASS SIMD functions accept double-precision arguments and return double-precision results. They are summarized in Table 20.

Table 20. MASS SIMD functions

Double-precision function	Single-precision function	Description	Double-precision function prototype	Single-precision function prototype
acosd2	acosf4	Computes the arc cosine of each element of vx.	vector double acosd2 (vector double vx);	vector float acosf4 (vector float vx);
acoshd2	acoshf4	Computes the arc hyperbolic cosine of each element of vx.	vector double acoshd2 (vector double vx);	vector float acoshf4 (vector float vx);
asind2	asinf4	Computes the arc sine of each element of vx.	vector double asind2 (vector double vx);	vector float asinf4 (vector float vx);
asinhd2	asinhf4	Computes the arc hyperbolic sine of each element of vx.	vector double asinhd2 (vector double vx);	vector float asinhf4 (vector float vx);
atand2	atanf4	Computes the arc tangent of each element of vx.	vector double atand2 (vector double vx);	vector float atanf4 (vector float vx);
atan2d2	atan2f4	Computes the arc tangent of each element of vx/vy.	vector double atan2d2 (vector double vx, vector double vy);	vector float atan2f4 (vector float vx, vector float vy);
atanhd2	atanhf4	Computes the arc hyperbolic tangent of each element of vx.	vector double atanhd2 (vector double vx);	vector float atanhf4 (vector float vx);
cbrtd2	cbrtf4	Computes the cube root of each element of vx.	vector double cbrtd2 (vector double vx);	vector float cbrtf4 (vector float vx);
cosd2	cosf4	Computes the cosine of each element of vx.	vector double cosd2 (vector double vx);	vector float cosf4 (vector float vx);
coshd2	coshf4	Computes the hyperbolic cosine of each element of vx.	vector double coshd2 (vector double vx);	vector float coshf4 (vector float vx);
cosisind2	cosisinf4	Computes the cosine and sine of each element of x, and stores the results in y and z as follows: cosisind2 (x,y,z) sets y and z to {cos(x1), sin(x1)} and {cos(x2), sin(x2)} where x={x1,x2}. cosisinf4 (x,y,z) sets y and z to {cos(x1), sin(x1), cos(x2), sin(x2)} and {cos(x3), sin(x3), cos(x4), sin(x4)} where x={x1,x2,x3,x4}.	void cosisind2 (vector double x, vector double *y, vector double *z)	void cosisinf4 (vector float x, vector float *y, vector float *z)

Table 20. MASS SIMD functions (continued)

Double-precision function	Single-precision function	Description	Double-precision function prototype	Single-precision function prototype
divd2	divf4	Computes the quotient vx/vy .	vector double divd2 (vector double vx, vector double vy);	vector float divf4 (vector float vx, vector float vy);
erfcd2	erfcf4	Computes the complementary error function of each element of vx.	vector double erfcd2 (vector double vx);	vector float erfcf4 (vector float vx);
erfd2	erff4	Computes the error function of each element of vx.	vector double erfd2 (vector double vx);	vector float erff4 (vector float vx);
expd2	expf4	Computes the exponential function of each element of vx.	vector double expd2 (vector double vx);	vector float expf4 (vector float vx);
exp2d2	exp2f4	Computes 2 raised to the power of each element of vx.	vector double exp2d2 (vector double vx);	vector float exp2f4 (vector float vx);
expm1d2	expm1f4	Computes (the exponential function of each element of vx) - 1.	vector double expm1d2 (vector double vx);	vector float expm1f4 (vector float vx);
exp2m1d2	exp2m1f4	Computes (2 raised to the power of each element of vx) -1.	vector double exp2m1d2 (vector double vx);	vector float exp2m1f4 (vector float vx);
hypotd2	hypotf4	For each element of vx and the corresponding element of vy, computes $\sqrt{x*x+y*y}$.	vector double hypotd2 (vector double vx, vector double vy);	vector float hypotf4 (vector float vx, vector float vy);
lgammad2	lgammaf4	Computes the natural logarithm of the absolute value of the Gamma function of each element of vx .	vector double lgammad2 (vector double vx);	vector float lgammaf4 (vector float vx);
logd2	logf4	Computes the natural logarithm of each element of vx.	vector double logd2 (vector double vx);	vector float logf4 (vector float vx);
log2d2	log2f4	Computes the base-2 logarithm of each element of vx.	vector double log2d2 (vector double vx);	vector float log2f4 (vector float vx);
log10d2	log10f4	Computes the base-10 logarithm of each element of vx.	vector double log10d2 (vector double vx);	vector float log10f4 (vector float vx);
log1pd2	log1pf4	Computes the natural logarithm of each element of (vx +1).	vector double log1pd2 (vector double vx);	vector float log1pf4 (vector float vx);
log21pd2	log21pf4	Computes the base-2 logarithm of each element of (vx +1).	vector double log21pd2 (vector double vx);	vector float log21pf4 (vector float vx);

Table 20. MASS SIMD functions (continued)

Double-precision function	Single-precision function	Description	Double-precision function prototype	Single-precision function prototype
powd2	powf4	Computes each element of vx raised to the power of the corresponding element of vy.	vector double powd2 (vector double vx, vector double vy);	vector float powf4 (vector float vx, vector float vy);
qdrtd2	qdrtf4	Computes the quad root of each element of vx.	vector double qdrtd2 (vector double vx);	vector float qdrtf4 (vector float vx);
rcbrtd2	rcbrtf4	Computes the reciprocal of the cube root of each element of vx.	vector double rcbrtd2 (vector double vx);	vector float rcbrtf4 (vector float vx);
recipd2	recipf4	Computes the reciprocal of each element of vx.	vector double recipd2 (vector double vx);	vector float recipf4 (vector float vx);
rqdrtd2	rqdrtf4	Computes the reciprocal of the quad root of each element of vx.	vector double rqdrtd2 (vector double vx);	vector float rqdrtf4 (vector float vx);
rsqrtd2	rsqrtf4	Computes the reciprocal of the square root of each element of vx.	vector double rsqrtd2 (vector double vx);	vector float rsqrtf4 (vector float vx);
sincosd2	sincosf4	Computes the sine and cosine of each element of vx.	void sincosd2 (vector double vx, vector double *vs, vector double *vc);	void sincosf4 (vector float vx, vector float *vs, vector float *vc);
sind2	sinf4	Computes the sine of each element of vx.	vector double sind2 (vector double vx);	vector float sinf4 (vector float vx);
sinhd2	sinhf4	Computes the hyperbolic sine of each element of vx.	vector double sinhd2 (vector double vx);	vector float sinhf4 (vector float vx);
sqrtd2	sqrtf4	Computes the square root of each element of vx.	vector double sqrtd2 (vector double vx);	vector float sqrtf4 (vector float vx);
tand2	tanf4	Computes the tangent of each element of vx.	vector double tand2 (vector double vx);	vector float tanf4 (vector float vx);
tanhd2	tanhf4	Computes the hyperbolic tangent of each element of vx.	vector double tanhd2 (vector double vx);	vector float tanhf4 (vector float vx);

Compiling and linking a program with MASS

To compile an application that calls the functions in the following MASS libraries, specify the corresponding library names on the `-l` link option.

Table 21. The scalar, vector, and SIMD MASS library

MASS library	Library name
Scalar library	<code>mass</code>
Vector library	<code>massv</code> , <code>massvp8</code> , or <code>massvp9</code>

Table 21. The scalar, vector, and SIMD MASS library (continued)

MASS library	Library name
SIMD library	mass_simdp8 or mass_simdp9

For example, if the MASS libraries are installed in the default directory, you can use one of the following commands:

Link object file prog.c with scalar library libmass.a and vector library libmassv.a

```
xlc prog.c -o prog -lmass -lmassv
```

Link object file prog.c with SIMD library libmass_simdp9.a

```
xlc prog.c -o prog -lmass_simdp9
```

Using libmass.a with the math system library

If you want to use the libmass.a scalar library for some functions and the normal math library libm.a for other functions, follow this procedure to compile and link your program:

1. Find the object file name of the function that you want to use from libmass.a. The file name starts with the name of the function and has a .o suffix.

For example, to find the tanf function, you can use the following command if XL C/C++ is installed to the default location:

```
nm /opt/ibm/xlmass/8.1.6/lib/libmass.a | egrep "^tan.*\.o"
```

The output is as follows:

```
tand2.p8xs.s64.o:
tand2.p8xs.s64.o:
tanf.s64.o:
tanhf.s64.o:
```

Among the listed object files, the object file name for the tanf function is tanf.s64.o.

2. Use the **ar** command to extract the object files of the wanted functions from libmass.a.

For example, to extract the object file for the tanf function, use the following command:

```
ar -x tanf.s64.o /opt/ibm/xlmass/8.1.6/lib/libmass.a
```

3. Archive the extracted object files into another library:

```
ar -qv libfasttanf.a tanf.s64.o
ranlib libfasttanf.a
```

4. Create the final executable using **xlc**, specifying **-lfasttanf** instead of **-lmass**:

```
xlc sample.c -o sample -Ldir_containing_libfasttanf -lfasttanf
```

This links only the tanf function from MASS (now in libfasttanf.a) and the remainder of the math functions from the standard system library.

Using the Basic Linear Algebra Subprograms - BLAS

Four Basic Linear Algebra Subprograms (BLAS) functions are shipped with the XL C/C++ compiler in the libxlopt library. The functions consist of the following:

- **sgemv** (single-precision) and **dgemv** (double-precision), which compute the matrix-vector product for a general matrix or its transpose

- `sgemm` (single-precision) and `dgemm` (double-precision), which perform combined matrix multiplication and addition for general matrices or their transposes

Because the BLAS routines are written in Fortran, all parameters are passed to them by reference and all arrays are stored in column-major order.

Note: Some error-handling code has been removed from the BLAS functions in `libxlopt`, and no error messages are emitted for calls to these functions.

“BLAS function syntax” describes the prototypes and parameters for the XL C/C++ BLAS functions. The interfaces for these functions are similar to those of the equivalent BLAS functions shipped in IBM's Engineering and Scientific Subroutine Library (ESSL); for more information and examples of usage of these functions, see *Engineering and Scientific Subroutine Library Guide and Reference*, available at the Engineering and Scientific Subroutine Library (ESSL) and Parallel ESSL web page.

“Linking the libxlopt library” on page 95 describes how to link to the XL C/C++ `libxlopt` library if you are also using a third-party BLAS library.

BLAS function syntax

The prototypes for the `sgemv` and `dgemv` functions are as follows:

```
void sgemv(const char *trans, int *m, int *n, float *alpha,
           void *a, int *lda, void *x, int *incx,
           float *beta, void *y, int *incy);
void dgemv(const char *trans, int *m, int *n, double *alpha,
           void *a, int *lda, void *x, int *incx,
           double *beta, void *y, int *incy);
```

The parameters are as follows:

trans

is a single character indicating the form of input matrix *a*, where:

- 'N' or 'n' indicates that *a* is to be used in computation
- 'T' or 't' indicates that the transpose of *a* is to be used in computation

m represents:

- the number of rows in input matrix *a*
- the length of vector *y*, if 'N' or 'n' is used for the *trans* parameter
- the length of vector *x*, if 'T' or 't' is used for the *trans* parameter

The number of rows must be greater than or equal to zero, and less than the leading dimension of matrix *a* (specified in *lda*)

n represents:

- the number of columns in input matrix *a*
- the length of vector *x*, if 'N' or 'n' is used for the *trans* parameter
- the length of vector *y*, if 'T' or 't' is used for the *trans* parameter

The number of columns must be greater than or equal to zero.

alpha

is the scaling constant for matrix *a*

a is the input matrix of float (for `sgemv`) or double (for `dgemv`) values

lda

is the leading dimension of the array specified by *a*. The leading dimension

must be greater than zero. The leading dimension must be greater than or equal to 1 and greater than or equal to the value specified in m .

x is the input vector of float (for sgemv) or double (for dgemv) values.

$incx$

is the stride for vector x . It can have any value.

$beta$

is the scaling constant for vector y

y is the output vector of float (for sgemv) or double (for dgemv) values.

$incy$

is the stride for vector y . It must not be zero.

Note: Vector y must have no common elements with matrix a or vector x ; otherwise, the results are unpredictable.

The prototypes for the sgemm and dgemm functions are as follows:

```
void sgemm(const char *transa, const char *transb,
           int *l, int *n, int *m, float *alpha,
           const void *a, int *lda, void *b, int *ldb,
           float *beta, void *c, int *ldc);
void dgemm(const char *transa, const char *transb,
           int *l, int *n, int *m, double *alpha,
           const void *a, int *lda, void *b, int *ldb,
           double *beta, void *c, int *ldc);
```

The parameters are as follows:

$transa$

is a single character indicating the form of input matrix a , where:

- 'N' or 'n' indicates that a is to be used in computation
- 'T' or 't' indicates that the transpose of a is to be used in computation

$transb$

is a single character indicating the form of input matrix b , where:

- 'N' or 'n' indicates that b is to be used in computation
- 'T' or 't' indicates that the transpose of b is to be used in computation

l represents the number of rows in output matrix c . The number of rows must be greater than or equal to zero, and less than the leading dimension of c .

n represents the number of columns in output matrix c . The number of columns must be greater than or equal to zero.

m represents:

- the number of columns in matrix a , if 'N' or 'n' is used for the $transa$ parameter
- the number of rows in matrix a , if 'T' or 't' is used for the $transa$ parameter

and:

- the number of rows in matrix b , if 'N' or 'n' is used for the $transb$ parameter
- the number of columns in matrix b , if 'T' or 't' is used for the $transb$ parameter

m must be greater than or equal to zero.

alpha

is the scaling constant for matrix *a*

a is the input matrix *a* of float (for sgemm) or double (for dgemm) values

lda

is the leading dimension of the array specified by *a*. The leading dimension must be greater than zero. If *transa* is specified as 'N' or 'n', the leading dimension must be greater than or equal to 1. If *transa* is specified as 'T' or 't', the leading dimension must be greater than or equal to the value specified in *m*.

b is the input matrix *b* of float (for sgemm) or double (for dgemm) values.

ldb

is the leading dimension of the array specified by *b*. The leading dimension must be greater than zero. If *transb* is specified as 'N' or 'n', the leading dimension must be greater than or equal to the value specified in *m*. If *transa* is specified as 'T' or 't', the leading dimension must be greater than or equal to the value specified in *n*.

beta

is the scaling constant for matrix *c*

c is the output matrix *c* of float (for sgemm) or double (for dgemm) values.

ldc

is the leading dimension of the array specified by *c*. The leading dimension must be greater than zero. If *transb* is specified as 'N' or 'n', the leading dimension must be greater than or equal to 0 and greater than or equal to the value specified in *l*.

Note: Matrix *c* must have no common elements with matrices *a* or *b*; otherwise, the results are unpredictable.

Linking the libxlopt library

By default, the `libxlopt` library is linked with any application that you compile with the XL C/C++ compiler. However, if you are using a third-party BLAS library but want to use the BLAS routines shipped with `libxlopt`, you must specify the `libxlopt` library before any other BLAS library on the command line at link time. For example, if your other BLAS library is called `libblas.a`, you would compile your code with the following command:

```
xlc app.c -lxlopt -lblas
```

The compiler will call the `sgemv`, `dgemv`, `sgemm`, and `dgemm` functions from the `libxlopt` library and all other BLAS functions in the `libblas.a` library.

Chapter 10. Using vector technology

Vector technology is a PowerPC technology for accelerating the performance-driven, high-bandwidth communications and computing applications. You can use the vector technology to get dramatic performance improvement for your applications.

There are two ways of using the vector technology:

- Hand coding
- Automatic vectorization

Automatic vectorization often brings the best performance when you write the code in the right way, but appropriate hand coding can provide additional performance improvement.

Using vector technology with hand coding

The following table lists the information about using the vector technology with hand coding and provides the links to the detailed information in different documents.

Table 22. Language features for using vector technology with hand coding:

Information you need	Sections you can read
Intrinsic data types	Vector types (IBM extension) in <i>XL C/C++ Language Reference</i>
Vector built-in functions	Vector built-in functions in <i>XL C/C++ Compiler Reference</i>
Using the vector libraries	Using the vector libraries

Using vector technology with auto-vectorization

The following table lists the information about compiler options for auto-vectorization and provides the links to the detailed information in different documents.

Table 23. Information about compiler options for auto-vectorization

To do...	Read...
Enable automatic generation of vector instructions for processors that support them.	-qsimd in <i>XL C/C++ Compiler Reference</i>
Perform high-order transformations (HOT) during optimization.	-qhot in <i>XL C/C++ Compiler Reference</i>
Produce listing files and understand how sections of code have been optimized.	<ul style="list-style-type: none">• -qlistfmt in <i>XL C/C++ Compiler Reference</i>• -qreport in <i>XL C/C++ Compiler Reference</i>• Using compiler reports to diagnose optimization opportunities• Parsing compiler reports with development tools

Table 23. Information about compiler options for auto-vectorization (continued)

To do...	Read...
Ensure that optimizations done by default, do not alter certain program semantics related to strict IEEE floating-point conformance.	-qstrict in <i>XL C/C++ Compiler Reference</i>
Tuning for your target architecture using -qarch and -qtune.	Tuning for your system architecture

The following table lists the directive and compiler option that you can use to prohibit auto-vectorization and provides the links to the detailed information in different documents.

Table 24. Directive and compiler option for auto-vectorization

To do...	Read...
Disable auto-vectorization.	-qsimd in <i>XL C/C++ Compiler Reference</i>

Some optimization processes are related to auto-vectorization, you can use compiler options to control these optimizations. The following table lists these optimization processes and provides the links to the detailed information in different documents.

Table 25. Optimizations related to auto-vectorization

To learn about...	Read...
The High-order transformation (HOT)	<ul style="list-style-type: none"> Using high-order loop analysis and transformations An intermediate step: adding -qhot suboptions at level 3
The Interprocedural analysis (IPA)	The IPA process

Chapter 11. Parallelizing your programs

The compiler offers you the following methods of implementing shared memory program parallelization:

- Automatic parallelization of countable program loops, which are defined in “Countable loops.” An overview of the compiler's automatic parallelization capabilities is provided in “Enabling automatic parallelization” on page 101.
- Explicit parallelization of C and C++ program code using pragma directives compliant to the OpenMP Application Program Interface specification. An overview of the OpenMP directives is provided in “Using OpenMP directives” on page 103.

All methods of automatic program parallelization and optimization are enabled when the **-qsmp** option is in effect while the **omp** suboption is not. You can parallelize only program code that is explicitly annotated with OpenMP directives with the **-qsmp=omp** compiler option, but doing so disables automatic parallelization.

Parallel regions of program code are executed by multiple threads, possibly running on multiple processors. The number of threads created is determined by environment variables and calls to library functions. Work is distributed among available threads according to scheduling algorithms specified by the environment variables. For any of the methods of parallelization, you can use the XLSMPOPTS environment variable and its suboptions to control thread scheduling; for more information about this environment variable, see *XLSMPOPTS* in the *XL C/C++ Compiler Reference*. If you are using OpenMP constructs, you can use the OpenMP environment variables to control thread scheduling; for information about OpenMP environment variables, see *OpenMP environment variables for parallel processing* in the *XL C/C++ Compiler Reference*. For more information about OpenMP built-in functions, see *Built-in functions for parallel processing* in the *XL C/C++ Compiler Reference*.

For details about the OpenMP constructs, environment variables, and runtime routines, refer to the *OpenMP Application Program Interface Specification*, available at <http://www.openmp.org>.

Related information:

“Using shared-memory parallelism (SMP)” on page 31

Related information in the *XL C/C++ Compiler Reference*

 XLSMPOPTS

 OpenMP environment variables for parallel processing

Related external information

 OpenMP Application Program Interface Language Specification, available at <http://www.openmp.org>

Countable loops

Loops are considered to be countable if they take any of the following forms:

- $iteration_variable = iteration_variable - increment$

where *increment* is a loop-invariant signed integer expression. The value of the expression is known at run time and is not 0. *increment* cannot reference external or static variables, pointers or pointer expressions, function calls, or variables that have their address taken.

Enabling automatic parallelization

The compiler can automatically locate and parallelize all countable loops where possible in your program code. A loop is considered to be countable if it has any of the forms shown in “Countable loops” on page 99, and:

- There is no branching into or out of the loop.
- The *increment_expression* is not within a critical section.

In general, a countable loop is automatically parallelized only if all of the following conditions are met:

- The order in which loop iterations start or end does not affect the results of the program.
- The loop does not contain I/O operations.
- Floating point reductions inside the loop are not affected by round-off error, unless the **-qnostrict** option is in effect.
- The **-qnostrict_induction** compiler option is in effect.
- The **-qsmp=auto** compiler option is in effect.

Data sharing attribute rules

The rules of data sharing attributes determine the attributes of variables that are referenced in `parallel` and `task` directives, and `worksharing` regions.

Data sharing attribute rules for variables referenced in a construct

The data sharing attributes of variables that are referenced in a construct can be classified into the following categories:

- Predetermined data sharing attributes
- Explicitly determined data sharing attributes
- Implicitly determined data sharing attributes

Specifying a variable in a `firstprivate`, `lastprivate`, or `reduction` clause of an enclosed construct initiates an implicit reference to the variable in the enclosing construct. Such implicit references also follow the data sharing attribute rules.

Some variables and objects have predetermined data sharing attributes as follows:

- Variables that are specified in `threadprivate` directives are `threadprivate`.
- Variables with automatic storage duration that are declared in a scope inside the construct are `private`.
- Objects with dynamic storage duration are `shared`.
- Static data members are `shared`.
- The loop iteration variables in the associated `for` loops of a `for` or `parallel for` construct are `private`.
- Variables with `const`-qualified types are `shared` if they have no mutable member.

- For variables with static storage duration, if they are declared in a scope inside the construct, they are shared.

Variables with predetermined data sharing attributes cannot be specified in data sharing attribute clauses. However, in the following situations, specifying a predetermined variable in a data sharing attribute clause is allowed and overrides the predetermined data sharing attributes of the variable.

- The loop iteration variables in the associated for loops of a `for` or `parallel for` construct can be specified in a `private` or `lastprivate` clause.
- For variables with `const`-qualified type, if they have no mutable member, they can be specified in a `firstprivate` clause.

Variables that meet the following conditions have explicitly determined data sharing attributes:

- The variables are referenced in a construct.
- The variables are specified in a data sharing attribute clause on the construct.

Variables that meet all the following conditions have implicitly determined data sharing attributes:

- The variables are referenced in a construct.
- The variables do not have predetermined data sharing attributes.
- The variables are not specified in a data sharing attribute clause on the construct.

For variables that have implicitly determined data sharing attributes, the rules are as follows:

- In a `parallel` or `task` construct, the data sharing attributes of the variables are determined by the `default` clause, if present.
- In a `parallel` construct, if no `default` clause is present, the variables are shared.
- For constructs other than `task`, if no `default` clause is present, the variables inherit their data sharing attributes from the enclosing context.
- In a `task` construct, if no `default` clause is present, variables that are determined to be shared in the enclosing context by all implicit tasks bound to the current team are shared.
- In a `task` construct, if no `default` clause is present, variables whose data sharing attributes are not determined by the rules above are `firstprivate`.

Data sharing attribute rules for variables referenced in a region but not in a construct

The data sharing attributes of variables that are referenced in a region, but not in a construct, are determined as follows:

- If variables with static storage duration are declared in called routines in the region, the variables are shared.
- Variables with `const`-qualified types are shared if they have no mutable member and are declared in called routines.
- File-scope or namespace-scope variables referenced in called routines in the region are shared unless they are specified in a `threadprivate` directive.
- Objects with dynamic storage duration are shared.
- Static data members are shared unless they are specified in a `threadprivate` directive.

- The formal arguments of called routines in the region that are passed by reference inherit the data sharing attributes of the associated actual arguments.
- Other variables declared in called routines in the region are private.

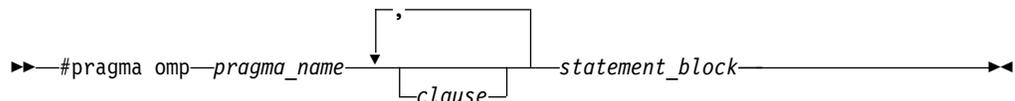
Using OpenMP directives

OpenMP directives exploit shared memory parallelism by defining various types of parallel regions. Parallel regions can include both iterative and non-iterative segments of program code.

The purposes of the **#pragma omp** pragmas fall into these general categories:

1. Defines parallel regions in which work is done by threads in parallel (**#pragma omp parallel**). Most of the OpenMP directives either statically or dynamically bind to an enclosing parallel region.
2. Defines how work is distributed or shared across the threads in a parallel region (**#pragma omp sections**, **#pragma omp for**, **#pragma omp single**, **#pragma omp task**).
3. Controls synchronization among threads (**#pragma omp atomic**, **#pragma omp master**, **#pragma omp barrier**, **#pragma omp critical**, **#pragma omp flush**, **#pragma omp ordered**).
4. Defines the scope of data visibility across parallel regions within the same thread (**#pragma omp threadprivate**).
5. Controls synchronization (**#pragma omp taskwait**, **#pragma omp barrier**).
6. Controls data or computation that is on another computing device.

OpenMP directive syntax



Adding certain clauses to the **#pragma omp** pragmas can fine tune the behavior of the parallel or work-sharing regions. For example, a `num_threads` clause can be used to control a parallel region pragma.

The **#pragma omp** pragmas generally appear immediately before the section of code to which they apply. The following example defines a parallel region in which iterations of a for loop can run in parallel:

```
#pragma omp parallel
{
    #pragma omp for
    for (i=0; i<n; i++)
        ...
}
```

This example defines a parallel region in which two or more non-iterative sections of program code can run in parallel:

```
#pragma omp parallel
{
    #pragma omp sections
    {
        #pragma omp section
        structured_block_1
        ...
    }
}
```

```

#pragma omp section
  structured_block_2
  ...
  ...
}
}

```

For a pragma-by-pragma description of the OpenMP directives, refer to *Pragma directives for parallel processing* in the *XL C/C++ Compiler Reference*.

Related information in the *XL C/C++ Compiler Reference*

-  Pragma directives for parallel processing
-  OpenMP built-in functions
-  OpenMP environment variables for parallel processing

Shared and private variables in a parallel environment

Some OpenMP clauses let you specify visibility context for selected data variables. A brief summary of data scope attribute clauses are listed below:

Data scope attribute clause	Description
private	The private clause declares the variables in the list to be private to each thread in a team.
firstprivate	The firstprivate clause provides a superset of the functionality provided by the private clause. The private variable is initialized by the original value of the variable when the parallel construct is encountered.
lastprivate	The lastprivate clause provides a superset of the functionality provided by the private clause. The private variable is updated after the end of the parallel construct.
shared	The shared clause declares the variables in the list to be shared among all the threads in a team. All threads within a team access the same storage area for shared variables.
reduction	The reduction clause performs a reduction on the scalar variables that appear in the list, with a specified operator.
default	The default clause allows the user to affect the data-sharing attribute of the variables appeared in the parallel construct.

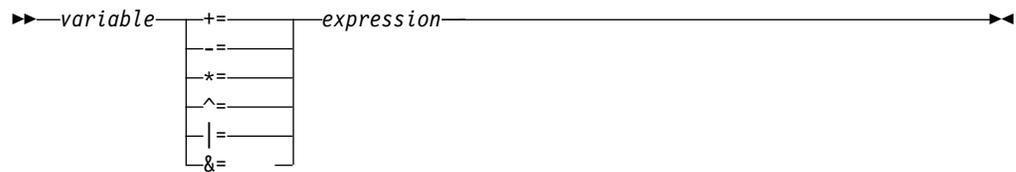
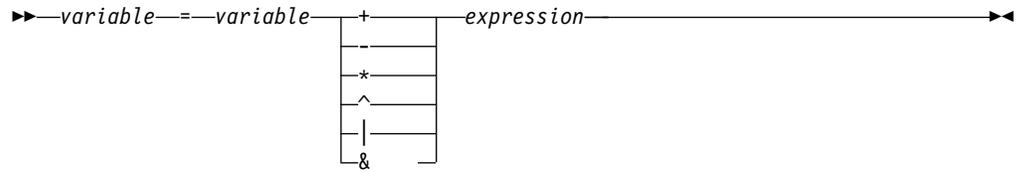
For more information, see the OpenMP directive descriptions in "Pragma directives for parallel processing" in the *XL C/C++ Compiler Reference*. You can also refer to the *OpenMP Application Program Interface Language Specification*, which is available at <http://www.openmp.org>.

Related information in the *XL C/C++ Compiler Reference*

-  Pragma directives for parallel processing

Reduction operations in parallelized loops

The compiler can recognize and properly handle most reduction operations in a loop during both automatic and explicit parallelization. In particular, it can handle reduction statements that have either of the following forms:



where:

variable

is an identifier designating an automatic or register variable that does not have its address taken and is not referenced anywhere else in the loop, including all loops that are nested. For example, in the following code, only `S` in the nested loop is recognized as a reduction:

```
int i,j, S=0;
for (i= 0 ;i < N; i++) {
    S = S+ i;
    for (j=0;j< M; j++) {
        S = S + j;
    }
}
```

expression

is any valid expression.

OpenMP directives provide you with mechanisms to specify reduction variables explicitly.

Chapter 12. Offloading computations to the NVIDIA GPUs

The combination of the IBM POWER processors and the NVIDIA GPUs provides a platform for heterogeneous high-performance computing that can run several technical computing workloads efficiently. The computational capability is built on top of massively parallel and multithreaded cores within the NVIDIA GPUs and the IBM POWER processors. You can offload parallel operations within applications, such as data analysis or high-performance computing workloads, to GPUs.

System prerequisites for offloading to the NVIDIA GPUs

To compile and link programs that contain code to be offloaded to the NVIDIA GPUs with IBM XL C/C++ for Linux, you must ensure these operating system, hardware, and software requirements are met.

- Use any IBM Power Systems™ server that has one or more NVIDIA GPUs installed and is supported by your Linux operating system distribution and the NVIDIA CUDA Toolkit.
- Use the supported little endian operating system.
- Install NVIDIA CUDA Toolkit 9.0 or 9.1.

For more information, see topic System prerequisites to offload computations to the NVIDIA GPUs in the *XL C/C++ Installation Guide*.

Programming with OpenMP device constructs

IBM XL C/C++ for Linux, V13.1.6 partially supports the OpenMP Application Program Interface Version 4.5 specification. You can offload compute-intensive parts of an application and associated data to the NVIDIA GPUs by using these supported device constructs.

Supported device constructs

- **omp target data**
- **omp target enter data**
- **omp target exit data**
- **omp target**
- **omp target update**
- **omp declare target**
- **omp teams**
- **omp distribute**
- **omp distribute parallel for**
- **omp distribute parallel for simd**
- **omp distribute simd**

For example, you can use the **omp target** directive to define a target region, which is a block of computation that operates within a distinct data environment and is intended to be offloaded onto a parallel computation device during execution. For more information about the OpenMP directives, see Pragma directives for parallel processing in the *XL C/C++ Compiler Reference*.

You can also use other OpenMP constructs with these OpenMP device constructs to exert finer control on parallelization, such as the combined constructs that are listed in Combined constructs in the *XL C/C++ Compiler Reference*.

GPU-related compiler options

You must specify the **-qoffload** option to enable the support for offloading OpenMP target regions to NVIDIA GPUs. For **-qoffload** to take effect, you must also specify the **-qsmp** option to enable support for OpenMP target regions. For more information, see **-qoffload** in the *XL C/C++ Compiler Reference*.

You can also use the **-qtgtarch** option to specify real or virtual GPU architectures where the code may run. For more information, see **-qtgtarch** in the *XL C/C++ Compiler Reference*.

GPU-related environment variables

You can control the computations offloaded to target devices by using the following environment variables:

- **XLSMPOPTS=target={mandatory | default | disabled}** controls which device to execute target regions on.
- **XLSMPOPTS=cudamemcheckfriendly={off | on}** controls whether to disable the check for pinned memory in the runtime and allow the program to be executed under the `cuda-memcheck` tool from the NVIDIA CUDA Toolkit.

For more information, see **XLSMPOPTS** in the *XL C/C++ Compiler Reference*.

GPU-related runtime functions

You can use the supported runtime functions, for example, to query the target environment or to manage device memory.

Table 26. Some useful OpenMP runtime functions for offloading computations to the NVIDIA GPUs

To query the target environment	To manage device memory
<ul style="list-style-type: none"> • <code>omp_get_default_device</code> • <code>omp_get_initial_device</code> • <code>omp_get_num_devices</code> • <code>omp_get_num_teams</code> • <code>omp_get_team_num</code> • <code>omp_is_initial_device</code> 	<ul style="list-style-type: none"> • <code>omp_target_alloc</code> • <code>omp_target_associate_ptr</code> • <code>omp_target_disassociate_ptr</code> • <code>omp_target_free</code> • <code>omp_target_is_present</code> • <code>omp_target_memcpy</code>

For more information about OpenMP runtime functions, see OpenMP runtime functions for parallel processing in the *XL C/C++ Compiler Reference*.

Extensions from OpenMP Technical Reports

IBM XL C/C++ for Linux, V13.1.6 also partially supports the OpenMP Technical Reports augmenting the OpenMP Application Program Interface Version 4.5 specification.

Support for implicit declare target

If a function is called and defined in the same translation unit, you do not need to explicitly declare it with the **omp declare target** directive. Such functions are

treated as if they were explicitly declared. This extension simplifies the use of functions defined in header files and used within a target region.

See the following code snippets for example.

```
// myprogram.h
int func1(int i) { return i; }
int func2(int i) { return func1(i); }
...
// myprogram.c
int main() {
    int rc = 0;
    #pragma omp target map(rc)
    {
        rc = func2(2);
    }
    return rc;
}
```

In this example, the `func2` function is used within a target region and defined in the same translation unit; the compiler implicitly treats the function as declared with the **omp declare target** directive. Furthermore, as function `func1` is called by `func2`, therefore `func1` is also treated as if it was declared using the **omp declare target** directive.

Pointer mapping for aggregate members

Mapping of a pointer is extended to support mapping a pointer that is a member of a struct and the storage that the pointer points to.

```
#include <stdio.h>
#include <stdlib.h>
struct S {
    int *p;
    int x;
};

void compute(struct S *s)
{
    for (int i=0; i<10; i++) s->p[i]++;
}

int main()
{
    struct S s;

    s.p = (int *) malloc(sizeof(int)*10);
    for (int i=0; i<10; i++) s.p[i] = i;

    #pragma omp target enter data map(to: s)
    #pragma omp target map(s.p[:10])
    {
        compute(&s); }
    for (int i=0; i<10; i++) printf("%d ", s.p[i]);
}
```

In this example, the structure `s` including the pointer `p` is mapped to the device in the target enter data construct. On entering the target region, by having mapped the storage that `p` points to, the `map(s.p[:10])` clause establishes the attachment between `s.p` on the device and the storage (`s.p[:10]`) being mapped to the device. As a result, in the target region, `s.p` points to the corresponding device storage (`s.p[0:10]`).

Using IBM XL C/C++ for Linux with NVCC

You can use IBM XL C/C++ for Linux as the host compiler with NVCC.

The NVIDIA CUDA C++ compiler (NVCC) from the NVIDIA CUDA Toolkit partitions C/C++ source code into host and device portions. You can use IBM XL C/C++ for Linux as the host compiler for the POWER processor with NVCC 9.0 or 9.1.

Related information

NVIDIA CUDA on IBM POWER8: Technical overview, software installation, and application development, downloadable from <http://www.redbooks.ibm.com/redpapers/pdfs/redp5169.pdf>

Limitations

Note these limitations when you offload computations to the NVIDIA GPUs with IBM XL C/C++ for Linux, V13.1.6.

Workaround to get debugging support on GPUs

To debug an OpenMP application, you need to compile it with `-g -qsmp=noopt`.

The reason is that due to a limitation in NVVM, IBM XL C/C++ for Linux, V13.1.6 cannot generate debug information when GPU runtime inlining is enabled. To cope with this case, you can turn off GPU runtime inlining by specifying `-qsmp=noopt`. GPU runtime inlining is enabled at all other optimization levels.

Limitation related to compute mode

If the compute mode of a target device is not set to default, a CUDA context created by your code might cause the OpenMP runtime to fail. The reason is that the target device cannot create a second context, which might be required depending on which OpenMP thread tries to initialize the device. You can check the compute mode to verify whether such case happens when you suspect a segment fault or deadlock.

Performance of nested parallelism on Volta

On Volta, nested level-1 parallel region and nested SIMD region run sequentially when only a subset of threads in a warp are active at the point of reaching the nested region. This is to cope with the independent thread scheduling of Volta to ensure forward progress in the presence of thread divergence.

Limitation caused by warp divergence and thread dependency

When the control path among two or more CUDA threads in the same warp diverges, these CUDA threads are serialized. IBM XL C/C++ for Linux generates one CUDA thread for each OpenMP thread. As a result, suppose thread A and B are two OpenMP threads, the program might hang if the progress of thread A depends on thread B performing some action, and if thread A executes first while thread B stalls because of *warp divergence*.

Warp divergence occurs when two threads of the same warp diverge in their execution due to a branch instruction, where one thread branches and the other

does not. This leads to serialization of the two threads by the CUDA hardware until their execution path converges again.

The following two examples demonstrate the problem:

Example 1

```
int omp_get_thread_num();

int main() {
    #pragma omp target
    {
        int i = 0;
        #pragma omp parallel
        {
            if (omp_get_thread_num() == 0) { // threadIdx.x 0 takes this path.
                #pragma omp atomic
                i++;
            }
            else { // The other threads take this path.
                int local_i;
                do {
                    #pragma omp atomic read
                    local_i = i;
                } while (local_i == 0);
            }
        }
    }
}
```

Example 2

```
int main() {
    #pragma omp target
    {
        int i = 0;
        #pragma omp parallel
        #pragma omp sections
        {
            #pragma omp section // threadIdx.x 0 executes this section.
            #pragma omp atomic
            i++;

            #pragma omp section // threadIdx.x 1 executes this section.
            {
                int local_i;
                do {
                    #pragma omp atomic read
                    local_i = i;
                } while (local_i == 0);
            }
        }
    }
}
```

Chapter 13. Vector element order toggling

To consistently use the instructions generated by vector built-in functions, users need to make all existing Vector Multimedia Extension (VMX) and Vector Scalar Extension (VSX) load and store built-in functions operate on the vectors in registers in the same vector element order, either little endian or big endian element order.

Vector element order

The `-maltivec` (`-qaltivec`) option affects the vector element order only in registers when the vectors are operated by a specific set of functions. In registers, the vector layout differs when the computer loads the vector in either big endian element order or little endian element order.

Note: To use the built-in functions enabled by the `-maltivec` (`-qaltivec`) option, you must explicitly code `#include <altivec.h>` in your program.

Big endian element order

Vectors are laid out in vector registers from left to right, so that element 0 is the leftmost element in the register.

Little endian element order

Vectors are laid out in vector registers from right to left, so that element 0 is the rightmost element in the register.

For more information, see “Example: Vector layout in the memory and register” on page 114.

Rules for vector element order toggling

The vector element order is toggled in registers by following these rules:

- The `-maltivec` (`-qaltivec`) option does not affect the vector element order in memory, where the vector elements are always stored in big endian element order.

For example, in memory, the `-maltivec` (`-qaltivec`) option does not affect the vector initialization. The vectors initialized by the union with the non-vectors (such as arrays) are always in big endian element order in memory. When the initialized vector is loaded to registers, the vector element order is always reversed to little endian element order in registers even when `-maltivec=be` (`-qaltivec=be`). However, if the vector loading is realized by using the vector built-in function, the vector element order is arranged with respect to the `-maltivec` (`-qaltivec`) option.

The previous rules apply to the vector literals as well. The vector literals are stored always in big endian element order in the memory. When the vector literals are loaded to registers, the vector element order is always reversed to little endian element order in registers.

- When `-maltivec=le` (`-qaltivec=le`) is in effect, the behaviours of functions are as follows:
 - The VMX and VSX load built-in functions load vectors to registers in little endian element order.
 - The VMX and VSX store built-in functions assume that the vectors to be stored are in little endian element order in registers.

- The nonload and nonstore built-in functions assume that vectors are loaded in registers in little endian element order.
- When `-maltivec=be` (`-qaltivec=be`) is in effect, these functions operate on the vectors in an opposite way of `-maltivec=le` (`-qaltivec=le`). The vectors in registers are in big endian element order.
- Regardless of the `-maltivec` (`-qaltivec`) option, the `vec_xl_be` function loads vectors to registers always in big endian element order and the `vec_xst_be` function assumes that vectors to be stored are always in big endian element order in registers.

For more information, see “Example: The vector built-in functions affected by the `-maltivec` (`-qaltivec`) option” on page 115 and “Example: The vector initialization by using the union with arrays” on page 115.

Example: Vector layout in the memory and register

The following example gets the first element of vector `va` by calling the `vec_extract` function. The function returning value is different based on the `-maltivec` (`-qaltivec`) option that determines whether `vec_extract` arranges the vector elements in big endian or little endian element order.

```
int get_first_element(va)
{
    vector signed int va;
    printf("%i\n", vec_extract(va, 0));
    //vec_extract is affected by the -qaltivec option
}
```

The following tables show the vector layout in the memory and the register.

Table 27. Vector layout in the memory.

Vector element value	E0	E1	E2	E3
----------------------	----	----	----	----

- When `-maltivec=be` (`-qaltivec=be`), the vector elements are loaded to registers in big endian element order and vector layout looks as follows.

Table 28. Vector layout in big endian element order.

Vector element number	0	1	2	3
Vector element value	E0	E1	E2	E3

The elements of vector `va` are ordered from the first to last, and stored from the left of registers. The `get_first_element` function gets the first element E0 from the left of registers.

- When `-maltivec=le` (`-qaltivec=le`), the vector elements are loaded to registers in little endian element order and vector layout looks as follows.

Table 29. Vector layout in little endian element order.

Vector element number	3	2	1	0
Vector element value	E3	E2	E1	E0

The elements of vector `va` are ordered from the last to first, and also stored from the left of registers. The `get_first_element` function gets the first element `E0` from the right of registers.

Example: The vector built-in functions affected by the `-maltivec (-qaltivec)` option

The following program `vec_xlw4.C` shows that the `vec_xlw4` function loads the vector elements in registers in the order specified by the `-maltivec (-qaltivec)` option.

```
int main()
{
    vector signed int a4;
    int c4[4] = {0,1,2,3};
    a4 = vec_xlw4(0, c4);
    //vec_xlw4 is affected by the -qaltivec option
    printf("%i %i %i %i\n", a4[0], a4[1], a4[2], a4[3]);
}
```

- Compile the program with `-qaltivec=be` by running the following command:

```
xlc vec_xlw4.C -qaltivec=be
```

Output:

```
0 1 2 3
```

- Compile the program with `-qaltivec=le` by running the following command:

```
xlc vec_xlw4.C -qaltivec=le
```

Output:

```
0 1 2 3
```

Example: The vector initialization by using the union with arrays

The following program example `vec_equiv.C` contains the vectors initialization by using the union with arrays. The vector loading is not affected by the `-maltivec (-qaltivec)` option and is loaded to registers always in little endian element order. Therefore, the vector elements extracted by `vec_extract` function are different between `-maltivec=be (-qaltivec=be)` and `-maltivec=le (-qaltivec=le)`.

```
int main()
{
    union {
        vector signed int a4;
        int c4[4];
    };

    //In the memory, the vector initialization (by using the union)
    //is not affected by the -qaltivec option and the vector is stored in
    //big endian element order. Then, the initialized vector is loaded
    //in registers by being reversed to the little endian element order.

    c4[0] = 0; c4[1] = 1; c4[2] = 2; c4[3] = 3;
    for (int i=0;i<4;i++)
        printf("%i ", vec_extract(a4,i));
    //vect_extract is affected by the -qaltivec option

    printf("\n");
}
```

- Compile the codes with `-qaltivec=le` by running the following command:

```
xlc vec_equiv.C -qaltivec=le
```

Output:

0 1 2 3

- Compile the program with `-qaltivec=be` by running the following command:

```
xlc vec_equiv.C -qaltivec=be
```

Output:

3 2 1 0

The compilation result is different from that of compilation with `-qaltivec=le`.

Related information:



`-qaltivec`



Supported GCC options



Migrating program containing vector built-in functions from big endian systems

Notices

Programming interfaces: Intended programming interfaces allow the customer to write programs to obtain the services of IBM XL C/C++ for Linux.

This information was developed for products and services offered in the U.S.A. IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those

websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who want to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

Intellectual Property Dept. for Rational Software
IBM Corporation
5 Technology Park Drive
Westford, MA 01886
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating

platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© (your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs. © Copyright IBM Corp. 1998, 2017.

PRIVACY POLICY CONSIDERATIONS:

IBM Software products, including software as a service solutions, ("Software Offerings") may use cookies or other technologies to collect product usage information, to help improve the end user experience, or to tailor interactions with the end user, or for other purposes. In many cases no personally identifiable information is collected by the Software Offerings. Some of our Software Offerings can help enable you to collect personally identifiable information. If this Software Offering uses cookies to collect personally identifiable information, specific information about this offering's use of cookies is set forth below.

This Software Offering does not use cookies or other technologies to collect personally identifiable information.

If the configurations deployed for this Software Offering provide you as customer the ability to collect personally identifiable information from end users via cookies and other technologies, you should seek your own legal advice about any laws applicable to such data collection, including any requirements for notice and consent.

For more information about the use of various technologies, including cookies, for these purposes, see IBM's Privacy Policy at <http://www.ibm.com/privacy> and IBM's Online Privacy Statement at <http://www.ibm.com/privacy/details> in the section entitled "Cookies, Web Beacons and Other Technologies," and the "IBM Software Products and Software-as-a-Service Privacy Statement" at <http://www.ibm.com/software/info/product-privacy>.

Trademarks

IBM, the IBM logo, and [ibm.com](http://www.ibm.com) are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at "Copyright and trademark information" at <http://www.ibm.com/legal/copytrade.shtml>.

Adobe is a registered trademark of Adobe Systems Incorporated in the United States, other countries, or both.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

NVIDIA and CUDA are either registered trademarks or trademarks of NVIDIA Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Index

Numerics

64-bit mode
optimization 57

A

advanced optimization
overview 24
alignment
aggregates 8
bit-fields 9
data 7
modes 7
modifiers 10
architecture
optimization 28
arrays
Fortran 4
attribute
aligned 10
packed 10
automatic parallelization 101

B

basic optimization
overview 22
bit-fields
alignment 9
BLAS library
overview 92

C

C++
static objects 18
templates 58
C++11
delegating constructors 58
rvalue references 59
target constructors 58
character variables
Fortran 4
constants
folding 14
rounding 14
correspondence
data types 2

D

data types
alignment 7
correspondence 2
size 7
debugging
optimized code 49

E

errors
detecting 50
floating-point 16
exceptions
floating-point 16

F

floating-point
exceptions 16
folding 14
formats 13
IEEE conformance 14
floating-point numbers
precision 13
range 13
folding
constants 14
floating-point 14
Fortran
arrays 4
character variables 4
function calls 4
function pointers 5
identifiers 1
function calls
Fortran 4
optimization 53
function cloning
interprocedural analysis 33
tuning 28
functions
linear algebra 92
matrix multiplication 92

I

identifiers
Fortran 1
IEEE conformance
floating-point 14
input/output
optimization 53
interprocedural analysis (IPA)
overview 33
IPA
processes 27

L

libmass library 82
libraries
BLAS 92
dynamic linking 17
shared 17
static 17
vector 84
xlopt 92

library
MASS 81
scalar 82
loops
optimization 99
loops analysis
optimization 29

M

MASS libraries
overview 81
scalar functions 82
SIMD functions 88
vector functions 84
memory
management 55
multiply-and-add operations
qfloat=nomaf 13
multithreading
parallelization 99
shared memory parallelism (SMP) 31

O

O0
overview 22
O2
overview 23
O3
overview 25
O4
overview 26
O5
overview 27
OpenMP
clauses 104
conformance 31
directives 103
optimization
64-bit mode 57
applications 21
diagnostics 45
errors 50
function calls 53
hardware 28
input/output 53
issues
debugging 49
levels
O0 22
O2 23
O3 25
O4 26
O5 27
loop analysis 29
loops 99
math functions 81
options
debugging 51

- optimization (*continued*)
 - options (*continued*)
 - overview 47
 - program units 33
 - programming techniques 53
 - results 50
 - strings 56
 - trade-offs
 - O3 25
 - O4 27
 - O5 28
- optimization and tuning
 - optimizing 21
 - tuning 21

P

- parallelization
 - data sharing 101
 - loops 101
 - multithreading 99
 - OpenMP directives 103
- parsing
 - compiler reports 47
- perfect forwarding
 - rvalue
 - move semantics 59
- performance tuning
 - coding 53
- pointers
 - Fortran 5
- pragma
 - omp 103
 - pack 10
- precision
 - floating-point 13
- priority
 - static objects 18
- profile-directed feedback (PDF)
 - cleanpdf 40
 - executable levels 41
 - mergepdf 37
 - object level 42
 - overview 36
 - qpdf1 36
 - recompiling 40
 - showpdf 36
 - standard PDF files 38
 - training 37
- profiling
 - qpdf1 36
 - qpdf2 40

Q

- qdatalocal
 - examples 44
- qhot
 - suggestions 30
- qipa
 - suggestions 35
- qsmp
 - parallelizing 99
 - SMP 31
 - suggestions 32

R

- range
 - floating-point 13
- reduction operations
 - parallelized loops 105
- rounding
 - constants 14
 - floating-point 14

S

- scalar MASS library 82
- shared library
 - compilation 17
- shared memory parallelism (SMP) 31
- static library
 - compilation 17
- static objects
 - initialization orders 18

T

- target machine options
 - mcpu (qarch) 28
 - mtune (qtune) 28
 - qcache 28
- templates
 - models 58
- type qualifiers
 - __align 10

V

- variables
 - imported 43
 - local 43
- vector element orders
 - toggling 113
- visibility attributes
 - propagation 69
 - types 61



Product Number: 5765-J08; 5725-C73

Printed in USA

SC27-6560-05

