

Index

Symbols

/ (divide) operator 6-9 to 6-10
– (minus) operator 6-6 to 6-7
!< (not less than) operator 6-4
!<= (not less than or equal) operator 6-4
!<> (not less or greater than) operator 6-4
!<>= (unordered) operator 6-4
!= (not equal) operator 6-4
!> (not greater than) operator 6-4
!>= (not greater than or equal) operator 6-4
* (multiply) operator 6-8
+ (plus) operator 6-5 to 6-6
< (less than) operator
 assembler 12-7
 defined 6-4
<= (less than or equal to) operator 6-4
<> (less or greater than) operator 6-4
<>= (ordered) operator 6-4
== (equal to) operator
 assembler 12-7
 defined 6-4
> (greater than) operator
 assembler 12-7
 defined 6-4
>= (greater than or equal to) operator 6-4
∞. *See* Infinities

Numerals

±0. *See* zero
680x0-based Macintosh computers
 numerics environment 1-13
 porting from A-1 to A-10
8087 coprocessor B-3

A

absolute value 4-5
 assembler 14-7
 compiler 10-11 to 10-12
accessing the environment
 assembler instructions 12-14 to 12-15
 C functions 8-9 to 8-13
 C functions, prerequisite D-1 to D-2

accuracy
 of basic arithmetic operations 1-4
 decimal to binary conversions 5-7 to 5-8
acos function 10-36 to 10-37
acosh function 10-45 to 10-46
addition 6-5 to 6-6
 assembler 14-4
 invalid exception, generating 4-5
address mode 11-5
AINT B-1
annuity function 10-52 to 10-54
ANSI X3J11.1 1-12 to 1-13
antilog functions. *See* exponential functions
APDA xix
arc cosine 10-36 to 10-37
arc cosine, hyperbolic 10-45 to 10-46
arc sine 10-37 to 10-39
arc sine, hyperbolic 10-47 to 10-48
arc tangent 10-39 to 10-40, 10-40 to 10-41
arc tangent, hyperbolic 10-48 to 10-50
argument reduction 6-11, 10-33
arithmetic assembler instructions 14-4 to 14-5
arithmetic operations 6-5 to 6-14
 addition 6-5 to 6-6
 assembler 14-4 to 14-7
 automatic type conversions 3-10
 division 6-9 to 6-10
 multiplication 6-8
 remainder 6-11 to 6-13
 round-to-integer 6-13 to 6-14
 square root 6-10 to 6-11
 subtraction 6-6 to 6-7
arithmetic, IEEE standard 1-3 to 1-13, 6-5 to 6-14
asin function 10-37 to 10-39
asinh function 10-47 to 10-48
assembler 11-3 to 14-8
 conversions 13-3 to 13-6
 data formats 11-3
 environmental access 12-3 to 12-15
 operations supported 14-3 to 14-8
atan function 10-39 to 10-40
atan2 function 10-40 to 10-41
atanh function 10-48 to 10-50
atomic operations 8-13
auxiliary functions 6-14 to 6-15
 assembler 14-8
 exponent field, return 10-29 to 10-30
 nan function 7-5
 nextafter functions 10-60 to 10-62

scaling 10-20 to 10-21
 sign manipulation 10-10 to 10-11

B

base 2 exponential 10-13 to 10-14
 BASIC B-1
 beq assembler instruction 12-6
 bge assembler instruction 12-6
 bgt assembler instruction 12-6
 bias of exponents 2-5
 binary logarithm 10-28 to 10-29
 binary to decimal conversions 5-7 to 5-12
 C functions 9-17 to 9-19
 double-double format 5-9 to 5-10
 strings 5-12
 structures 5-10 to 5-11, 9-13 to 9-15
 ble assembler instruction 12-6
 blt assembler instruction 12-6
 bne assembler instruction 12-6
 bng assembler instruction 12-6
 bnl assembler instruction 12-6
 bnu assembler instruction 12-6
 branch assembler instructions 12-6
 bta assembler instruction 12-6
 bun assembler instruction 12-6

C

C language
 compilers, FPCE recommendations for D-1 to D-9
 conformance to IEEE 754 1-12 to 1-13
 constants, floating-point D-3, D-5 to D-7
 conversions 9-3 to 9-25
 data types, new 7-3 to 7-8
 double type. *See* double format
 environmental controls 8-3 to 8-15
 expression evaluation D-3 to D-9
 float type. *See* single format
 function calls, conversions during 3-8
 long double type. *See* double-double format
 transcendental functions 10-3 to 10-67
 CDC computers B-2
 ceil function 9-6 to 9-7
 classcomp SANE function A-6
 classdouble SANE function A-6
 classes of floating-point numbers 2-5 to 2-11
 assembler 12-7 to 12-9
 compiler 7-4 to 7-5
 classextended SANE function A-6
 classfloat SANE function A-6

common logarithm 10-25 to 10-26
 comp data type (porting) A-4
 comparison functions 10-3 to 10-9
 comparison operations. *See* comparisons
 comparison operators 6-3 to 6-5
 comparisons 6-3 to 6-5
 assembler (branch instructions) 12-6
 assembler instructions 14-3 to 14-4
 C functions 10-3 to 10-9
 invalid exception, generating 4-5
 involving Infinities 6-3
 involving NaNs 6-3
 compatibility across architectures A-9 to A-10
 compiler optimizations
 and evaluation of floating-point constant
 expressions D-5
 and floating-point environment D-1 to D-2
 and widest-need evaluation D-5
 complementary error function 10-56 to 10-57
 compound function 10-50 to 10-52
 computer approximation of real numbers 1-3
 Condition Register 11-4, 12-5 to 12-6
 constants, floating-point
 evaluation D-5 to D-7
 hexadecimal D-3
 contraction operators D-2 to D-3
 controlling the environment
 assembler instructions 12-3 to 12-15
 C functions 8-3 to 8-15
 conversions 5-3 to 5-12
 accuracy of decimal to binary 5-7 to 5-8
 assembler 13-3 to 13-6
 between decimal formats 5-10, 9-19 to 9-23
 between floating-point formats 5-5 to 5-7, 9-13, 13-5
 binary to decimal 5-7 to 5-12, 9-13 to 9-19
 C functions 9-3 to 9-25
 ceil function 9-6 to 9-7
 decimal to binary 5-7 to 5-12
 C functions 9-13 to 9-19
 double-double format 5-9 to 5-10
 double-double to decimal 5-9 to 5-10
 during expression evaluation 3-3 to 3-11
 floating-point to integer 5-3 to 5-5, 6-13 to 6-14, 9-3
 to 9-11, 13-4 to 13-5
 floor function 9-7 to 9-8
 inexact exception 5-4, 5-5, 5-7
 integer to floating-point 5-3 to 5-5, 9-12, 13-3 to 13-4
 invalid exception 4-5, 5-4
 nearbyint function 9-9 to 9-10
 overflow exception 5-5, 5-7
 rint function 6-13 to 6-14
 rinttol function 9-3 to 9-4
 round function 9-10 to 9-11
 roundtol function 9-5 to 9-6
 SANE A-1 to A-2

trunc function 9-11 to 9-12
 underflow exception 5-5, 5-7
 copysign function 10-10 to 10-11
 invalid exception 4-5
 SANE A-5
 copysignl function 10-10 to 10-11
 cos function 10-33 to 10-34
 cosh function 10-42 to 10-43
 cosine 10-33 to 10-34
 cosine, hyperbolic 10-42 to 10-43
 CR. *See* Condition Register
 Cray computers B-2
 current rounding direction 4-3 to 4-4
 nearbyint function 9-9 to 9-10
 rint function 6-13 to 6-14
 rinttol function 9-3 to 9-4

D

data formats 2-3 to 2-17
 assembler 11-3
 choosing 2-16
 classes of numbers 2-5 to 2-11
 assembler 12-7 to 12-9
 compiler 7-4 to 7-5
 compiler 7-3 to 7-8
 converting between 5-5 to 5-7, 9-13, 13-5
 diagrams 2-11 to 2-15
 diagrams, symbols used in 2-11
 double format 2-13 to 2-14
 double-double format 2-14 to 2-15
 expression evaluation format 3-3
 minimum evaluation format 3-3 to 3-5, D-4
 precision of 2-16 to 2-17
 range of 2-16 to 2-17
 SANE A-1, A-4 to A-5
 semantic type 3-3
 single format 2-11 to 2-12
 widening for efficiency 7-3 to 7-4, A-9
 dec2f function 9-16 to 9-17
 dec2l function 9-16 to 9-17
 dec2num function 9-16 to 9-17
 dec2numl function 9-16 to 9-17
 dec2s function 9-16 to 9-17
 dec2str function 9-19 to 9-21
 decform structure 5-11
 definition 9-14 to 9-15
 digits field 9-14 to 9-15, 9-18, 9-20
 style field 9-14 to 9-15
 decimal data, reading and writing 5-8 to 5-10
 decimal formatting structure 5-11, 9-14 to 9-15
 decimal fractions 1-3
 decimal output

 fixed-style 9-15
 floating-style 9-14 to 9-15
 decimal strings 5-12
 decimal structure 5-10 to 5-11
 decimal structure 5-10 to 5-11
 definition 9-13 to 9-14
 exp field 9-13 to 9-14, 9-15, 9-17, 9-18
 sgn field 9-13 to 9-14, 9-15
 sig field 9-14, 9-16 to 9-17, 9-18, 9-20
 decimal to binary conversions 5-7 to 5-12
 C functions 9-16 to 9-17
 double-double format 5-9 to 5-10
 strings 5-12
 structures 5-10 to 5-11, 9-13 to 9-15
 decimal to decimal conversions 5-10, 9-19 to 9-23
 DECIMAL_DIG constant A-10
 default environment 4-4
 default rounding direction 4-3
 denormalized numbers 2-6 to 2-7
 density of 2-6
 double-double format 2-15
 SANE A-2
 DENORMALNUM SANE constant A-6
 density of denormalized numbers 2-6
 density of single-precision numbers 2-5
 difference operation
 assembler 14-4
 defined 6-6 to 6-7
 difference, positive function 10-4 to 10-5
 DIVBYZERO SANE constant A-7
 / (divide) operator 6-9 to 6-10
 divide-by-zero exception
 assembler 12-11
 defined 4-6
 division 6-9 to 6-10
 assembler 14-4
 invalid exception, generating 4-5
 by zero 1-9
 double format 2-13 to 2-14
 compiler 2-4, 7-3
 converting from double-double format 5-7
 converting from single format
 assembler 13-5
 defined 5-5
 converting to double-double format 5-7
 converting to single format
 assembler 13-5 to 13-6
 defined 5-5
 diagram 2-13
 diagram, symbols used in 2-11
 as minimum evaluation format D-4
 precision 2-16
 range 2-14
 representation of values 2-13
 double type. *See* double format

DOUBLE_SIZE macro A-10
double_t typedef 7-3 to 7-4
 for compatibility A-9
 in transcendental function declarations A-4
double-double format 2-14 to 2-15
 compared to extended format 2-3 to 2-4
 compiler 2-4, 7-3
 converting from double format 5-7
 converting from single format 5-5 to 5-7
 converting to decimal 5-9 to 5-10
 converting to double format 5-7
 converting to single format 5-5 to 5-7
 diagram 2-14
 diagram, symbols used in 2-11
 interpretation of values 2-14 to 2-15
 as minimum evaluation format D-4, D-5
 precision 2-14 to 2-15, 2-16
 range 2-15
downward rounding
 defined 4-3
 floor function 9-7 to 9-8
DOWNWARD SANE constant A-7

E

elementary functions. *See* transcendental functions
environment 4-3 to 4-6
 accessing
 assembler instructions 12-14 to 12-15
 C functions 8-9 to 8-13
 C functions prerequisite D-1 to D-2
 assembler 12-3 to 12-15
 C functions, types 8-3 to 8-15
 default 4-4
 ignoring D-2
 restoring
 assembler 12-14 to 12-15
 compiler 8-11 to 8-12, 8-12 to 8-13
 SANE A-3, A-7 to A-8
 saving
 assembler 12-14 to 12-15
 compiler 8-10, 8-10 to 8-11
 setting (compiler) 8-11 to 8-12
 use B-3
environment SANE type A-7
environmental access switch
 defined D-1 to D-2
 purpose, note on 8-3
environmental controls 4-3 to 4-6
 assembler instructions 12-3 to 12-15
 C functions 8-3 to 8-15
 SANE A-3, A-7 to A-8
== (equal to) operator

 assembler 12-7
 defined 6-4
erf function 10-55 to 10-56
erfc function 10-56 to 10-57
error functions 10-55 to 10-60
evaluation format 3-3
 minimum 3-3, D-4
 widest need 3-5 to 3-7
evaluation rules B-2
exception handling 1-7 to 1-9
exception SANE type A-7
exceptional events 1-6 to 1-9
exceptions 1-6 to 1-9
 assembler instructions 12-10 to 12-13
C functions 8-5 to 8-9
clearing
 assembler 12-11
 compiler 8-6, 8-10 to 8-11
in Condition Register 12-6
descriptions of 4-4 to 4-6
divide-by-zero 4-6
enabling and disabling (assembler) 12-12
inexact 4-6
invalid 4-5
overflow 4-5
preserving
 assembler 12-14 to 12-15
 compiler 8-10 to 8-11, 8-12 to 8-13
raising
 assembler 12-11
 compiler 8-7 to 8-8
restoring (compiler) 8-8
saving
 assembler 12-14 to 12-15
 compiler 8-7, 8-10 to 8-11
setting
 assembler 12-11
 compiler 8-7 to 8-8, 8-12 to 8-13
spurious 8-13
testing
 assembler 12-12 to 12-13
 compiler 8-8 to 8-9
underflow 4-5
exp function 10-12 to 10-13
exp1 SANE function A-6
exp2 function 10-13 to 10-14
expm1 function 10-14 to 10-15
exponent
 defined 2-5
 determining value of 10-21 to 10-22, 10-29 to 10-30
exponential functions 10-12 to 10-21
 base 2 exponential 10-13 to 10-14
 natural exponential 10-12 to 10-13
 natural exponential – 1 10-14 to 10-15
expression evaluation format 3-3

expression evaluation methods 3-3 to 3-11
 compared 3-8 to 3-11
 compiler D-3 to D-9
 examples 3-8 to 3-11
 floating-point constants D-5 to D-7
 minimum evaluation format only 3-3 to 3-5, D-4
 SANE A-2
 widest-need evaluation 3-5 to 3-6, D-5
 extended data type A-5
 compared to double-double format 2-3 to 2-4
 in definitions of `float_t` and `double_t` 7-4
 in transcendental function declarations A-4

F

`fabs` assembler instruction 14-7
`fabs` function 4-5, 10-11 to 10-12
`fabsl` function 10-11 to 10-12
`fadd` assembler instruction 14-4 to 14-5
`fcmpo` assembler instruction 14-3 to 14-4
`fcmpl` assembler instruction 14-3 to 14-4
`ftw` assembler instruction 13-4 to 13-5
`ftwz` assembler instruction 13-4 to 13-5
`fdim` function 10-4 to 10-5
`fdiv` assembler instruction 14-4 to 14-5
`FE_ALL_EXCEPT` constant 8-6
`FE_DFL_ENV` constant 8-10
`FE_DIVBYZERO` constant 8-6
`FE_DOWNWARD` constant 8-3
`FE_INEXACT` constant 8-6
`FE_INVALID` constant 8-6
`FE_OVERFLOW` constant 8-6
`FE_TONEAREST` constant 8-3
`FE_TOWARDZERO` constant 8-3
`FE_UNDERFLOW` constant 8-6
`FE_UPWARD` constant 8-3
`feclearexcept` function 8-6
`fegetenv` function
 definition 8-10
 difference from `feholdexcept` function 8-11
`fegetexcept` function
 definition 8-7
 with `fesetexcept` function 8-8
`fegetround` function
 definition 8-3 to 8-4
 with `fesetround` function 8-4, 8-5
`feholdexcept` function 8-10 to 8-11
`fenv_access` pragma option D-1 to D-2
`fenv_t` type 8-10
`fenv.h` file 8-3 to 8-15, C-12 to C-13
`feraiseexcept` function 8-7 to 8-8
`fesetenv` function 8-11 to 8-12
`fesetexcept` function 8-8

`fesetround` function 8-4 to 8-5
`fetestexcept` function 8-8 to 8-9
`feupdateenv` function
 definition 8-12 to 8-13
 with `feholdexcept` function 8-11
`fexcept_t` type 8-6
 financial functions 10-50 to 10-54
`float` type. *See* single format
`float_t` typedef 7-3 to 7-4, A-9
 floating-point constants
 evaluation D-5 to D-7
 hexadecimal D-3
 floating-point data formats. *See* data formats
 floating-point environment. *See* environment
 floating-point exceptions. *See* exceptions
 floating-point expressions, evaluating 3-3 to 3-11, D-3
 to D-9
 floating-point numbers
 classes of 2-5 to 2-11
 assembler 12-7 to 12-9
 compiler 7-4 to 7-5
 converting to integer 6-13 to 6-14
 integers, converting to 5-3 to 5-5
 assembler 13-4 to 13-5
 compiler 9-3 to 9-11
 truncating 4-3
 splitting 10-30 to 10-31
 floating-point registers 11-3
 floating-point result flags 12-7
 Floating-Point Status and Control Register (FPSCR).
See FPSCR
 floating-point values, interpreting 2-4 to 2-11
 floating-point variables, initialization D-7
`floor` function 9-7 to 9-8
 flush-to-zero systems 2-6
`fmadd` assembler instruction 14-6 to 14-7
`fmax` function 10-5 to 10-6
`fmin` function 10-6 to 10-7
`fmod` function 6-11 to 6-13
`fmr` assembler instruction 14-7
`fmsub` assembler instruction 14-6 to 14-7
`fmul` assembler instruction 14-4 to 14-5
`fnabs` assembler instruction 14-7
`fneg` assembler instruction 14-7
`fnmadd` assembler instruction 14-6 to 14-7
`fnmsub` assembler instruction 14-6 to 14-7
 format conventions for this book xviii to xix
 formats. *See* data formats
 formatters, numeric 9-19 to 9-21
 formatting output
 fixed-style decimal 9-15
 floating-style decimal 9-14 to 9-15
 Fortran B-1, B-2, B-3
`__FP__` macro A-10
`fp_contract` pragma D-2 to D-3

FPCE technical report 1-12 to 1-13
 compiler, recommendations for D-1 to D-9
 conversions 9-3 to 9-25
 data types 7-3
 environmental access 8-3 to 8-15
 expression evaluation D-3 to D-9
 transcendental functions 10-3 to 10-67
 fpclassify macro 7-4
 fp.h file C-1 to C-11
 functions 9-3 to 9-25, 10-3 to 10-67
 porting to A-4 to A-8
 FPSCR 11-4
 exception bits 12-10 to 12-11
 format 12-3 to 12-5
 manipulation 12-3 to 12-15
 result flags 12-7
 rounding direction 12-9 to 12-10
 fp_wide_function_parameters pragma D-9
 fp_wide_function_returns pragma D-8
 fp_wide_variables pragma D-9
 fraction field
 defined 2-3
 determining value of 10-21 to 10-22
 frexp function 10-21 to 10-22
 frsp assembler instruction 13-5
 fsub assembler instruction 14-4 to 14-5
 functions 6-3 to 6-15
 auxiliary 6-14 to 6-15
 comparison 10-3 to 10-9
 error 10-55 to 10-60
 exponential 10-12 to 10-21
 financial 10-50 to 10-54
 gamma 10-55 to 10-60
 hyperbolic 10-42 to 10-50
 logarithmic 10-21 to 10-31
 sign manipulation 10-9 to 10-12
 trigonometric 10-31 to 10-41

G

gamma function 10-57 to 10-58
 gamma functions 10-55 to 10-60
 getenv SANE function A-8
 getround SANE function A-7
 gradual underflow 2-7
 > (greater than) operator
 assembler 12-7
 defined 6-4
 >= (greater than or equal to) operator 6-4

H

hexadecimal floating-point constants in C D-3
 HP Spectrum quad format B-2
 hyperbolic functions 10-42 to 10-50
 hypot function 10-62 to 10-63
 hypotenuse 10-62 to 10-63

I

IBM Q format B-2
 IEEE arithmetic
 advantages 1-3 to 1-9
 operations 6-5 to 6-14
 IEEE data formats 2-3 to 2-4
 . *See also* single format, double format
 IEEE standard xvii
 advantages 1-3 to 1-13
 arithmetic operations 6-5 to 6-14
 auxiliary functions 6-14 to 6-15
 C language 1-12 to 1-13
 comparisons 6-4
 conversions required 5-3
 data formats 2-3 to 2-4
 exceptions 4-4 to 4-6
 rounding direction modes 4-3 to 4-4, 5-4
 . *See also* rounding direction
 rounding precision modes 4-4
 IEEE Standard 754. *See* IEEE standard
 IEEE Standard 854 1-3
 logb function 10-29
 nearbyint function 9-9
 IEEE standard arithmetic. *See* IEEE arithmetic
 IEEEDEFAULTENV SANE constant A-7
 inexact exception 4-6
 assembler 12-11
 conversions 5-4, 5-5, 5-7
 INEXACT SANE constant A-7
 INFINITE SANE constant A-6
 Infinities 2-7 to 2-8
 as alternative to stopping 1-7, 1-8 to 1-9
 comparisons 6-3
 converting to decimal 9-18
 converting to floating-point 9-17
 converting to integer 5-4
 converting to string 9-20
 double-double format 2-15
 negative 2-8
 positive 2-8
 SANE A-2
 INFINITY constant 7-5
 initialization of floating-point variables D-7
 instant rounding B-2

INT B-1
integer types 2-8
integers, converting 5-3 to 5-5
 assembler 13-3 to 13-4
 compiler 9-12
 rounding 4-3
 truncating 4-3
interpreting floating-point values 2-4 to 2-11
interval arithmetic 1-5
invalid exception 4-5
 assembler 12-10
 conversions 5-4
 signaling NaN, result of 2-8
invalid operation flag B-3
INVALID_SANE constant A-7
invalid-operation exception. *See* invalid exception
inverse operations 1-5 to 1-6
ipower SANE function A-6
isfinite macro 7-4
isnan macro 7-4
isnormal macro 7-4

L

ldexp function 10-16 to 10-17
<> (less or greater than) operator 6-4
< (less than) operator
 assembler 12-7
 defined 6-4
<= (less than or equal to) operator 6-4
lfd assembler instruction 11-6
lfdl assembler instruction 11-6
lfdx assembler instruction 11-7
lfdx assembler instruction 11-7
lfs assembler instruction 11-6, 13-5
lfsu assembler instruction 11-6, 13-5
lfsux assembler instruction 11-7, 13-5
lfsx assembler instruction 11-7, 13-5
lgamma function 10-59 to 10-60
load assembler instructions 11-5 to 11-7
 as conversion operations 13-5
 formats 11-5 to 11-6
log function 10-23 to 10-25
log1 SANE function A-6
log10 function 10-25 to 10-26
log1p function 10-26 to 10-27
log2 function 10-28 to 10-29
logarithmic functions 10-21 to 10-31
 binary 10-28 to 10-29
 common 10-25 to 10-26
 log of gamma 10-59 to 10-60
 natural 10-23 to 10-25, 10-26 to 10-27
logb function 10-29 to 10-30

long double type. *See* double-double format
LONG_DOUBLE_SIZE macro A-10

M

MathLib 1-12 to 1-13
 conversions 9-3 to 9-25
 data types, new 7-3 to 7-8
 environmental controls 8-3 to 8-15
 expression evaluation extensions D-8 to ??, D-8, ??
 to D-9
 porting to A-4 to A-8
 transcendental functions 10-3 to 10-67
maximum function 10-5 to 10-6
MC68881 coprocessor B-3
mcrfs assembler instruction 12-9, 12-12
mffs assembler instruction 12-14
_MIN_EVAL_FORMAT macro D-8
minimum evaluation format 3-3 to 3-5
 compared to widest-need evaluation 3-8 to 3-11
 compiler recommendations D-4
 examples 3-8 to 3-11
minimum function 10-6 to 10-7
– (minus) operator 6-6 to 6-7
mixed formats B-2
modf function 10-30 to 10-31
modulo function 6-12
move assembler instructions 14-7
mtfsb0 assembler instruction 12-11, 12-12
mtfsb1 assembler instruction 12-11, 12-12
mtfsf assembler instruction 12-14
mtfsfi assembler instruction 12-10, 12-12
multiplication 6-8
 assembler 14-4
 invalid exception, generating 4-5
* (multiply) operator 6-8
multiply-add assembler instructions 14-6 to 14-7
 enabling and disabling D-2 to D-3
 format 14-6

N

NAN constant 7-5
nan function
 PowerPC Numerics 7-5
 SANE A-6
NaNs 2-8 to 2-10
 as alternative to stopping 1-7, 1-8
 comparisons 6-3
 converting to decimal 9-18
 converting to floating-point 9-17

- converting to integer 5-4
- converting to string 9-20
- creating 7-5
- double-double format 2-15
- porting programs B-3
- quiet 2-8 to 2-10, 4-5
- SANE A-2
- signaling 2-8 to 2-10, 4-5, 6-4
- natural exponential 10-12 to 10-13
- natural exponential minus 1 10-14 to 10-15
- natural logarithm 10-23 to 10-25, 10-26 to 10-27
- NCEG 1-12 to 1-13
- nearbyint function 9-9 to 9-10
- negative Infinity. *See* Infinities
- negative zero. *See* zero
- nextafter functions
 - PowerPC Numerics 10-60 to 10-62
 - SANE A-6
- normalized numbers 2-5 to 2-6
 - compared to denormalized numbers 2-6
 - double-double format 2-15
- NORMALNUM SANE constant A-6
- != (not equal) operator 6-4
- !> (not greater than) operator 6-4
- !>= (not greater than or equal) operator 6-4
- !<> (not less or greater than) operator 6-4
- !< (not less than) operator 6-4
- !<= (not less than or equal) operator 6-4
- !<>= (unordered) operator 6-4
- not unordered comparison 6-4
- Not-a-Number. *See* NaNs
- num2dec function
 - definition 9-17 to 9-19
 - with dec2str function 9-21
- numbers, classes of 2-5 to 2-11
 - assembler 12-7 to 12-9
 - compiler 7-4 to 7-5
- numclass SANE type A-6
- Numerical C Extensions Group 1-12 to 1-13

O

- operations 6-3 to 6-15
 - arithmetic
 - assembler 14-4 to 14-7
 - defined 6-5 to 6-14
 - assembler 14-3 to 14-8
 - comparison
 - assembler 12-6, 14-3 to 14-4
 - defined 6-3 to 6-5
 - compiler 6-3 to 6-15
 - conversion
 - assembler 13-3 to 13-6

- compiler 9-3 to 9-25
 - SANE A-2 to A-3
 - subject to arithmetic conversions 3-4
- optimizations
 - and evaluation of floating-point constant expressions D-5
 - and floating-point environment D-1 to D-2
 - and widest-need evaluation D-5
- ordered comparison
 - assembler 14-3
 - defined 6-4
- <>= (ordered) operator 6-4
- output
 - fixed-style decimal 9-15
 - floating-style decimal 9-14 to 9-15
- overflow 4-5
 - assembler 12-11
 - conversions 5-5, 5-7
- OVERFLOW SANE constant A-7

P

- Pascal B-1
- PDP-11C B-3
- pi constant 10-33
- pi SANE function A-6
- + (plus) operator 6-5 to 6-6
- porting programs
 - from SANE A-3 to A-10
 - from non-Macintosh computers B-1 to B-3
- positive difference function 10-4 to 10-5
- positive Infinity. *See* Infinities
- positive zero. *See* zero
- pow function
 - PowerPC Numerics 10-17 to 10-20
 - SANE A-6
- power function 10-17 to 10-20
- PowerPC floating-point architecture 11-3 to 14-8
 - conversions 13-3 to 13-6
 - data formats 11-3
 - environmental access 12-3 to 12-15
 - operations supported 14-3 to 14-8
- PowerPC Numerics xvii
 - advantages 1-3 to 1-9
 - conversions supported 5-3 to 5-12
 - data formats 2-3 to 2-17
 - environmental controls 4-3 to 4-6
 - expression evaluation 3-3 to 3-11
 - functions supported 6-3 to 6-15
 - operations supported 6-3 to 6-15
 - SANE, compared to 1-13, A-1 to A-10
 - SANE, porting from A-3 to A-10
- pragmas

`fenv_access` D-1 to D-2
`fp_contract` D-2 to D-3
`fp_wide_function_parameters` D-8 to D-9
`fp_wide_function_returns` D-8 to D-9
`fp_wide_variables` D-8 to D-9
precision 1-4
 of data formats 2-16 to 2-17
 of expression evaluation 3-3 to 3-11
`procentry` SANE function A-8
`procexit` SANE function A-8

Q

`QNAN` SANE constant A-6
quiet NaNs 2-8 to 2-10, 4-5

R

random number generator 10-63 to 10-64
`randomx` function 10-63 to 10-64
range of data formats 2-16 to 2-17
real numbers
 computer approximation 1-3
 order of 6-3
recommendations, FPCE for compilers D-1 to D-9
registers
 Condition Register 11-4, 12-5 to 12-6
 floating-point 11-3
 FPSCR 11-4, 12-3 to 12-15
 special-purpose 11-4
`relation` function 10-8 to 10-9
relational operators 6-3 to 6-5
remainder function
 defined 6-11 to 6-13
 invalid exception, generating 4-5
`remquo` function 6-11 to 6-13
result flags 12-7
result, tiny 4-5
`rint` function 6-13 to 6-14
`rinttol` function 9-3 to 9-4
`round` function 9-10 to 9-11
round to integer operation 6-13 to 6-14
`rounddir` SANE type A-7
rounding
 defined 1-5 to 1-6
 instant B-2
rounding direction 4-3 to 4-4
 assembler 12-9 to 12-10
 compiler 8-3 to 8-5
 control 1-5
 current 6-13 to 6-14, 9-3 to 9-4, 9-9 to 9-10

 default 4-3
 downward 4-3
 saving (compiler) 8-3 to 8-4
 setting
 assembler 12-9 to 12-10
 compiler 8-4 to 8-5
 to nearest 4-3
 toward zero 4-3
 upward 4-3
rounding downward
 defined 4-3
 `floor` function 9-7 to 9-8
rounding modes. *See* rounding direction
rounding precision modes 4-4
rounding to integer 4-3
rounding to nearest value 4-3
rounding toward zero
 defined 4-3
 `trunc` function 9-11 to 9-12
rounding upward
 `ceil` function 9-6 to 9-7
 defined 4-3
 example 8-5
roundoff error with denormalized numbers 2-6
`roundtol` function 9-5 to 9-6

S

SANE xvii
 compared to PowerPC Numerics 1-13, A-1 to A-10
 conversions A-1 to A-2
 data formats A-1
 denormalized numbers A-2
 environment A-3, A-7 to A-8
 expression evaluation A-2
 Infinities A-2
 NaNs A-2
 operations A-2 to A-3
 porting programs from A-3 to A-10
 transcendental functions A-3, A-5 to A-6
`__SANE__` macro A-10
`sane.h` file A-4 to A-8
`scalb` function
 PowerPC Numerics 10-20 to 10-21
 SANE A-6
scaling functions
 `ldexp` function 10-16 to 10-17
 `scalb` function 10-20 to 10-21
scanners 9-21 to 9-23
semantic type 3-3
`setenvironment` SANE function A-8
`setexception` SANE function A-7
`setround` SANE function A-7

sign bit 2-3, 2-4
 sign manipulation functions 10-9 to 10-12
 copysign 10-10 to 10-11
 fabs function 10-11 to 10-12
 sign of zero 2-10 to 2-11
 SIGN(A) B-1
 SIGN(A,B) B-1
 signaling NaNs 2-8 to 2-10
 comparisons 6-4
 invalid exception 4-5
signbit macro 7-4
 significand 2-4
signnum SANE function A-6
sin function 10-34 to 10-35
 sine 10-34 to 10-35
 sine, hyperbolic 10-43 to 10-44
 single format 2-11 to 2-12
 compiler 2-4, 7-3
 converting from double format
 assembler 13-5 to 13-6
 defined 5-5
 converting from double-double format 5-5 to 5-7
 converting to double format
 assembler 13-5
 defined 5-5
 converting to double-double format 5-5 to 5-7
 diagram 2-12
 diagram, symbols used in 2-11
 as minimum evaluation format D-4
 precision 2-16
 range 2-12
 representation of values 2-12
 single-precision numbers, density of 2-5
sinh function 10-43 to 10-44
 small values
 and error analysis 2-7
 representing 2-6 to 2-7
 SNAN SANE constant A-6
 special-purpose registers 11-4
 spurious exceptions 8-13
sqrt function 6-10 to 6-11
 square root operation
 defined 6-10 to 6-11
 invalid exception, generating 4-5
 Standard Apple Numerics Environment (SANE). *See*
 SANE
stfd assembler instruction 11-6
stfdu assembler instruction 11-6
stfdx assembler instruction 11-7
stfdx assembler instruction 11-7
stfs assembler instruction 11-6, 13-5
stfsu assembler instruction 11-6, 13-5
stfsux assembler instruction 11-7, 13-5
stfsx assembler instruction 11-7, 13-5
 stopping program B-3

store assembler instructions 11-5 to 11-7
 as conversion operations 13-5 to 13-6
 formats 11-5 to 11-6
str2dec function 9-21 to 9-23
 string conversions 5-12
 subtraction operation
 assembler 14-4
 defined 6-6 to 6-7
 symbols in format diagrams 2-11

T

tagp parameter 7-5
tan function 10-35 to 10-36
 tangent 10-35 to 10-36
 tangent, hyperbolic 10-44 to 10-45
tanh function 10-44 to 10-45
testexception SANE function A-7
 tiny result 4-5
 to-nearest rounding 4-3
 TONEAREST SANE constant A-7
 toward $+\infty$ rounding. *See* upward rounding
 toward $-\infty$ rounding. *See* downward rounding
 toward-zero rounding
 defined 4-3
 trunc function 9-11 to 9-12
 TOWARDZERO SANE constant A-7
 transcendental functions 10-3 to 10-67
 assembler 14-8
 defined 1-12 to 1-13, 6-15
 SANE A-3, A-5 to A-6
 transported code B-3
 trigonometric functions 10-31 to 10-41
 trigonometric functions, hyperbolic 10-42 to 10-50
Trunc function B-1
trunc function 9-11 to 9-12
 truncating floating-point to integer 4-3, 9-11 to 9-12
 types. *See* data formats

U

underflow 4-5
 assembler 12-11
 conversions 5-5, 5-7
 gradual 2-7
 UNDERFLOW SANE constant A-7
 unordered (comparison)
 assembler 12-7
 defined 6-4
 upward rounding 4-3
 ceil function 9-6 to 9-7

example 8-5
UPWARD SANE constant A-7

V

values, interpreting 2-4 to 2-11
variable types. *See* data formats
VAX H format B-2

W

widening for efficiency 7-3 to 7-4, A-9
_WIDEST_NEED_EVAL macro D-8
widest-need evaluation 3-5 to 3-6, D-5
 compared to minimum evaluation 3-8 to 3-11
 examples 3-8 to 3-11

Z

zero
 division by 1-9
 double-double format 2-15
 -0 as a result 2-10
 rounding toward 4-3, 9-11 to 9-12
 sign of 2-10 to 2-11
ZERONUM SANE constant A-6