Accelerator Programming

- Simple introductory program
- Programming model
- Building Accelerator programs
- High-level Programming
- Interpreting Compiler Feedback
- Tips and Techniques
- Performance Tuning

Jacobi Relaxation

```
change = tolerance + 1.0
while change > tolerance
    change = 0
    do i = 2, m-1
        do j = 2, n-1
            newa[i,j] = w0*a[i,j] + &
                w1 * (a[i-1,j] + a[i,j-1] + &
                    a[i+1,j] + a[i,j+1]) + &
                w2 * (a[i-1,j-1] + a[i-1,j+1] + &
                    a[i+1,j-1] + a[i+1,j+1])
            change = max(change, abs(newa[i,j] - a[i,j]))
        enddo
    enddo
for i = 1, m-1: ++i
    for j = 1, n-1: ++j
        a[i,j] = newa[i,j]
    enddo
wend
```
Jacobi Relaxation

change = tolerance + 1.0
!$omp parallel shared(change)
do while(change > tolerance)
  change = 0
!$omp do reduction(max:change) private(i,j)
do i = 2, m-1
  do j = 2, n-1
    newa(i,j) = w0*a(i,j) + w1 * (a(i-1,j) + a(i,j-1) + a(i+1,j) + a(i,j+1)) + w2 * (a(i-1,j-1) + a(i-1,j+1) + a(i+1,j-1) + a(i+1,j+1))
    change = max(change,abs(newa(i,j)-a(i,j)))
  enddo
  change = max(change,abs(newa(1,1)-a(1,1)))
enddo
a(2:m-1,2:n-1) = newa(2:m-1,2:n-1)
enddo

#pragma acc data region local(newa[0:n-1][0:m-1])
  copy(a[0:n-1][0:m-1])
{ do{
  change = 0;
  #pragma acc region
  { for( i = 1; i < m-1; ++i )
    for( j = 1; j < n-1; ++j )
      newa[j][i] = w0*a[j][i] + w1 * (a[j][i-1] + a[j][i+1] + a[j+1][i] + a[j-1][i-1] + a[j-1][i+1] + a[j+1][i-1] + a[j+1][i+1]);
    change = fmax(change,fabs(newa[j][i]-a[j][i]));
  }
  for( i = 1; i < m-1; ++i )
  for( j = 1; j < n-1; ++j )
      a[j][i] = newa[j][i];
} while( change > tolerance ); }

Abstracted x64+Tesla Accelerator Architecture
Why use Accelerator Directives?

- **Productivity**
  - Higher level programming model
  - a la OpenMP

- **Portability**
  - Ignore directives, portable to the host
  - Portable to other accelerators
  - Performance portability

- **Performance feedback**

- **Downsides**
  - It's not as easy as inserting a few directives
  - A good host algorithm is not necessarily a good GPU algorithm

Basic Syntactic Concepts

- **Fortran accelerator directive syntax**
  - !$acc directive [clause]...
  - & continuation
  - Fortran-77 syntax rules
    - !$acc or $acc in columns 1-5
    - continuation with nonblank in column 6

- **C accelerator directive syntax**
  - #pragma acc directive [clause]... eol
  - Continue to next line with backslash

Region

- **region is single-entry/single-exit region**
  - In Fortran, delimited by begin/end directives
  - In C, a single statement, or {...} region
  - No jumps into/out of region, no return

- **compute region contains loops to send to GPU**
  - Loop iterations translated to GPU threads
  - Loop indices become threadidx/blockidx indices

- **data region encloses compute regions**
  - Data moved at region boundaries
### Appropriate Algorithm

- Nested parallel loops
  - Iterations map to threads
  - Parallelism means threads are independent
  - Nested loops mean lots of parallelism
- Regular array indexing
  - Allows for stride-1 array fetches

```c
#pragma acc data region local(newa[0:n-1][0:m-1])
copy(a[0:n-1][0:m-1])
{
do{
  change = 0;
  #pragma acc region
  {
    for (i = 1; i < m-1; ++i)
      for (j = 1; j < n-1; ++j)
      {
        newa[j][i] = w0*a[j][i] +
                    w1 * (a[j][i-1] + a[j][i+1] +
                          a[j-1][i] + a[j+1][i]) +
                    w2 * (a[j-1][i-1] + a[j-1][i+1] +
                          a[j+1][i-1] + a[j+1][i+1]);
        change = fmax(change,fabs(newa[j][i]-a[j][i]));
      }
  }
  for (i = 1; i < m-1; ++i)
    for (j = 1; j < n-1; ++j)
      a[j][i] = newa[j][i];
}while( change > tolerance );
```

### Behind the Scenes

- Compiler determines parallelism
- Compiler generates thread code
  - Split up the iterations into threads, thread groups
  - Inserts code to use software data cache
  - Accumulate partial sum
  - Second kernel to combine final sum
- Compiler also inserts data movement
  - Compiler or user determines what data to move
  - Data moved at boundaries of data/compute region

- Virtualization penalties
  - Fine grain control
  - Thread scheduling
  - Shared memory usage
  - Loop unrolling
Building Accelerator Programs

- pgfortran -ta=nvidia a.f90
- pgcc -ta=nvidia a.c

Other options:
- -ta=nvidia[,cc10|cc11|cc12|cc13|cc20]
  - default in siterc file:
    - set COMPUTECAP=13;
- -ta=nvidia[,cuda2.3|cuda3.1|cuda3.2]
  - default in siterc file:
    - set DEFCUDAVERSION=3.1;
- -ta=nvidia,time
- -ta=nvidia,host

Enable compiler feedback with -Minfo or -Minfo=accel

Performance Goals

- Data movement between Host and Accelerator
  - minimize amount of data
  - minimize number of data moves
  - minimize frequency of data moves
  - optimize data allocation in device memory

- Parallelism on Accelerator
  - Lots of MIMD parallelism to fill the multiprocessors
  - Lots of SIMD parallelism to fill cores on a multiprocessor
  - Lots more MIMD parallelism to fill multithreading parallelism

- Data movement between device memory and cores
  - minimize frequency of data movement
  - optimize strides – stride-1 in vector dimension
  - optimize alignment – 16-word aligned in vector dimension
  - store array blocks in data cache

- Other goals?
  - minimize register usage?
  - small kernels vs. large kernels?
  - minimize instruction count?
  - minimize synchronization points?
Program Execution Model

- **Host**
  - executes most of the program
  - allocates accelerator memory
  - initiates data copy from host memory to accelerator
  - sends kernel code to accelerator
  - queues kernels for execution on accelerator
  - waits for kernel completion
  - initiates data copy from accelerator to host memory
  - deallocates accelerator memory

- **Accelerator**
  - executes kernels, one after another
  - concurrently, may transfer data between host and accelerator

Compute Region

- **C**
  ```
  #pragma acc region
  { .... }
  ```

- **Fortran**
  ```
  !$acc region
  ....
  !$acc end region
  ```

PGI Accelerator Compilers

- Host x64 asm File
  - **compile**
  - Auto-generated GPU code
    - `...` no change to existing makefiles, scripts, IDEs, programming environment, etc.

- Unified a.out
  - `...` execute
Compute Region Clauses

C
#pragma acc region if(n > 100)
{
    for( i = 0; i < n; ++i ) r[i] = a[i]*2.0f;
}

Fortran
!$acc region if( n.gt.100 )
do i = 1,n
    r(i) = a(i) * 2.0
enddo
!$acc end region

Compute Region Clauses

C
#pragma acc region copyin(a[0:n-1])
{
    for( i = 0; i < n; ++i ) r[i] = a[i]*2.0f;
}

Fortran
!$acc region copyin(a)
do i = 1,n
    r(i) = a(i) * 2.0
enddo
!$acc end region

Compute Region Clauses

C
#pragma acc region copyin(a[0:n-1]) copyout(r)
{
    for( i = 0; i < n; ++i ) r[i] = a[i]*2.0f;
}

Fortran
!$acc region copyin(a[1:n]) copyout(r)
do i = 1,n
    r(i) = a(i) * 2.0
enddo
!$acc end region
Compute Region Clauses

C

#pragma acc region copyin(a[0:n-1]) \ 
    copyout(r[0:n-1])
{
    for( i = 0; i < n; ++i ) x[i] = a[i]*2.0f;
}

Fortran

!$acc region copyin(a(1:n)) copyout(r(1:n))
    do i = 1,n
        r(i) = a(i) * 2.0  
    enddo  
!$acc end region

Compute Region Clauses

C

#pragma acc region copyin(a[0:n-1][0:m-1]) \ 
    copy(r[0:n-1][0:m-1])
{
    for( i = 1; i < n-1; ++i )
        for( j = 1; j < m-1; ++j )
            r[i][j] = a[i][j]*2.0f;
}

Fortran

!$acc region copyin(a(1:m,1:n))
    do i = 2,n-1  
        do j = 2,m-1  
            r(j,i) = a(j,i) * 2.0  
        enddo  
    enddo  
!$acc end region

Compute Region Clauses

C

#pragma acc region copyin(a[0:n-1][0:m-1]) \ 
    copy(r[0:n-1][0:m-1])
{
    for( i = 1; i < n-1; ++i )
        for( j = 1; j < m-1; ++j )
            r[i][j] = a[i][j]*2.0f;
}

Fortran

!$acc region copyin(a(1:m,1:n)) local(r)
    do times = 1,niters  
        do i = 2,n-1  
            do j = 2,m-1  
                r(j,i) = 0.25*(a(j-1,i)+a(j,i-1)+a(j+1,i)+a(j,i+1))  
            enddo  
        enddo  
    enddo  
!$acc end region

Compute Region Clauses

C

#pragma acc region copyin(a[0:n-1][0:m-1]) \ 
    copy(r[0:n-1][0:m-1])
{
    for( i = 1; i < n-1; ++i )
        for( j = 1; j < m-1; ++j )
            r[i][j] = a[i][j]*2.0f;
}

Fortran

!$acc region copyin(a(1:m,1:n))
    do i = 2,n-1  
        do j = 2,m-1  
            r(j,i) = a(j,i) * 2.0  
        enddo  
    enddo  
!$acc end region

© Copyright 2009-2011, The Portland Group, Inc.
Compute Region Clauses

C
#pragma acc region copyin(a[0:n-1][0:m-1])
/* data copied to Accelerator here */
{
  for( i = 1; i < n-1; ++i )
    for( j = 1; j < m-1; ++j )
      r[i][j] = A[i][j]*2.0f;
} /* data copied to Host here */

Fortran
!$acc region copyin(a(1:m,1:n))
! data copied to Accelerator here
do i = 2,n-1
  do j = 2,m-1
    r(j,i) = a(j,i) * 2.0
  enddo
enddo
! data copied to Host here
!$acc end region
Data Region

- C
  
  ```c
  #pragma acc data region
  
  ....
  ```

- Fortran
  
  ```fortran
  !$acc data region
  ....
  !$acc end data region
  ```

- May be nested and may contain compute regions
- May not be nested within a compute region

Data Region Clauses

- Data allocation clauses
  - copy(list)
  - copyin(list)
  - copyout(list)
  - local(list)
  - data in the lists must be distinct (data in only one list)
  - may not be in a data allocate clause for an enclosing data region

- Data update clauses
  - updatein(list) or update_device(list)
  - updateout(list) or update_host(list)
  - data must be in a data allocate clause for an enclosing data region

Data Region Update Directives

- update_host(list)
- update_device(list)
  - data must be in a data allocate clause for an enclosing data region
  - both may be on a single line
  - update_host(list) device(list)
Passing Device Copies

```fortran
subroutine sub(a, b)
  real :: a(:), b(:)
  !$acc reflected(a)
  !$acc region copyin(b)
  do i = 1, n
    a(i) = a(i) * b(i)
  enddo
  !$acc end region
end subroutine
```

```fortran
subroutine bus(x, y)
  real :: x(:), y(:)
  !$acc data region copy(x)
  call sub(x, y)
end subroutine
```

- Reflected clause only available in Fortran
- List of argument arrays
- Caller must have a visible device copy at the call site
- Subprogram interface must be explicit
  - interface block or module
- Compiler will pass a hidden argument corresponding to the device copy

How did we make Vectors Work?

Compiler-to-Programmer Feedback – a classic “Virtuous Cycle”

- Directives, Options, Restructuring
- Vectorization

This Feedback Loop Unique to Compilers!

We can use this same methodology to enable effective migration of applications to Multi-core and Accelerators
Compiler-to-Programmer Feedback

- Directives, Options, RESTRUCTURING
- HPC Code
- PGI Compiler
- HPC User
- Performance
- x64 + Acc
- Trace
- PGPROF

Restructuring for Accelerators will be More Difficult

HPC Code

Compiler-to-User Feedback

% pgfortran -fast -tnvidia -Minfo mm.F90
mm1:
6. Generating copyout(a(1:m,1:m))
Generating copyin(c(1:m,1:m))
Generating copyin(b(1:m,1:m))
7. Loop is parallelizable
8. Loop is parallelizable
Accelerator kernel generated
7. !$acc do parallel, vector(16)
8. !$acc do parallel, vector(16)
11. Loop carried reuse of 'a' prevents parallelization
12. Loop is parallelizable
Accelerator kernel generated
7. !$acc do parallel, vector(16)
11. !$acc do seq
Cached references to size [16x16] block of 'b'
Cached references to size [16x16] block of 'c'
12. !$acc do parallel, vector(16)
Using register for 'a'

Common Compiler Feedback Format
http://www.pgroup.com/ccff

Source File
Compiler
Linker
pgextract
Object File
 executable File
 pgprof
pgprof.out File

PGPROF with CCFF Messages

© Copyright 2009-2011, The Portland Group, Inc.
General Compiler Feedback

- How the function was compiled
- Interprocedural optimizations
- Profile-feedback runtime data
  - Block execution counts
  - Loop counts, range of counts
- Compiler optimizations, missed opportunities
  - Vectorization, parallelization
  - Altcode, re-ordering of loops, inlining
  - X64+GPU code generation, GPU kernel mapping, data movement
- Compute intensity – important for GPUs & Multi-core

Accelerator Compiler Feedback

- Generating copyin(b(1:n,1:m))
- Generating copyout(b(2:n-1,2:m-1))
- Generating copy(a(1:n,1:m))
- Generating local(c(1:n,1:m))
- Loop is parallelizable
- Accelerator kernel generated
- No parallel kernels found, accelerator region ignored
- Loop carried dependence due to exposed use of ... prevents parallelization
- Parallelization would require privatization of array ...

Accelerator Compiler Feedback

- hindrances
  - live out variables, non-private arrays
  - loop-carried dependences
  - unsupported datatype, unsupported operation
  - unknown array bounds
- data movement
  - copyin, copyout, local
- optimizations performed, not performance
  - cached data
  - non-stride-1 accesses (non-coalesced)
- loop schedule
  - mapping of loop iterations to thread/block indices

What can appear in a Compute Region?

- Arithmetic
  - C: int, float, double, struct
  - F: integer, real, double precision, complex, derived types
  - Loops, Ifs
  - Kernel loops must be rectangular: trip count is invariant
- Obstacles with C
  - unbound pointers – use restrict keyword, or -Msaferptr, or -Mipa=fast
  - default is double – use float constants (0.0f), or -Mfcon, and float intrinsics
- Obstacles with Fortran
  - Fortran pointer attribute is not supported
C Intrinsics

- C: #include <accelmath.h>

<table>
<thead>
<tr>
<th>acos</th>
<th>asin</th>
<th>atan</th>
<th>atan2</th>
</tr>
</thead>
<tbody>
<tr>
<td>cos</td>
<td>cosh</td>
<td>exp</td>
<td>fabs</td>
</tr>
<tr>
<td>fmax</td>
<td>fmin</td>
<td>log</td>
<td>log10</td>
</tr>
<tr>
<td>pow</td>
<td>sin</td>
<td>sinh</td>
<td>sqrt</td>
</tr>
<tr>
<td>tan</td>
<td>tanh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>acosf</td>
<td>asinf</td>
<td>atanf</td>
<td>atan2f</td>
</tr>
<tr>
<td>cosf</td>
<td>coshf</td>
<td>expf</td>
<td>fabsf</td>
</tr>
<tr>
<td>fmaxf</td>
<td>fminf</td>
<td>logf</td>
<td>log10f</td>
</tr>
<tr>
<td>powf</td>
<td>sinf</td>
<td>sinhf</td>
<td>sqrtf</td>
</tr>
<tr>
<td>tanf</td>
<td>tanhf</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fortran Intrinsics

- Abs
- Acos
- Asin
- Atan
- Atan2
- Cos
- Cosh
- Dble
- Exp
- Iand
- Ior
- Int
- Log
- Log10
- Max
- Min
- Mod
- Not
- Rea
- Sign
- Sin
- Sinh
- Sqrt
- Tan
- Tanh

Other functions

- Libm routines
  - use libm
  - #include <accelmath.h>
- Device built-in routines
  - use cudadevice
  - #include <cudadevice.h>

Obstacles to Parallelization

- Unknown dependences
  - use simple array subscripts
- Unstructured Accumulations
  - rewrite to accumulate to local element
- Scalars whose last value is needed after the loop
  - declare local, or private, or rename in the loop
  - global variables, Fortran dummy arguments, are always live-out
Unstructured Accumulation

```fortran
do i = 1, n
  x = someveryexpensivecomputation(i)
  ni = nbridx(i)
  do nbr = 1, nbrcnt(i)
    j = nbr(ni+nbr)
    flow(j) = flow(j) + x
  enddo
enddo
```

Restructured Accumulation

```fortran
do i = 1, n
  x(i) = someveryexpensivecomputation(i)
enddo

do j = 1, n
  nj = invnbridx(i)
  do nbr = 1, invnbrcnt(j)
    i = invnbr(nj+nbr)
    flow(j) = flow(j) + x(i)
  enddo
enddo
```

Loop Schedules

- vector loops correspond to threadidx indices
- parallel loops correspond to blockidx indices
- this schedule has a CUDA schedule
  `<<< dim3(ceil(N/16),ceil(M/16)), dim3(16,16) >>>`
- Compiler strip-mines to protect against very long loop limits

Common Loop Schedules

- parallel
- vector
- vector(n)
- parallel, vector(n)
- seq
- host
Loop Directive

- C
  ```c
  #pragma acc for clause...
  for( i = 0; i < n; ++i ){
      ....
  }
  ```

- Fortran
  ```fortran
  !$acc do clause...
  do i = 1, n
  ```

Loop Scheduling Clauses

- !$acc do parallel
  - runs in 'parallel' mode only (blockIdx)
  - does not declare that the loop is in fact parallel (use independent)

- !$acc do parallel(32)
  - runs in 'parallel' mode only with gridDim == 32 (32 blocks)

- !$acc do vector(128)
  - runs in 'vector' mode (threadIdx) with blockDim == 128 (128 threads)
  - vector size, if present, must be compile-time constant

- !$acc do parallel vector(128)
  - strip mines loop
  - inner loop runs in vector mode, 128 threads (threadIdx)
  - outer loop runs across thread blocks (blockIdx)

Loop Scheduling Clauses

- Want stride-1 loop to be in 'vector' mode (threadIdx)
  - look at -Minfo messages!

- Want lots of parallelism

Kernel Clause

```fortran
!$acc region
  do i = 1, n
      do j = 2, m-1
          a(i,j) = 0.5*(b(i,j-1) + b(i,j+1))
      enddo
      do j = 2, m-1
          b(i,j) = a(i,j)
      enddo
  enddo
!$acc end region
```
Kernel Clause

!$acc region
  !$acc do vector
do i = 1, n
  !$acc do parallel
do j = 2, m-1
  a(i,j) = 0.5*(b(i,j-1) + b(i,j+1))
  enddo
dodo
  !$acc do vector
do i = 1, n
  !$acc do parallel
do j = 2, m-1
  b(i,j) = a(i,j)
  enddo
enddo
!$acc end region

Kernel Clause

!$acc region
  !$acc do parallel vector(128)
do i = 1, n
  !$acc do seq
do j = 2, m-1
  a(i,j) = 0.5*(b(i,j-1) + b(i,j+1))
  enddo
  !$acc do seq
  do j = 2, m-1
  b(i,j) = a(i,j)
  enddo
enddo
!$acc end region

Kernel Clause

!$acc region
  !$acc do parallel vector(128)
  kernel
do i = 1, n
  do j = 2, m-1
  a(i,j) = 0.5*(b(i,j-1) + b(i,j+1))
  enddo
  do j = 2, m-1
  b(i,j) = a(i,j)
  enddo
enddo
!$acc end region

Loop Directive Clauses

- Scheduling Clauses
  - vector or vector(n)
  - parallel or parallel(n)
  - seq or seq(n)
  - host or host(n)
  - unroll(n) (after parallel, vector, seq)
- kernel
  - useful when you want a large outer loop body to be the kernel
- independent
  - use with care, overrides compiler analysis for dependence, private variables
- private(list)
  - private data for each iteration of the loop
  - different from local (how?)

© Copyright 2009-2011, The Portland Group, Inc.
Compiler Feedback Messages

- Data related
  - Generating copyin(b(1:n,1:m))
  - Generating copyout(b(2:n-1,2:m-1))
  - Generating copy(a(1:n,1:n))
  - Generating local(c(1:n,1:n))

- Loop or kernel related
  - Loop is parallelizable
  - Accelerator kernel generated

- Barriers to GPU code generation
  - No parallel kernels found, accelerator region ignored
  - Loop carried dependence due to exposed use of ... prevents parallelization
  - Parallelization would require privatization of array ...

Compiler Messages Continued

- Memory optimization related
  - Cached references to size [18x18] block of 'b'
  - Non-stride-1 memory accesses for 'a'

Time for a Live Demo

- Himeno benchmark
  - Effects of using clauses on parallel schedule / performance
  - Effects of padding
Obstacles to Parallelization

- Computed Index (linearization, look-up)
- While Loops
- Triangular Loops
- "live-out" Variables
- Privatization of Local Arrays
- Function calls
- Device Runtime Errors
- Compiler Errors

Computed Index – Linearization

```plaintext
!$acc region
do i = 1, M
  do j = 1, N
    idx = ((i-1)*M)+j
    A(idx) = B(i,j)
  enddo
enddo
!$acc end region
```

To fix,
Remove the linearization or Use the "independent" clause

```plaintext
!$acc region
!$acc do independent
do i = 1, M
  do j = 1, N
    idx = ((i-1)*M)+j
    A(idx) = B(i,j)
  enddo
enddo
!$acc end region
```

Computed Index – Look-up

```plaintext
!$acc region
do i = 1, M
  idx = lookup(i)
do j = 1, N
  A(idx,j) = ((i-1)*M)+j
enddo
!$acc end region
```

To fix,
Remove the linearization or Use the "independent" clause

```plaintext
!$acc region
!$acc do independent
  do i = 1, M
    idx = lookup(i)
    A(idx,j) = ((i-1)*M)+j
  enddo
!$acc end region
```

While Loops

```plaintext
!$acc region
  i = 0
  do, while (.not.found)
    i = i + 1
    if (A(i) .eq. 102) then
      found = i
    endif
  enddo
!$acc end region
```

To fix,
Change to a rectangular loop

```plaintext
!$acc region
  do i = 1, N
    if (A(i) .eq. 102) then
      found(i) = i
    else
      found(i) = 0
    endif
  enddo
!$acc end region
```

The Independent or parallel clauses could be used to force parallelization but is not recommended

```
print *, 'Found at ', maxval(found)
```
**Triangular Loops**

```
!$acc region copyout(A)
do i = 1, M 
do j = i, N 
A(i,j) = i+j 
enddo
enddo
!$acc end region
```

All loop schedules must be rectangular. For triangular loops, the compiler will either serialize the inner loop or make the inner loop rectangular and add an implicit if statement to skip the lower part of the triangle.

Problem: The compiler will copy out the entire array A. The lower triangle contains garbage since it was not initialized. Use "copy(A)" to initialize the values.

---

**“live-out” Variables**

```
!$acc region
  do i = 1, M 
  do j = 1, N 
    idx = i+j 
    A(i,j) = idx 
  enddo 
enddo
!$acc end region
print *, idx, A(1,1), A(M,N)
```

```
% pgf90 -ta=nvidia,time -Minfo=accel liveout.f90
liveout:
11, Generating copyout(a(1:1024,1:1024))
12, Loop is parallelizable
  Accelerator kernel generated
13, Inner sequential loop scheduled on accelerator
14, Accelerator restriction: induction variable live-out from loop: idx
15, Accelerator restriction: induction variable live-out from loop: idx
```

Privatize the scalar

```
!$acc region
  do i = 1, M 
  !$acc do private(idx)
  do j = 1, N 
    idx = i+j 
    A(i,j) = idx 
  enddo 
enddo
!$acc end region
print *, idx, A(1,1), A(M,N)
```

```
% pgf90 -ta=nvidia,time -Minfo=accel liveout2.f90
liveout2:
10, Generating copyout(a(1:1024,1:1024))
11, Loop is parallelizable
13, Loop is parallelizable
```

**Privatization of Local Arrays**

```
!$acc region
  do i = 1, M 
  do j = 1, N 
  do jj = 1, 10 
    tmp(jj) = jj 
  enddo 
  A(i,j) = sum(tmp) 
enddo
!$acc end region
```

```
% pgf90 -ta=nvidia,time -Minfo=accel private.f90
privatearr:
10, Generating copyout(tmp(1:10))
Generating compute capability 1.0 binary
Generating compute capability 1.3 binary
11, Parallelization would require privatization of array 'tmp(1:10)'
Sequential loop scheduled on host
14, Loop is parallelizable
  Accelerator kernel generated
15, !$acc do parallel, vector(10)
```

```
% pgf90 -ta=nvidia,time -Minfo=accel private2.f90
privatearr2:
10, Generating copyout(tmp(1:10))
Generating compute capability 1.0 binary
Generating compute capability 1.3 binary
11, Loop is parallelizable
13, Loop is parallelizable
```

**Privatization of Local Arrays - cont.**

```
!$acc region
  do i = 1, M 
  !$acc do private(tmp)
  do j = 1, N 
  do jj = 1, 10 
    tmp(jj) = jj 
  enddo 
  A(i,j) = sum(tmp) 
enddo
!$acc end region
```

```
% pgf90 -ta=nvidia,time -Minfo=accel private2.f90
privatearr2:
10, Generating copyout(a(1:1024,1:1024))
Generating compute capability 1.0 binary
Generating compute capability 1.3 binary
11, Loop is parallelizable
13, Loop is parallelizable
```

Need to privatize local temporary arrays. Default is to assume that they are shared.
Function Calls

- Function calls are not allowed within a compute region.
- Restriction is due to lack of a device linker and hardware support.
- Functions must be inlined, either manually or by the compiler with –Minline or –Mipa=inline.

Device Errors

call to cuMemcpyDtoH returned error 700: Launch failed
- Typically occurs when the device kernel gets an execution error, such as an out-of-bounds or other memory access violation.

call to cuMemcpy2D returned error 1: Invalid value
- Occurs if there is an error when copying data to/from the device

Compiler Errors

- No software is without bugs, including the compiler
- If you encounter a problem that seems to be the fault of the compiler, please send a report to PGI Customer Service (trs@pgroup.com)

Example of an Internal Compiler Error (ICE)

```
% pgf90 -ta=nvidia -c bug.f90
http://pgaccRVzoy6dyF/gpu(20): error: identifier "a0" is undefined
1 error detected in the compilation of "http://pgnvdASVc89Pmxo9p.nv0".
PGF90-F-2006-Internal compiler error. pgnvd job exited with nonzero status code 0 (bug.f90: 22)
```

Combined Directives

- C
  ```
  #pragma acc region for copyin(a[0:n-1])
  parallel vector(32)
  for( i = 0; i < n; ++i ) r[i] = a[i]*2.0f;
  ```

- Fortran
  ```
  !$acc region do copyin(a(1:n)) parallel vector(32)
  do i = 1,n
    r(i) = a(i) * 2.0
  enddo
  ```
Selecting Device

- C
  ```c
#include <accel.h>
  n = acc_get_num_devices(acc_device_nvidia);
  ...
  acc_set_device_num( 1, acc_device_nvidia);
  ```

- Fortran
  ```fortran
use accel_lib
  n = acc_get_num_devices( acc_device_nvidia )
  ...
  call acc_set_device_num( 0, acc_device_nvidia)
  ```

- Environment Variable
  ```
setenv ACC_DEVICE_NUM 1
export ACC_DEVICE_NUM=1
  ```

PGI Unified Binary

- One binary that executes on GPU or on host
  - default: auto-detect presence of GPU, use GPU if present
  - override default with environment variable `ACC_DEVICE`
  - override within program with call to `acc_set_device`

Building PGI Unified Binary

- `pgfortran -ta=nvidia,host -fast program.f90`

Running PGI Unified Binary

- `a.out`
- `setenv ACC_DEVICE host ; a.out`
- `setenv ACC_DEVICE nvidia ; a.out`

before first region
- `call acc_set_device( acc_device_nvidia )`
- `call acc_set_device( acc_device_host )`

PGI Unified Binary Interactions

- PGI Unified Binary for multiple host types:
  ```
pforeach -tp=barcelona-64,core2-64
  up to H optimized copies of each routine with H host types
  ```

- PGI Unified Binary for multiple accelerator types:
  ```
pforeach -ta=nvidia,host
  up to two optimized copies of each routine (with accelerator regions)
  ```

- PGI Unified Binary for both multiple host and accelerator types:
  ```
pforeach -tp=barcelona-64,core2-64 -ta=nvidia,host
  with H host types and A accelerator types:

  - up to H optimized copies for each of H hosts, plus
  - up to A accelerator-enabled copies, optimized for “generic” host
  ```

© Copyright 2009-2011, The Portland Group, Inc.
Performance Profiling

- TAU (Tuning and Analysis Utilities, University of Oregon)
  - collects performance information
- cudaprof (NVIDIA)
  - gives a trace of kernel execution
- pgfortran -ta=nvidia,time (on link line)
  - dump of region-level and kernel-level performance
  - upload/download data movement time
  - kernel execution time
- pgcollect a.out
- PGI_ACC_PROFILE environment variable
  - enables profile data collection for accelerator regions
- ACC_NOTIFY environment variable
  - prints one line for each kernel invocation

Performance Profiling

Accelerator Kernel Timing data
f3.f90
smooth
- 24: region entered 1 time
time(us): total=1116701 init=1115986 region=715
  kernels=22 data=693
  w/o init: total=715 max=715 min=715 avg=715
- 27: kernel launched 5 times
grid: [7x7] block: [16x16]
time(us): total=17 max=10 min=1 avg=3
- 34: kernel launched 5 times
grid: [7x7] block: [16x16]
time(us): total=5 max=1 min=1 avg=1

Profiling an Accelerator Model Program

```
$ make
pgf90 -fast -Minfo=ccff ../mm2.f90 ../mmdriv.f90 -o mm2 -ta=nvidia
../mm2.f90:
../mmdriv.f90:
$ pgcollect -time ./mm2 < ../in
[...program output...]
target process has terminated, writing profile data
$
$ pgprof -exe ./mm2
```

- Build as usual
  - Here we add -Minfo=ccff to enhance profile data
- Just invoke with pgcollect
  - Currently collects data similar to -ta=time
- Invoke the PGPROF performance profiler

Dynamic Profile Data Collection (11.4)

```
acc_enable_time( acc_device_nvidia );
...
long etime = acc_exec_time( acc_device_nvidia );
long ttime = acc_total_time( acc_device_nvidia );
acc_disable_time( acc_device_nvidia );

unsigned int allocs = acc_allocs();
unsigned int frees = acc_frees();
unsigned int copyins = acc_copyins();
unsigned int copyouts = acc_copyouts();
unsigned long bytesalloc = acc_bytesalloc();
unsigned long bytesin = acc_bytesin();
unsigned long bytesout = acc_bytesout();
unsigned int kernels = acc_kernels();
unsigned int regions = acc_regions();
unsigned long totmem = acc_get_memory();
unsigned long freemem = acc_get_free_memory();
```
Directives Summary

- acc region
  - if(condition)
  - copy(list) copyin(list) copyout(list) local(list)
  - updatein(list) updateout(list)
- acc data region
  - copy(list) copyin(list) copyout(list) local(list)
  - updatein(list) updateout(list)
- acc do or acc for
  - host(width) parallel(width) seq(width) vector(width)
  - unroll(width)
  - kernel
  - independent
  - private(list)

Directives Summary

- update host(list)
- update device(list)
- _ACCEL macro
  #ifdef _ACCEL
  call acc_init( acc_device_nvidia )
  #endif

Behind the Scenes

- Live variable analysis and Array region analysis
  - augments information in region / loop directive clauses
  - determines in/out sets for the region
  - programmer feedback
- Dependence and parallelism analysis
  - determine loops that can be executed in parallel
  - programmer feedback
- Map loops onto hardware parallelism
  - parallel / vector (MIMD/SIMD), including strip mining
  - programmer feedback
- Kernel extraction, GPU code generation
  - kernel code optimization
  - same code generator as CUDA Fortran
- Host code generation
  - allocate memory, copy data, upload kernel code, launch kernel, copy results, deallocate memory
  - for unified binary, keep a host version of the loop

time for a live demo (6)
cudaprof
-ta=nvidia,time
pgcollect and pgprof
ACC_NOTIFY

© Copyright 2009-2011, The Portland Group, Inc.
**Runtime Summary**

- `acc_get_num_devices(acc_device_nvidia)`
- `acc_set_device_num(n, acc_device_nvidia)`
- `acc_set_device(acc_device_nvidia | acc_device_host)`
- `acc_get_device()`
- `acc_init(acc_device_nvidia | acc_device_host)`

**Environment Variable Summary**

- `ACC_DEVICE_NUM`
- `ACC_DEVICE`
- `ACC_NOTIFY`

**Command Line Summary**

- `-ta=nvidia`
- `-ta=nvidia,time`
- `-ta=nvidia,nofma`
- `-ta=nvidia,fastmath`
- `-ta=nvidia,mul24` (Tesla only)
- `-ta=nvidia,[cc10|cc11|cc12|cc13|cc20]` (multiple allowed)
- `-ta=nvidia,maxregcount:n`
- `-ta=nvidia,host`

**Multiple Devices**

- Nest Accelerator Region in OpenMP Parallel
- Each thread chooses one NVIDIA device
- Each thread computes its own loop limits
  - Not the same as `&omp for`
- One GPU per host thread (CUDA limitation)
- All threads/GPUs share PCI bus
Multiple Devices

```c
!$omp parallel private(ilo,ihi,i)
   num_threads(2)
   call acc_set_device(omp_get_thread_num())
ilo = omp_get_thread_num()*(N+1)/2 + 1
ihi = min(N,ilo+(N+1)/2)
!$acc region do
   do i = ilo,ihi
      a(i) = b(i) + c(i)
   enddo
!$omp end region
```

Performance Tuning

- Performance Measurement
- Choose an appropriately parallel algorithm
- Optimize data movement between host and GPU
  - frequency, volume, regularity
- Optimize device memory accesses
  - strides, alignment
  - use data cache
- Optimize kernel code
- Optimize compute intensity
- Linux only: pgcudainit
- Windows: must be on the console
- OSX: two GPUs doesn't mean you can use the idle one

Writing Accelerator Programs

- Step 1: Appropriate algorithm: lots of parallelism
  - Lots of MIMD parallelism to fill the multiprocessors
  - Lots of SIMD parallelism to fill cores on a multiprocessor
  - Lots more MIMD parallelism to fill multithreading parallelism
  - High compute intensity
  - Insert directives, read feedback for parallelism hindrances
    - Dependence relations, pointers
    - Scalars live out from the loop
    - Arrays that need to be private
    - Iterate

- Step 2: Tune data movement between Host and Accelerator
  - read compiler feedback about data moves
  - minimize amount of data moved
  - minimize frequency of data moves
  - minimize noncontiguous data moves
  - optimize data allocation in device memory
  - optimize allocation in host memory (esp. for C)
  - Insert local, copyin, copyout clauses
  - use data regions, update clauses
Manually managing memory (CUDA)

- CUDA Fortran device arrays
  - CUDA Fortran device arrays can be used in accelerator regions
  - compile with .cuf suffix or -Mcuda in addition to -ta=nvidia
- CUDA C device arrays
  - no attribute to tell the compiler where a pointer points
  - new in PGI 11.4
    - #pragma acc [data] region deviceptr(ap)
    - tells the compiler that ap was allocated on the GPU
    - ap = acc_malloc( nbytes ); ... acc_free( ap );
    - allocates / frees ap on the current device
    - You may also use data allocated by cudaMalloc or cuMemAlloc

Tuning Accelerator Programs

- Step 3: Tune data movement between device memory and cores
  - minimize frequency of data movement
  - optimize strides – stride-1 in vector dimension
  - optimize alignment – 16-word aligned in vector dimension
  - look at cache feedback from compiler

Tuning Accelerator Programs

- Step 4: Tune kernel code
  - profile the code (cudaprof, pgcollect, -ta=nvidia, time)
  - experiment with kernel schedule using loop directives
    - unroll clauses
    - enable experimental optimizations (-ta=nvidia,O3)
    - single precision vs. double precision
    - 24-bit multiply for indexing (-ta=nvidia,mul24)
    - low precision transcendental (-ta=nvidia,fastmath)

unroll clauses

- Consider each vector space as a ‘tile’ of iterations
  - !$acc do parallel unroll(2)
    - each thread block does two tiles
  - !$acc do parallel unroll(4) does four tiles
  - !$acc do vector(128) unroll(2)
    - tile size is 128 in this dimension
    - use thread block of ½ tile size (64) to do this tile
    - !$acc do vector(128) unroll(4) uses ¼ tile size (32)
  - !$acc do parallel unroll(2) vector(128) unroll(4)
    - strip mines loop to tile size 128
    - each thread block does two tiles
    - thread block size is (32 in this dimension)
- unroll width must be compile-time constant
Near Term Future Additions

- passing device data to subroutines (available now in Fortran)
- more code optimization
  - aggressive procedure inlining in compute regions
  - instruction count optimizations
  - post-optimal low-level code improvements
- automatic private array detection
- improvements to kernel schedule selection
- Direct calling of CUDA routines inside region
- Asynchronous data movement
- Asynchronous compute regions

Longer Term Evolution

- C++
- New targets
  - multicore, Larrabee?, ATI?, other?
- Overlap compute / data movement
  - pipelined loops
- More tools support
- Libraries of device routines
- Programming model evolution
  - multiple GPUs
  - standardization
  - Concurrent kernels, a la OpenCL / Fermi

PGI Accelerator vs CUDA/OpenCL

The PGI Accelerator programming model is a high-level implicit programming model for x64+GPU systems, similar to OpenMP for multi-core x64. The PGI Accelerator model:

- Enables offloading of compute-intensive loops and code regions from a host CPU to a GPU accelerator using simple compiler directives
- Implements directives as Fortran comments and C pragmas, so programs can remain 100% standard-compliant and portable
- Makes GPGPU programming and optimization incremental and accessible to application domain experts
- Is supported in both the PGF95 and PGCC C99 compilers

CUDA is a lower-level explicit programming model with substantial runtime library components that give expert programmers direct control of:

- Splitting up of source code into host CPU code and GPU compute kernel code in appropriately defined functions and subprograms
- Allocation of page-locked host memory, GPU device main memory, GPU constant memory and GPU shared memory
- All data movement between host main memory and the various types of GPU memory
- Definition of thread/block grids and launching of compute kernels
- Synchronization of threads within a CUDA thread group
- Asynchronous launch of GPU compute kernels, synchronization with host CPU

OpenCL is similar to CUDA, adds some functionality, even lower level
Why Not Just Use OpenMP?

- Caveat: PGI is an active member of the OpenMP ARB and language committee
  - OpenMP is looking at expanding its model to GPUs starting with the PGI Accelerator model
- OpenMP model
  - shared memory with weak memory model
  - long-lived threads + work sharing
  - defined synchronization points
  - dynamic scheduling, nested parallelism
  - model is prescriptive, not descriptive

Why Not Just Use OpenMP?

- Preserve all of OpenMP
  - simulate OpenMP thread behavior on GPU
  - simulate scheduling, synchronization, critical sections
- Only allow a subset of OpenMP
  - define a subset that maps well
  - disallow dynamic schedules, synchronization
- Break OpenMP, define new behavior
  - perhaps define master thread as the host, worker threads as on the GPU

OpenMP Accelerator Subcommitee

- !$omp acc_region [clause]...
  #pragma omp acc_region [clause]...
  - acc_shared( list )
  - acc_copy( list )
  - acc_copyin( list )
  - acc_copyout( list )
  - private( list )
  - present( list )
  - num_pes( depth: num )
  - host_shared( list )
  - device( expression )

OpenMP Accelerator Subcommitee

- !$omp acc_data [clause]...
  #pragma omp acc_data [clause]...
  - acc_shared( list )
  - acc_copy( list )
  - acc_copyin( list )
  - acc_copyout( list )
  - present( list )
  - host_shared( list )
  - device( expression )
OpenMP Accelerator Subcommitee

- !$omp acc_loop [clause]...
  #pragma omp acc_loop [clause]...
  - cache( obj:depth )
  - collapse( n )
  - level( n )
  - num_pes( depth:n )
  - reduction( operator:name )
  - schedule( sched-name )
  - hetero( n, width )

PGI’s Dilemma

- Use a standard, but one that fits badly
  - some programs will run well, most will not
  - does not take advantage of accelerator strengths
- OR, define a new model that fits well
  - multidimensional parallelism
  - multivariate scheduling options
  - defined data movement points
  - use compiler feedback to drive porting / tuning

Availability and Additional Information

- PGI Accelerator Programming Model – is supported for x64+NVIDIA targets in the PGI Fortran and C compilers, available now
- Other GPU and Accelerator Targets – are being studied by PGI, and may be supported in the future as the necessary low-level software infrastructure (e.g. OpenCL) becomes more widely available
- Further Information – see www.pgroup.com/accelerate for a detailed specification of the PGI Accelerator model, an FAQ, and related articles and white papers

Where to get help

- PGI Customer Support - trs@pgroup.com
Copyright Notice

© Contents copyright 2009-2011, The Portland Group, Inc. This material may not be reproduced in any manner without the expressed written permission of The Portland Group.