PROGRAMMING GPU-ACCELERATED POWER SYSTEMS WITH OPENACC

GPU TECHNOLOGY CONFERENCE 2019

20 March 2019 | Andreas Herten | Forschungszentrum Jülich *Handout Version*

Member of the Helmholtz Association

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

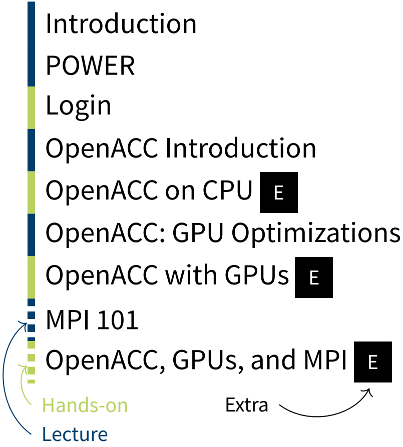
Overview, Outline

What you will learn today

- What's special about GPU-equipped POWER systems
- Parallelization strategies with OpenACC
- OpenACC on CPU, GPU, (GPUs)
- *All in 120 minutes*

What you will not learn today

- Analyze program in-detail
- Strategies for complex programs
- Using multiple GPUs



Introduction

POWER

Login

OpenACC Introduction

OpenACC on CPU E

OpenACC: GPU Optimizations

OpenACC with GPUs E

MPI 101

OpenACC, GPUs, and MPI E

Hands-on

Lecture

Extra

Overview, Outline

What you will learn today

- What's special about GPU-equipped POWER systems
- Parallelization strategies with OpenACC
- OpenACC on CPU, GPU, (GPUs)
- *All in 120 minutes*

What you will not learn today

- Analyze program in-detail
- Strategies for complex programs
- Using multiple GPUs

- Introduction
- OpenPOWER
  - Newell, POWER9
  - Using Nimbix
- OpenACC Introduction
  - About OpenACC
  - Modus Operandi
  - OpenACC's Models
  - Parallelization Workflow
- First Steps in OpenACC
  - Example Program
  - Identify Parallelism
  - Parallelize Loops
    - `parallel`
    - `loops`
    - `kernels`
- OpenACC on the GPU
  - Compiling on GPU
  - Data Locality
    - `copy`
    - `data`
    - `enter data`
- OpenACC on Multiple GPUs
  - MPI 101
  - Jacobi MPI Strategy
  - Asynchronous
- Conclusions, Summary
- Appendix
- List of Tasks

Jülich

Jülich Supercomputing Centre

- Forschungszentrum Jülich: One of largest research centers in Europe
- Jülich Supercomputing Centre (JSC): Host of and research in supercomputers
  - JUWELS Large-scale Intel x86 system; some GPUs
  - JURECA Multi-purpose Intel x86 system; some GPUs, many KNLs
- Me: Physicist, now working within JSC’s *NVIDIA Application Lab*

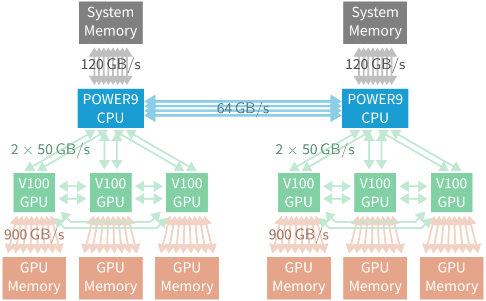
OpenPOWER Foundation

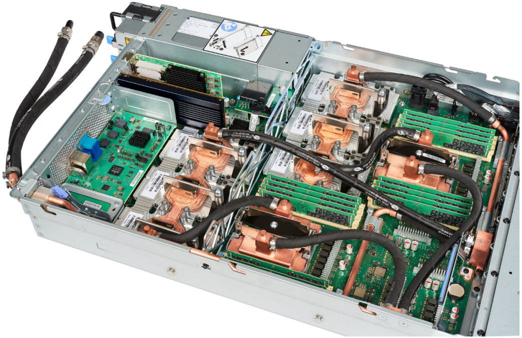
- Platform for collaboration around POWER processor architecture
- Started by IBM, NVIDIA, many more (now $>$ 250 members)
- Objectives
  - Licensing of processor architecture to partners
  - Collaborate on system extension
  - Open-Source Software
- Example technology: NVLink, fast GPU-CPU interconnect; CAPI

→ `https://openpowerfoundation.org/`

Newell

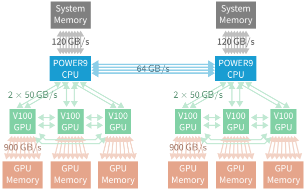
- POWER9 processors available in IBM servers AC922, codename *Newell*
- 2 IBM POWER9 CPUs, 3 or 2 NVIDIA Volta V100 GPUs, full NVLink 2





→ Appendix 1, 2

System Core Numbers



POWER9 CPU

- 2 sockets, each 16 (22) cores, each 4× SMT
- 2.3 GHz to 3.8 GHz; 8 FLOP/Cycle/Core
- 512 GB memory (120 GB/s)
- NVLink: 2 × 25 GB/s per GPU, per dir.
- L3, L2, L1 $ per core: 10 MB, 512 kB, 64 kB

0.5 TFLOP/s

NVLink2 (50 GB/s)

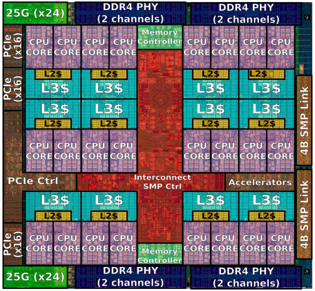
V100 GPU

- 84 Streaming Multiprocessors (SMs)
- SM: 64 INT32, 64 FP32, 32 FP64, 8 TC
- 16 GB (32) memory (900 GB/s)
- L2 $: 6 MB
- Shared Memory: $\leq$96 kB

8 TFLOP/s

System Core Numbers

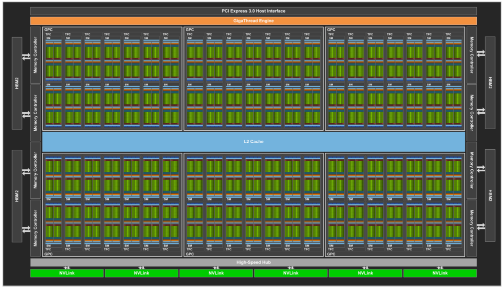
POWER9 CPU



0.5 TFLOP/s

NVLink2 (50 GB/s)

V100 GPU



8 TFLOP/s

Summit

- Latest supercomputer at Oak Ridge National Lab
- 4600 Newell-like nodes
- $>$ 200 PFLOP/s performance
- World’s fastest supercomputer!
- Also: Sierra at Lawrence Livermore National Laboratory, Top500 #2

Nimbix Platform

- Cloud platform of this lab: **Nimbix** *(not NVIDIA's DLI platform!)*
- Specialty: HPC-grade computing equipment of various flavors
- **POWER**, FPGAs, InfiniBand, etc.
- »HPC in the Cloud«, only light virtualization; Est. 2010
- Access through *JARVICE*
- Sponsor this tutorial with free access to their systems! Thank you!

Using Nimbix

A gentle start

TASK 1

Task 1: JARVICE

- Website of Lab: `http://bit.ly/gtc19-openacc`
- Log in to JARVICE at `https://platform.jarvice.com/`
  login provided on slip of paper
- Spin up the lab's Cloud Card named *GTC19 POWER+OpenACC Lab*
- Select POWER9 or POWER8 resources
- Connect to server via Jupyter Notebook
  - ”Click here to connect” → username: `nimbix`; password is shown
  - Launch Notebook of Task
  - Solutions are always given (Notebook and sources)! You decide when to look
  - Edit files with Jupyter's source code editor (just open `.c` file)

? How many cores are on a compute node? How many CUDA cores? See `README.md`

Using Nimbix

A gentle start

TASK 1

Task 1: JARVICE

- Website of Lab: `http://bit.ly/gtc19-openacc`
- Log in to JARVICE at `https://platform.jarvice.com/`
  login provided on slip of paper
- Spin up the lab's Cloud Card named *GTC19 POWER+OpenACC Lab*
- Select POWER9 or POWER8 resources
- Connect to server via Jupyter Notebook
  - ”Click here to connect” → username: `nimbix`; password is shown
  - Launch Notebook of Task
  - Solutions are always given (Notebook and sources)! You decide when to look
  - Edit files with Jupyter's source code editor (just open `.c` file)

? How many cores are on a compute node? How many CUDA cores? See `README.md`

Using Nimbix

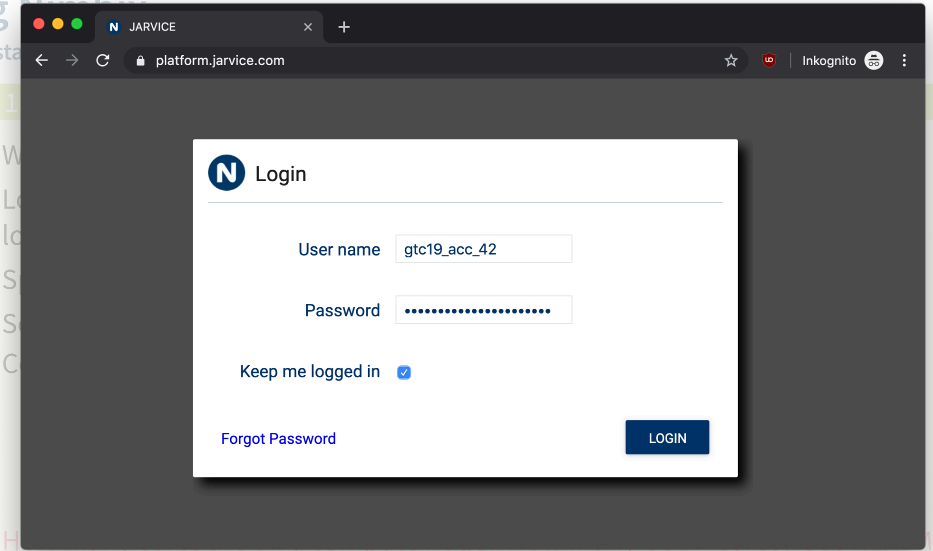
A gentle sta

TASK 1

Task 1

- W
- L lo
- S
- S
- C

? H ME.md



JARVICE

platform.jarvice.com

Inkognito

Login

User name: gtc19_acc_42

Password: ••••••••••••••••••••

Keep me logged in ☑

Forgot Password

LOGIN

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Using Nimbix

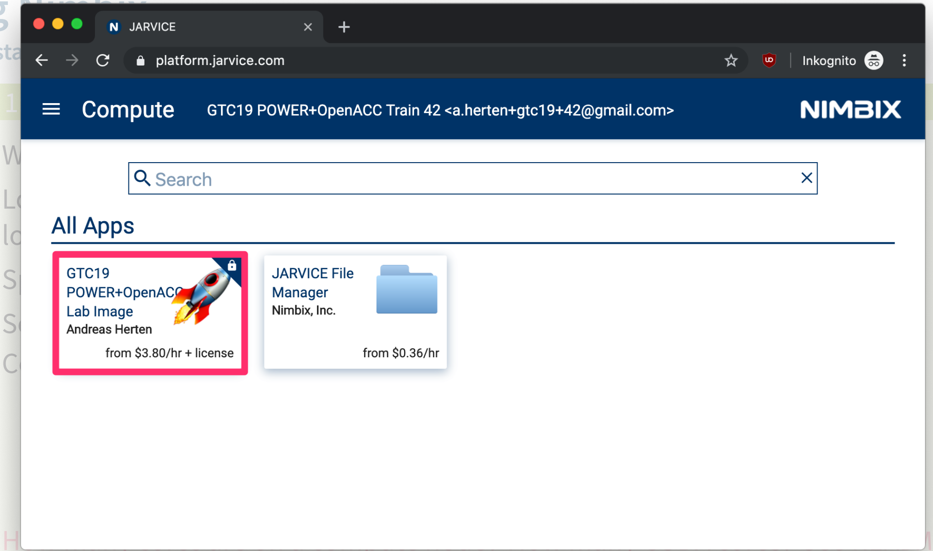
A gentle sta

TASK 1

Task 1

- W
- L
  lo
- S
- S
- C

? H ME.md



GTC19 POWER+OpenACC Train 42 <a.herten+gtc19+42@gmail.com>

All Apps

GTC19 POWER+OpenACC Lab Image
Andreas Herten
from $3.80/hr + license

JARVICE File Manager
Nimbix, Inc.
from $0.36/hr

Using Nimbix

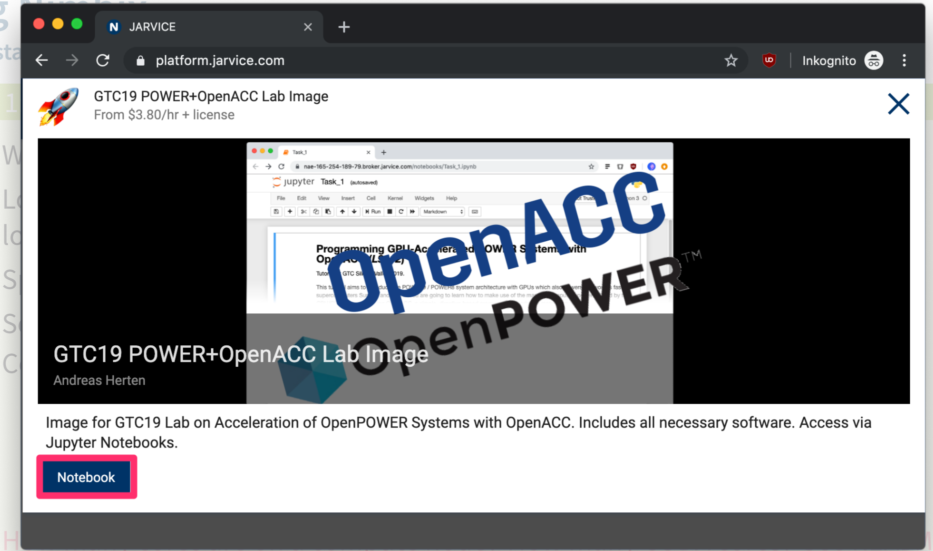
A gentle sta

TASK 1

Task 1

- W
- L
  lo
- S
- S
- C

? H E.md



GTC19 POWER+OpenACC Lab Image

From $3.80/hr + license

Image for GTC19 Lab on Acceleration of OpenPOWER Systems with OpenACC. Includes all necessary software. Access via Jupyter Notebooks.

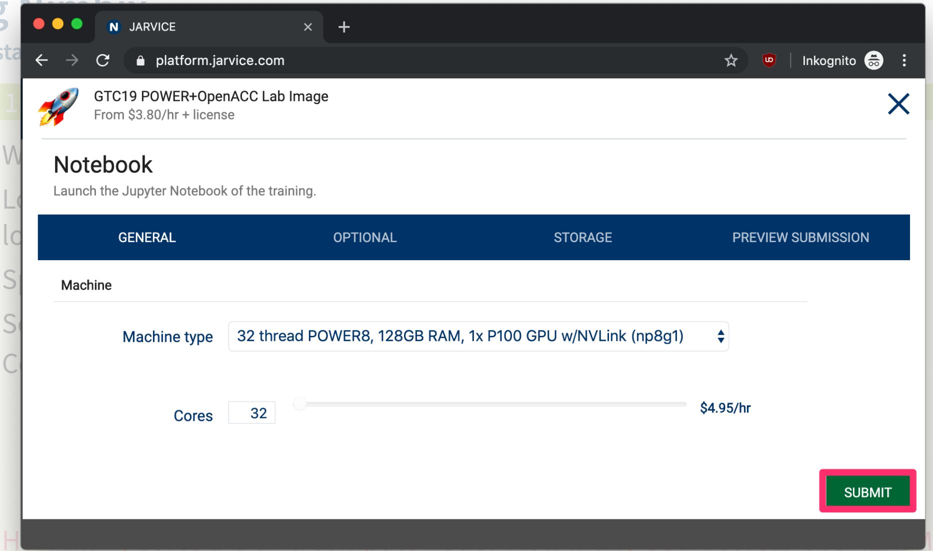
Using Nimbix

A gentle sta

TASK 1

Task 1

- W
- L
  lo
- S
- S
- C



JARVICE

platform.jarvice.com

Inkognito

GTC19 POWER+OpenACC Lab Image

From $3.80/hr + license

Notebook

Launch the Jupyter Notebook of the training.

Machine

Machine type: 32 thread POWER8, 128GB RAM, 1x P100 GPU w/NVLink (np8g1)

Cores: 32 $4.95/hr

? H ME.md

Using Nimbix

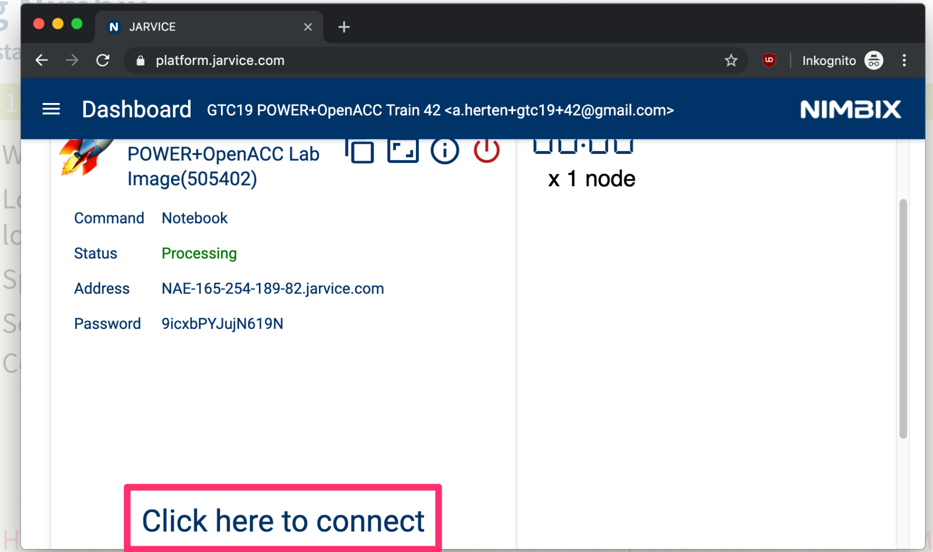
A gentle sta

TASK 1

Task 1

- W
- L
  lo
- S
- S
- C

? H ME.md



GTC19 POWER+OpenACC Train 42 <a.herten+gtc19+42@gmail.com>

POWER+OpenACC Lab Image(505402)

Command	Notebook
Status	Processing
Address	NAE-165-254-189-82.jarvice.com
Password	9icxbPYJujN619N

Click here to connect

Using Nimbix

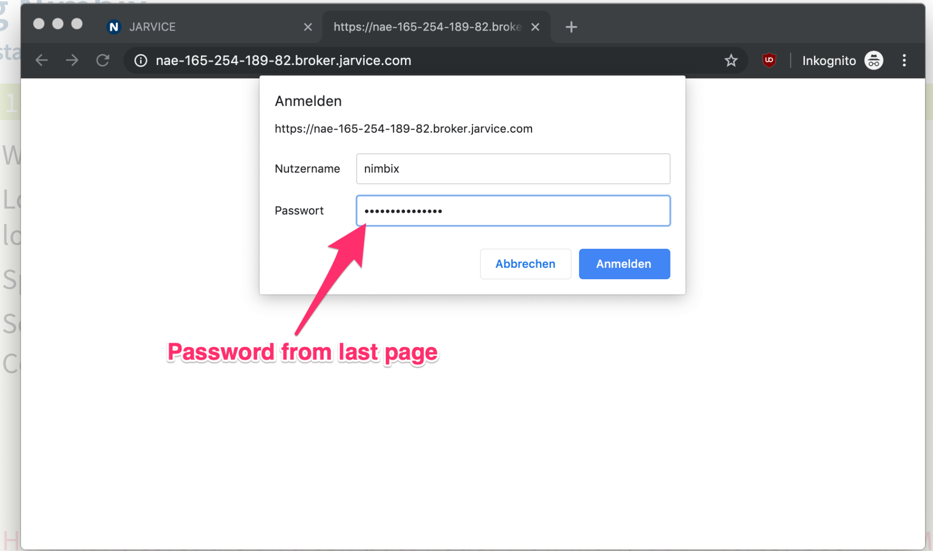
A gentle sta

TASK 1

Task 1

- W
- L
  lo
- S
- S
- C

? H ME.md



JARVICE
https://nae-165-254-189-82.broke
nae-165-254-189-82.broker.jarvice.com
Inkognito

Anmelden

https://nae-165-254-189-82.broker.jarvice.com

Nutzername: nimbix

Passwort: •••••••••••••••

Password from last page

Using N

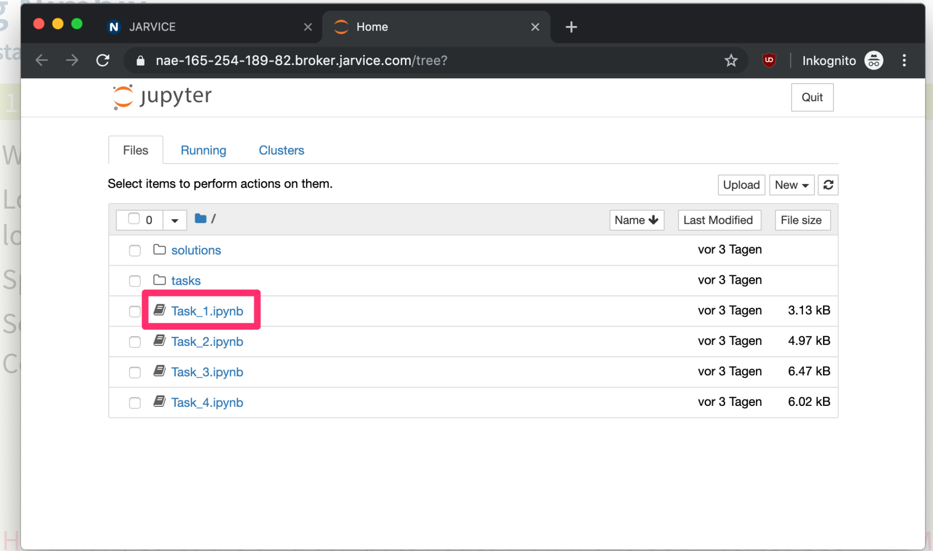
A gentle sta

TASK 1

Task 1

- W
- L
  lo
- S
- S
- C

? H ME.md



nae-165-254-189-82.broker.jarvice.com/tree?

jupyter

Files | Running | Clusters

Select items to perform actions on them.

Name	Last Modified	File size
solutions	vor 3 Tagen	
tasks	vor 3 Tagen	
Task_1.ipynb	vor 3 Tagen	3.13 kB
Task_2.ipynb	vor 3 Tagen	4.97 kB
Task_3.ipynb	vor 3 Tagen	6.47 kB
Task_4.ipynb	vor 3 Tagen	6.02 kB

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

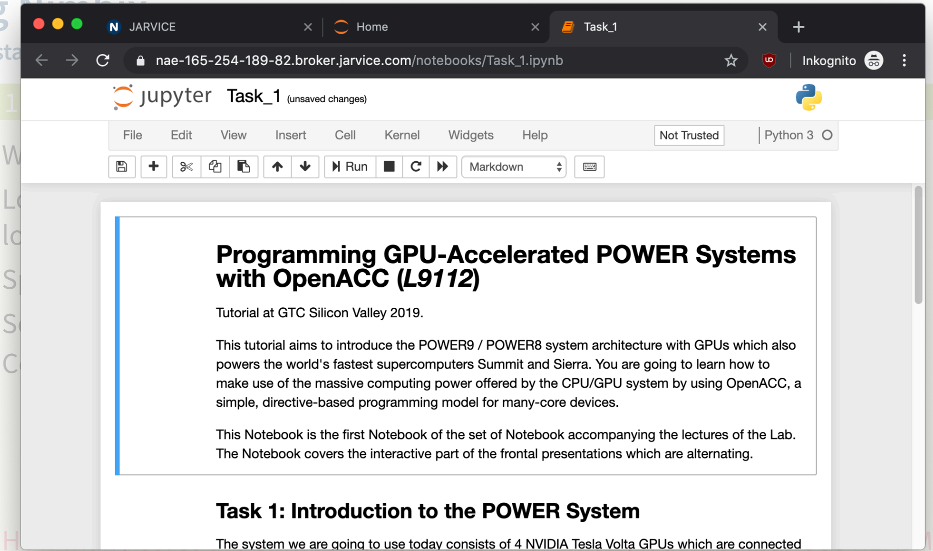
Using Nimbix

A gentle sta

TASK 1

Task 1

- W
- L
  lo
- S
- S
- C



Programming GPU-Accelerated POWER Systems with OpenACC (*L9112*)

Tutorial at GTC Silicon Valley 2019.

This tutorial aims to introduce the POWER9 / POWER8 system architecture with GPUs which also powers the world's fastest supercomputers Summit and Sierra. You are going to learn how to make use of the massive computing power offered by the CPU/GPU system by using OpenACC, a simple, directive-based programming model for many-core devices.

This Notebook is the first Notebook of the set of Notebook accompanying the lectures of the Lab. The Notebook covers the interactive part of the frontal presentations which are alternating.

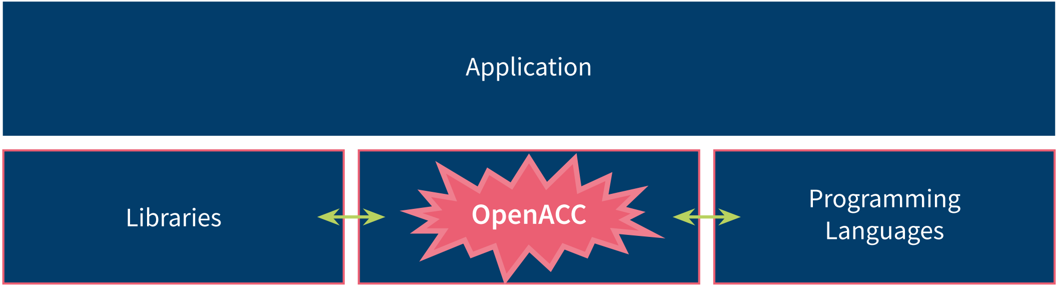
Task 1: Introduction to the POWER System

The system we are going to use today consists of 4 NVIDIA Tesla Volta GPUs which are connected

? H ME.md

OpenACC Introduction

Primer on GPU Computing



Application

Libraries

OpenACC

Programming Languages

Drop-in Acceleration

Easy Acceleration

Flexible Acceleration

About OpenACC

History

2011 OpenACC 1.0 specification is released
NVIDIA, Cray, PGI, CAPS

2013 OpenACC 2.0: More functionality, portability

2015 OpenACC 2.5: Enhancements, clarifications

2017 OpenACC 2.6: Deep copy, …

2018 OpenACC 2.7: Clarifications, more host, …

→ `https://www.openacc.org/` (*see also: Best practice guide*)

Support

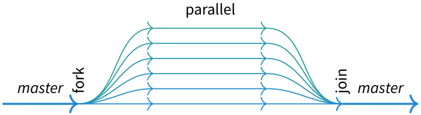
- Compiler: PGI, GCC, Clang, *Sunway*
- Languages: C/C++, Fortran

Open{MP↔ACC}

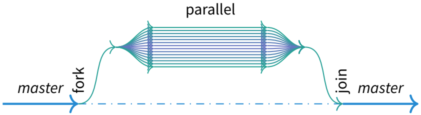
Everything's connected

- OpenACC modeled after OpenMP …
- … but specific for accelerators
- OpenACC more descriptive, OpenMP more prescriptive
- Basic principle same: Fork/join model

  Master thread launches parallel child threads; merge after execution



OpenMP



OpenACC

Modus Operandi

Three-step program

1. Annotate code with directives, indicating parallelism
2. OpenACC-capable compiler generates accelerator-specific code
3. $uccess

1 Directives

pragmatic

- Compiler directives state intend to compiler

C/C++

```
#pragma acc kernels
for (int i = 0; i < 23; i++)
// ...
```

Fortran

```
!$acc kernels
do i = 1, 24
! ...
!$acc end kernels
```

- Ignored by compiler which does not understand OpenACC
- High level programming model for many-core machines, especially accelerators
- OpenACC: Compiler directives, library routines, environment variables
- Portable across host systems and accelerator architectures

2 Compiler

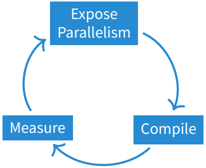
Simple and abstracted

- Compiler support
  - PGI *Best performance, great support, free*
  - GCC *Actively performance-improved, OSS*
  - Clang *First alpha version*
- Trust compiler to generate intended parallelism; always check status output!
- No need to know ins'n'outs of accelerator; leave it to expert compiler engineers*
- One code can target different accelerators: GPUs, or even multi-core CPUs → **Portability**

*: *Eventually you want to tune for device; but that's possible*

3 $uccess

Iteration is key



- Serial to parallel: fast
- Serial to fast parallel: more time needed
- Start simple → refine

⇒ **Productivity**

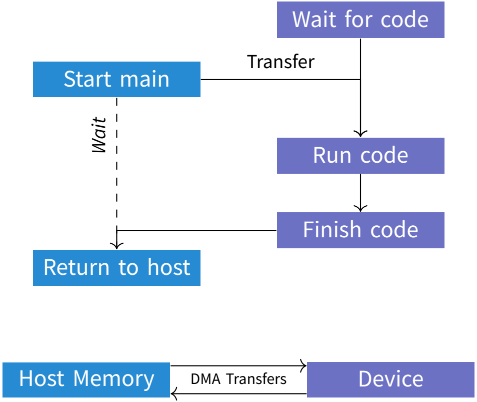
- Because of *generalness*: Sometimes not last bit of hardware performance accessible
- But: Use OpenACC together with other accelerator-targeting techniques (CUDA, libraries, …)

OpenACC Accelerator Model

For computation and memory spaces

- Main program executes on host
- Device code is transferred to accelerator
- Execution on accelerator is started
- Host waits until return (except: async)

- Two separate memory spaces; data transfers back and forth
  - Transfers hidden from programmer
  - Memories not coherent!
  - Compiler helps; GPU runtime helps



OpenACC Programming Model

A binary perspective

- OpenACC interpretation needs to be activated as compile flag

PGI	`pgcc -acc [-ta=tesla|-ta=multicore]`
GCC	`gcc -fopenacc`
→	Ignored by incapable compiler!

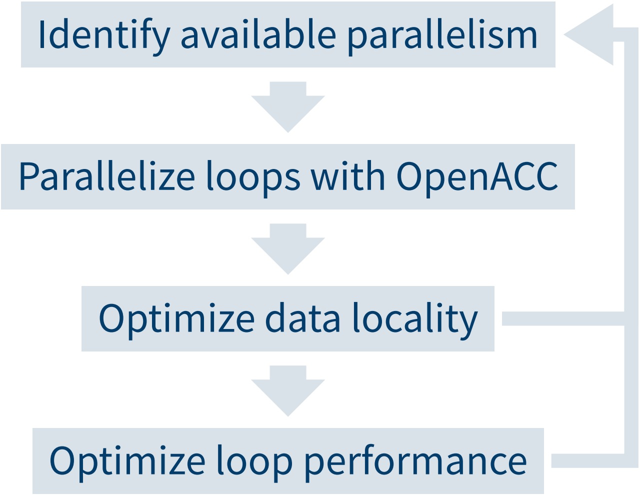
- Additional flags possible to improve/modify compilation

`-ta=tesla:cc70`	Use compute capability 7.0
`-ta=tesla:lineinfo`	Add source code correlation into binary
`-ta=tesla:managed`	Use unified memory
`-fopenacc-dim=geom`	Use *geom* configuration for threads

A Glimpse of OpenACC

```
#pragma acc data copy(x[0:N],y[0:N])
#pragma acc parallel loop
{
    for (int i=0; i<N; i++) {
        x[i] = 1.0;
        y[i] = 2.0;
    }
    for (int i=0; i<N; i++) {
        y[i] = i*x[i]+y[i];
    }
}
```

Parallelization Workflow



Identify available parallelism

Parallelize loops with OpenACC

Optimize data locality

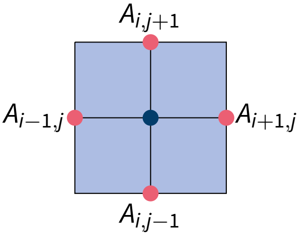
Optimize loop performance

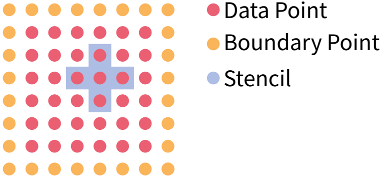
First Steps in OpenACC

Jacobi Solver

Algorithmic description

- Example for acceleration: **Jacobi solver**
- Iterative solver, converges to correct value
- Each iteration step: compute average of neighboring points
- Example: 2D Poisson equation: $\nabla^2 A(x,y) = B(x,y)$





$$A_{k+1}(i,j) = -\frac{1}{4}\left(B(i,j) - (A_k(i-1,j) + A_k(i,j+1), +A_k(i+1,j) + A_k(i,j-1))\right)$$

Jacobi Solver

Source code

```
while ( error > tol && iter < iter_max ) {
    error = 0.0;
    for (int ix = ix_start; ix < ix_end; ix++) {
        for (int iy = iy_start; iy < iy_end; iy++) {
            Anew[iy*nx+ix] = -0.25 * (rhs[iy*nx+ix] -
                (  A[iy*nx+ix+1] + A[iy*nx+ix-1]
               + A[(iy-1)*nx+ix] + A[(iy+1)*nx+ix]));
            error = fmaxr(error, fabsr(Anew[iy*nx+ix]-A[iy*nx+ix]));
    }}
    for (int iy = iy_start; iy < iy_end; iy++) {
        for( int ix = ix_start; ix < ix_end; ix++ ) {
            A[iy*nx+ix] = Anew[iy*nx+ix];
    }}
    for (int ix = ix_start; ix < ix_end; ix++) {
            A[0*nx+ix]      = A[(ny-2)*nx+ix];
            A[(ny-1)*nx+ix] = A[1*nx+ix];
    }
    // same for iy
    iter++;
}
```

Iterate until converged

Iterate across matrix elements

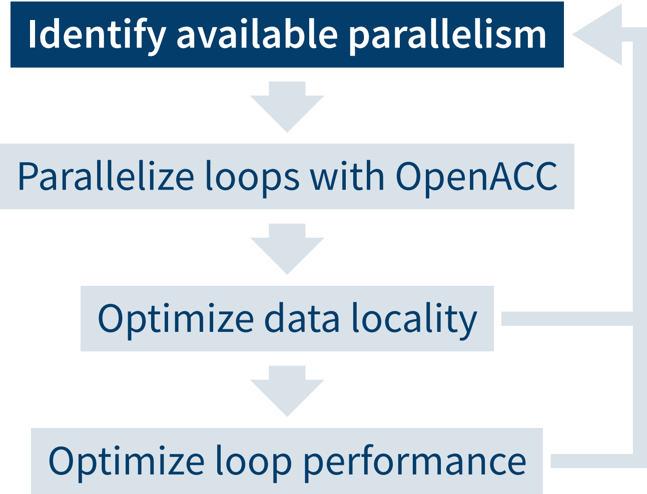
Calculate new value from neighbors

Accumulate error

Swap input/output

Set boundary conditions

Parallelization Workflow



Identify available parallelism

Parallelize loops with OpenACC

Optimize data locality

Optimize loop performance

Profiling

Profile

> *[…] premature optimization is the root of all evil.*
> ***Yet we should not pass up our [optimization] opportunities […]***
>
> *– Donald Knuth [10]*

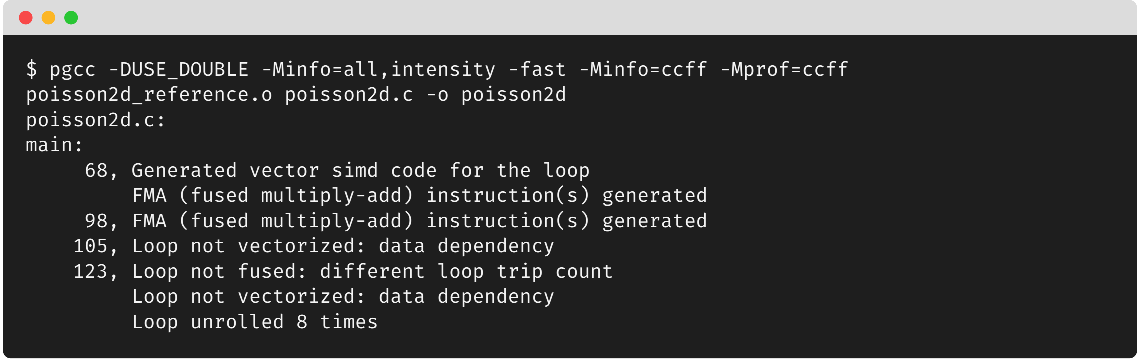
- Investigate hot spots of your program!

→ Profile!

- Many tools, many levels: perf, PAPI, Score-P, Intel Advisor, NVIDIA Visual Profiler, …
- Here: Examples from PGI

Profile of Application

Info during compilation

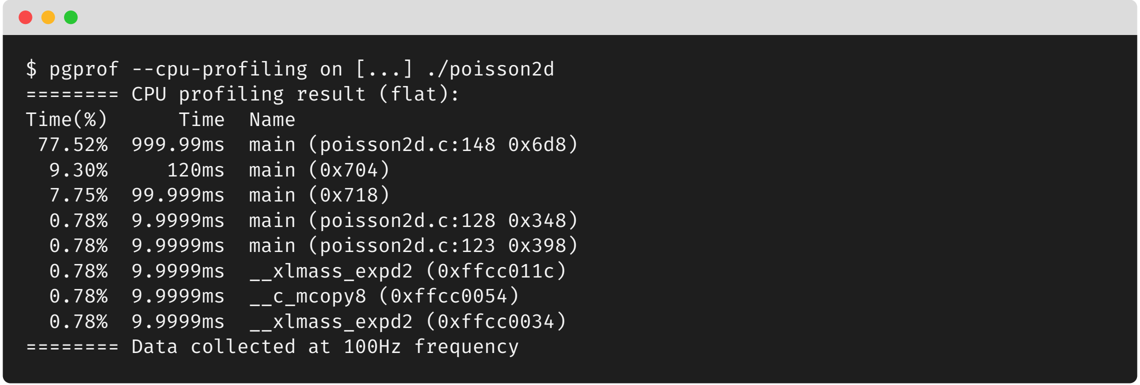


```
$ pgcc -DUSE_DOUBLE -Minfo=all,intensity -fast -Minfo=ccff -Mprof=ccff
poisson2d_reference.o poisson2d.c -o poisson2d
poisson2d.c:
main:
     68, Generated vector simd code for the loop
         FMA (fused multiply-add) instruction(s) generated
     98, FMA (fused multiply-add) instruction(s) generated
    105, Loop not vectorized: data dependency
    123, Loop not fused: different loop trip count
         Loop not vectorized: data dependency
         Loop unrolled 8 times
```

- Automated optimization of compiler, due to `-fast`
- Vectorization, FMA, unrolling

Profile of Application

Info during run



```
$ pgprof --cpu-profiling on [...] ./poisson2d
======== CPU profiling result (flat):
Time(%)      Time  Name
 77.52%  999.99ms  main (poisson2d.c:148 0x6d8)
  9.30%     120ms  main (0x704)
  7.75%  99.999ms  main (0x718)
  0.78%  9.9999ms  main (poisson2d.c:128 0x348)
  0.78%  9.9999ms  main (poisson2d.c:123 0x398)
  0.78%  9.9999ms  __xlmass_expd2 (0xffcc011c)
  0.78%  9.9999ms  __c_mcopy8 (0xffcc0054)
  0.78%  9.9999ms  __xlmass_expd2 (0xffcc0034)
======== Data collected at 100Hz frequency
```

- 78 % in `main()`
- Since everything is in `main` – limited helpfulness
- Let’s look into `main`!

Code Independency Analysis

Independence is key

```
while ( error > tol && iter < iter_max ) {
    error = 0.0;
    for (int ix = ix_start; ix < ix_end; ix++) {
        for (int iy = iy_start; iy < iy_end; iy++) {
            Anew[iy*nx+ix] = -0.25 * (rhs[iy*nx+ix] -
                (  A[iy*nx+ix+1] + A[iy*nx+ix-1]
               + A[(iy-1)*nx+ix] + A[(iy+1)*nx+ix]));
            error = fmaxr(error, fabsr(Anew[iy*nx+ix]-A[iy*nx+ix]));
    }}
    for (int iy = iy_start; iy < iy_end; iy++) {
        for( int ix = ix_start; ix < ix_end; ix++ ) {
            A[iy*nx+ix] = Anew[iy*nx+ix];
    }}
    for (int ix = ix_start; ix < ix_end; ix++) {
            A[0*nx+ix]      = A[(ny-2)*nx+ix];
            A[(ny-1)*nx+ix] = A[1*nx+ix];
    }
    // same for iy
    iter++;
}
```

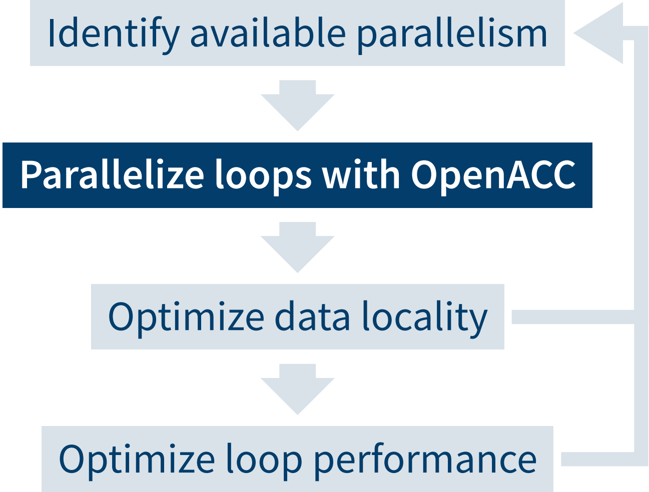
Data dependency between iterations

Independent loop iterations

Independent loop iterations

Independent loop iterations

Parallelization Workflow



Identify available parallelism

Parallelize loops with OpenACC

Optimize data locality

Optimize loop performance

Parallel Loops: Parallel

Maybe the second most important directive

- Programmer identifies block containing parallelism
  → compiler generates parallel code (*kernel*)
- Program launch creates *gangs* of parallel threads on parallel device
- Implicit barrier at end of parallel region
- Each gang executes same code sequentially

OpenACC: `parallel`

```
#pragma acc parallel [clause, [, clause] ...] newline
{structured block}
```

Parallel Loops: Parallel

Clauses

Diverse clauses to augment the parallel region

`private(var)` A copy of variables `var` is made for each gang

`firstprivate(var)` Same as `private`, except `var` will initialized with value from host

`if(cond)` Parallel region will execute on accelerator only if `cond` is true

`reduction(op:var)` Reduction is performed on variable `var` with operation op; supported: `+ * max min …`

`async[(int)]` No implicit barrier at end of parallel region

Parallel Loops: Loops

Maybe the third most important directive

- Programmer identifies loop eligible for parallelization
- Directive must be directly before loop
- Optional: Describe type of parallelism

OpenACC: `loop`

```
#pragma acc loop [clause, [, clause] ...] newline
{structured block}
```

Parallel Loops: Loops

Clauses

`independent` Iterations of loop are data-independent (implied if in `parallel` region (and no `seq` or `auto`))

`collapse(int)` Collapse `int` tightly-nested loops

`seq` This loop is to be executed sequentially (not parallel)

`tile(int[,int])` Split loops into loops over tiles of the full size

`auto` Compiler decides what to do

Parallel Loops: Parallel Loops

Maybe the most important directive

- Combined directive: shortcut

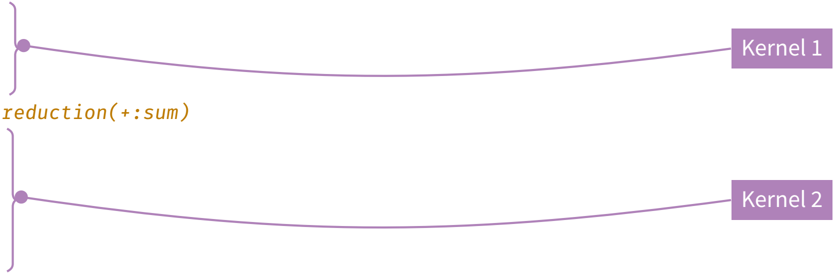
  Because its used so often
- Any clause that is allowed on `parallel` or `loop` allowed
- Restriction: May not appear in body of another parallel region

OpenACC: `parallel loop`

```
#pragma acc parallel loop [clause, [, clause] ...]
```

Parallel Loops Example

```
double sum = 0.0;
#pragma acc parallel loop
for (int i=0; i<N; i++) {
    x[i] = 1.0;
    y[i] = 2.0;
}
#pragma acc parallel loop reduction(+:sum)
{
for (int i=0; i<N; i++) {
    y[i] = i*x[i]+y[i];
    sum+=y[i];
}
}
```



Kernel 1

Kernel 2

Parallel Jacobi

TASK 2

Add parallelism

- Add OpenACC parallel region to main loop in Jacobi solver source code (CPU parallelism)

→ Congratulations, you are a parallel developer!

Task 2: A First Parallel Loop

- Open `Task_2.ipynb` Notebook and follow instructions
- Change number of CPU threads via `$ACC_NUM_CORES` or `$OMP_NUM_THREADS`

? What's your speed-up? What's the best configuration for cores?

E Compare it to OpenMP

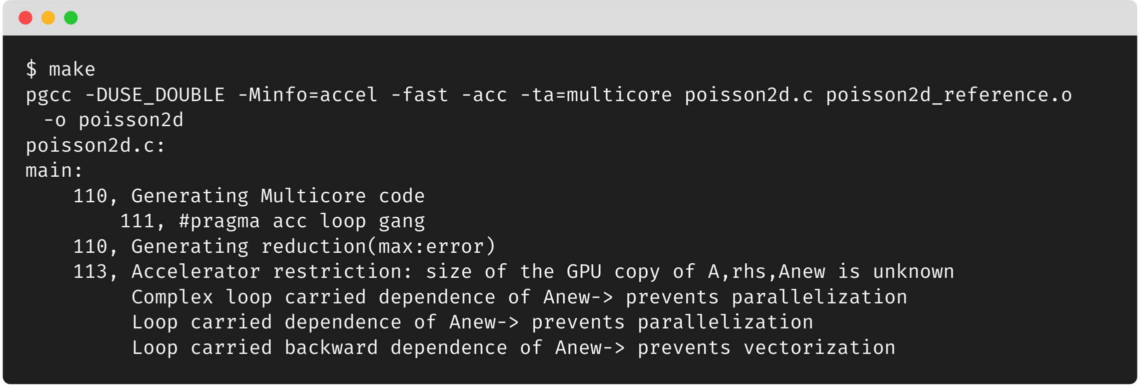
Parallel Jacobi

Source Code

```
#pragma acc parallel loop reduction(max:error)
for (int ix = ix_start; ix < ix_end; ix++)
{
    for (int iy = iy_start; iy < iy_end; iy++)
    {
        Anew[iy*nx+ix] = -0.25 * (rhs[iy*nx+ix] - ( A[iy*nx+ix+1] +
        ↪   A[iy*nx+ix-1]
                                               + A[(iy-1)*nx+ix] +
                                               ↪   A[(iy+1)*nx+ix] ));
        error = fmaxr( error, fabsr(Anew[iy*nx+ix]-A[iy*nx+ix]));
    }
}
```

Parallel Jacobi

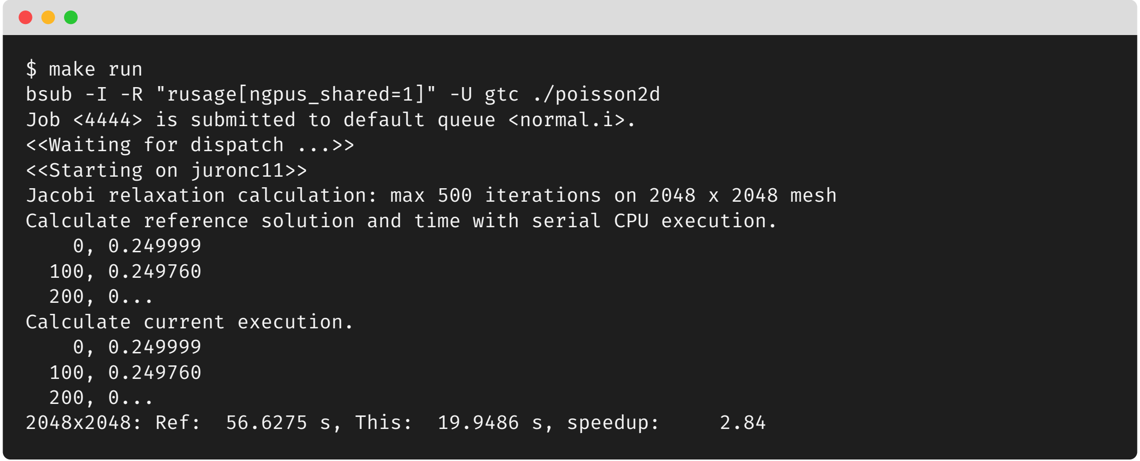
Compilation result



```
$ make
pgcc -DUSE_DOUBLE -Minfo=accel -fast -acc -ta=multicore poisson2d.c poisson2d_reference.o
  -o poisson2d
poisson2d.c:
main:
    110, Generating Multicore code
        111, #pragma acc loop gang
    110, Generating reduction(max:error)
    113, Accelerator restriction: size of the GPU copy of A,rhs,Anew is unknown
         Complex loop carried dependence of Anew-> prevents parallelization
         Loop carried dependence of Anew-> prevents parallelization
         Loop carried backward dependence of Anew-> prevents vectorization
```

Parallel Jacobi

Run result



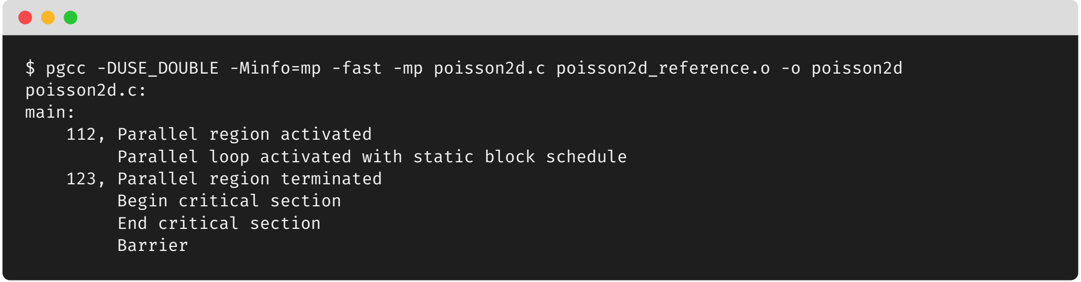
```
$ make run
bsub -I -R "rusage[ngpus_shared=1]" -U gtc ./poisson2d
Job <4444> is submitted to default queue <normal.i>.
<<Waiting for dispatch ...>>
<<Starting on juronc11>>
Jacobi relaxation calculation: max 500 iterations on 2048 x 2048 mesh
Calculate reference solution and time with serial CPU execution.
    0, 0.249999
  100, 0.249760
  200, 0...
Calculate current execution.
    0, 0.249999
  100, 0.249760
  200, 0...
2048x2048: Ref:  56.6275 s, This:  19.9486 s, speedup:     2.84
```

Parallel Jacobi: OpenMP

- OpenMP pragma is quite similar

```
#pragma acc parallel loop reduction(max:error)
#pragma omp parallel for  reduction(max:error)
for (int ix = ix_start; ix < ix_end; ix++) { ... }
```

- PGI’s compiler output is a bit different (but states the same)



```
$ pgcc -DUSE_DOUBLE -Minfo=mp -fast -mp poisson2d.c poisson2d_reference.o -o poisson2d
poisson2d.c:
main:
    112, Parallel region activated
         Parallel loop activated with static block schedule
    123, Parallel region terminated
         Begin critical section
         End critical section
         Barrier
```

- Run time should be very similar!

More Parallelism: Kernels

More freedom for compiler

- Kernels directive: second way to expose parallelism
- Region may contain parallelism
- Compiler determines parallelization opportunities

→ More freedom for compiler

- Rest: Same as for `parallel`

OpenACC: `kernels`

```
#pragma acc kernels [clause, [, clause] ...]
```

Kernels Example

```
double sum = 0.0;
#pragma acc kernels
{
for (int i=0; i<N; i++) {
    x[i] = 1.0;
    y[i] = 2.0;
}
for (int i=0; i<N; i++) {
    y[i] = i*x[i]+y[i];
    sum+=y[i];
}
}
```

Kernels created here

`kernels` vs. `parallel`

- Both approaches equally valid; can perform equally well
- `kernels`
  - Compiler performs parallel analysis
  - Can cover large area of code with single directive
  - Gives compiler additional leeway
- `parallel`
  - Requires parallel analysis by programmer
  - Will also parallelize what compiler may miss
  - More explicit
  - Similar to OpenMP
- Both regions may not contain other `kernels`/`parallel` regions
- No braunching into or out
- Program must not depend on order of evaluation of clauses
- At most: One `if` clause

OpenACC on the GPU

Changes for GPU-OpenACC

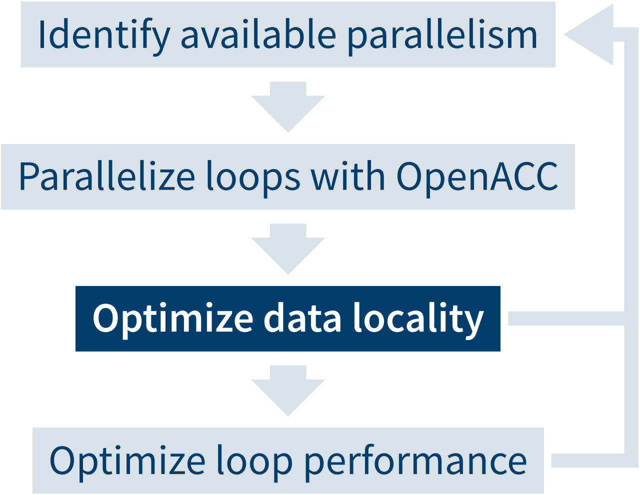
Immensely complicated changes

- Necessary for previous code to run on GPU: `-ta=tesla` instead of `-ta=multicore`

⇒ **That’s it!**

- But we can optimize!

Parallelization Workflow



Identify available parallelism

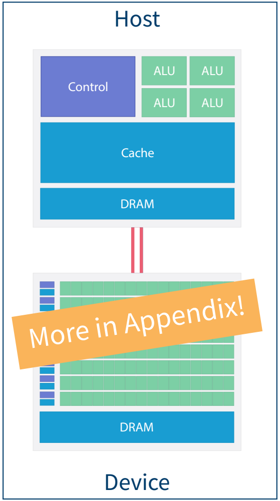
Parallelize loops with OpenACC

Optimize data locality

Optimize loop performance

Automatic Data Transfers

- Up to now: We did not care about **data transfers**
- Compiler and runtime care
- CPU data can be copied automatically to GPU via Managed Memory
- Magic keyword: `-ta=tesla:managed`
- Be more explicit for full portability and full performance



Copy Clause

- Explicitly inform OpenACC compiler about data intentions
- Use data which is already on GPU; only copy parts of it; …

OpenACC: `copy`

```
#pragma acc parallel copy(A[start:end])
```

Also: `copyin(B[s:e])` `copyout(C[s:e])` `present(D[s:e])` `create(E[s:e])`

Data Regions

To manually specify data locations: `data` construct

- Defines region of code in which data remains on device
- Data is shared among all kernels in region
- Explicit data transfers

OpenACC: `data`

```
#pragma acc data [clause, [, clause] ...]
```

Data Regions

Clauses

Clauses to augment the data regions

`copy(var)` Allocates memory of `var` on GPU, copies data to GPU at beginning of region, copies data to host at end of region
Specifies size of `var`: `var[`*`lowerBound:size`*`]`

`copyin(var)` Allocates memory of `var` on GPU, copies data to GPU at beginning of region

`copyout(var)` Allocates memory of `var` on GPU, copies data to host at end of region

`create(var)` Allocates memory of `var` on GPU

`present(var)` Data of `var` is not copies automatically to GPU but considered present

Data Region Example

```
#pragma acc data copyout(y[0:N]) create(x[0:N])
{
double sum = 0.0;
#pragma acc parallel loop
for (int i=0; i<N; i++) {
    x[i] = 1.0;
    y[i] = 2.0;
}
#pragma acc parallel loop
for (int i=0; i<N; i++) {
    y[i] = i*x[i]+y[i];
}
}
```

Data Regions II

Looser regions: `enter data` directive

- Define data regions, but not for structured block
- Closest to `cudaMemcpy()`
- Still, explicit data transfers

OpenACC: `enter data`

```
#pragma acc enter data [clause, [, clause] ...]
#pragma acc exit data [clause, [, clause] ...]
```

Parallel Jacobi II

TASK 3

More parallelism, Data locality

- Add OpenACC parallelism to other loops of `while` (L:123 – L:141)
  Use either `kernels` or `parallel`
- Add data regions such that all data resides on device during iterations

Task 3: More Parallel Loops

- Open `Task_3.ipynb` Notebook
- Follow instructions of sub-tasks!

? What's your speed-up?

E Change order of `for` loop!

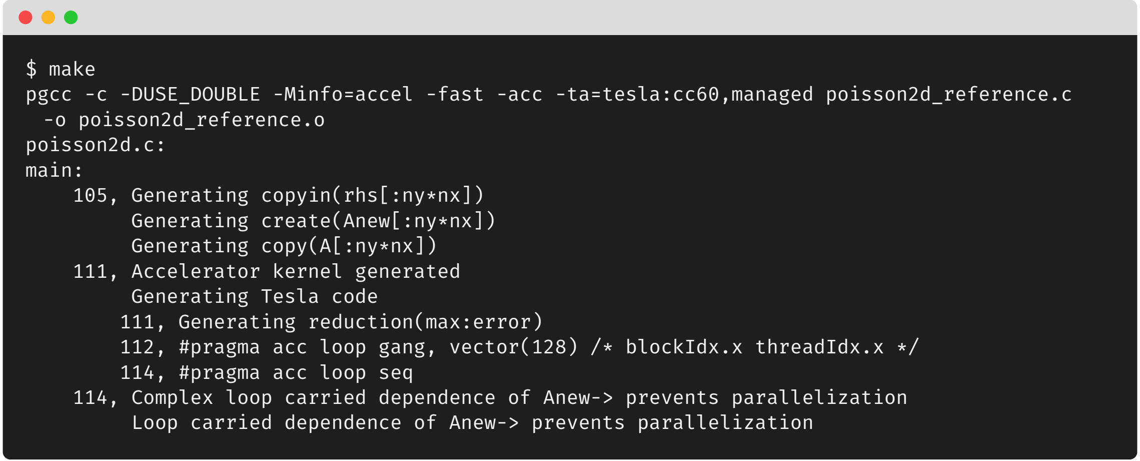
Parallel Jacobi II

Source Code

```
#pragma acc data copy(A[0:nx*ny]) copyin(rhs[0:nx*ny]) create(Anew[0:nx*ny])
while ( error > tol && iter < iter_max )
{
    error = 0.0;

    // Jacobi kernel
    #pragma acc parallel loop reduction(max:error)
    for (int ix = ix_start; ix < ix_end; ix++)
    {
        for (int iy = iy_start; iy < iy_end; iy++)
        {
            Anew[iy*nx+ix] = -0.25 * (rhs[iy*nx+ix] - ( A[iy*nx+ix+1] + A[iy*nx+ix-1]
                                                   + A[(iy-1)*nx+ix] + A[(iy+1)*nx+ix] ));
            error = fmaxr( error, fabsr(Anew[iy*nx+ix]-A[iy*nx+ix]));
        }
    }

    // A <-> Anew
    #pragma acc parallel loop
    for (int iy = iy_start; iy < iy_end; iy++)
    // …
}
```

Parallel Jacobi II

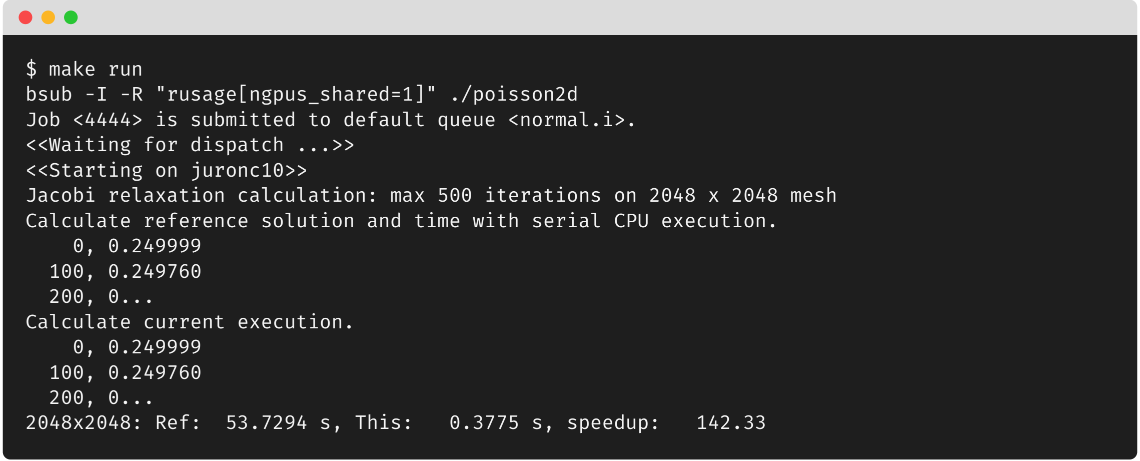
Compilation result



```
$ make
pgcc -c -DUSE_DOUBLE -Minfo=accel -fast -acc -ta=tesla:cc60,managed poisson2d_reference.c
  -o poisson2d_reference.o
poisson2d.c:
main:
    105, Generating copyin(rhs[:ny*nx])
         Generating create(Anew[:ny*nx])
         Generating copy(A[:ny*nx])
    111, Accelerator kernel generated
         Generating Tesla code
        111, Generating reduction(max:error)
        112, #pragma acc loop gang, vector(128) /* blockIdx.x threadIdx.x */
        114, #pragma acc loop seq
    114, Complex loop carried dependence of Anew-> prevents parallelization
         Loop carried dependence of Anew-> prevents parallelization
```

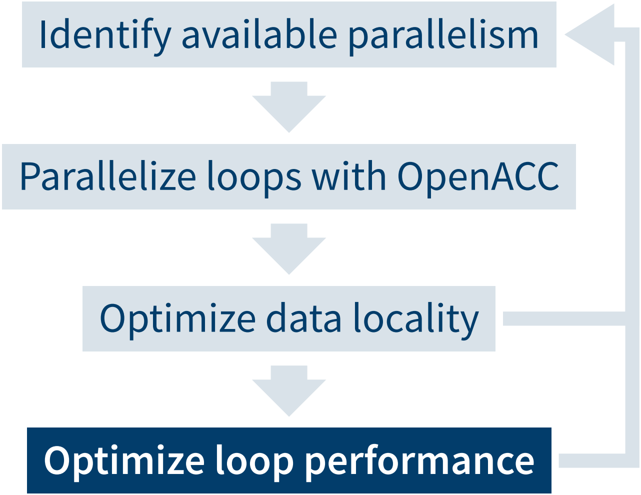
Parallel Jacobi II

Run result



```
$ make run
bsub -I -R "rusage[ngpus_shared=1]" ./poisson2d
Job <4444> is submitted to default queue <normal.i>.
<<Waiting for dispatch ...>>
<<Starting on juronc10>>
Jacobi relaxation calculation: max 500 iterations on 2048 x 2048 mesh
Calculate reference solution and time with serial CPU execution.
    0, 0.249999
  100, 0.249760
  200, 0...
Calculate current execution.
    0, 0.249999
  100, 0.249760
  200, 0...
2048x2048: Ref:  53.7294 s, This:   0.3775 s, speedup:   142.33
```

Parallelization Workflow



Identify available parallelism

Parallelize loops with OpenACC

Optimize data locality

Optimize loop performance

Parallel Jacobi II+

Expert Task

E

Improve memory access pattern: Loop order in main loop

```
#pragma acc parallel loop reduction(max:error)
for (int ix  = ix_start; ix < ix_end; ix++) {
    #pragma acc loop vector
    for (int iy  = iy_start; iy < iy
        Anew[ iy*nx  +
        ↪  (rhs[iy*nx+
           (  A[iy*nx+
          + A[(iy-1)*n
        //...
```

ix Outer r dex; accesses

memory locations

ex; accesses offset

tions

nge order to optimize pattern ✓

More on OpenACC thread configuration in Appendix!

Parallel Jacobi II+

E

Expert Task

Improve memory access pattern: Loop order in main loop

```
#pragma acc parallel loop reduction(max:error)
for (int iy = iy_start; iy < iy_end; iy++) {
    #pragma acc loop vector
    for (int ix = ix_start; ix < i
        Anew[ iy*nx +
        ↪  (rhs[iy*nx+
            (   A[iy*nx+
          + A[(iy-1)*n
        //...
```

ix Outer dex; accesses

memory locations

ex; accesses offset

tions

nge order to optimize pattern ✓

More on OpenACC thread configuration in Appendix!

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

Parallel Jacobi II+

E

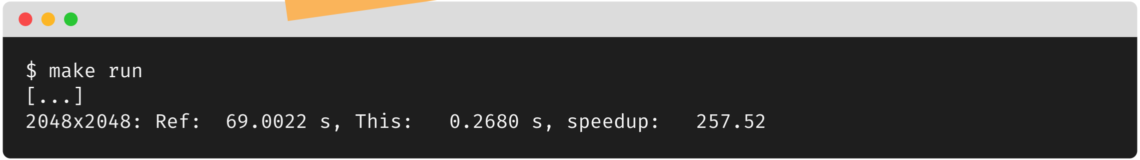
Expert Task

Improve memory access pattern: Loop order in main loop

```
#pragma acc parallel loop reduction(max:error)
for (int iy = iy_start; iy < iy_end; iy++) {
    #pragma acc loop vector
    for (int ix = ix_start; ix < i
        Anew[ iy*nx +
        ↪   (rhs[iy*nx+
            (  A[iy*nx+
          + A[(iy-1)*n
        //...
```

- ix Outer r dex; accesses memory locations
- ex; accesses offset tions
- ange order to optimize pattern ✓

More on OpenACC thread configuration in Appendix!



```
$ make run
[...]
2048x2048: Ref:   69.0022 s, This:   0.2680 s, speedup:   257.52
```

Parallel Jacobi II+

E

Expert Task

Improve memory access pattern: Loop order in main loop

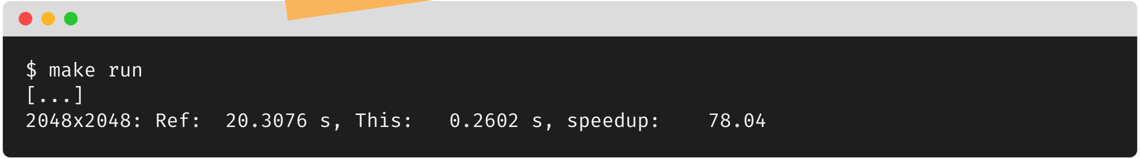
```
#pragma acc parallel loop reduction(max:error)
for (int iy = iy_start; iy < iy_end; iy++) {
    #pragma acc loop vector
    for (int ix = ix_start; ix < i
        Anew[ iy*nx + 
        ↪  (rhs[iy*nx+
            (  A[iy*nx+
          + A[(iy-1)*n
        //...
```

ix Outer … dex; accesses … memory locations

… ex; accesses offset … tions

… nge order to optimize pattern ✓

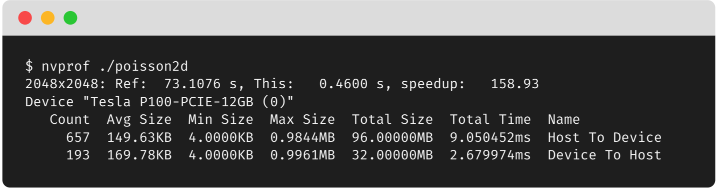
More on OpenACC thread configuration in Appendix!



```
$ make run
[...]
2048x2048: Ref:  20.3076 s, This:   0.2602 s, speedup:    78.04
```

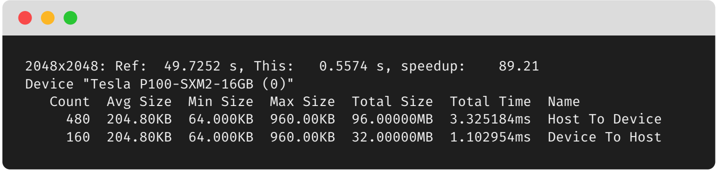
Aside: Data Transfer with NVLink

- One feature of POWER not showcased in tutorial: NVLink between CPU and GPU
- Task 3 with PCI-E:

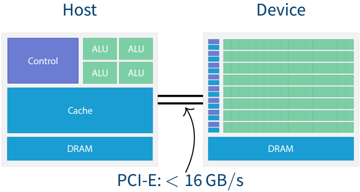


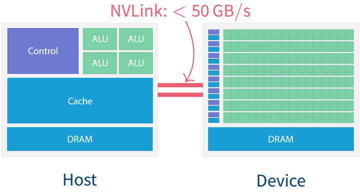
```
$ nvprof ./poisson2d
2048x2048: Ref:  73.1076 s, This:   0.4600 s, speedup:   158.93
Device "Tesla P100-PCIE-12GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     657  149.63KB  4.0000KB  0.9844MB  96.00000MB  9.050452ms  Host To Device
     193  169.78KB  4.0000KB  0.9961MB  32.00000MB  2.679974ms  Device To Host
```

- Task 3 with NVLink:

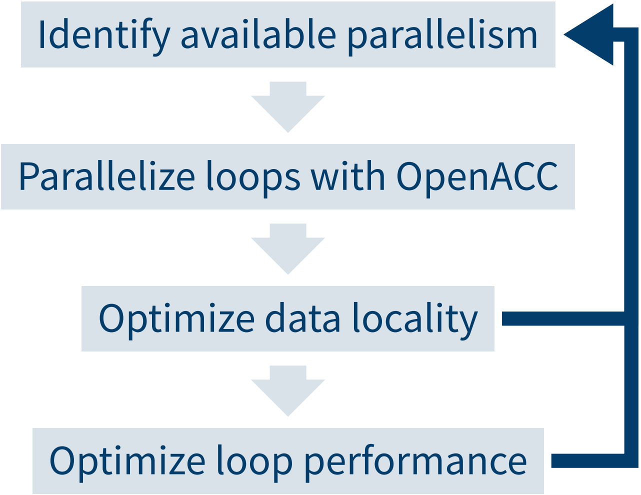


```
2048x2048: Ref:  49.7252 s, This:   0.5574 s, speedup:    89.21
Device "Tesla P100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     480  204.80KB  64.000KB  960.00KB  96.00000MB  3.325184ms  Host To Device
     160  204.80KB  64.000KB  960.00KB  32.00000MB  1.102954ms  Device To Host
```





Parallelization Workflow



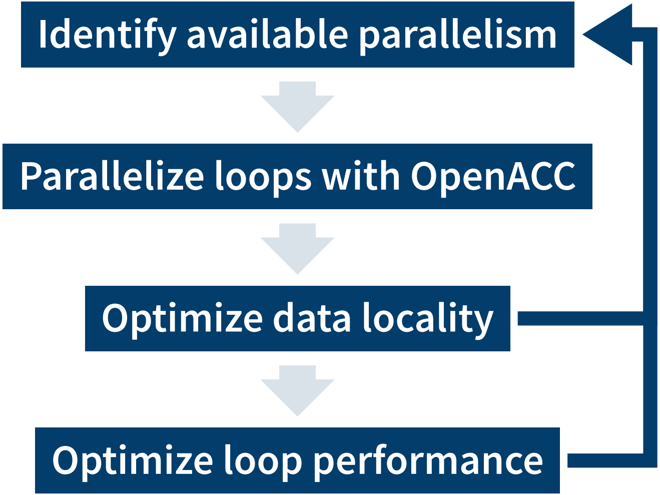
Identify available parallelism

Parallelize loops with OpenACC

Optimize data locality

Optimize loop performance

Parallelization Workflow



Identify available parallelism

Parallelize loops with OpenACC

Optimize data locality

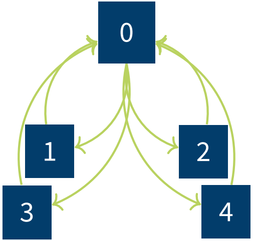
Optimize loop performance

OpenACC on Multiple GPUs

Message Passing Interface Introduction

- **MPI**: Message Passing Interface
- Standardized API to communicate data across processes and nodes; compilers
- Various implementations: OpenMPI, MPICH, MVAPICH, Vendor-specific versions
- Standard in parallel and distributed High Performance Computing
- Unrelated to OpenACC, but works well together!

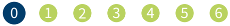
→ www.open-mpi.org/doc/



MPI API Examples

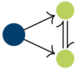
- Configuration calls

`MPI_Comm_size()`	Get number of total processes
`MPI_Comm_rank()`	Get current process number



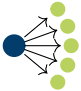
- Point-to-point routines

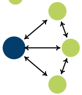
`MPI_Send()`	Send data to other process
`MPI_Recv()`	Receive data from other process
`MPI_Sendrecv()`	Do both in one call



- Collective routines

`MPI_Bcast()`	Broadcast data from one process to all others
`MPI_Reduce()`	Reduce (e.g. sum) values on all processes
`MPI_Allgather()`	Gathers data from all processes, distributes to all





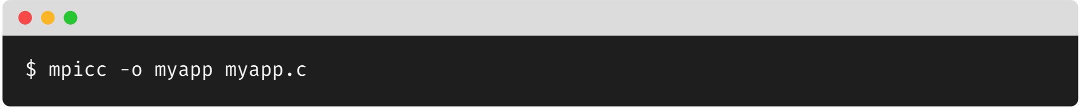
- *And many, many more!*

MPI Skeleton

```c
#include <mpi.h>
int main(int argc, char *argv[]) {
    // Initialize MPI
    MPI_Init(&argc, &argv);

    int rank, size;
    // Get current rank ID
    MPI_Comm_rank(MPI_COMM_WORLD ,&rank);
    // Get total number of ranks
    MPI_Comm_size(MPI_COMM_WORLD, &size);

    // Do something (call MPI routines, ...)
    ...

    // Shutdown MPI
    MPI_Finalize();
    return 0;
}
```

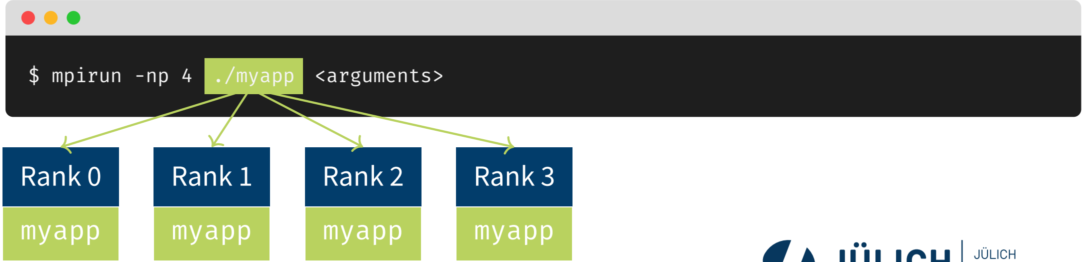
Using MPI

- Compile with MPI compiler (wrapper around usual compiler)



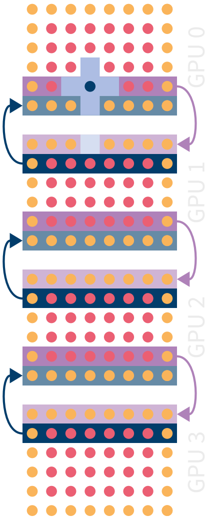
```
$ mpicc -o myapp myapp.c
```

- Run with MPI launcher `mpirun` (takes care about configuration, `$VARS`, …)



```
$ mpirun -np 4 ./myapp <arguments>
```

MPI Strategy for Jacobi Solver



- Goal: Extend parallelization from GPU threads to multiple GPUs
- Distribute grid of points to GPUs
- Halo points ● need special consideration
  That's what makes things interesting here
  - Evaluated point ● needs data from neighboring points ■
  - At border: Data might be on different GPU → Halos! ■
  - For every iteration step: Update halo from other GPU device
    ⇒ Regular MPI communications to top ■ and from top ■

```
MPI_Sendrecv( A+iy_start*nx+1 , nx-2, MPI_DOUBLE, top, 0,
              A+iy_end*nx+1 , nx-2, MPI_DOUBLE, bottom, 0,
              MPI_COMM_WORLD, MPI_STATUS_IGNORE);

MPI_Sendrecv( A+(iy_end-1)*nx+1 , nx-2, MPI_DOUBLE, top, 0,
              A+(iy_start-1)*nx+1 , nx-2, MPI_DOUBLE, bottom, 0,
              MPI_COMM_WORLD, MPI_STATUS_IGNORE);
```

Determining GPU ID

Affinity on nodes with multiple GPUs

- Problem: Usually, nodes have more than one GPU
- How would MPI know how to distribute the load?
- Select active GPU with

  ```
  #pragma acc set device_num(ID)
  ```
- Alternative and more in appendix

Parallel Jacobi III

TASK 4

Multi-GPU parallelism, asynchronous execution

- Implement domain decomposition for multiple GPUs

Task 4: Multi-GPU Usage

- *Note: Re-launch your cloud card with a **multi-GPU instance***
- Launch `Task_4.ipynb` Notebook
- Read text carefully and work on the implementations.

? What's your speed-up?

E Implement asynchronous halo communication; see `README.md` in `Task4`*`E`*`/`!

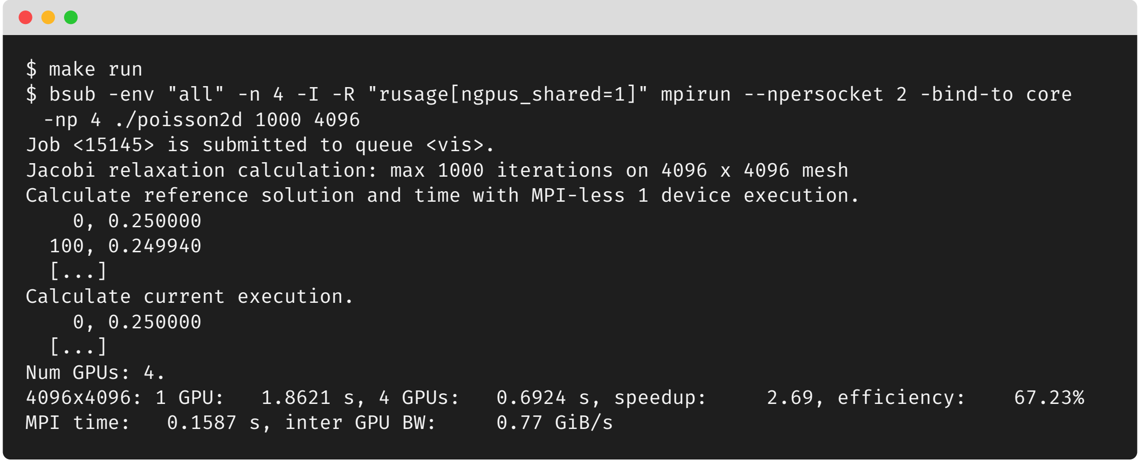
Parallel Jacobi III

Source Code

```
#pragma acc set device_num(rank)
// ...
int iy_start = rank * chunk_size;
int iy_end   = iy_start + chunk_size;
// ...
MPI_Sendrecv( A+iy_start*nx+ix_start, (ix_end-ix_start), MPI_REAL_TYPE, top   , 0,
              A+iy_end*nx+ix_start,   (ix_end-ix_start), MPI_REAL_TYPE, bottom, 0,
              MPI_COMM_WORLD, MPI_STATUS_IGNORE );
MPI_Sendrecv( A+(iy_end-1)*nx+ix_start,   (ix_end-ix_start), MPI_REAL_TYPE, bottom, 0,
              A+(iy_start-1)*nx+ix_start, (ix_end-ix_start), MPI_REAL_TYPE, top   , 0,
              MPI_COMM_WORLD, MPI_STATUS_IGNORE );
```

Parallel Jacobi III

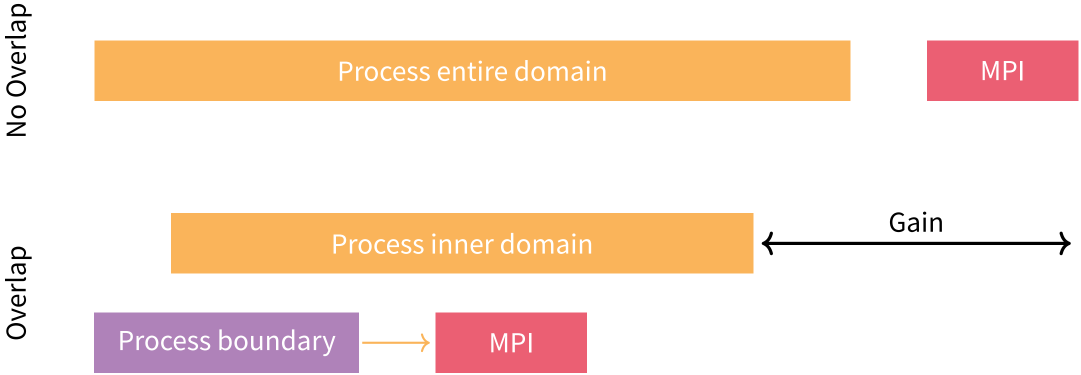
MPI run result



```
$ make run
$ bsub -env "all" -n 4 -I -R "rusage[ngpus_shared=1]" mpirun --npersocket 2 -bind-to core
  -np 4 ./poisson2d 1000 4096
Job <15145> is submitted to queue <vis>.
Jacobi relaxation calculation: max 1000 iterations on 4096 x 4096 mesh
Calculate reference solution and time with MPI-less 1 device execution.
    0, 0.250000
  100, 0.249940
  [...]
Calculate current execution.
    0, 0.250000
  [...]
Num GPUs: 4.
4096x4096: 1 GPU:   1.8621 s, 4 GPUs:   0.6924 s, speedup:     2.69, efficiency:    67.23%
MPI time:   0.1587 s, inter GPU BW:     0.77 GiB/s
```

Overlap Communication and Computation

Disentangling



Overlap Communication and Computation

OpenACC keyword

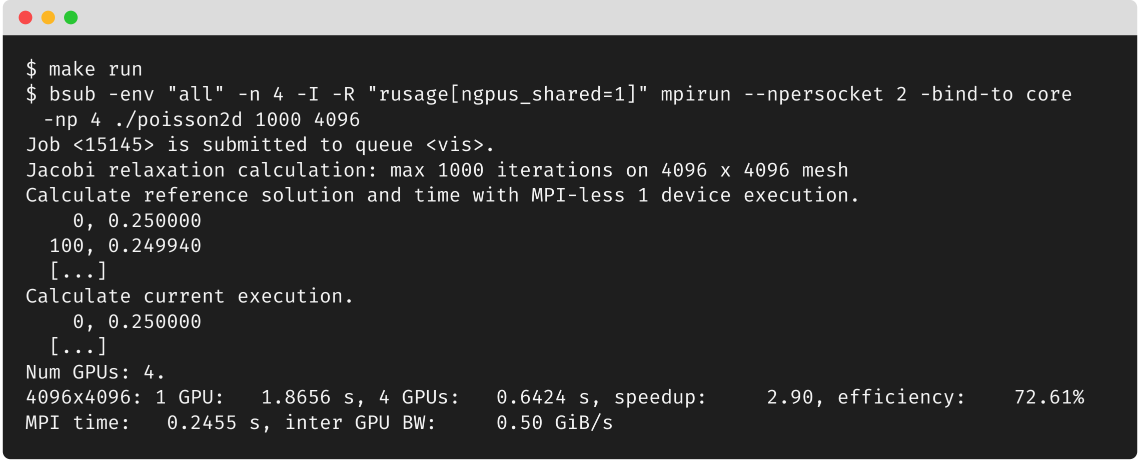
- OpenACC: Enable asynchronous execution with `async` keyword
- Runtime will execute `async`’ed region at same time
- Barrier: `wait`

```
#pragma acc parallel loop present(A, Anew)
for ( ... )  { } // Process boundary
#pragma acc parallel loop present(A, Anew) async
for ( ... )  { } // Process inner domain
#pragma acc host_data use_device (A) {
    MPI_Sendrecv(A+iy_start*nx+1, nx-2, MPI_DOUBLE, top, 0,
                 A+iy_end*nx+1, nx-2, MPI_DOUBLE, bottom, 0,
                 MPI_COMM_WORLD, MPI_STATUS_IGNORE);
    MPI_Sendrecv(A+(iy_end-1)*nx+1, nx-2, MPI_DOUBLE, bottom, 1,
                 A+(iy_start-1)*nx+1, nx-2, MPI_DOUBLE, top, 1,
                 MPI_COMM_WORLD, MPI_STATUS_IGNORE);
}
#pragma acc wait  // Wait for inner domain to finish processing
```

Parallel Jacobi III+

E

MPI *async* run result



```
$ make run
$ bsub -env "all" -n 4 -I -R "rusage[ngpus_shared=1]" mpirun --npersocket 2 -bind-to core
 -np 4 ./poisson2d 1000 4096
Job <15145> is submitted to queue <vis>.
Jacobi relaxation calculation: max 1000 iterations on 4096 x 4096 mesh
Calculate reference solution and time with MPI-less 1 device execution.
    0, 0.250000
  100, 0.249940
  [...]
Calculate current execution.
    0, 0.250000
  [...]
Num GPUs: 4.
4096x4096: 1 GPU:   1.8656 s, 4 GPUs:   0.6424 s, speedup:     2.90, efficiency:    72.61%
MPI time:   0.2455 s, inter GPU BW:     0.50 GiB/s
```

Conclusions, Summary

Conclusions & Summary

We’ve learned a lot today!

- **Newell** nodes are fat nodes:
  2 POWER9 CPUs (2 × 20 cores), 4 V100 GPUs (4 × 84 SMs)
- **OpenACC** can be used to efficiently exploit parallelism
- … on the CPU, similar to OpenMP,
- … on the GPU, for which it is specially designed for,
- … on multiple GPUs, working well together with MPI.

- There are still many more tuning possibilities and keywo …)

→ Great online resources to **deepen your knowledge** (see a

Thank you
for your attention!
a.herten@fz-juelich.de

APPENDIX

Appendix

- List of Tasks
- Supplemental: POWER9 Structure Diagrams
- Supplemental: NVIDIA GPU Memory Spaces
- Supplemental: Leveraging OpenACC Threads
- Supplemental: MPI
- Further Reading
- Glossary
- References

List of Tasks

Task 1: JARVICE

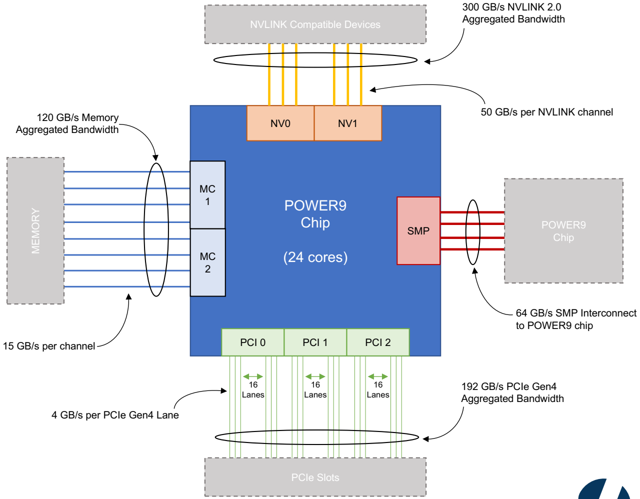
Task 2: A First Parallel Loop

Task 3: More Parallel Loops

Task 4: Multi-GPU Usage

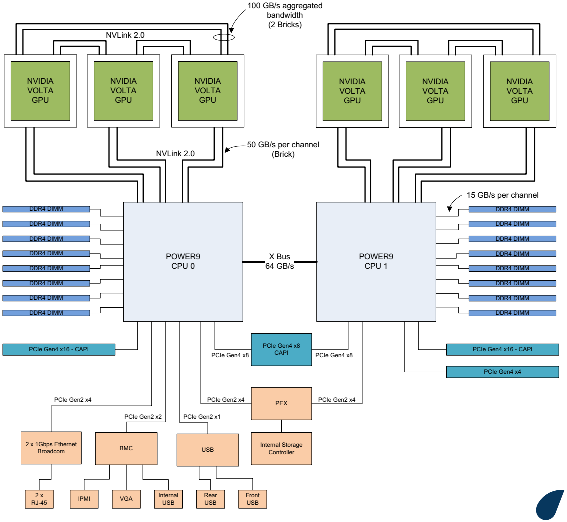
Supplemental: POWER9 Structure Diagrams

POWER9 Structure Diagram



Bicas Caldeira [7]

Newell Structure Diagram



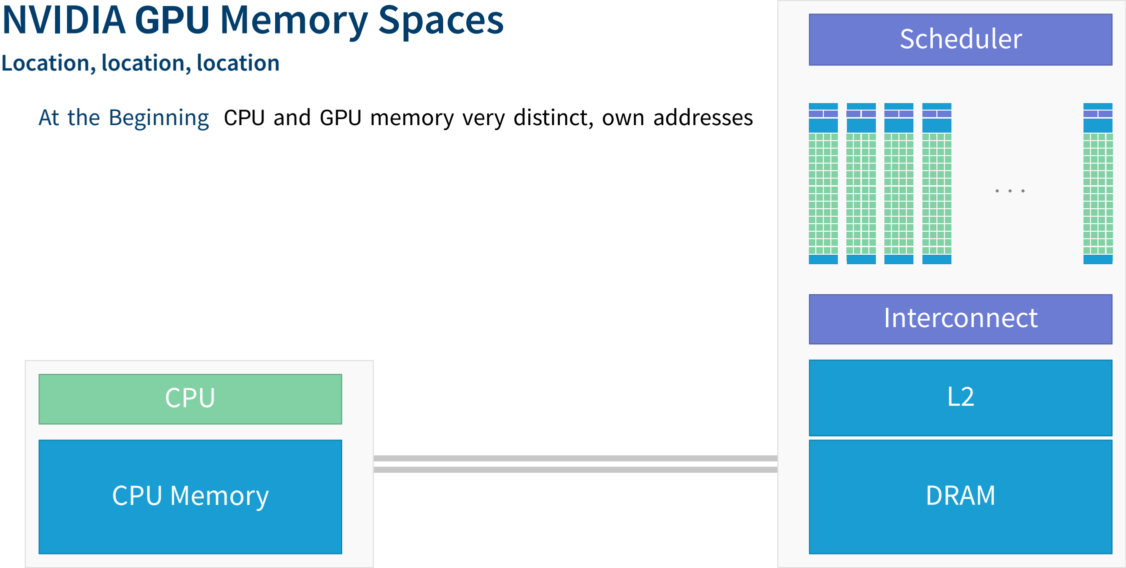
Bicas Caldeira [7]

Supplemental: NVIDIA GPU Memory Spaces

NVIDIA GPU Memory Spaces

Location, location, location

At the Beginning CPU and GPU memory very distinct, own addresses



NVIDIA GPU Memory Spaces

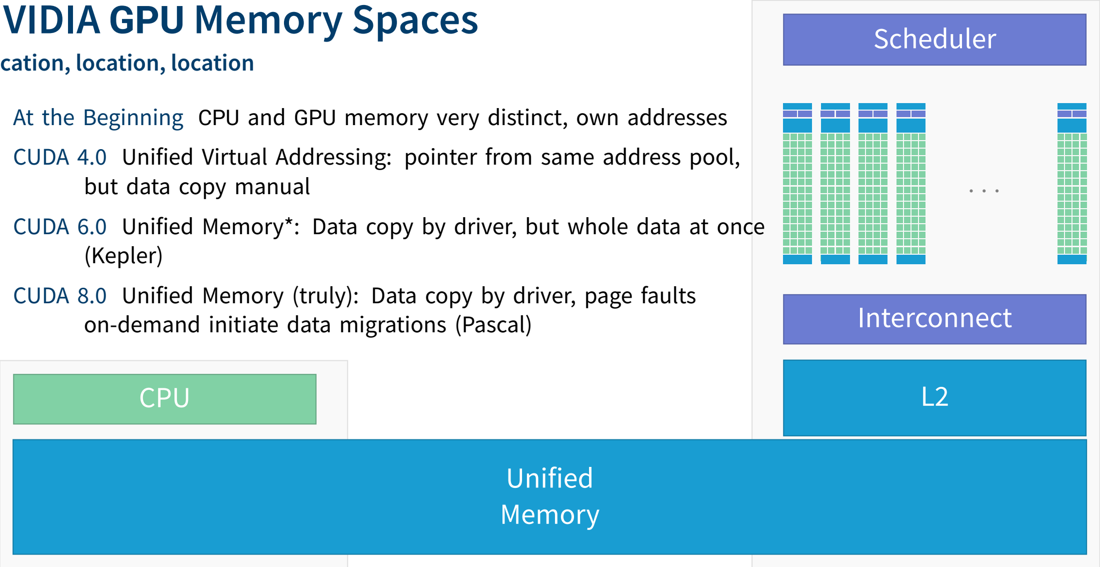
Location, location, location

At the Beginning CPU and GPU memory very distinct, own addresses

CUDA 4.0 Unified Virtual Addressing: pointer from same address pool, but data copy manual

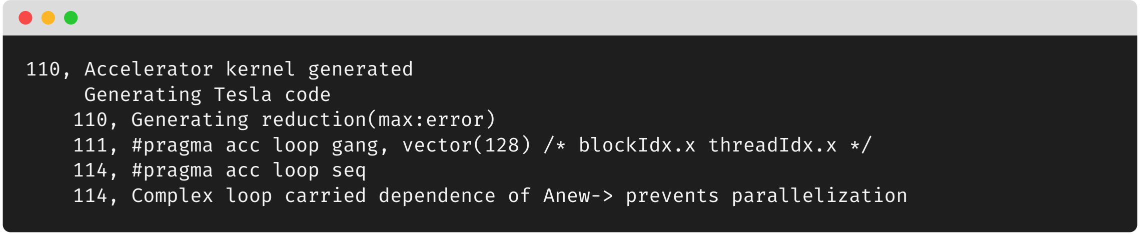
CUDA 6.0 Unified Memory*: Data copy by driver, but whole data at once (Kepler)

CUDA 8.0 Unified Memory (truly): Data copy by driver, page faults on-demand initiate data migrations (Pascal)



Supplemental: Leveraging OpenACC Threads

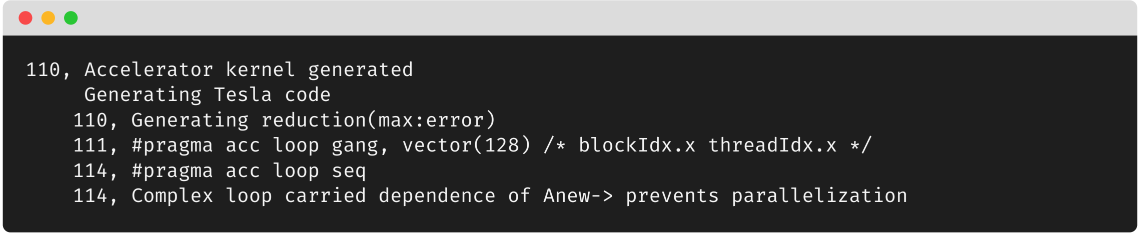
Understanding Compiler Output



```
110, Accelerator kernel generated
     Generating Tesla code
    110, Generating reduction(max:error)
    111, #pragma acc loop gang, vector(128) /* blockIdx.x threadIdx.x */
    114, #pragma acc loop seq
    114, Complex loop carried dependence of Anew-> prevents parallelization
```

```
110  #pragma acc parallel loop reduction(max:error)
111  for (int ix = ix_start; ix < ix_end; ix++)
112  {
113      // Inner loop
114      for (int iy = iy_start; iy < iy_end; iy++)
115      {
116          Anew[iy*nx+ix] = -0.25 * (rhs[iy*nx+ix] - ( A[iy*nx+ix+1] +  A[iy*nx+ix-1] +
             ↪  A[(iy-1)*nx+ix] + A[(iy+1)*nx+ix] ));
117          error = fmaxr( error, fabsr(Anew[iy*nx+ix]-A[iy*nx+ix]));
118      }
119  }
```

Understanding Compiler Output

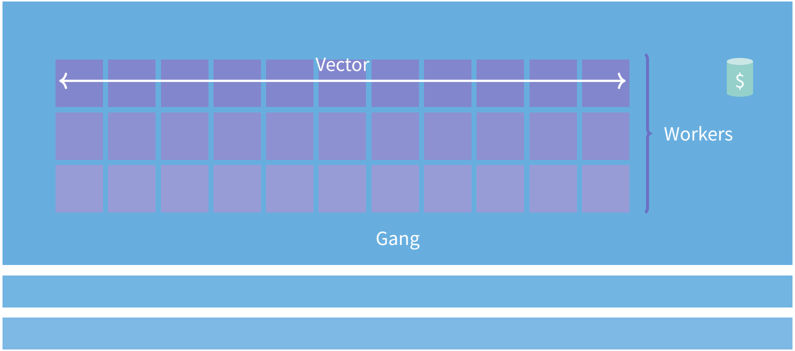


```
110, Accelerator kernel generated
     Generating Tesla code
    110, Generating reduction(max:error)
    111, #pragma acc loop gang, vector(128) /* blockIdx.x threadIdx.x */
    114, #pragma acc loop seq
    114, Complex loop carried dependence of Anew-> prevents parallelization
```

- Outer loop: Parallelism with gang and `vector`
- Inner loop: Sequentially per thread (`#pragma acc loop seq`)
- Inner loop was never parallelized!
- Rule of thumb: **Expose as much parallelism as possible**

OpenACC Parallelism

3 Levels of Parallelism



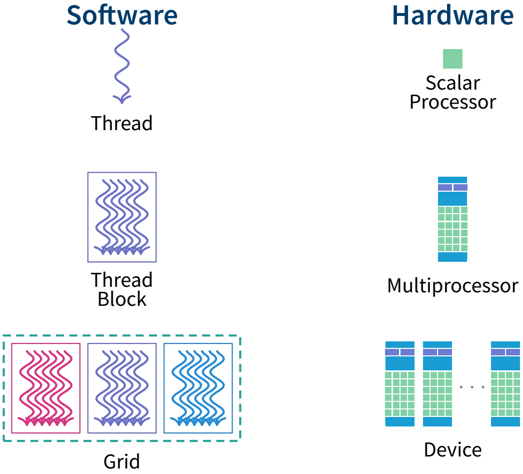
Vector
Vector threads work in lockstep (SIMD/SIMT parallelism)

Worker
Has 1 or more vector; workers share common resource (*cache*)

Gang
Has 1 or more workers; multiple gangs work independently from each other

CUDA Parallelism

CUDA Execution Model



Software

Thread

Thread Block

Grid

Hardware

Scalar Processor

Multiprocessor

Device

- **Threads** executed by scalar processors (*CUDA cores*)

- Thread **blocks**: Executed on multiprocessors (*SM*)
- Do not migrate
- Several concurrent thread blocks can reside on multiprocessor
  Limit: Multiprocessor resources (register file; shared memory)

- Kernel launched as **grid** of thread blocks
- Blocks, grids: Multiple dimensions

From OpenACC to CUDA

$\texttt{map}(||_{\texttt{acc}}, ||_{\texttt{<<<>>>}})$

- In general: Compiler free to do what it thinks is best
- Usually

gang	Mapped to blocks *(coarse grain)*
worker	Mapped to threads *(fine grain)*
vector	Mapped to threads *(fine SIMD/SIMT)*
seq	*No parallelism; sequential*

- Exact mapping compiler dependent
- Performance tips
    - Use vector size divisible by 32
    - Block size: `num_workers` × `vector_length`

Declaration of Parallelism

Specify configuration of threads

- Three **clauses** of parallel region (`parallel`, `kernels`) for changing distribution/configuration of group of threads
- Presence of keyword: Distribute using this level
- Optional size: Control size of parallel entity

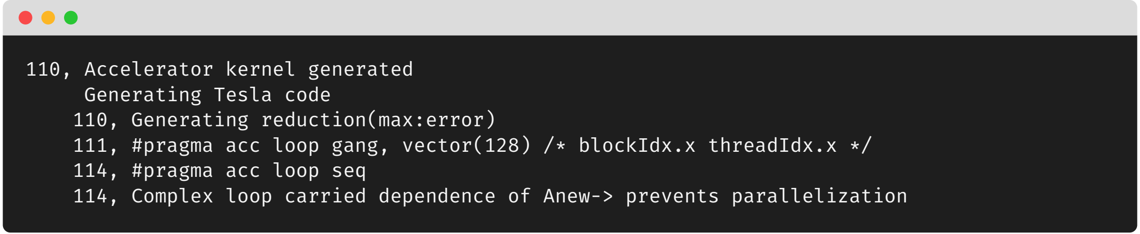
OpenACC: `gang worker vector`

```
#pragma acc parallel loop gang vector
```

Also: `worker`

Size: `num_gangs(n)`, `num_workers(n)`, `vector_length(n)`

Understanding Compiler Output II



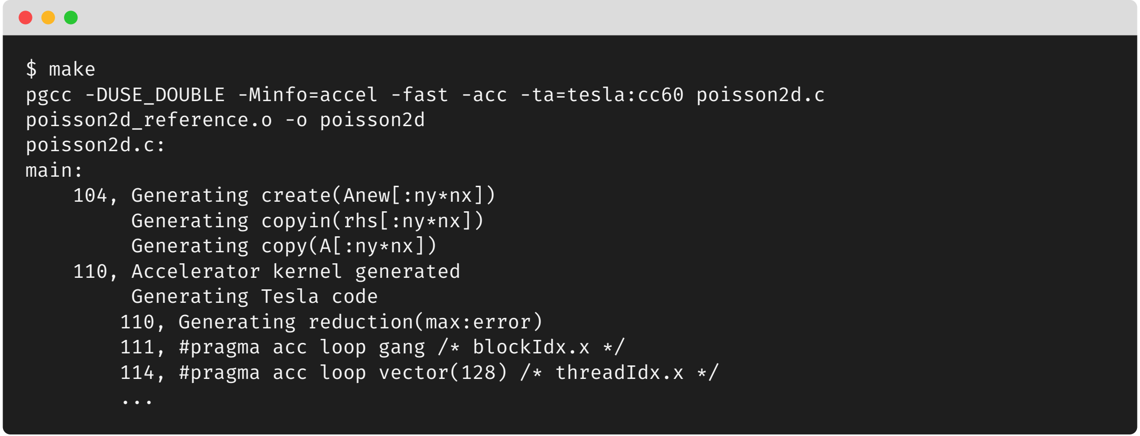
```
110, Accelerator kernel generated
     Generating Tesla code
    110, Generating reduction(max:error)
    111, #pragma acc loop gang, vector(128) /* blockIdx.x threadIdx.x */
    114, #pragma acc loop seq
    114, Complex loop carried dependence of Anew-> prevents parallelization
```

- Compiler reports configuration of parallel entities
  - **Gang** mapped to `blockIdx.x`
  - **Vector** mapped to `threadIdx.x`
  - **Worker** not used
- Here: 128 threads per block; as many blocks as needed

  128 seems to be default for Tesla/NVIDIA

More Parallelism

Compiler Output

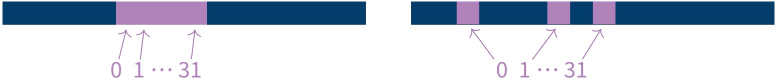


```
$ make
pgcc -DUSE_DOUBLE -Minfo=accel -fast -acc -ta=tesla:cc60 poisson2d.c
poisson2d_reference.o -o poisson2d
poisson2d.c:
main:
    104, Generating create(Anew[:ny*nx])
         Generating copyin(rhs[:ny*nx])
         Generating copy(A[:ny*nx])
    110, Accelerator kernel generated
         Generating Tesla code
        110, Generating reduction(max:error)
        111, #pragma acc loop gang /* blockIdx.x */
        114, #pragma acc loop vector(128) /* threadIdx.x */
        ...
```

Memory Coalescing

Memory in batch

- Coalesced access *good*
  - Threads of warp (group of 32 contiguous threads) access adjacent words
  - Few transactions, high utilization
- Uncoalesced access *bad*
  - Threads of warp access scattered words
  - Many transactions, low utilization
- Best **performance**: `threadIdx.x` should access contiguously



Supplemental: MPI

Handling Multi-GPU Hosts

The alternative

- Use OpenACC API to select GPU

```
#if _OPENACC
acc_device_t device_type = acc_get_device_type();  // Get dev type
int ngpus = acc_get_num_devices(device_type);  // Get number of devs
int devicenum = rank%ngpus;  // Compute active dev number based on rank
acc_set_device_num(devicenum, device_type);
#endif /*_OPENACC*/
```

- Get rank ID
  - MPI API: `MPI_Comm_rank()`
  - Environment variables (`int rank = atoi(getenv(...))`)

OpenMPI	`$OMPI_COMM_WORLD_LOCAL_RANK`
MVAPICH2	`$MV2_COMM_WORLD_LOCAL_RANK`

Further Reading

Further Resources on OpenACC

- `www.openacc.org`: Official home page of OpenACC
- `developer.nvidia.com/openacc-courses`: OpenACC courses, upcoming (live) and past (recorded)
- `https://nvidia.qwiklab.com/quests/3`: Qwiklabs for OpenACC; various levels
- Book: **Chandrasekaran and Juckeland** *OpenACC for Programmers: Concepts and Strategies* `https://www.amazon.com/OpenACC-Programmers-Strategies-Sunita-Chandrasekaran/dp/0134694287` [11]
- Book: **Farber** *Parallel Programming with OpenACC* `https://www.amazon.com/Parallel-Programming-OpenACC-Rob-Farber/dp/0124103979` [12]

Glossary I

API A programmatic interface to software by well-defined functions. Short for application programming interface. 78, 79, 112

CUDA Computing platform for GPUs from NVIDIA. Provides, among others, CUDA C/C++. 28, 99, 100, 106

GCC The GNU Compiler Collection, the collection of open source compilers, among others for C and Fortran. 27, 30

MPI The Message Passing Interface, a API definition for multi-node computing. 78, 79, 80, 81, 82, 83, 86, 89, 91, 93, 111, 112

NVIDIA US technology company creating GPUs. 4, 5, 6, 23, 93, 98, 99, 100, 115, 116, 117

Glossary II

NVLink NVIDIA's communication protocol connecting CPU ↔ GPU and GPU ↔ GPU with high bandwidth. 5, 6, 7, 8, 74, 117

OpenACC Directive-based programming, primarily for many-core machines. 2, 3, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 36, 41, 42, 44, 46, 48, 53, 57, 58, 60, 61, 64, 65, 69, 70, 71, 72, 73, 75, 76, 78, 88, 91, 93, 101, 104, 106, 107, 112, 114

OpenMP Directive-based programming, primarily for multi-threaded machines. 24, 48, 52, 55, 91

PAPI The Performance API, a C/C++ API for querying performance counters. 37

Pascal GPU architecture from NVIDIA (announced 2016). 99, 100

perf Part of the Linux kernel which facilitates access to performance counters; comes with command line utilities. 37

Glossary III

PGI Compiler creators. Formerly *The Portland Group, Inc.*; since 2013 part of NVIDIA. 27, 30, 37, 52

POWER CPU architecture from IBM, earlier: PowerPC. See also POWER8. 2, 3, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 74, 91, 93, 95, 96, 117

POWER8 Version 8 of IBM's POWERprocessor, available also under the OpenPOWER Foundation. 117

V100 A large GPU with the Volta architecture from NVIDIA. It employs NVLink 2 as its interconnect and has fast *HBM2* memory. Additionally, it features *Tensorcores* for Deep Learning and Independent Thread Scheduling. 7, 8, 91

Volta GPU architecture from NVIDIA (announced 2017). 6, 117

CPU Central Processing Unit. 2, 3, 5, 6, 7, 8, 27, 48, 59, 72, 74, 91, 99, 100, 116, 117

Glossary IV

GPU Graphics Processing Unit. 2, 3, 4, 5, 6, 7, 8, 22, 27, 29, 57, 59, 60, 62, 74, 82, 83, 84, 91, 93, 98, 99, 100, 112, 115, 116, 117

SM Streaming Multiprocessor. 7, 91

SMT Simultaneous Multithreading. 7

References I

[6] The Next Platform. *Power9 To The People*. POWER9 Performance Data. URL: `https://www.nextplatform.com/2017/12/05/power9-to-the-people/`.

[7] Alexandre Bicas Caldeira. *IBM Power System AC922: Introduction and Technical Overview*. IBM Redbooks. URL: `http://www.redbooks.ibm.com/redpieces/pdfs/redp5472.pdf` (pages 6, 96, 97).

[10] Donald E. Knuth. “Structured Programming with Go to Statements”. In: *ACM Comput. Surv.* 6.4 (Dec. 1974), pp. 261–301. ISSN: 0360-0300. DOI: `10.1145/356635.356640`. URL: `http://doi.acm.org/10.1145/356635.356640` (page 37).

References II

[11] Sunita Chandrasekaran and Guido Juckeland. *OpenACC for Programmers: Concepts and Strategies*. Addison-Wesley Professional, 2017. ISBN: 0134694287. URL: https://www.amazon.com/OpenACC-Programmers-Strategies-Sunita-Chandrasekaran/dp/0134694287 (page 114).

[12] Rob Farber. *Parallel Programming with OpenACC*. Morgan Kaufmann, 2016. ISBN: 0124103979. URL: https://www.amazon.com/Parallel-Programming-OpenACC-Rob-Farber/dp/0124103979 (page 114).

References: Images, Graphics I

[1] SpaceX. *SpaceX Launch*. Freely available at Unsplash. URL: `https://unsplash.com/photos/uj3hvdfQujI`.

[2] Forschungszentrum Jülich. *Hightech made in 1960: A view into the control room of DIDO*. URL: `http://historie.fz-juelich.de/60jahre/DE/Geschichte/1956-1960/Dekade/_node.html` (page 4).

[3] Forschungszentrum Jülich. *Forschungszentrum Bird's Eye*. (Page 4).

[4] Forschungszentrum Jülich. *JUQUEEN Supercomputer*. URL: `http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html` (page 4).

References: Images, Graphics II

[5] Rob984 via Wikimedia Commons. *Europe orthographic Caucasus Urals boundary (with borders)*. URL: https://commons.wikimedia.org/wiki/File:Europe_orthographic_Caucasus_Urals_boundary_(with_borders).svg (page 4).

[8] Wikichip. *POWER9 Scale-Out Die (Annotated)*. URL: https://en.wikichip.org/wiki/File:power9_so_die_(annotated).png.

[9] Setyo Ari Wibowo. *Ask*. URL: https://thenounproject.com/term/ask/1221810.