Parallel Programming

(Lectures on High-performance Computing for Economists XI)

Jesús Fernández-Villaverde, Pablo Guerrón, and David Zarruk Valencia

November 21, 2021

1University of Pennsylvania
2Boston College
3ITAM
Motivation
Why parallel?

- Moore’s Law (1965): transistor density of semiconductor chips would double roughly every 18 months.

- Problems when transistor size falls by a factor $x$:
  1. Electricity consumption goes up by $x^4$.
  2. Heat goes up.
  3. Manufacturing costs go up.

- Inherent limits on serial machines imposed by the speed of light (30 cm/ns) and transmission limit of copper wire (9 cm/ns): virtually impossible to build a serial Teraflop machine with current approach.

- Furthermore, real bottleneck is often memory access (RAM latency has only improved around 10% a year).

- Alternative: having more processors!
Cray-1, 1975
Parallel programming

• Main idea ⇒ divide a complex problem into easier parts:

  1. Numerical computation ⇒ matrix multiplication.

  2. Data handling (MapReduce and Spark) ⇒ computing moments.

• Two issues:

  1. Algorithms.

  2. Coding.
Some references

- *Introduction to High Performance Computing for Scientists and Engineers* by Georg Hager and Gerhard Wellein.

- *Parallel Computing for Data Science: With Examples in R, C++ and CUDA*, by Norman Matloff.


- *An Introduction to Parallel Programming* by Peter Pacheco.

- *Principles of Parallel Programming* by Calvin Lin and Larry Snyder.

- *Structured Parallel Programming: Patterns for Efficient Computation* by Michael McCool, James Reinders, and Arch Robison.
When do we parallelize? I

• Scalability:

  1. Strongly scalable: problems that are inherently easy to parallelize.
  2. Weakly scalable: problems that are not.

• Granularity:

  1. Coarse: more computation than communication.
  2. Fine: more communication.

• Overheads and load balancing.
When do we parallelize? II

- Whether or not the problem is easy to parallelize may depend on the way you set it up.

- Taking advantage of your architecture.

- Trade off between speed up and coding time.

- Debugging and profiling may be challenging.

- You will need a good IDE, debugger, and profiler.
Example I: value function iteration

\[
V(k) = \max_{k'} \{ u(c) + \beta V(k') \}
\]
\[
c = k^\alpha + (1 - \delta) k - k'
\]

1. We have a grid of capital with 100 points, \( k \in [k_1, k_2, ..., k_{100}] \).
2. We have a current guess \( V^n(k) \).
3. We can send the problem:

\[
\max_{k'} \{ u(c) + \beta V^n(k') \}
\]
\[
c = k_1^\alpha + (1 - \delta) k_1 - k'
\]

to processor 1 to get \( V^{n+1}(k_1) \).
4. We can send similar problem for each \( k \) to each processor.
5. When all processors are done, we gather the \( V^{n+1}(k_1) \) back.
Example II: random walk Metropolis-Hastings

• Draw $\theta \sim P(\cdot)$

• How?
  1. Given a state of the chain $\theta_{n-1}$, we generate a proposal:

$$\theta^* = \theta_{n-1} + \lambda \varepsilon, \; \varepsilon \sim \mathcal{N}(0, 1)$$

  2. We compute:

$$\alpha = \min \left\{ 1, \frac{P(\theta^*)}{P(\theta_{n-1})} \right\}$$

  3. We set:

$$\begin{align*}
\theta_n &= \theta^* \text{ w.p. } \alpha \\
\theta_n &= \theta_{n-1} \text{ w.p. } 1 - \alpha
\end{align*}$$

• Problem: to generate $\theta^*$ we need to $\theta_{n-1}$.

• No obvious fix (parallel chains violate the asymptotic properties of the chain).
The Model
Life-cycle model

- Households solve:

\[ V(t, e, x) = \max_{\{c, x'\}} \left\{ \frac{c^{1-\sigma}}{1-\sigma} + \beta \mathbb{E} V(t+1, e', x') \right\} \]

\[ s.t. \]

\[ c + x' \leq (1 + r)x + ew \]

\[ \mathbb{P}(e' | e) = \Gamma(e) \]

\[ x' \geq 0 \]

\[ t \in \{1, \ldots, T\} \]
Computing the model

1. Choose grids for assets \( X = \{ x_1, \ldots, x_{n_x} \} \) and shocks \( E = \{ e_1, \ldots, e_{n_e} \} \).
Computing the model

1. Choose grids for assets $X = \{x_1, \ldots, x_{n_x}\}$ and shocks $E = \{e_1, \ldots, e_{n_e}\}$.

2. Backwards induction:
1. Choose grids for assets $X = \{x_1, \ldots, x_{n_x}\}$ and shocks $E = \{e_1, \ldots, e_{n_e}\}$.

2. Backwards induction:
   
   2.1 For $t = T$ and every $x_i \in X$ and $e_j \in E$, solve the static problem:

   $$V(t, e_j, x_i) = \max_{\{c\}} u(c) \quad s.t. \quad c \leq (1 + r)x_i + e_jw$$
Computing the model

1. Choose grids for assets $X = \{x_1, \ldots, x_{n_x}\}$ and shocks $E = \{e_1, \ldots, e_{n_e}\}$.

2. Backwards induction:

   2.1 For $t = T$ and every $x_i \in X$ and $e_j \in E$, solve the static problem:
   
   $$V(t, e_j, x_i) = \max_{\{c\}} u(c) \quad s.t. \quad c \leq (1 + r)x_i + e_j w$$

   2.2 For $t = T - 1, \ldots, 1$, use $V(t + 1, e_j, x_i)$ to solve:

   $$V(t, e_j, x_i) = \max_{\{c, x' \in X\}} u(c) + \beta \mathbb{E} V(t + 1, e', x') \quad s.t. \\
   c + x' \leq (1 + r)x_i + e_j w \\
   \mathbb{P}(e' \in E | e_j) = \Gamma(e_j)$$
for(age = 1:-1:T)
    for(ix = 1:nx)
        for(ie = 1:ne)
            VV = -10^3;
            for(ixp = 1:nx)
                expected = 0.0;
                if(age < T)
                    for(iep = 1:ne)
                        expected = expected + P[ie, iep]*V[age+1, ixp, iep];
                    end
                end
                cons = (1+r)*xgrid[ix] + egrid[ie]*w - xgrid[ixp];
                utility = (cons^(-1/sigma))/(1-sigma) + bbeta*expected;
                if(cons <= 0)
                    utility = -10^5;
                end
                if(utility >= VV)
                    VV = utility;
                end
            end
        end
    end
end
In parallel

1. Set $t = T$.
2. Given $t$, the computation of $V(t, e_j, x_i)$ is independent of the computation of $V(t, e_{j'}, x_{i'})$, for $i \neq i'$, $j \neq j'$.
3. One processor can compute $V(t, e_j, x_i)$ while another processor computes $V(t, e_{j'}, x_{i'})$.
4. When the different processors are done at computing $V(t, e_j, x_i)$, $\forall x_i \in X$ and $\forall e_j \in E$, set $t = t - 1$.
5. Go to 1.

Note that the problem is not parallelizable on $t$. The computation of $V(t, e, x)$ depends on $V(t + 1, e, x)$!
Computational features of the model

1. The simplest life-cycle model.

2. **Three** state variables:
   2.1 Age.
   2.2 Assets.
   2.3 Productivity shock.

3. Parallelizable only on assets and shock, **not** on age.

4. May become infeasible to estimate:
   4.1 With more state variables:
      - Health.
      - Housing.
      - Money.
      - Different assets.
   4.2 If embedded in a general equilibrium.
In parallel

**SCATTER**

- Thread 1: 1 2
- Thread 2: 3 4
- Thread 3: 5 6

**GATHER**

- Thread 1: 1 2
- Thread 2: 3 4
- Thread 3: 5 6
Parallel execution of the code
Many workers instead of one

**Figure 1:** 1 Core Used for Computation

**Figure 2:** 8 Cores Used for Computation
Parallelization limits
Costs of parallelization

- Amdahl’s Law: the speedup of a program using multiple processors in parallel computing is limited by the time needed for the sequential fraction of the program.

- Costs:
  - Starting a thread or a process/worker.
  - Transferring shared data to workers.
  - Synchronizing.

- Load imbalance: for large machines, it is often difficult to use more than 10% of its computing power.
Parallelization limits on a laptop

- Newest processors have plenty of processor.
- For example, in these slides, we used 4 physical cores + 4 virtual cores = 8 logical cores.
Multi-core processors

Intel® Core™ X-Series Processors

- High-performance desktops
- First 18-core processor
- Extreme gaming, mega-tasking, and high-end content creation
• Spend some time getting to know your laptop’s limits and the problem to parallelize.

• In our life-cycle problem with many grid points, parallelization improves performance almost linearly, up to the number of physical cores.

• Parallelizing over different threads of the same physical core does not improve speed if each thread uses 100% of core capacity.

• For computationally heavy problems, adding more threads than cores available may even reduce performance.
Your laptop is not the limit!

- Cluster servers.

- Amazon Web Services - EC2 at https://aws.amazon.com/ec2/:
  - Almost as big as you want!
  - Replace a large initial capital cost for a variable cost (use-as-needed).
  - Check: https://aws.amazon.com/ec2/pricing/
    - 8 processors with 32Gb, general purpose: $0.332 per hour.
    - 64 processors with 256Gb, compute optimized: $3.20 per hour.
Running an instance on AWS

- Go to: https://console.aws.amazon.com/
- Click on EC2.
- Click on Launch Instance and follow the window links (for example, Ubuntu Server 18.04).
- Public key:
  - Create a new key pair.
  - Download key.
  - Store it in a secure place (usually ~/.ssh/).
- Run instance.
Working on AWS instance

On Unix/Linux terminal:

- Transfer folder from local to instance with scp:

```bash
$ scp -i "/path/"Harvard_Spring_2018.pem" -r "/pathfrom/FOLDER/" ec2-user@ec2-34-226-147-93.compute-1.amazonaws.com:~
```

- Make sure key is not publicly available:

```bash
$ chmod 400 "Harvard_Spring_2018.pem"
```

- Connect to instance with ssh:

```bash
$ ssh -i "Harvard_Spring_2018.pem"
ec2-user@ec2-34-226-147-93.compute-1.amazonaws.com
```
Parallelization
• More common in economics.

1. Packages/libraries/toolboxes within languages:
   
   1.1 Julia.
   
   1.2 Python.
   
   1.3 R.
   
   1.4 Matlab.

2. Explicit parallelization:
   
   2.1 OpenMP.
   
   2.2 MPI.
   
   2.3 GPU programming: CUDA, OpenCL, and OpenACC.
Programming modes II

- Less common in economics.

1. Automatic parallelization: AutoParInGCC, Intel compilers.

2. Partitioned Global Address Space Languages (PGAS):
   - 2.1 Coarray Fortran.
   - 2.2 UPC.
   - 2.3 X10.
   - 2.4 Chapel.

3. Pthreads (POSIX threads).

4. TPUs (Tensor processing units).

5. FPGAs (field programmable gate arrays).

6. Hybrids.
Flynn’s taxonomy

- SISD
- MISD
- SIMD
- MIMD

MPMD

SPMD
Two ways of parallelizing

1. for loop:
   - Adding a statement before a for loop that wants to be parallelized.

2. Map and reduce:
   - Create a function that depends on the state variables over which the problem can be parallelized:
     - In our example, we have to create a function that computes the value function for a given set of state variables.
   - Map computes in parallel the function at a vector of states.
   - Reduce combines the values returned by map in the desired way.
Julia
Parallelization in Julia - \textit{for} loops

- Parallelization of \textit{for} loops is worth for "small tasks."

- "Small task" \textit{==} "few computations on each parallel iteration":
  - Few control variables.
  - Few grid points on control variables.
  - Our model is a "small task."
Parallelization in Julia - *for* loops

1. Load distributed module
   
   ```
   using Distributed
   ```

2. Set number of workers:
   
   ```
   addprocs(5)
   ```

3. Remove workers:
   
   ```
   rmprocs(2,3,5)
   ```

4. Checking workers:
   
   ```
   workers()
   ```
Parallelization in Julia - *for* loops

1. Load distributed and SharedArrays modules
   ```julia
   using Distributed
   using SharedArrays
   ```

2. Declare variables used inside the parallel *for* loop that are not modified inside parallel iterations to be `@everywhere`:
   ```julia
   @everywhere nx = 1500;
   ```

3. Declare variables used inside the parallel *for* loop that are modified inside parallel iterations as `SharedArray`:
   ```julia
   tempV = SharedArray{Float64}(ne*nx);
   ```
Parallelization in Julia - *for* loops

4. Data structure of state and exogenous variables

```julia
@everywhere struct modelState
    ind::Int64
    ne::Int64
    nx::Int64
    T::Int64
    age::Int64
    P::Array{Float64,2}
    xgrid::Vector{Float64}
    egrid::Vector{Float64}
    ssigma::Float64
    bbeta::Float64
    V::Array{Float64,2}
    w::Float64
    r::Float64
end
```
5. Define a function that computes value function for a given state:

```julia
@everywhere function value(currentState::modelState)
    ind = currentState.ind
    age = currentState.age
    # ...
    VV = -10.0^3;
    ixpopt = 0;
    for ixp = 1:nx
        # ...
    end
    return(VV);
end
```
6. For parallelizing a `for` loop, add `@distributed` before the `for` statement:

```julia
@distributed for ind = 1:(ne*nx)
    # ...
end
```

7. To synchronize before the code continues its execution, add `@sync` before the `@distributed for` statement:

```julia
@sync @distributed for ind = 1:(ne*nx)
    # ...
end
```
Parallelization in Julia - *for* loops

- Choose appropriately the dimension(s) to parallelize:

```julia
nx = 350;
ne = 9;
for(ie = 1:ne)
    @sync @distributed for(ix = 1:nx)
        # ...
    end
end
```

- The first one is *much* faster, as there is less communication.
OR convert the problem so all state variables are computed by iterating over a one-dimensional loop:

```julia
@sync @distributed for ind = 1:(ne*nx)
    ix  = convert(Int,ceil(ind/ne));
    ie  = convert(Int,floor(mod(ind-0.05, ne))+1);
    # ...
end
```

Communication time is minimized!
Parallelization in Julia - Performance

**Figure 3:** Julia - 1 core used for computation

**Figure 4:** Julia - 8 cores used for computation
Parallelization in Julia - Performance

Figure 5: Computing time (s)
Parallelization in Julia - *for* loops

- Speed decreases with the number of global variables used.
- Very sensible to the use of large `SharedArray` objects.
- Can be faster without parallelization than with large shared objects.
- See code on github
• Problems with more computations per iteration.

• Value function/life-cycle models with more computations per state:
  
  • Many control variables.
  
  • Discrete choice (marry-not marry, accept-reject work offer, default-repay, etc.).

• If problem is “small”, using map for parallelization is slower.

• See examples 3 and 4 on github.
1. Most of the code as as in the for case.

2. The function \texttt{pmap(f,s)} computes the function \texttt{f} at every element of \texttt{s} in parallel:

```julia
for(age = T:-1:1)
    pars = [modelState(ix, age, ..., w, r) for ix in 1:nx];
    s = pmap(value,pars);
    for(ind = 1:nx)
        V[age, ix, ie] = s[ix];
    end
end
```
Parallelization in Julia - Performance

Figure 6: Computing time (s)
• Assess size of problem, but usually problem grows as paper evolves!

• Wrapping value function computation for every state might significantly increase speed (even more than parallelizing).
Python
1. Use `joblib` package

```python
from joblib import Parallel, delayed
import multiprocessing
```

2. Define a parameter structure for value function computation:

```python
class modelState(object):
    def __init__(self, age, ix, ...):
        self.age = age
        self.ix = ix
        # ...
```
3. Define a function that computes value for a given input `states` of type `modelState`:

```python
def value_func(states):
    nx = states.nx
    age = states.age
    # ...
    VV = math.pow(-10, 3)
    for ixp in range(0,nx):
        # ...
    return [VV];
```
4. The function `Parallel`:

```python
results = Parallel(n_jobs=num_cores)(delayed(value_func)(modelState(ix, age, ..., w, r)) for ind in range(0,nx*ne))
```

maps the function `value_func` at every element of `modelState(ix, age, ..., w, r)` in parallel using `num_cores` cores.
5. Life-cycle model:

```python
for age in reversed(range(0,T)):
    results = Parallel(n_jobs=num_cores)(delayed(value_func)
        (modelState(ix, age, ..., w, r)) for ix in range(0,nx))
    for ix in range(0,nx):
        V[age, ix] = results[ix][0];
```
Parallelization in Python - Performance

Figure 7: Computing time (s)
R
1. Use package `parallel`:

   ```r
   library("parallel")
   ```

2. Create the structure of parameters for the function that computes the value for a given state as a `list`:

   ```r
   states = lapply(1:nx, function(x) list(age=age, ix=x, ..., r=r))
   ```
3. Create the function that computes the value for a given state:

```r
value = function(x){
  age   = x$age
  ix    = x$ix
  ...
  VV = -10^3;
  for(ixp in 1:nx){
    # ...
  }
  return(VV);
}
```
4. Define the cluster with desired number of cores:

```r
cl <- makeCluster(no_cores)
```

5. Use function `parLapply(cl, states, value)` to compute `value` at every state in `states` with `cl` cores:

```r
for(age in T:1){
    states = lapply(1:nx, ...
    for(ix in 1:nx){
        V[age, ix] = s[[ix]][1]
    }
}
```
Parallelization in R - Performance

![Graph showing the relationship between the number of cores and computing time (s). The graph indicates a decrease in time as the number of cores increases, reaching a plateau at around 4 cores. The x-axis represents the number of cores (1 to 8), and the y-axis represents time (s) with a minimum of 0 and a maximum of 1200. A red dashed line indicates the physical core limit.](image-url)
Matlab
Using the **parallel toolbox**:

1. Initialize number of workers with `parpool()`:

   ```matlab
   parpool(6)
   ```

2. Replace the `for` loop with `parfor`:

   ```matlab
   for age = T:-1:1
     parfor ie = 1:1:ne
       % ...
     end
   end
   ```
Parallelization in Matlab - Performance

Figure 9: Computing time (s)
Parallelization in Matlab

- Extremely easy.
- Also simple to extend to GPU.
- There is no free lunch $\implies$ very poor performance.
OpenMP
OpenMP I

- Open specifications for multi-processing.
- It has been around for two decades. Current version 4.5.
- Official web page: http://openmp.org/wp/
- Tutorial: https://computing.llnl.gov/tutorials/openMP/
- *Using OpenMP: Portable Shared Memory Parallel Programming* by Barbara Chapman, Gabriele Jost, and Ruud van der Pas.
- Fast to learn, reduced set of instructions, easy to code, but you need to worry about contention and cache coherence.
• API for multi-processor/core, shared memory machines defined by a group of major computer hardware and software vendors.

• C++ and Fortran. Extensions to other languages.

• For example, you can have OpenMP in Mex files in Matlab.

• Supported by major compilers (GCC) and IDEs (Clion).

• Thus, it is usually straightforward to start working with it.
• Multithreading with fork-join.

• Rule of thumb: One thread per processor.

• Job of the user to remove dependencies and synchronize data.

• Heap and stack (LIFO).

• Race conditions: you can impose fence conditions and/or make some data private to the thread.

• Remember: synchronization is expensive and loops suffer from overheads.
Fork-join

- Master Thread
  - Thread 1
  - Thread 2
  - Thread 3
  - Parallel region

- Thread 1
  - Thread 2
  - Thread 3
  - Thread 4
  - Parallel region

- Master Thread
  - Thread 1
  - Thread 2
  - Parallel region
1. At compilation, add flag:

```
-fopenmp
```

2. Set environmental variable `OMP_NUM_THREADS`:

```
export OMP_NUM_THREADS=32
```

3. Add line before loop:

```
#pragma omp parallel for shared(V, ...) private(VV, ...)
  for(int ix=0; ix<nx; ix++){
    // ...
  }
```

4. We can always recompile without the flag and compiler directives are ignored.

5. Most implementations (although not the standard!) allow for nested parallelization and dynamic thread changes.
Parallelization in C++ using OpenMP - Performance

![Graph showing the computing time (s) for different numbers of cores. The y-axis represents time in seconds, ranging from 0.0 to 6.0, and the x-axis represents the number of cores, ranging from 1 to 8. The graph shows a decrease in time as the number of cores increases, with a peak around 4 cores. The physical cores are indicated by a red dashed line.]
1. Write your code in C++, adding the parallelization statement

```cpp
#pragma omp parallel for shared(...) private(...)```

2. In the C++ code, add the following line to any function that you want to import from R:

```cpp
// [[Rcpp::export]]```

3. In R, load the `Rcpp` package:

```r
library("Rcpp")```
4. Set the environmental variable `OMP_NUM_THREADS` using the `Sys.setenv()` function:

```r
Sys.setenv("OMP_NUM_THREADS"="8")
```

5. Add the `-fopenmp` flag using `Sys.setenv()` function:

```r
Sys.setenv("PKG_CXXFLAGS"=" -fopenmp")
```

6. Compile and import using `sourceCpp`:

```r
sourceCpp("my_file.cpp")
```
MPI
• Message Passing Interface (MPI) is a standardized and portable message-passing system based on the consensus of the MPI Forum.

• Official web page (and for downloads): http://www.open-mpi.org/

• Tutorial: https://computing.llnl.gov/tutorials/mpi/

• A couple of references:
  

  2. *MPI: The Complete Reference - Volumes 1 and 2*, by several authors.
• MPI is organized as a library performed with routine calls.

• Bindings for C++ and Fortran. Also for Python, Julia, R, and other languages.

• For example, you can have MPI in Mex files in Matlab.

• Harder to learn (MPI 3.0 standard has more than 440 routines) and code, but extremely powerful ⇒ used for state-of-the-art computations.

• Multiple processes (thread with its own controller).

• Thus, better for coarse parallelization.
In particular, MPI can broadcast (transfers an array from the master to every thread), scatter (transfers different parts of an array from the master node to every thread; note the difference with broadcasting), and gather information (transfers an array from each of the threads to a single array in the master node). Figure 17 illustrates the ideas of broadcasting, scattering, and gathering of information among workers.
Next, we define the portion of state space that each thread must execute, according to the thread id, \( tid \):

\[
\text{int loop_min} = \lfloor \text{tid} \times \lceil \frac{\text{nx} \times \text{ne}}{\text{nthreads}} \rceil \rfloor
\]

\[
\text{int loop_max} = \lfloor (\text{tid} + 1) \times \lceil \frac{\text{nx} \times \text{ne}}{\text{nthreads}} \rceil \rfloor
\]

The `MPI_Bcast` function instructs the master thread to send copies of the same information to every worker in the parallel pool. In our model, at every iteration \( t \), we need the master thread to send a copy of the value function at \( t + 1 \). For this reason, we must employ the `MPI_Bcast` function at the beginning of the loop:

\[
\text{MPI_Bcast} \left( \text{Value}, \text{T} \times \text{ne} \times \text{nx}, \text{MPI_FLOAT}, 0, \text{MPI_COMM_WORLD} \right);
\]

Given that every thread will execute the same task, each one for a different portion of the state space, we must make this explicit on the `for`:
• Invoked with a compiler wrapper

```bash
mpic++ -o ClassMPI ClassMPI.cpp
```

• Plenty of libraries (PLAPACK, Boost.MPI).

• Parallel I/O features.
```cpp
#include "mpi.h"
#include <iostream>

int main( int argc, char *argv[] ){
    int rank, size;
    MPI::Init(argc, argv);
    rank = MPI::COMM_WORLD.Get_rank();
    size = MPI::COMM_WORLD.Get_size();
    std::cout << "I am " << rank << " of " << size << std::endl;
    MPI::Finalize();
    return 0;
}
```
Routines

• Communication:
  1. Send and receive: between two processors.
  2. Broadcast, scatter, and gather data on all processors.
  3. Compute and move (sum, product, max of, ...) data on many processors.

• Synchronization.

• Enquiries:
  1. How many processes?
  2. Which process is this one?
  3. Are all messages here?
• MPI predefines its primitive data types:
  1. MPI_CHAR
  2. MPI_DOUBLE_PRECISION
  3. MPI_C_DOUBLE_COMPLEX

• Also for structs and vectors.

• Particularly important for top performance.
Parallelization in C++ using MPI - Performance

![Graph showing the computing time (s) vs. the number of cores.](image)

Physical cores

Figure 11: Computing time (s)
GPUs
Big difference

- Latency: amount of time required to complete a unit of work.

- Throughput: amount of work completed per unit of time.

- Latency devices: CPU cores.

- Throughput devices: GPU cores.

- Intermediate: Coprocessors.

- Nature of your application?
GPU
Throughput Oriented Cores

Chip

Compute Unit

Cache/Local Mem

Registers

SIMD Unit

Threading
Floating-point operations per second for the CPU and GPU
When to go to the GPU?

1. Problem is easily scalable because computation is massively parallel.

2. Much more time spent on computation than on communication.

Remember: a GPU is attached to the CPU via a PCI (Peripheral Component Interconnect) Express bus.
CUDA, OpenCL, and OpenACC

- Three approaches to code in the GPU:
  
  1. **CUDA** (Compute Unified Device Architecture):
  
  2. **OpenCL** (Open Computing Language).
  
  3. **OpenACC**.

- Also, using some of the packages/libraries of languages such as R or Matlab.
• Tapping the supercomputer under your desk: Solving dynamic equilibrium models with graphics processors.


• *Heterogeneous Computing with OpenCL 2.0* by David R. Kaeli and Perhaad Mistry.

• *OpenACC for Programmers: Concepts and Strategies* by Sunita Chandrasekaran and Guido Juckeland.
CUDA
CUDA (ComputeUnifiedDeviceArchitecture) was created by Nvidia to facilitate GPU programming.

- It is based on C/C++ with a set of extensions to enable heterogenous programming.

- Introduced in 2007, it is being actively developed (current version 10).

- Many toolboxes: cuBlas, curand, cuSparse, thrust.

- It can be access from other languages such as Fortran, Matlab, Python.

- Widely used in data mining, computer vision, medical imaging, bio-informatics.
CUDA: advantages

- Massive acceleration for parallelizable problems.
- Brings C++11, since version 7, and C++14, since version 9, language features, albeit only a subset available in Device.
- Fast shared memory that can be accessed by threads.
- Rapidly expanding third-party libraries: OpenCV machine learning, CULA linear algebra, HIPLAR linear algebra for R.
- Enter Thrust:
  1. Library of parallel algorithms and data structures.
  2. Flexible, high-level interface for GPU programming.
  3. A few lines of code to perform GPU-accelerated sort, scan, transform, and reduction operations.
// Functions to be executed only from GPU
__device__ float utility(float consumption, float ssigma){
    float utility = pow(consumption, 1-ssigma) / (1-ssigma);
    // ...
    return utility;
}

// Functions to be executed from CPU and GPU
__global__ float value(parameters params, float* V, ...){
    // ...
}
CUDA: disadvantages

- Runs only in Nvidia devices.

- High startup cost. Tricky to program even for experienced programmers.

- Tracking host and device codes.

- Demands knowledge of architecture: grid, blocks, threats. Memory management.

- Copying between host and device may reduce speed gains.

- Not all applications benefit from parallelization.

- Limited community, most information comes from Nvidia and third-party developers.
Some additional books and references for CUDA programming.

• Books:

1. *CUDA by Example*, by Jason Sanders and Edward Kandrot.

• References:

Thrust
Thrust brings the power of GPUs to the masses (at least those familiar with C++).

Thrust is a parallel algorithms library in the spirit of C++’s Standard Template Library.

Thrust’s main goal is to solve problems that

1. “can be implemented efficiently without a detailed mapping to the target architecture,” and
2. “don’t merit or won’t receive (for whatever reason) significant optimization attention from the programmer.”

The idea is that the programmer spends more time on the problem, rather than on the implementation of the algorithms solving the problem.
• Low-level customization and easy interaction with CUDA, OpenMP, or TBB.\textsuperscript{1}

• Thrust has two main features.

  1. An STL-style vector container for host and device, and
  2. A set of high-level algorithms for copying, merging, sorting, transforming.

• Thrust can be used for parallel computing for multicore CPUs.

• Thrust incorporates tuned implementation for each backend: CUDA, OpenMP, and TBB.

• This results in portability across parallel frameworks and hardware architecture without losing performance.

\textsuperscript{1}Intel’s TBB – Threading Building Blocks – is a C++ template library for task parallelism.
• Of course, Thrust has limitations.

• No multidimensional data structures libraries.

• Thrust is “entirely defined in header files.” Hence, each modification in code requires recompilation.

• Thrust is not for situations in which performance, customization are crucial.

• Documentation is limited and mostly based on examples. But it has improved over the years.

• Although the last release, version 1.8.1, dates back to 2015, Nvidia seems to be working on an update.²

²https://www.reddit.com/r/cpp/comments/7erub1/anybody_still_using_thrust/
```c++
#include <thrust/host_vector.h>
#include <thrust/device_vector.h>
#include <thrust/generate.h>
#include <thrust/reduce.h>
#include <thrust/functional.h>
#include <algorithm>
#include <cstdlib>

int main(void){
    // generate random data serially
    thrust::host_vector<int> h_vec(100);
    std::generate(h_vec.begin(), h_vec.end(), rand);
    // transfer to device and compute sum
    thrust::device_vector<int> d_vec = h_vec;
    int x = thrust::reduce(d_vec.begin(), d_vec.end(), 0,
                           thrust::plus<int__() );
    return 0;
}
```
Thrust IV

- Let's take a more detailed peek at some of Thrust's capabilities.

- Thrust provides two vector containers:

  1. **host_vector** stored in the CPU's memory

     ```cpp
     thrust::host_vector<int> hexample(10,1) // host vector with 10 elements set to 1
     ```

  2. **device_vector** resides in the GPU's device memory.

     ```cpp
     thrust::device_vector<int> dexample(hexample.begin(),hexample.begin()+5) // device vector with first 5 elements of hexample
     ```

- Some algorithms that operate on vectors are: **thrust::fill()**, **thrust::copy()**, **thrust::sequence()**.

- Last algorithm creates a sequence of equally spaced values.
Transformations are “algorithms that apply an operation to each element in a set of input ranges and stores result in destination range.”

- Compute $Y = -X$:
  
  ```cpp
  thrust::transform(X.begin(), X.end(), Y.begin(), thrust::negate<int>());
  ```

- Compute $Y = X \mod 2$
  
  ```cpp
  thrust::transform(X.begin(), X.end(), Z.begin(), Y.begin(), thrust::modulus<int>());
  ```
Thrust V – Algorithms

- **Reduction** uses “a binary operation to reduce an input sequence to a single value.”

- Sum elements in device vector \( Y \):

  ```cpp
  int sum = thrust::reduce(Y.begin(), Y.end(), (int) 0, thrust::plus<int>());
  ```

- Thrust includes other reduction operations:
  1. **thrust::count** number of instances of specific value,
  2. **thrust::min_element** // find minimum in vector,
  3. **thrust::max_element**, 
  4. **thrust::inner_product**, 

Thrust offers many more algorithms to, for example, reordering, sorting, and prefix-sums.

Another important feature in thrust is the **fancy iterators**.

1. `thrust::constant_iterator` iterator returns same value when dereference it,
2. `transform_iterator`,
3. `permutation_iterator` fuse, gather, and scatter operations with thrust algorithms,
4. `zip_iterator` takes multiple input sequences and yields a sequence of tuples.

`zip_iterator` can be used to create “a virtual array of 3d vectors” that can be fed to other algorithms.
Additional resources

Some books and references for thrust programming.

- Books:
  1. Sorry! No books that we are aware of.

- References:
OpenACC
OpenACC

• Like Thrust, OpenACC tries to bring heterogenous HPC to the masses.

• Its motto is “More Science, Less Programming.”

• OpenACC is a “user-driven directive-based performance-portable parallel programming model.”

• Main idea is to take existing serial code, say C++, and give hints to compiler to what should be parallelized.

• OpenACC is a model designed to allow parallel programming across different computer architectures with minimum effort by the developer. Portability means that the code should be independent of hardware/compiler.

• OpenACC specification supports C/C++ and Fortran and runs in CPUs and GPUs.
OpenACC II

- OpenACC is built around a very simple set of directives, very similar in design to OpenMP: OpenACC uses the fork-join paradigm.

- The same program can be compiled to be executed in parallel using the CPU or the GPU (or mixing them), depending on the hardware available.

- Communication between the master and worker threads in the parallel pool is automatically handled, although the user can state directives to grant explicit access to variables, and to transfer objects from the CPU to the GPU when required.

- The OpenACC website describes multiple compilers, profilers, and debuggers for OpenACC.

- We use the PGI Community Edition compiler. The PGI compiler can be used with the CPUs and with NVIDIA Tesla GPUs. In this way, it suffices to select different flags at compilation time to execute the code in the CPU or the GPU.
Using OpenACC I: Analyze

- Use a profiler to check where your code spends lots of time. Example of bottlenecks are loops.

- Check if there is an optimized library that implements some of your code: cuBlas, Armadillo.
Using OpenACC II: Parallelize

- Expose your code to parallelism starting with functions/operations that are time consuming on CPU.

- To initiate parallel execution:
  ```
  # pragma acc parallel
  ```

- To execute a kernel:
  ```
  # pragma acc parallel kernel
  ```

- To parrellelize a loop:
  ```
  # pragma acc parallel for
  ```
Example code of parallelizing a loop in C++.

Basic parallel loop:

```cpp
#pragma acc parallel loop
for(int ix=0; ix<nx; ix++){
    // ...
}
```

Ensuring copy of relevant objects:

```cpp
#pragma acc data copy(....)
#pragma acc parallel loop
for(int ix = 0; ix<nx; ix++){
    //...
}
```
Example code II

- By choosing the appropriate compilation flag, we can compile the code to be executed in parallel only in the CPU or in the GPU.

- To compile the code to be executed by the CPU, we must include the \(-\text{ta=}\text{multicore}\) flag at compilation. In addition, the \(-\text{acc}\) flag must be added to tell the compiler this is OpenACC code:

  \[
  \text{pgc++ Cpp\_main\_OpenACC.cpp -o Cpp\_main -acc -ta=multicore}
  \]

- If, instead, we want to execute the program with an NVIDIA GPU, we rely on the \(-\text{ta=}\text{nvidia}\) flag:

  \[
  \text{pgc++ Cpp\_main\_OpenACC.cpp -o Cpp\_main -acc -ta=nvidia}
  \]
Using OpenACC III: Optimize

- Give info to compiler of parts that can be optimize: data management (minimize copying between host and device).

- Instruct compiler how to parallelize loops.

- Step 3 is not trivial and maybe involved, limiting OpenACC’s applicability.
Some books and references for OpenACC programming.

- Books:

- References:
Comparisons
• All results are specific to our life-cycle model example.

• There are other ways to improve speed on each language:
  
  • Function wrapping in Julia.
  
  • Vectorizing in Matlab.
  
  • Etc.
The comparisons regarding parallelization are specific to the packages used on these slides:

<table>
<thead>
<tr>
<th></th>
<th>Community</th>
<th>Speed</th>
<th>Parallelization</th>
<th>Time to program</th>
<th>Debug</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matlab</td>
<td>Large</td>
<td>Medium</td>
<td>Easy</td>
<td>Low</td>
<td>Fast</td>
</tr>
<tr>
<td>Julia</td>
<td>Very small</td>
<td>Fast</td>
<td>Medium</td>
<td>Medium</td>
<td>Fast</td>
</tr>
<tr>
<td>R</td>
<td>Large</td>
<td>Slow</td>
<td>Medium</td>
<td>Medium</td>
<td>Fast</td>
</tr>
<tr>
<td>Python</td>
<td>Large</td>
<td>Slow</td>
<td>Medium</td>
<td>Medium</td>
<td>Fast</td>
</tr>
<tr>
<td>C++</td>
<td>Large</td>
<td>Fast</td>
<td>Easy</td>
<td>High</td>
<td>Slow</td>
</tr>
</tbody>
</table>
Advice

- **Short-run:**
  - MATLAB, Python, Julia, or R

- **Medium-run:**
  - Rcpp

- **Long-run:**
  - C++ with OpenMP, MPI, or GPUs
The graph shows the time (in seconds) taken by various parallel computing frameworks as the number of cores increases. The frameworks include:

- **OpenMP**
- **MPI**
- **Julia Par**
- **Rcpp**
- **OpenAcc**
- **CUDA**

The x-axis represents the number of cores, ranging from 1 to 8. The y-axis represents time, ranging from 0 to 10 seconds. The physical cores are indicated by a red dashed line, showing the optimal number of cores for each framework. The graph illustrates how effectively each framework scales with an increase in the number of cores.