Multi-processing and Distributed Computing

An implementation of distributed memory parallel computing is provided by module `Distributed` as part of the standard library shipped with Julia.

Most modern computers possess more than one CPU, and several computers can be combined together in a cluster. Harnessing the power of these multiple CPUs allows many computations to be completed more quickly. There are two major factors that influence performance: the speed of the CPUs themselves, and the speed of their access to memory. In a cluster, it's fairly obvious that a given CPU will have fastest access to the RAM within the same computer (node). Perhaps more surprisingly, similar issues are relevant on a typical multicore laptop, due to differences in the speed of main memory and the cache. Consequently, a good multiprocessing environment should allow control over the "ownership" of a chunk of memory by a particular CPU. Julia provides a multiprocessing environment based on message passing to allow programs to run on multiple processes in separate memory domains at once.

Julia's implementation of message passing is different from other environments such as MPI[1]. Communication in Julia is generally "one-sided", meaning that the programmer needs to explicitly manage only one process in a two-process operation. Furthermore, these operations typically do not look like "message send" and "message receive" but rather resemble higher-level operations like calls to user functions.

Distributed programming in Julia is built on two primitives: *remote references* and *remote calls*. A remote reference is an object that can be used from any process to refer to an object stored on a particular process. A remote call is a request by one process to call a certain function on certain arguments on another (possibly the same) process.

Remote references come in two flavors: `Future` and `RemoteChannel`.

A remote call returns a `Future` to its result. Remote calls return immediately; the process that made the call proceeds to its next operation while the remote call happens somewhere else. You can wait for a remote call to finish by calling `wait` on the returned `Future`, and you can obtain the full value of the result using `fetch`.

On the other hand, `RemoteChannel` s are rewritable. For example, multiple processes can co-ordinate their processing by referencing the same remote `Channel`.

Each process has an associated identifier. The process providing the interactive Julia prompt always has

an `id` equal to 1. The processes used by default for parallel operations are referred to as "workers". When there is only one process, process 1 is considered a worker. Otherwise, workers are considered to be all processes other than process 1. As a result, adding 2 or more processes is required to gain benefits from parallel processing methods like `pmap`. Adding a single process is beneficial if you just wish to do other things in the main process while a long computation is running on the worker.

Let's try this out. Starting with `julia -p n` provides `n` worker processes on the local machine. Generally it makes sense for `n` to equal the number of CPU threads (logical cores) on the machine. Note that the `-p` argument implicitly loads module `Distributed`.

```
$ ./julia -p 2

julia> r = remotecall(rand, 2, 2, 2)
Future(2, 1, 4, nothing)

julia> s = @spawnat 2 1 .+ fetch(r)
Future(2, 1, 5, nothing)

julia> fetch(s)
2×2 Array{Float64,2}:
 1.18526  1.50912
 1.16296  1.60607
```

The first argument to `remotecall` is the function to call. Most parallel programming in Julia does not reference specific processes or the number of processes available, but `remotecall` is considered a low-level interface providing finer control. The second argument to `remotecall` is the `id` of the process that will do the work, and the remaining arguments will be passed to the function being called.

As you can see, in the first line we asked process 2 to construct a 2-by-2 random matrix, and in the second line we asked it to add 1 to it. The result of both calculations is available in the two futures, `r` and `s`. The `@spawnat` macro evaluates the expression in the second argument on the process specified by the first argument.

Occasionally you might want a remotely-computed value immediately. This typically happens when you read from a remote object to obtain data needed by the next local operation. The function `remotecall_fetch` exists for this purpose. It is equivalent to `fetch(remotecall(...))` but is more efficient.

```
julia> remotecall_fetch(getindex, 2, r, 1, 1)
0.18526337335308085
```

Remember that `getindex(r,1,1)` is equivalent to `r[1,1]`, so this call fetches the first element of the

future `r`.

To make things easier, the symbol `:any` can be passed to `[@spawnat]`, which picks where to do the operation for you:

```
julia> r = @spawnat :any rand(2,2)
Future(2, 1, 4, nothing)

julia> s = @spawnat :any 1 .+ fetch(r)
Future(3, 1, 5, nothing)

julia> fetch(s)
2×2 Array{Float64,2}:
 1.38854  1.9098
 1.20939  1.57158
```

Note that we used `1 .+ fetch(r)` instead of `1 .+ r`. This is because we do not know where the code will run, so in general a `fetch` might be required to move `r` to the process doing the addition. In this case, `@spawnat` is smart enough to perform the computation on the process that owns `r`, so the `fetch` will be a no-op (no work is done).

(It is worth noting that `@spawnat` is not built-in but defined in Julia as a macro. It is possible to define your own such constructs.)

An important thing to remember is that, once fetched, a `Future` will cache its value locally. Further `fetch` calls do not entail a network hop. Once all referencing `Future`s have fetched, the remote stored value is deleted.

`@async` is similar to `@spawnat`, but only runs tasks on the local process. We use it to create a "feeder" task for each process. Each task picks the next index that needs to be computed, then waits for its process to finish, then repeats until we run out of indices. Note that the feeder tasks do not begin to execute until the main task reaches the end of the `@sync` block, at which point it surrenders control and waits for all the local tasks to complete before returning from the function. As for v0.7 and beyond, the feeder tasks are able to share state via `nextidx` because they all run on the same process. Even if `Tasks` are scheduled cooperatively, locking may still be required in some contexts, as in asynchronous I/O. This means context switches only occur at well-defined points: in this case, when `remotecall_fetch` is called. This is the current state of implementation and it may change for future Julia versions, as it is intended to make it possible to run up to N `Tasks` on M `Process`, aka M:N Threading. Then a lock acquiring\releasing model for `nextidx` will be needed, as it is not safe to let multiple processes read-write a resource at the same time.

Code Availability and Loading Packages

Your code must be available on any process that runs it. For example, type the following into the Julia prompt:

```
julia> function rand2(dims...)
           return 2*rand(dims...)
       end

julia> rand2(2,2)
2×2 Array{Float64,2}:
 0.153756  0.368514
 1.15119   0.918912

julia> fetch(@spawnat :any rand2(2,2))
ERROR: RemoteException(2, CapturedException(UndefVarError(Symbol("#rand2"))
Stacktrace:
[...]
```

Process 1 knew about the function `rand2`, but process 2 did not.

Most commonly you'll be loading code from files or packages, and you have a considerable amount of flexibility in controlling which processes load code. Consider a file, `DummyModule.jl`, containing the following code:

```
module DummyModule

export MyType, f

mutable struct MyType
    a::Int
end

f(x) = x^2+1

println("loaded")

end
```

In order to refer to `MyType` across all processes, `DummyModule.jl` needs to be loaded on every process. Calling `include("DummyModule.jl")` loads it only on a single process. To load it on every process, use the `@everywhere` macro (starting Julia with `julia -p 2`):

```
julia> @everywhere include("DummyModule.jl")
loaded
      From worker 3:    loaded
      From worker 2:    loaded
```

As usual, this does not bring `DummyModule` into scope on any of the process, which requires `using` or `import`. Moreover, when `DummyModule` is brought into scope on one process, it is not on any other:

```
julia> using .DummyModule

julia> MyType(7)
MyType(7)

julia> fetch(@spawnat 2 MyType(7))
ERROR: On worker 2:
UndefVarError: MyType not defined
⋮

julia> fetch(@spawnat 2 DummyModule.MyType(7))
MyType(7)
```

However, it's still possible, for instance, to send a `MyType` to a process which has loaded `DummyModule` even if it's not in scope:

```
julia> put!(RemoteChannel(2), MyType(7))
RemoteChannel{Channel{Any}}(2, 1, 13)
```

A file can also be preloaded on multiple processes at startup with the `-L` flag, and a driver script can be used to drive the computation:

```
julia -p <n> -L file1.jl -L file2.jl driver.jl
```

The Julia process running the driver script in the example above has an `id` equal to 1, just like a process providing an interactive prompt.

Finally, if `DummyModule.jl` is not a standalone file but a package, then `using DummyModule` will *load* `DummyModule.jl` on all processes, but only bring it into scope on the process where `using` was called.

Starting and managing worker processes

The base Julia installation has in-built support for two types of clusters:

- A local cluster specified with the `-p` option as shown above.
- A cluster spanning machines using the `--machine-file` option. This uses a passwordless `ssh` login to start Julia worker processes (from the same path as the current host) on the specified machines.

Functions `addprocs`, `rmprocs`, `workers`, and others are available as a programmatic means of adding, removing and querying the processes in a cluster.

```
julia> using Distributed

julia> addprocs(2)
2-element Array{Int64,1}:
 2
 3
```

Module `Distributed` must be explicitly loaded on the master process before invoking `addprocs`. It is automatically made available on the worker processes.

Note that workers do not run a `~/.julia/config/startup.jl` startup script, nor do they synchronize their global state (such as global variables, new method definitions, and loaded modules) with any of the other running processes. You may use `addprocs(exeflags="--project")` to initialize a worker with a particular environment, and then `@everywhere using <modulename>` or `@everywhere include("file.jl")`.

Other types of clusters can be supported by writing your own custom `ClusterManager`, as described below in the ClusterManagers section.

Data Movement

Sending messages and moving data constitute most of the overhead in a distributed program. Reducing the number of messages and the amount of data sent is critical to achieving performance and scalability. To this end, it is important to understand the data movement performed by Julia's various distributed programming constructs.

`fetch` can be considered an explicit data movement operation, since it directly asks that an object be moved to the local machine. `@spawnat` (and a few related constructs) also moves data, but this is not as obvious, hence it can be called an implicit data movement operation. Consider these two approaches to constructing and squaring a random matrix:

Method 1:

```
julia> A = rand(1000,1000);
```

```
julia> Bref = @spawnat :any A^2;

[...]

julia> fetch(Bref);
```

Method 2:

```
julia> Bref = @spawnat :any rand(1000,1000)^2;

[...]

julia> fetch(Bref);
```

The difference seems trivial, but in fact is quite significant due to the behavior of `@spawnat`. In the first method, a random matrix is constructed locally, then sent to another process where it is squared. In the second method, a random matrix is both constructed and squared on another process. Therefore the second method sends much less data than the first.

In this toy example, the two methods are easy to distinguish and choose from. However, in a real program designing data movement might require more thought and likely some measurement. For example, if the first process needs matrix `A` then the first method might be better. Or, if computing `A` is expensive and only the current process has it, then moving it to another process might be unavoidable. Or, if the current process has very little to do between the `@spawnat` and `fetch(Bref)`, it might be better to eliminate the parallelism altogether. Or imagine `rand(1000,1000)` is replaced with a more expensive operation. Then it might make sense to add another `@spawnat` statement just for this step.

Global variables

Expressions executed remotely via `@spawnat`, or closures specified for remote execution using `remotecall` may refer to global variables. Global bindings under module `Main` are treated a little differently compared to global bindings in other modules. Consider the following code snippet:

```
A = rand(10,10)
remotecall_fetch(()->sum(A), 2)
```

In this case `sum` MUST be defined in the remote process. Note that `A` is a global variable defined in the local workspace. Worker 2 does not have a variable called `A` under `Main`. The act of shipping the closure `()->sum(A)` to worker 2 results in `Main.A` being defined on 2. `Main.A` continues to exist on worker 2 even after the call `remotecall_fetch` returns. Remote calls with embedded global references (under

`Main` module only) manage globals as follows:

- New global bindings are created on destination workers if they are referenced as part of a remote call.
- Global constants are declared as constants on remote nodes too.
- Globals are re-sent to a destination worker only in the context of a remote call, and then only if its value has changed. Also, the cluster does not synchronize global bindings across nodes. For example:

```
A = rand(10,10)
remotecall_fetch(()->sum(A), 2) # worker 2
A = rand(10,10)
remotecall_fetch(()->sum(A), 3) # worker 3
A = nothing
```

  Executing the above snippet results in `Main.A` on worker 2 having a different value from `Main.A` on worker 3, while the value of `Main.A` on node 1 is set to `nothing`.

As you may have realized, while memory associated with globals may be collected when they are reassigned on the master, no such action is taken on the workers as the bindings continue to be valid. `clear!` can be used to manually reassign specific globals on remote nodes to `nothing` once they are no longer required. This will release any memory associated with them as part of a regular garbage collection cycle.

Thus programs should be careful referencing globals in remote calls. In fact, it is preferable to avoid them altogether if possible. If you must reference globals, consider using `let` blocks to localize global variables.

For example:

```
julia> A = rand(10,10);

julia> remotecall_fetch(()->A, 2);

julia> B = rand(10,10);

julia> let B = B
           remotecall_fetch(()->B, 2)
       end;

julia> @fetchfrom 2 InteractiveUtils.varinfo()
name           size summary
```

```
--------- --------- --------------------
A         800 bytes 10×10 Array{Float64,2}
Base                Module
Core                Module
Main                Module
```

As can be seen, global variable `A` is defined on worker 2, but `B` is captured as a local variable and hence a binding for `B` does not exist on worker 2.

Parallel Map and Loops

Fortunately, many useful parallel computations do not require data movement. A common example is a Monte Carlo simulation, where multiple processes can handle independent simulation trials simultaneously. We can use `@spawnat` to flip coins on two processes. First, write the following function in `count_heads.jl`:

```
function count_heads(n)
    c::Int = 0
    for i = 1:n
        c += rand(Bool)
    end
    c
end
```

The function `count_heads` simply adds together `n` random bits. Here is how we can perform some trials on two machines, and add together the results:

```
julia> @everywhere include_string(Main, $(read("count_heads.jl", String)), "count_he

julia> a = @spawnat :any count_heads(100000000)
Future(2, 1, 6, nothing)

julia> b = @spawnat :any count_heads(100000000)
Future(3, 1, 7, nothing)

julia> fetch(a)+fetch(b)
100001564
```

This example demonstrates a powerful and often-used parallel programming pattern. Many iterations run independently over several processes, and then their results are combined using some function. The combination process is called a *reduction*, since it is generally tensor-rank-reducing: a vector of

numbers is reduced to a single number, or a matrix is reduced to a single row or column, etc. In code, this typically looks like the pattern `x = f(x,v[i])`, where `x` is the accumulator, `f` is the reduction function, and the `v[i]` are the elements being reduced. It is desirable for `f` to be associative, so that it does not matter what order the operations are performed in.

Notice that our use of this pattern with `count_heads` can be generalized. We used two explicit `@spawnat` statements, which limits the parallelism to two processes. To run on any number of processes, we can use a *parallel for loop*, running in distributed memory, which can be written in Julia using `@distributed` like this:

```
nheads = @distributed (+) for i = 1:200000000
    Int(rand(Bool))
end
```

This construct implements the pattern of assigning iterations to multiple processes, and combining them with a specified reduction (in this case `(+)`). The result of each iteration is taken as the value of the last expression inside the loop. The whole parallel loop expression itself evaluates to the final answer.

Note that although parallel for loops look like serial for loops, their behavior is dramatically different. In particular, the iterations do not happen in a specified order, and writes to variables or arrays will not be globally visible since iterations run on different processes. Any variables used inside the parallel loop will be copied and broadcast to each process.

For example, the following code will not work as intended:

```
a = zeros(100000)
@distributed for i = 1:100000
    a[i] = i
end
```

This code will not initialize all of `a`, since each process will have a separate copy of it. Parallel for loops like these must be avoided. Fortunately, Shared Arrays can be used to get around this limitation:

```
using SharedArrays

a = SharedArray{Float64}(10)
@distributed for i = 1:10
    a[i] = i
end
```

Using "outside" variables in parallel loops is perfectly reasonable if the variables are read-only:

```
a = randn(1000)
@distributed (+) for i = 1:100000
    f(a[rand(1:end)])
end
```

Here each iteration applies `f` to a randomly-chosen sample from a vector `a` shared by all processes.

As you could see, the reduction operator can be omitted if it is not needed. In that case, the loop executes asynchronously, i.e. it spawns independent tasks on all available workers and returns an array of `Future` immediately without waiting for completion. The caller can wait for the `Future` completions at a later point by calling `fetch` on them, or wait for completion at the end of the loop by prefixing it with `@sync`, like `@sync @distributed for`.

In some cases no reduction operator is needed, and we merely wish to apply a function to all integers in some range (or, more generally, to all elements in some collection). This is another useful operation called *parallel map*, implemented in Julia as the `pmap` function. For example, we could compute the singular values of several large random matrices in parallel as follows:

```
julia> M = Matrix{Float64}[rand(1000,1000) for i = 1:10];

julia> pmap(svdvals, M);
```

Julia's `pmap` is designed for the case where each function call does a large amount of work. In contrast, `@distributed for` can handle situations where each iteration is tiny, perhaps merely summing two numbers. Only worker processes are used by both `pmap` and `@distributed for` for the parallel computation. In case of `@distributed for`, the final reduction is done on the calling process.

Remote References and AbstractChannels

Remote references always refer to an implementation of an `AbstractChannel`.

A concrete implementation of an `AbstractChannel` (like `Channel`), is required to implement `put!`, `take!`, `fetch`, `isready` and `wait`. The remote object referred to by a `Future` is stored in a `Channel{Any}(1)`, i.e., a `Channel` of size 1 capable of holding objects of `Any` type.

`RemoteChannel`, which is rewritable, can point to any type and size of channels, or any other implementation of an `AbstractChannel`.

The constructor `RemoteChannel(f::Function, pid)()` allows us to construct references to channels holding more than one value of a specific type. `f` is a function executed on `pid` and it must return an `AbstractChannel`.

For example, `RemoteChannel(()->Channel{Int}(10), pid)`, will return a reference to a channel of type `Int` and size 10. The channel exists on worker `pid`.

Methods `put!`, `take!`, `fetch`, `isready` and `wait` on a `RemoteChannel` are proxied onto the backing store on the remote process.

`RemoteChannel` can thus be used to refer to user implemented `AbstractChannel` objects. A simple example of this is provided in `dictchannel.jl` in the Examples repository, which uses a dictionary as its remote store.

Channels and RemoteChannels

- A `Channel` is local to a process. Worker 2 cannot directly refer to a `Channel` on worker 3 and vice-versa. A `RemoteChannel`, however, can put and take values across workers.
- A `RemoteChannel` can be thought of as a *handle* to a `Channel`.
- The process id, `pid`, associated with a `RemoteChannel` identifies the process where the backing store, i.e., the backing `Channel` exists.
- Any process with a reference to a `RemoteChannel` can put and take items from the channel. Data is automatically sent to (or retrieved from) the process a `RemoteChannel` is associated with.
- Serializing a `Channel` also serializes any data present in the channel. Deserializing it therefore effectively makes a copy of the original object.
- On the other hand, serializing a `RemoteChannel` only involves the serialization of an identifier that identifies the location and instance of `Channel` referred to by the handle. A deserialized `RemoteChannel` object (on any worker), therefore also points to the same backing store as the original.

The channels example from above can be modified for interprocess communication, as shown below.

We start 4 workers to process a single `jobs` remote channel. Jobs, identified by an id (`job_id`), are written to the channel. Each remotely executing task in this simulation reads a `job_id`, waits for a random amount of time and writes back a tuple of `job_id`, time taken and its own `pid` to the results channel. Finally all the `results` are printed out on the master process.

```
julia> addprocs(4); # add worker processes

julia> const jobs = RemoteChannel(()->Channel{Int}(32));

julia> const results = RemoteChannel(()->Channel{Tuple}(32));

julia> @everywhere function do_work(jobs, results) # define work function everywhere
           while true
```

```
            job_id = take!(jobs)
            exec_time = rand()
            sleep(exec_time) # simulates elapsed time doing actual work
            put!(results, (job_id, exec_time, myid()))
        end
    end

julia> function make_jobs(n)
        for i in 1:n
            put!(jobs, i)
        end
    end;

julia> n = 12;

julia> @async make_jobs(n); # feed the jobs channel with "n" jobs

julia> for p in workers() # start tasks on the workers to process requests in paral]
        remote_do(do_work, p, jobs, results)
    end

julia> @elapsed while n > 0 # print out results
        job_id, exec_time, where = take!(results)
        println("$job_id finished in $(round(exec_time; digits=2)) seconds on wor
        global n = n - 1
    end
1 finished in 0.18 seconds on worker 4
2 finished in 0.26 seconds on worker 5
6 finished in 0.12 seconds on worker 4
7 finished in 0.18 seconds on worker 4
5 finished in 0.35 seconds on worker 5
4 finished in 0.68 seconds on worker 2
3 finished in 0.73 seconds on worker 3
11 finished in 0.01 seconds on worker 3
12 finished in 0.02 seconds on worker 3
9 finished in 0.26 seconds on worker 5
8 finished in 0.57 seconds on worker 4
10 finished in 0.58 seconds on worker 2
0.055971741
```

Remote References and Distributed Garbage Collection

Objects referred to by remote references can be freed only when *all* held references in the cluster are deleted.

The node where the value is stored keeps track of which of the workers have a reference to it. Every time a `RemoteChannel` or a (unfetched) `Future` is serialized to a worker, the node pointed to by the reference is notified. And every time a `RemoteChannel` or a (unfetched) `Future` is garbage collected locally, the node owning the value is again notified. This is implemented in an internal cluster aware serializer. Remote references are only valid in the context of a running cluster. Serializing and deserializing references to and from regular `IO` objects is not supported.

The notifications are done via sending of "tracking" messages–an "add reference" message when a reference is serialized to a different process and a "delete reference" message when a reference is locally garbage collected.

Since `Future`s are write-once and cached locally, the act of `fetch`ing a `Future` also updates reference tracking information on the node owning the value.

The node which owns the value frees it once all references to it are cleared.

With `Future`s, serializing an already fetched `Future` to a different node also sends the value since the original remote store may have collected the value by this time.

It is important to note that *when* an object is locally garbage collected depends on the size of the object and the current memory pressure in the system.

In case of remote references, the size of the local reference object is quite small, while the value stored on the remote node may be quite large. Since the local object may not be collected immediately, it is a good practice to explicitly call `finalize` on local instances of a `RemoteChannel`, or on unfetched `Future`s. Since calling `fetch` on a `Future` also removes its reference from the remote store, this is not required on fetched `Future`s. Explicitly calling `finalize` results in an immediate message sent to the remote node to go ahead and remove its reference to the value.

Once finalized, a reference becomes invalid and cannot be used in any further calls.

Local invocations

Data is necessarily copied over to the remote node for execution. This is the case for both remotecalls and when data is stored to a `RemoteChannel` / `Future` on a different node. As expected, this results in a copy of the serialized objects on the remote node. However, when the destination node is the local node, i.e. the calling process id is the same as the remote node id, it is executed as a local call. It is usually(not always) executed in a different task - but there is no serialization/deserialization of data. Consequently, the call refers to the same object instances as passed - no copies are created. This behavior is highlighted below:

```
julia> using Distributed;

julia> rc = RemoteChannel(()->Channel(3));   # RemoteChannel created on local node

julia> v = [0];

julia> for i in 1:3
           v[1] = i                          # Reusing `v`
           put!(rc, v)
       end;

julia> result = [take!(rc) for _ in 1:3];

julia> println(result);
Array{Int64,1}[[3], [3], [3]]

julia> println("Num Unique objects : ", length(unique(map(objectid, result))));
Num Unique objects : 1

julia> addprocs(1);

julia> rc = RemoteChannel(()->Channel(3), workers()[1]);   # RemoteChannel created c

julia> v = [0];

julia> for i in 1:3
           v[1] = i
           put!(rc, v)
       end;

julia> result = [take!(rc) for _ in 1:3];

julia> println(result);
Array{Int64,1}[[1], [2], [3]]

julia> println("Num Unique objects : ", length(unique(map(objectid, result))));
Num Unique objects : 3
```

As can be seen, `put!` on a locally owned `RemoteChannel` with the same object `v` modifed between calls results in the same single object instance stored. As opposed to copies of `v` being created when the node owning `rc` is a different node.

It is to be noted that this is generally not an issue. It is something to be factored in only if the object is both being stored locally and modifed post the call. In such cases it may be appropriate to store a

`deepcopy` of the object.

This is also true for remotecalls on the local node as seen in the following example:

```
julia> using Distributed; addprocs(1);

julia> v = [0];

julia> v2 = remotecall_fetch(x->(x[1] = 1; x), myid(), v);     # Executed on local r

julia> println("v=$v, v2=$v2, ", v === v2);
v=[1], v2=[1], true

julia> v = [0];

julia> v2 = remotecall_fetch(x->(x[1] = 1; x), workers()[1], v); # Executed on remot

julia> println("v=$v, v2=$v2, ", v === v2);
v=[0], v2=[1], false
```

As can be seen once again, a remote call onto the local node behaves just like a direct invocation. The call modifies local objects passed as arguments. In the remote invocation, it operates on a copy of the arguments.

To repeat, in general this is not an issue. If the local node is also being used as a compute node, and the arguments used post the call, this behavior needs to be factored in and if required deep copies of arguments must be passed to the call invoked on the local node. Calls on remote nodes will always operate on copies of arguments.

Shared Arrays

Shared Arrays use system shared memory to map the same array across many processes. While there are some similarities to a `DArray`, the behavior of a `SharedArray` is quite different. In a `DArray`, each process has local access to just a chunk of the data, and no two processes share the same chunk; in contrast, in a `SharedArray` each "participating" process has access to the entire array. A `SharedArray` is a good choice when you want to have a large amount of data jointly accessible to two or more processes on the same machine.

Shared Array support is available via module `SharedArrays` which must be explicitly loaded on all participating workers.

`SharedArray` indexing (assignment and accessing values) works just as with regular arrays, and is

efficient because the underlying memory is available to the local process. Therefore, most algorithms work naturally on `SharedArray`s, albeit in single-process mode. In cases where an algorithm insists on an `Array` input, the underlying array can be retrieved from a `SharedArray` by calling `sdata`. For other `AbstractArray` types, `sdata` just returns the object itself, so it's safe to use `sdata` on any `Array`-type object.

The constructor for a shared array is of the form:

```
SharedArray{T,N}(dims::NTuple; init=false, pids=Int[])
```

which creates an `N`-dimensional shared array of a bits type `T` and size `dims` across the processes specified by `pids`. Unlike distributed arrays, a shared array is accessible only from those participating workers specified by the `pids` named argument (and the creating process too, if it is on the same host). Note that only elements that are `isbits` are supported in a SharedArray.

If an `init` function, of signature `initfn(S::SharedArray)`, is specified, it is called on all the participating workers. You can specify that each worker runs the `init` function on a distinct portion of the array, thereby parallelizing initialization.

Here's a brief example:

```
julia> using Distributed

julia> addprocs(3)
3-element Array{Int64,1}:
 2
 3
 4

julia> @everywhere using SharedArrays

julia> S = SharedArray{Int,2}((3,4), init = S -> S[localindices(S)] = repeat([myid()
3×4 SharedArray{Int64,2}:
 2  2  3  4
 2  3  3  4
 2  3  4  4

julia> S[3,2] = 7
7

julia> S
3×4 SharedArray{Int64,2}:
 2  2  3  4
```

```
 2  3  3  4
 2  7  4  4
```

`SharedArrays.localindices` provides disjoint one-dimensional ranges of indices, and is sometimes convenient for splitting up tasks among processes. You can, of course, divide the work any way you wish:

```
julia> S = SharedArray{Int,2}((3,4), init = S -> S[indexpids(S):length(procs(S)):ler
3×4 SharedArray{Int64,2}:
 2  2  2  2
 3  3  3  3
 4  4  4  4
```

Since all processes have access to the underlying data, you do have to be careful not to set up conflicts. For example:

```
@sync begin
    for p in procs(S)
        @async begin
            remotecall_wait(fill!, p, S, p)
        end
    end
end
```

would result in undefined behavior. Because each process fills the *entire* array with its own `pid`, whichever process is the last to execute (for any particular element of `S`) will have its `pid` retained.

As a more extended and complex example, consider running the following "kernel" in parallel:

```
q[i,j,t+1] = q[i,j,t] + u[i,j,t]
```

In this case, if we try to split up the work using a one-dimensional index, we are likely to run into trouble: if `q[i,j,t]` is near the end of the block assigned to one worker and `q[i,j,t+1]` is near the beginning of the block assigned to another, it's very likely that `q[i,j,t]` will not be ready at the time it's needed for computing `q[i,j,t+1]`. In such cases, one is better off chunking the array manually. Let's split along the second dimension. Define a function that returns the `(irange, jrange)` indices assigned to this worker:

```
julia> @everywhere function myrange(q::SharedArray)
           idx = indexpids(q)
           if idx == 0 # This worker is not assigned a piece
               return 1:0, 1:0
```

```
            end
            nchunks = length(procs(q))
            splits = [round(Int, s) for s in range(0, stop=size(q,2), length=nchunks+
            1:size(q,1), splits[idx]+1:splits[idx+1]
        end
```

Next, define the kernel:

```
julia> @everywhere function advection_chunk!(q, u, irange, jrange, trange)
            @show (irange, jrange, trange)  # display so we can see what's happening
            for t in trange, j in jrange, i in irange
                q[i,j,t+1] = q[i,j,t] + u[i,j,t]
            end
            q
        end
```

We also define a convenience wrapper for a `SharedArray` implementation

```
julia> @everywhere advection_shared_chunk!(q, u) =
            advection_chunk!(q, u, myrange(q)..., 1:size(q,3)-1)
```

Now let's compare three different versions, one that runs in a single process:

```
julia> advection_serial!(q, u) = advection_chunk!(q, u, 1:size(q,1), 1:size(q,2), 1:
```

one that uses `@distributed`:

```
julia> function advection_parallel!(q, u)
            for t = 1:size(q,3)-1
                @sync @distributed for j = 1:size(q,2)
                    for i = 1:size(q,1)
                        q[i,j,t+1]= q[i,j,t] + u[i,j,t]
                    end
                end
            end
            q
        end;
```

and one that delegates in chunks:

```
julia> function advection_shared!(q, u)
            @sync begin
```

```
            for p in procs(q)
                @async remotecall_wait(advection_shared_chunk!, p, q, u)
            end
        end
        q
    end;
```

If we create `SharedArray`s and time these functions, we get the following results (with `julia -p 4`):

```
julia> q = SharedArray{Float64,3}((500,500,500));

julia> u = SharedArray{Float64,3}((500,500,500));
```

Run the functions once to JIT-compile and `@time` them on the second run:

```
julia> @time advection_serial!(q, u);
(irange,jrange,trange) = (1:500,1:500,1:499)
 830.220 milliseconds (216 allocations: 13820 bytes)

julia> @time advection_parallel!(q, u);
   2.495 seconds      (3999 k allocations: 289 MB, 2.09% gc time)

julia> @time advection_shared!(q,u);
        From worker 2:       (irange,jrange,trange) = (1:500,1:125,1:499)
        From worker 4:       (irange,jrange,trange) = (1:500,251:375,1:499)
        From worker 3:       (irange,jrange,trange) = (1:500,126:250,1:499)
        From worker 5:       (irange,jrange,trange) = (1:500,376:500,1:499)
 238.119 milliseconds (2264 allocations: 169 KB)
```

The biggest advantage of `advection_shared!` is that it minimizes traffic among the workers, allowing each to compute for an extended time on the assigned piece.

Shared Arrays and Distributed Garbage Collection

Like remote references, shared arrays are also dependent on garbage collection on the creating node to release references from all participating workers. Code which creates many short lived shared array objects would benefit from explicitly finalizing these objects as soon as possible. This results in both memory and file handles mapping the shared segment being released sooner.

ClusterManagers

The launching, management and networking of Julia processes into a logical cluster is done via cluster managers. A `ClusterManager` is responsible for

- launching worker processes in a cluster environment
- managing events during the lifetime of each worker
- optionally, providing data transport

A Julia cluster has the following characteristics:

- The initial Julia process, also called the `master`, is special and has an `id` of 1.
- Only the `master` process can add or remove worker processes.
- All processes can directly communicate with each other.

Connections between workers (using the in-built TCP/IP transport) is established in the following manner:

- `addprocs` is called on the master process with a `ClusterManager` object.
- `addprocs` calls the appropriate `launch` method which spawns required number of worker processes on appropriate machines.
- Each worker starts listening on a free port and writes out its host and port information to `stdout`.
- The cluster manager captures the `stdout` of each worker and makes it available to the master process.
- The master process parses this information and sets up TCP/IP connections to each worker.
- Every worker is also notified of other workers in the cluster.
- Each worker connects to all workers whose `id` is less than the worker's own `id`.
- In this way a mesh network is established, wherein every worker is directly connected with every other worker.

While the default transport layer uses plain `TCPSocket`, it is possible for a Julia cluster to provide its own transport.

Julia provides two in-built cluster managers:

- `LocalManager`, used when `addprocs()` or `addprocs(np::Integer)` are called
- `SSHManager`, used when `addprocs(hostnames::Array)` is called with a list of hostnames

`LocalManager` is used to launch additional workers on the same host, thereby leveraging multi-core and multi-processor hardware.

Thus, a minimal cluster manager would need to:

- be a subtype of the abstract `ClusterManager`
- implement `launch`, a method responsible for launching new workers
- implement `manage`, which is called at various events during a worker's lifetime (for example, sending an interrupt signal)

`addprocs(manager::FooManager)` requires `FooManager` to implement:

```
function launch(manager::FooManager, params::Dict, launched::Array, c::Condition)
    [...]
end

function manage(manager::FooManager, id::Integer, config::WorkerConfig, op::Symbol)
    [...]
end
```

As an example let us see how the `LocalManager`, the manager responsible for starting workers on the same host, is implemented:

```
struct LocalManager <: ClusterManager
    np::Integer
end

function launch(manager::LocalManager, params::Dict, launched::Array, c::Condition)
    [...]
end

function manage(manager::LocalManager, id::Integer, config::WorkerConfig, op::Symbol
    [...]
end
```

The `launch` method takes the following arguments:

- `manager::ClusterManager`: the cluster manager that `addprocs` is called with
- `params::Dict`: all the keyword arguments passed to `addprocs`
- `launched::Array`: the array to append one or more `WorkerConfig` objects to
- `c::Condition`: the condition variable to be notified as and when workers are launched

The `launch` method is called asynchronously in a separate task. The termination of this task signals that all requested workers have been launched. Hence the `launch` function MUST exit as soon as all the requested workers have been launched.

Newly launched workers are connected to each other and the master process in an all-to-all manner.

Specifying the command line argument `--worker[=<cookie>]` results in the launched processes initializing themselves as workers and connections being set up via TCP/IP sockets.

All workers in a cluster share the same cookie as the master. When the cookie is unspecified, i.e, with the `--worker` option, the worker tries to read it from its standard input. `LocalManager` and `SSHManager` both pass the cookie to newly launched workers via their standard inputs.

By default a worker will listen on a free port at the address returned by a call to `getipaddr()`. A specific address to listen on may be specified by optional argument `--bind-to bind_addr[:port]`. This is useful for multi-homed hosts.

As an example of a non-TCP/IP transport, an implementation may choose to use MPI, in which case `--worker` must NOT be specified. Instead, newly launched workers should call `init_worker(cookie)` before using any of the parallel constructs.

For every worker launched, the `launch` method must add a `WorkerConfig` object (with appropriate fields initialized) to `launched`

```
mutable struct WorkerConfig
    # Common fields relevant to all cluster managers
    io::Union{IO, Nothing}
    host::Union{AbstractString, Nothing}
    port::Union{Integer, Nothing}

    # Used when launching additional workers at a host
    count::Union{Int, Symbol, Nothing}
    exename::Union{AbstractString, Cmd, Nothing}
    exeflags::Union{Cmd, Nothing}

    # External cluster managers can use this to store information at a per-worker le
    # Can be a dict if multiple fields need to be stored.
    userdata::Any

    # SSHManager / SSH tunnel connections to workers
    tunnel::Union{Bool, Nothing}
    bind_addr::Union{AbstractString, Nothing}
    sshflags::Union{Cmd, Nothing}
    max_parallel::Union{Integer, Nothing}

    # Used by Local/SSH managers
    connect_at::Any

    [...]
end
```

Most of the fields in `WorkerConfig` are used by the inbuilt managers. Custom cluster managers would typically specify only `io` or `host` / `port`:

- If `io` is specified, it is used to read host/port information. A Julia worker prints out its bind address and port at startup. This allows Julia workers to listen on any free port available instead of requiring worker ports to be configured manually.
- If `io` is not specified, `host` and `port` are used to connect.
- `count`, `exename` and `exeflags` are relevant for launching additional workers from a worker. For example, a cluster manager may launch a single worker per node, and use that to launch additional workers.
    - `count` with an integer value `n` will launch a total of `n` workers.
    - `count` with a value of `:auto` will launch as many workers as the number of CPU threads (logical cores) on that machine.
    - `exename` is the name of the `julia` executable including the full path.

    - `exeflags` should be set to the required command line arguments for new workers.
- `tunnel`, `bind_addr`, `sshflags` and `max_parallel` are used when a ssh tunnel is required to connect to the workers from the master process.
- `userdata` is provided for custom cluster managers to store their own worker-specific information.

`manage(manager::FooManager, id::Integer, config::WorkerConfig, op::Symbol)` is called at different times during the worker's lifetime with appropriate `op` values:

- with `:register`/`:deregister` when a worker is added / removed from the Julia worker pool.
- with `:interrupt` when `interrupt(workers)` is called. The `ClusterManager` should signal the appropriate worker with an interrupt signal.
- with `:finalize` for cleanup purposes.

Cluster Managers with Custom Transports

Replacing the default TCP/IP all-to-all socket connections with a custom transport layer is a little more involved. Each Julia process has as many communication tasks as the workers it is connected to. For example, consider a Julia cluster of 32 processes in an all-to-all mesh network:

- Each Julia process thus has 31 communication tasks.
- Each task handles all incoming messages from a single remote worker in a message-processing loop.
- The message-processing loop waits on an `IO` object (for example, a `TCPSocket` in the default implementation), reads an entire message, processes it and waits for the next one.
- Sending messages to a process is done directly from any Julia task–not just communication tasks–again, via the appropriate `IO` object.

Replacing the default transport requires the new implementation to set up connections to remote workers and to provide appropriate `IO` objects that the message-processing loops can wait on. The manager-specific callbacks to be implemented are:

```
connect(manager::FooManager, pid::Integer, config::WorkerConfig)
kill(manager::FooManager, pid::Int, config::WorkerConfig)
```

The default implementation (which uses TCP/IP sockets) is implemented as `connect(manager::ClusterManager, pid::Integer, config::WorkerConfig)`.

`connect` should return a pair of `IO` objects, one for reading data sent from worker `pid`, and the other to write data that needs to be sent to worker `pid`. Custom cluster managers can use an in-memory `BufferStream` as the plumbing to proxy data between the custom, possibly non-`IO` transport and Julia's in-built parallel infrastructure.

A `BufferStream` is an in-memory `IOBuffer` which behaves like an `IO`–it is a stream which can be handled asynchronously.

The folder `clustermanager/0mq` in the Examples repository contains an example of using ZeroMQ to connect Julia workers in a star topology with a 0MQ broker in the middle. Note: The Julia processes are still all *logically* connected to each other–any worker can message any other worker directly without any awareness of 0MQ being used as the transport layer.

When using custom transports:

- Julia workers must NOT be started with `--worker`. Starting with `--worker` will result in the newly launched workers defaulting to the TCP/IP socket transport implementation.
- For every incoming logical connection with a worker, `Base.process_messages(rd::IO, wr::IO)()` must be called. This launches a new task that handles reading and writing of messages from/to the worker represented by the `IO` objects.
- `init_worker(cookie, manager::FooManager)` *must* be called as part of worker process initialization.
- Field `connect_at::Any` in `WorkerConfig` can be set by the cluster manager when `launch` is called. The value of this field is passed in all `connect` callbacks. Typically, it carries information on *how to connect* to a worker. For example, the TCP/IP socket transport uses this field to specify the `(host, port)` tuple at which to connect to a worker.

`kill(manager, pid, config)` is called to remove a worker from the cluster. On the master process, the corresponding `IO` objects must be closed by the implementation to ensure proper cleanup. The default implementation simply executes an `exit()` call on the specified remote worker.

The Examples folder `clustermanager/simple` is an example that shows a simple implementation using UNIX domain sockets for cluster setup.

Network Requirements for LocalManager and SSHManager

Julia clusters are designed to be executed on already secured environments on infrastructure such as local laptops, departmental clusters, or even the cloud. This section covers network security requirements for the inbuilt `LocalManager` and `SSHManager`:

- The master process does not listen on any port. It only connects out to the workers.
- Each worker binds to only one of the local interfaces and listens on an ephemeral port number assigned by the OS.
- `LocalManager`, used by `addprocs(N)`, by default binds only to the loopback interface. This means that workers started later on remote hosts (or by anyone with malicious intentions) are unable to connect to the cluster. An `addprocs(4)` followed by an `addprocs(["remote_host"])` will fail.

Some users may need to create a cluster comprising their local system and a few remote systems. This can be done by explicitly requesting `LocalManager` to bind to an external network interface via the `restrict` keyword argument: `addprocs(4; restrict=false)`.

- `SSHManager`, used by `addprocs(list_of_remote_hosts)`, launches workers on remote hosts via SSH. By default SSH is only used to launch Julia workers. Subsequent master-worker and worker-worker connections use plain, unencrypted TCP/IP sockets. The remote hosts must have passwordless login enabled. Additional SSH flags or credentials may be specified via keyword argument `sshflags`.
- `addprocs(list_of_remote_hosts; tunnel=true, sshflags=<ssh keys and other flags>)` is useful when we wish to use SSH connections for master-worker too. A typical scenario for this is a local laptop running the Julia REPL (i.e., the master) with the rest of the cluster on the cloud, say on Amazon EC2. In this case only port 22 needs to be opened at the remote cluster coupled with SSH client authenticated via public key infrastructure (PKI). Authentication credentials can be supplied via `sshflags`, for example `sshflags=`-i <keyfile>`.

  In an all-to-all topology (the default), all workers connect to each other via plain TCP sockets. The security policy on the cluster nodes must thus ensure free connectivity between workers for the ephemeral port range (varies by OS).

  Securing and encrypting all worker-worker traffic (via SSH) or encrypting individual messages can be done via a custom `ClusterManager`.
- If you specify `multiplex=true` as an option to `addprocs`, SSH multiplexing is used to create a tunnel between the master and workers. If you have configured SSH multiplexing on your own and the connection has already been established, SSH multiplexing is used regardless of `multiplex` option. If multiplexing is enabled, forwarding is set by using the existing connection (`-O forward` option in ssh). This is beneficial if your servers require password authentication; you can avoid authentication in Julia by logging in to the server ahead of `addprocs`. The control socket will be located at `~/.ssh/julia-%r@%h:%p` during the session unless the existing multiplexing connection is used. Note that bandwidth may be limited if you create multiple processes on a node and enable multiplexing, because in that case processes share a single multiplexing TCP connection.

Cluster Cookie

All processes in a cluster share the same cookie which, by default, is a randomly generated string on the master process:

- `cluster_cookie()` returns the cookie, while `cluster_cookie(cookie)()` sets it and returns the new cookie.
- All connections are authenticated on both sides to ensure that only workers started by the master are allowed to connect to each other.

- The cookie may be passed to the workers at startup via argument `--worker=<cookie>`. If argument `--worker` is specified without the cookie, the worker tries to read the cookie from its standard input (`stdin`). The `stdin` is closed immediately after the cookie is retrieved.
- `ClusterManager`s can retrieve the cookie on the master by calling `cluster_cookie()`. Cluster managers not using the default TCP/IP transport (and hence not specifying `--worker`) must call `init_worker(cookie, manager)` with the same cookie as on the master.

Note that environments requiring higher levels of security can implement this via a custom `ClusterManager`. For example, cookies can be pre-shared and hence not specified as a startup argument.

Specifying Network Topology (Experimental)

The keyword argument `topology` passed to `addprocs` is used to specify how the workers must be connected to each other:

- `:all_to_all`, the default: all workers are connected to each other.
- `:master_worker`: only the driver process, i.e. `pid` 1, has connections to the workers.
- `:custom`: the `launch` method of the cluster manager specifies the connection topology via the fields `ident` and `connect_idents` in `WorkerConfig`. A worker with a cluster-manager-provided identity `ident` will connect to all workers specified in `connect_idents`.

Keyword argument `lazy=true|false` only affects `topology` option `:all_to_all`. If `true`, the cluster starts off with the master connected to all workers. Specific worker-worker connections are established at the first remote invocation between two workers. This helps in reducing initial resources allocated for intra-cluster communication. Connections are setup depending on the runtime requirements of a parallel program. Default value for `lazy` is `true`.

Currently, sending a message between unconnected workers results in an error. This behaviour, as with the functionality and interface, should be considered experimental in nature and may change in future releases.

Noteworthy external packages

Outside of Julia parallelism there are plenty of external packages that should be mentioned. For example MPI.jl is a Julia wrapper for the `MPI` protocol, or DistributedArrays.jl, as presented in Shared Arrays. A mention must be made of Julia's GPU programming ecosystem, which includes:

1. Low-level (C kernel) based operations OpenCL.jl and CUDAdrv.jl which are respectively an OpenCL interface and a CUDA wrapper.

2. Low-level (Julia Kernel) interfaces like CUDAnative.jl which is a Julia native CUDA implementation.
3. High-level vendor-specific abstractions like CuArrays.jl and CLArrays.jl
4. High-level libraries like ArrayFire.jl and GPUArrays.jl

In the following example we will use both `DistributedArrays.jl` and `CuArrays.jl` to distribute an array across multiple processes by first casting it through `distribute()` and `CuArray()`.

Remember when importing `DistributedArrays.jl` to import it across all processes using `@everywhere`

```
$ ./julia -p 4

julia> addprocs()

julia> @everywhere using DistributedArrays

julia> using CuArrays

julia> B = ones(10_000) ./ 2;

julia> A = ones(10_000) .* π;

julia> C = 2 .* A ./ B;

julia> all(C .≈ 4*π)
true

julia> typeof(C)
Array{Float64,1}

julia> dB = distribute(B);

julia> dA = distribute(A);

julia> dC = 2 .* dA ./ dB;

julia> all(dC .≈ 4*π)
true

julia> typeof(dC)
DistributedArrays.DArray{Float64,1,Array{Float64,1}}

julia> cuB = CuArray(B);
```

```
julia> cuA = CuArray(A);

julia> cuC = 2 .* cuA ./ cuB;

julia> all(cuC .≈ 4*π);
true

julia> typeof(cuC)
CuArray{Float64,1}
```

Keep in mind that some Julia features are not currently supported by CUDAnative.jl[2] , especially some functions like `sin` will need to be replaced with `CUDAnative.sin` (cc: @maleadt).

In the following example we will use both `DistributedArrays.jl` and `CuArrays.jl` to distribute an array across multiple processes and call a generic function on it.

```
function power_method(M, v)
    for i in 1:100
        v = M*v
        v /= norm(v)
    end

    return v, norm(M*v) / norm(v)  # or  (M*v) ./ v
end
```

`power_method` repeatedly creates a new vector and normalizes it. We have not specified any type signature in function declaration, let's see if it works with the aforementioned datatypes:

```
julia> M = [2. 1; 1 1];

julia> v = rand(2)
2-element Array{Float64,1}:
0.40395
0.445877

julia> power_method(M,v)
([0.850651, 0.525731], 2.618033988749895)

julia> cuM = CuArray(M);

julia> cuv = CuArray(v);

julia> curesult = power_method(cuM, cuv);
```

```
julia> typeof(curesult)
CuArray{Float64,1}

julia> dM = distribute(M);

julia> dv = distribute(v);

julia> dC = power_method(dM, dv);

julia> typeof(dC)
Tuple{DistributedArrays.DArray{Float64,1,Array{Float64,1}},Float64}
```

To end this short exposure to external packages, we can consider `MPI.jl`, a Julia wrapper of the MPI protocol. As it would take too long to consider every inner function, it would be better to simply appreciate the approach used to implement the protocol.

Consider this toy script which simply calls each subprocess, instantiate its rank and when the master process is reached, performs the ranks' sum

```
import MPI

MPI.Init()

comm = MPI.COMM_WORLD
MPI.Barrier(comm)

root = 0
r = MPI.Comm_rank(comm)

sr = MPI.Reduce(r, MPI.SUM, root, comm)

if(MPI.Comm_rank(comm) == root)
   @printf("sum of ranks: %s\n", sr)
end

MPI.Finalize()
```

```
mpirun -np 4 ./julia example.jl
```

- 1 In this context, MPI refers to the MPI-1 standard. Beginning with MPI-2, the MPI standards committee introduced a new set of communication mechanisms, collectively referred to as Remote Memory Access (RMA). The motivation for adding rma to the MPI

standard was to facilitate one-sided communication patterns. For additional information on the latest MPI standard, see https://mpi-forum.org/docs.

- 2 Julia GPU man pages

« Multi-Threading Running External Programs »