Simulated MPI

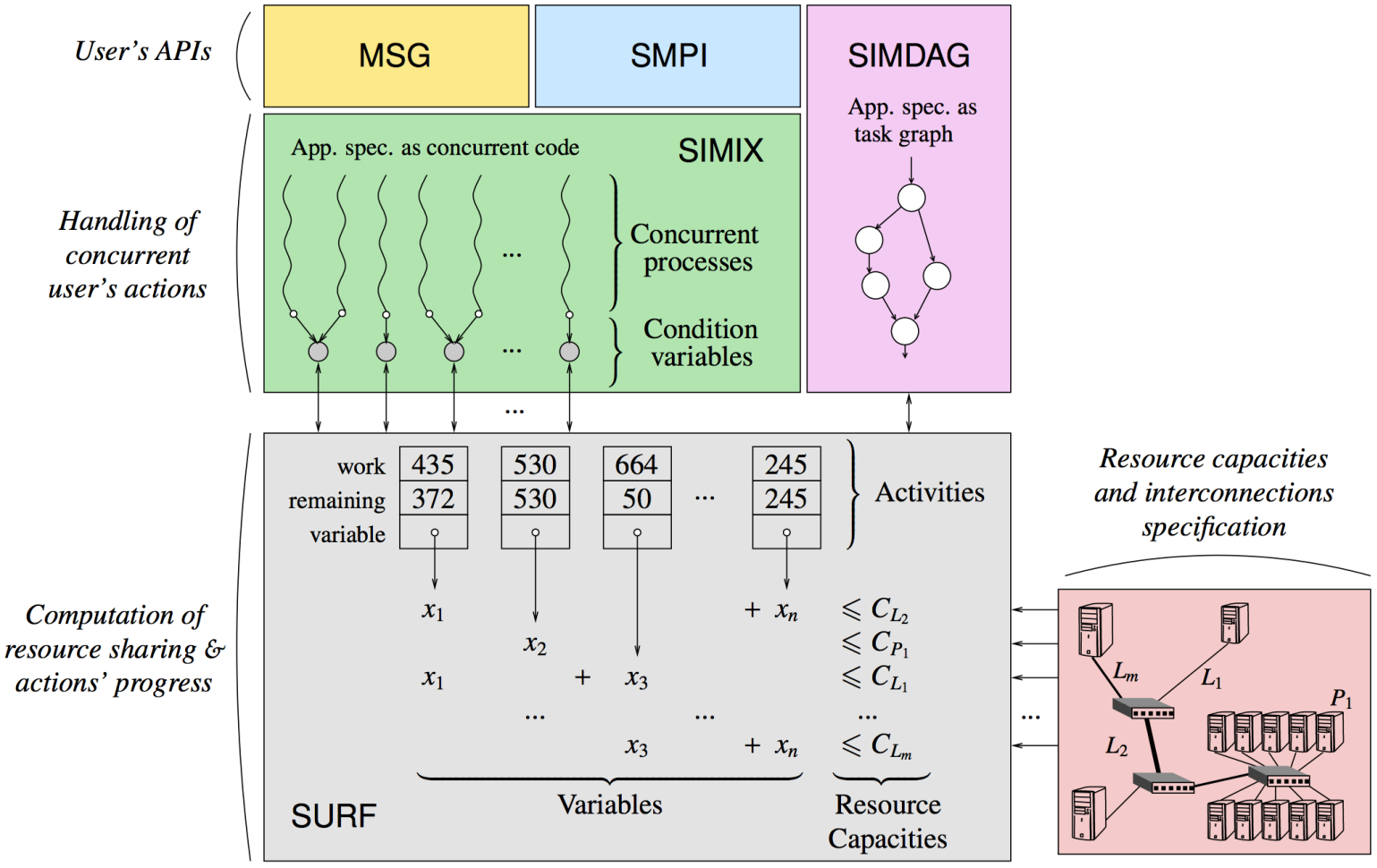
ICS632: Principles of High Performance Computing

Henri Casanova (henric@hawaii.edu)

Fall 2015

MPI in Simulation

- We use SimGrid: http://simgrid.gforge.inria.fr



Installing/Testing/Running SimGrid

- Installing SimGrid:
  - There is a lot of information on the SimGrid Web site
  - (some summarized on the ICS632 Web site)
- Using SMPI:
  - Compiling an SMPI Program: use `smpicc` just like you'd use `mpicc`
  - Running an SMPI program: use `smpirun` just like you'd use `mpirun`.... but for a few extra command-line arguments
  - NO BATCH SCHEDULER

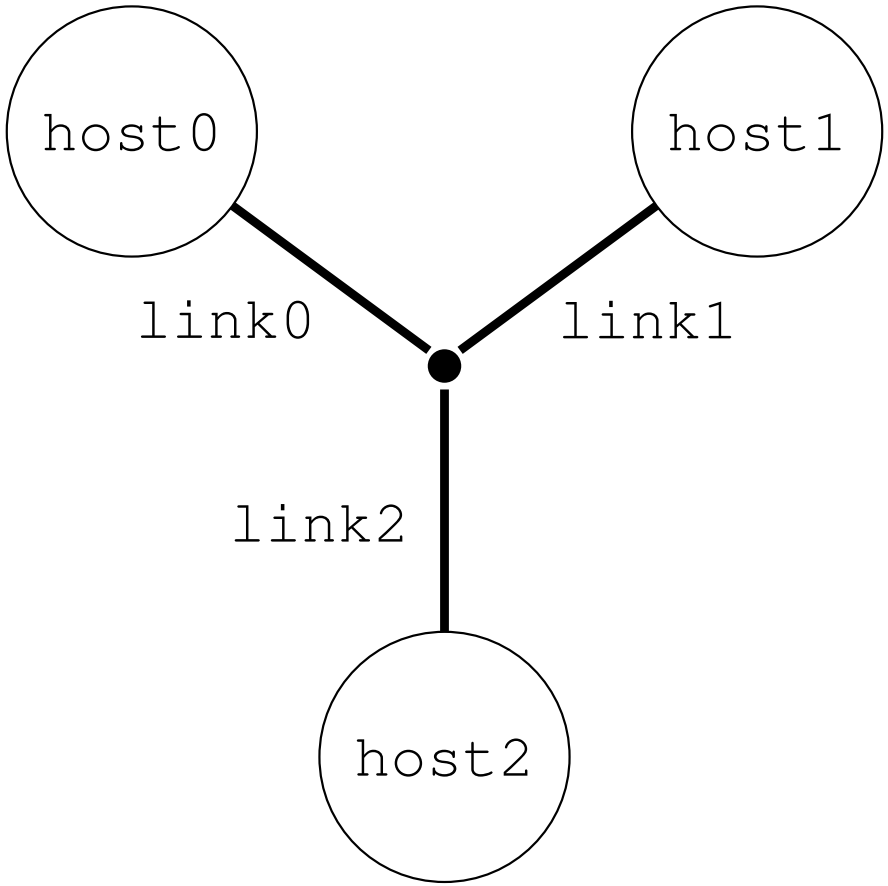
How does SMPI work?

- Should compile standard MPI code, unmodified
- Runs MPI processes as threads within a single process, and the threads run in round-robin fashion
- Application computational code is executed, but when an MPI call is encountered, a communication delay is simulated
- Application code runs on you machine, and you can specify to `smpirun` how fast your machine compute speed is compared to that of the platform to be simulated
- You also describe to `smpirun` the physical characteristics of the platform to be simulated as an XML file

A Note on Simulation Accuracy

- Simulation accuracy is a main concern of SimGrid
  - This seems like a given, but check out the SimGrid JPDC article to see how other "competitors" fare
- As a result, simulation models are complex and very much unlike the simple $\alpha + m\beta$ model from previous slides
  - Network protocol effects, MPI optimizations, ...
- So the simulation may appear to behave strangely when parameters are tuned, just like a real network
  - e.g., increasing message by 1 byte can have a large impact on data transfer time
  - e.g., very large latencies have odd side-effects
- Something to keep in mind when analyzing simulation results

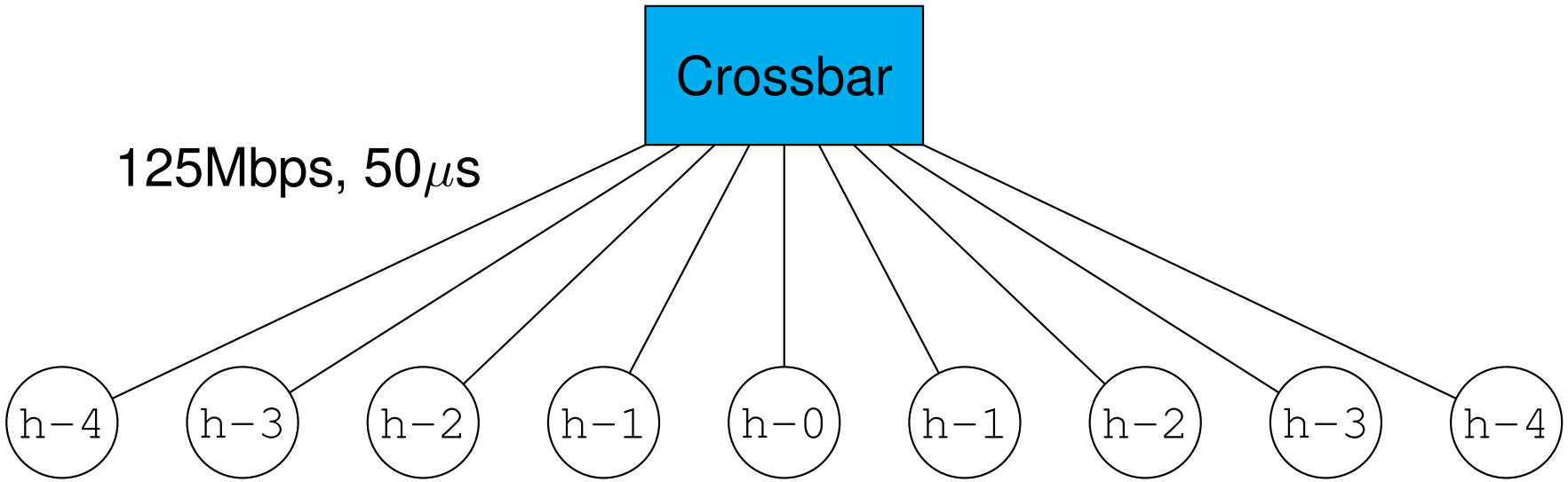
3 Interconnected Hosts



3 Interconnected Hosts

```
<?xml version='1.0'?>
<!DOCTYPE platform SYSTEM "http://simgrid.gforge.inria.fr/simgrid.dtd">
<platform version="3">
  <AS  id="AS0"  routing="Full">
    <host id="host0" power="1Gf"/> <host id="host1" power="2Gf"/>
    <host id="host2" power="40Gf"/>
    <link id="link0" bandwidth="125MBps" latency="100us"/>
    <link id="link1" bandwidth="50MBps"  latency="150us"/>
    <link id="link2" bandwidth="250MBps" latency="50us"/>
    <route src="host0" dst="host1"><link_ctn id="link0"/><link_ctn
       id="link1"/></route>
    <route src="host1" dst="host2"><link_ctn id="link1"/><link_ctn
       id="link2"/></route>
    <route src="host0" dst="host2"><link_ctn id="link0"/><link_ctn
       id="link2"/></route>
  </AS>
</platform>
```

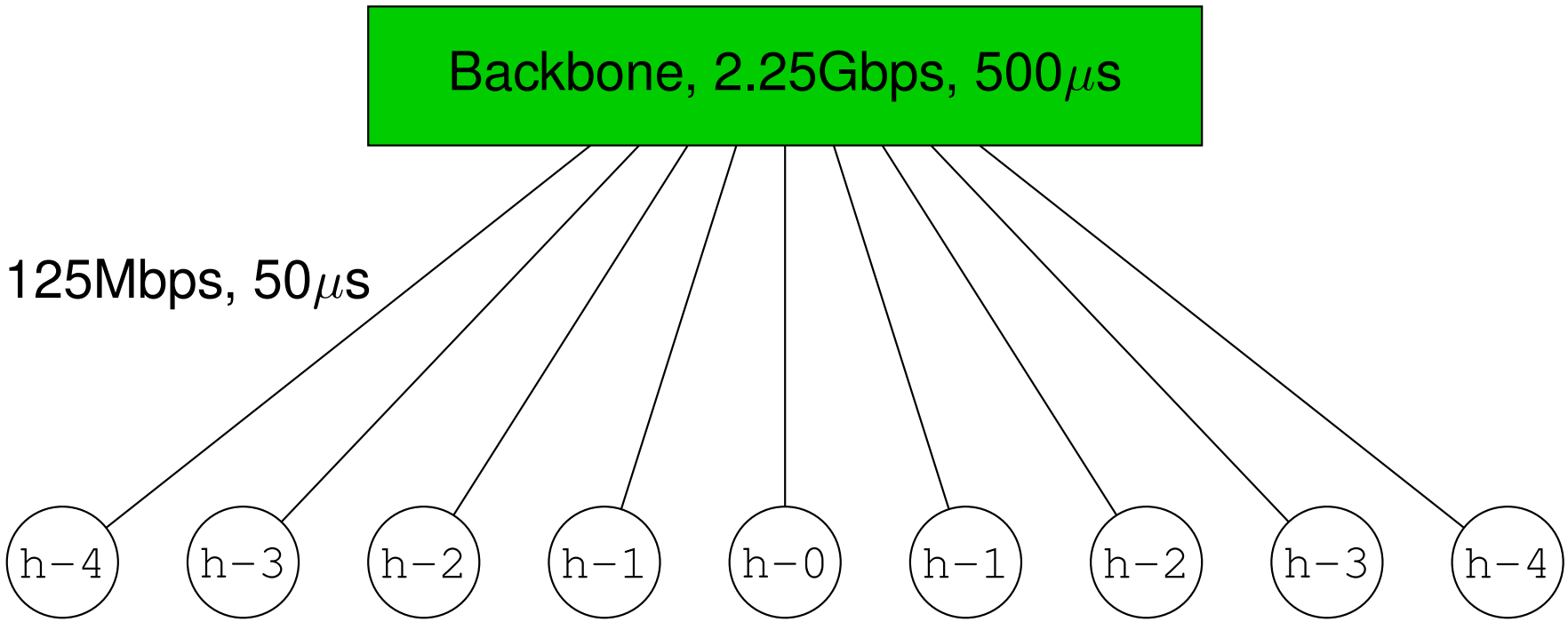
Homogeneous Cluster with Crossbar



Homogeneous Cluster with Crossbar

```
<?xml version='1.0'?>
<!DOCTYPE platform SYSTEM "http://simgrid.gforge.inria.fr/simgrid.dtd">
<platform version="3">
<AS id="AS0" routing="Full">
  <cluster id="my_cluster" prefix="host-" suffix=".hawaii.edu"
           radical="0-255" power="1Gf" bw="125MBps" lat="50us"/>
</AS>
</platform>
```

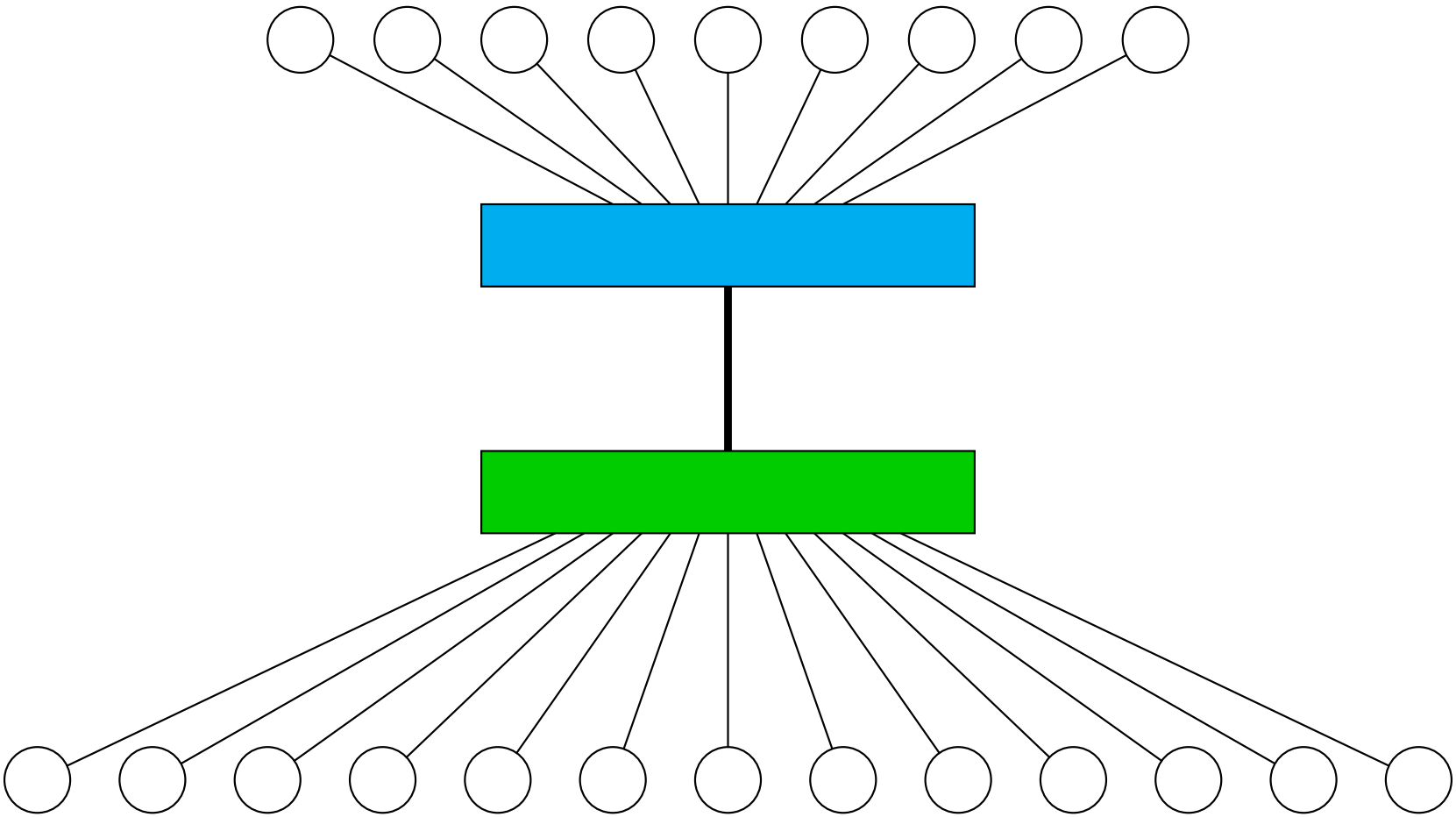
Homogeneous Cluster with Shared Backbone



Homogeneous Cluster with Shared Backbone

```
<?xml version='1.0'?>
<!DOCTYPE platform SYSTEM "http://simgrid.gforge.inria.fr/simgrid.dtd">
<platform version="3">
<AS id="AS0" routing="Full">
  <cluster id="my_cluster" prefix="node-" suffix=".hawaii.edu"
           radical="0-255" power="1Gf" bw="125MBps" lat="50us"
           bb_bw="2.25GBps" bb_lat="500us"/>
</AS>
</platform>
```

Two Clusters



Two Clusters

```
<?xml version='1.0'?>
<!DOCTYPE platform SYSTEM "http://simgrid.gforge.inria.fr/simgrid.dtd">
<platform version="3">
<AS id="AS0" routing="Full">
  <cluster id="my_cluster_1" prefix="C1-" suffix=".hawaii.edu"
      radical="0-15" power="1Gf" bw="125MBps" lat="50us"
      bb_bw="2.25GBps" bb_lat="500us" />
  <cluster id="my_cluster_2" prefix="C2-" suffix=".hawaii.edu"
      radical="0-31" power="2Gf" bw="125MBps" lat="50us" />
  <link id="internet_backbone" bandwidth="1.25GBps" latency="500us" />
  <ASroute src="my_cluster_1" dst="my_cluster_2"
      gw_src="C1-my_cluster_1_router.hawaii.edu"
      gw_dst="C2-my_cluster_2_router.hawaii.edu" symmetrical="YES">
    <link_ctn id="internet_backbone" />
  </ASroute>
</AS>
</platform>
```

Some Misc. Information

- Units:
  - `Bps`: bytes (`MBps`, `GBps`)
  - `bps`: bits (`Mbps`, `Gbps`)
- No need to take averages over multiple trials as long as your application doesn't have a random behavior
  - Simulations are reproducible exactly!
- One declares the route from A to B, and by default it is assumed that the same route is used from B to A
- See more information on the SimGrid Web site...

SMPI Execution Example

- Let us consider a simple MPI program, `roundtrip.c` in which the process of rank i sends a 10MiB message to process i+1, starting with process 0, stopping when the message has done a full round-trip
- We use an XML platform description file
- We create a hostfile that lists all the hostnames

```
% smpicc rountrip.c -o rountrip
% smpirun --cfg=smpi/running_power:1 -np 48 -hostfile ./hostfile -platform
    ./two_clusters.xml ./roundtrip
```

- Let's run examples and play around...

SMPI Limitations

- Don't use multi-threading in your (simulated) MPI processes
  - We simulate single-core systems or we "abstract" away multi-core systems as single-core systems
  - Full-fledge simulation of multi-core code with (simulated) OpenMP is currently not available in SMPI
- Unless you're ok with a slow simulation, SMPI can bypass some of the simulated code, but the simulation no longer computes anything useful
  - Data dependent behavior is lost (e.g., control flow based on computed values)
- Only Gigabit Ethernet networks are (accurately) simulated
- If you make your C/C++ code weird, `smpicc` may have a hard time compiling it

Conclusion

- It will be hard to make simulation results match with real results on the Cray
  - SIMGRID doesn't simulate Infiniband (yet)
- It will actually be interesting whether they match at all...
- Let's look at our programming assignment...